Bioinformatic tools and computational methods for mapping DNA methylation variability

Dissertation

zur Erlangung des Grades
des Doktors der Naturwissenschaften
der Naturwissenschaftlich-Technischen Fakultät III
Chemie, Pharmazie, Bio- und Werkstoffwissenschaften
der Universität des Saarlandes

von

Pavlo Lutsik, M.Sc.

Saarbrücken
2016

Tag des Kolloquiums:	15.02.2017
Dekan:	Prof. Dr. Guido Kickelbick
Berichterstatter:	Prof. Dr. Jörn Walter
	Prof. Dr. Matthias Hein
Vorsitz:	Prof. Dr. Volkhard Helms
Akad. Mitarbeiter:	Dr. Björn Becker

Abstract

DNA methylation is one of the key epigenetic marks extensively studied for its association with environmental exposures and human diseases. DNA methylation can be profiled by a range of methods which differ drastically in their genomic coverage, throughput and resolution. The present thesis encompasses a series of bioinformatic solutions for tackling related data analysis problems.

First, comprehensive and user-friendly tools were developed for processing and primary analysis of bisulfite-based DNA methylation data. The R-package RnBeads supports analysis of genome-scale profiles from Infinium microarrays and bisulfite sequencing, while BiQ Analyzer HT and HiMod enable complete and interactive analysis of deep locus-specific sequencing assays of 5-methylcytosine and its oxidative derivatives. Second, to address cellular heterogeneity in a genome-wide DNA methylation study of birth-weight we proposed an original approach for correcting the statistical analysis. Third, a novel deconvolution method MeDeCom was developed that facilitates data-driven exploration of heterogeneous DNA methylomes.

Collectively, the results of the present thesis comprise different data analysis facets of a large-scale DNA methylation study. Most of the presented bioinformatic solutions already facilitate epigenetic research in numerous life-science groups worldwide.

Kurzfassung

DNA Methylierung ist eine wichtige epigenetische Modifikation, die besonders in Hinblick auf ihre Assoziation mit Umwelteinflüssen und Krankheiten intensiv untersucht wird. DNA Methylierung kann durch verschidene von Methoden, die sich stark in Bezug auf ihre genomische Abdeckung, den Probendurchsatz sowie ihre Auflösung unterscheiden, ermittelt werden. Die vorliegende Arbeit umfasst eine Reihe bioinformatischer Lösungen, um relevante Probleme der Datenanalyze zu beheben:

Erstens, umfassende und benutzerfreundliche Werkzeuge zur Verarbeitung und primären Analyse von Bisulfit-basierten DNA Methylierungsdaten. Ein R-Paket RnBeads unterstützt die Analyze von genomweiten Infinium Bead Arrays und Bisulfitsequenzierungen. BiQ Analyzer HT und Himod ermöglichen eine volle und interaktive Analyse von lokus-spezifischer Bisulfit-Tiefensequenzierung von 5-Methylcytosin und seinen oxidativen Derivaten.

Zweitens, ein neues Verfaren zur Korrektur der statistischen Analyse, um das Problem der Zellularer Heterogenität des Methyloms in genomweiten DNA Methylierungsstudien zum Einfluss des Geburtsgewichtes zu lösen.

Drittens, eine neue Dekonvolutionsmethode “MeDeCom”, die die Referenz-freie Untersuchung heterogener Datensätze erlaubt.

Zusammengenommen umfassen die Ergebnisse der vorliegenden Arbeit verschiedene Aspekte der Datenanalyse im Rahmen einer großangelegten DNA Methylierungsstudie. Die hier dargestellten Lösungen vereinfachen die Arbeit von Biowissenschaftlern in vielen Forschungsgruppen weltweit.

Acknowledgements

This doctoral work would have not been possible without generous and sincere help of my colleagues, family and friends whom I owe a deep sense of gratitude.

Firstly, I would like to thank all my co-authors whose work enabled the contributions presented below. My most special co-author Nicole Souren combines many other qualities as my best friend, my wife and the mother of my wonderful son Joep. Nicole supported me enormously throughout the years as a PhD student, standing courageously by my side in all the hurdles and troubles. She granted me more time to work than any other partner would have done, and still had kindness to me even though I did not use it as good as I could. I am also thankful to my constantly worried parents and to my brother Petro, who provided me with helpful advice on broader mathematical issues and intricate subtleties of LaTeX. I enjoyed tremendous support from my family in law whom I want to deeply thank as well.

I am enormously grateful to my supervisor Joern Walter who accepted me as a PhD student in his group in spite of my mediocre wet-lab skills. Joern has given me an unseen freedom to pursue the research directions that, as I believed, were the most fruitful, and had stoic patience all the numerous times when they turned out not to be so. I also appreciate the daily assistance of my colleagues and co-authors for the great years spent in Joern's lab. Julia Arand was my collaborator and a real friend since the days I was working of my M.Sc. thesis. With Gilles Gasparoni, Karl Nordström and Abdulrahman Salhab we shared the joy of discovering something exciting in the data and frustration about the Linux-based computing infrastructure. Together with Mark Wossidlo, Konstantin Lepikhov, Sascha Tierling, Jie Lan and Pascal Giehr we fought the adverse effects of a sedentary lifestyle at table football. I am thankful to Kathrin Kattler for her kind help in preparing the German version of the Abstract.

I am indebted to Christoph Bock who introduced me to epigenetic research and thereby determined my future career for many years ahead. Together with Thomas Lengauer they have been very demanding and critical, yet fair supervisors of the software projects that I worked on, bringing them to a qualitatively higher level. Yassen Assenov and Fabian Müller were my partners in R coding endeavours from whom I learned so much.

I am grateful to Matthias Hein and Martin Slawski who helped me building a bridge between the fundamental computer science and the cutting edge biological research. Owing to the insightful discussions they had time for, I gained a lot of understanding of deep mathematical subjects which I had little chance to obtain elsewhere.

Finally, I thank my closest friends Stephan Neumann, Ayman Heidar and Hassan Soumsomani for providing such a necessary distraction in those rare days when the routine prevailed over the excitement of the scientific ventures. Although life brings us to different places I keep hoping our friendship will stand this trial.

Throughout a large period of my PhD studies I was supported by the EU Framework Programme 7, grant agreement No. 267038 (NOTOX).

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Preface

Epigenetics has greatly contributed to our understanding of how complex and versatile cell phenotypes arise through the interaction of a relatively constant genetic background with environmental influences. Among the so far known epigenetic mechanisms, the phenomenon of DNA methylation has for a long time been in the central focus.

The present thesis deals with methodological aspects of DNA methylation data analysis and provides concrete computational and bioinformatic solutions for its specific problems. Nevertheless, it appears important to first answer the following questions: What is DNA methylation and where is it found with respect to the genome? How are the DNA methylation patterns established and get diverse in different cells of an organism? How similar are methylation profiles of different individuals and what are the reasons behind the differences? What are the functions of DNA methylation and which relation does it have to the genetic information? How is DNA methylation linked to diseases and how can this association be studied? Which methods exist to map DNA methylation and what are their strong and weak sides? Which difficulties are they associated with, both wet and dry lab, and which bioinformatic solutions exist for these difficulties? What is understood under DNA methylation heterogeneity? When is it a problem and how can it be addressed experimentally and computationally?

The purpose of this introduction is to provide a minimal yet sufficient background for understanding the projects summarized in this thesis. The introduction is organized in three parts. Section 1.1 provides basic knowledge about DNA methylation, its genomic and intra- and inter-organismal patterns, functional role and link to diseases. Section 1.2 gives a detailed introduction into the methods for mapping DNA methylation, their limitations and data analysis aspects. Section 1.3 outlines the problem of methylome heterogeneity, introduces available experimental and computational methods for addressing it.

1.1 DNA Methylation

The four letter DNA alphabet has been most likely inherited by all known species from their last common ancestor [Szathmáry, 2003]. The canonical chemical structure of the the four nitrogenous bases is studied in sufficient detail (Figure 1.1) [Townsend, 2013]. DNA methylation refers to a covalent modification of the bases directly in DNA by the addition of methyl groups at strictly defined positions. The most widespread form of DNA methylation is methylation of cytosine at the fifths carbon atom of the pyrimidine ring. The resulting base variant is commonly known as 5-methylcytosine (the rightmost structure in Figure 1.1).

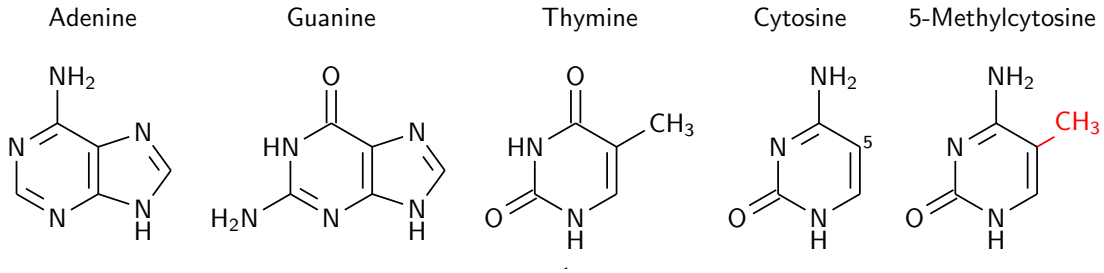


Figure 1.1: Bases of DNA and 5-methylcytosine.

1.1.1 Early observations and the cellular memory hypothesis

Strikingly, the presence of 5-methylcytosine in a living organism, bacteria *Mycobacterium tuberculosis*, was for the first time detected long before anything was known about DNA and its biological role [Ruppel, 1898]. Later it was also found in other prokaryotes [Dunn and Smith, 1955], mammals [Hotchkiss, 1948] and plants [Wyatt, 1951]. While the function of DNA methylation in prokaryotes as a mechanism in a restriction-modification host deference system was elucidated quite early [Arber and Dussoix, 1962; Smith *et al.*, 1972], little was known about what role it played in higher organisms. Even more puzzling was the complete or near-complete absence of DNA methylation in several important model species such as yeast *S. cerevisia* and the fruit fly *D. melanogaster* [Capuano *et al.*, 2014].

Observations of the early studies [Srinivasan and Borek, 1964] crystallized into a hypothesis stating that DNA methylation could be a mitotically heritable mechanism of cellular memory about gene activity states [Holliday and Pugh, 1975; Riggs, 1975]. Evidence supporting the cellular memory hypothesis came from four major directions: i) non-random patterns of 5-methylcytosine across the genome and dedicated molecular mechanisms to propagate these patterns between the cell generations; ii) diversity of methylation patterns across the cells of the same organism; iii) inter-individual differences; iv) participation in mechanisms controlling activity of the underlying genomic sequence. Each of these aspects are covered in detail below.

1.1.2 Genomic distribution, enzymatic setting and removal

CpGs and mitotic heritability of DNA methylation patterns

The cellular memory hypothesis took into account an earlier finding that in vertebrate genomes 5-methylcytosine was predominantly found to precede a guanine base, i.e. at dinucleotide CpG motifs or, shortly, *CpGs* [Doscocil *et al.*, 1962]. The fact that CpG is palindromic, i.e. repeats

itself in an anti-parallel fashion on both strands of the same DNA molecule, is directly linked to the propagation of the methylation patterns through the cell divisions. It was shown that after the duplication of methylated DNA the methylation marks were introduced into the newly synthesized DNA strands [Bird, 1978]. Importantly, methyl groups were transferred only to those CpGs which have already been methylated in the original DNA molecule, reproducing its methylation pattern [Bird, 1978; Pollack *et al.*, 1980; Stein *et al.*, 1982a]. These findings were advocating for the earlier suggested semi-conservative mechanism behind the mitotic inheritance of the DNA methylation [Holliday and Pugh, 1975], establishing the latter as a plausible cellular memory phenomenon.

Genome-wide picture: CpG-dense islands in the genomic sea of methylation

CpG dinucleotides are drastically underrepresented in vertebrate genomes, comprising approximately a quarter of the expected number [Swartz *et al.*, 1962] and have a very non-uniform distribution across the genome. Experiments with CpG-cutting restriction enzymes revealed that CpGs are clustering in short stretches of 1 to 2 kb where their frequency is an order of magnitude higher than average [Bird *et al.*, 1985; Cooper *et al.*, 1983]. These genomic regions were termed *CpG islands* [Bird, 1986].

It was also shown that most of the CpGs are methylated, i.e. contain 5-methylcytosine [Bird and Taggart, 1980; Ehrlich *et al.*, 1982]. Failure to repair deaminated methylcytosines in the germline could explain the observed CpG depletion taking place along the evolutionary history [Bird, 1980; Coulondre *et al.*, 1978]. The most likely reason behind the local enrichment of CpG sites was the fact that, unlike the rest of genomic CpGs, the majority of the island CpGs were unmethylated [Bird, 1986].

Since then CpG islands have occupied a central place in our view of DNA methylation landscapes. There were several attempts to devise a quantitative definition for a CpG island and its borders which would enable their automatic search and annotation [Gardiner-Garden and Frommer, 1987; Hackenberg *et al.*, 2006; Takai and Jones, 2003]. More recent genome-wide studies extended the perception of a CpG island. In particular, it turned out that DNA methylation is more dynamic in the regions surrounding the islands which obtained a term "CpG island shores" [Irizarry *et al.*, 2008]. CpG island shores are now defined as 2 kb regions flanking a CpG island upstream and downstream. Furthermore, 2 kb regions flanking the shores are now referred to as "CpG island shelves" [Bibikova *et al.*, 2011].

Setting and maintenance by DNA methyltransferases

An important and ubiquitous biological mechanism implies the presence of a dedicated enzymatic machinery. The cellular memory hypothesis suggested two classes of enzymes responsible for setting methylation marks, *de novo* methyltransferases, methylating previously unmodified DNA, and *maintenance* methyltransferases copying the established pattern to a newly synthesized DNA strand [Holliday and Pugh, 1975]. Subsequently, the first eucaryotic DNA methyltransferase (DNMT1) was discovered [Bestor and Ingram, 1983; Gruenbaum *et al.*, 1982] and subsequently cloned from the mouse [Bestor *et al.*, 1988] and human genomes [Yen *et al.*, 1992]. Due to a clear preference for a hemimethylated substrate, DNMT1 was assigned with the maintenance role. Later another family of DNMTs was cloned, which included two candidate *de novo* DNMTs, DNMT3A and DNMT3B [Okano *et al.*, 1999]. Further research added new dimensions to the initial simple model of the DNA methylation enzymes [Arand *et al.*, 2012], but the concepts of the maintenance and *de novo* activity are still at the core of our understanding about how methylation patterns are established [Jones and Liang, 2009].

Removal by TETs, oxidative DNA modifications

A more recent advance in DNA methylation research was the discovery of oxidative 5-methylcytosine varieties in mammalian brain [Kriaucionis and Heintz, 2009; Tahiliani *et al.*, 2009]. Simultaneously, the key players were identified to be the family of ten-eleven translocation (TET) enzymes, TET1, TET2 and TET3 [Tahiliani *et al.*, 2009]. TETs oxidize 5-methylcytosine to 5-hydroxymethylcytosine with the assistance of α-ketoglutarate and Fe^{2+} cations [Delatte *et al.*, 2014]. Furthermore, it was shown that TETs are capable of catalyzing further oxidation leading to 5-formyl- and 5-carboxycytosine [Ito *et al.*, 2011; Tahiliani *et al.*, 2009]. Among other, the evidence was collected that oxidation of 5-methylcytosine by TETs is the initial event of the demethylation cascade [Guo *et al.*, 2011]. The removal of the mark could proceed either as passive dilution due to inability of DNMT1 to use 5-hydroxymethylcytosine as substrate, or as an active elimination by base-excision repair machinery [Delatte *et al.*, 2014].

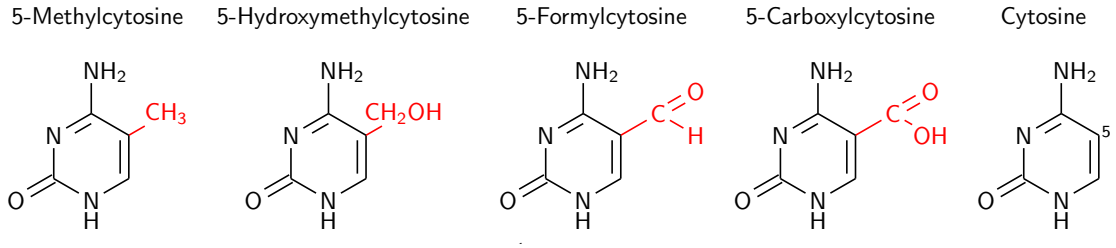


Figure 1.2: Oxidative derivatives of 5-methylcytosine.

Summary: the dynamic DNA methylation landscape

The DNA methylation landscape and its dynamics in a typical vertebrate genome looks roughly as follows. 5-methylcytosine is found with almost no exception within CpG dinucleotides, although in certain mammalian cell types, such as stem cells and adult neurons, abundant non-CpG methylation was reported [Lister *et al.*, 2013, 2009]. The genome as a whole has low CpG density and is hypermethylated, while the short CpG dense islands are predominantly unmethylated. Methylation patterns are set by *de novo* methyltransferases DNMT3A and DNMT3B and maintained by DNMT1. Furthermore, methylcytosine can be present in oxidized forms established by TETs. Oxidation by TETs usually precedes the removal of methylation. The genomic DNA methylation profile and its temporal changes have to be considered when studying it genome-wide.

1.1.3 Origins and cell type-specificity of DNA methylation patterns

In addition to a complex and, with a high probability, functionally relevant distribution of 5-methylcytosine across the genome it was also noticed that the extent and patterns of DNA methylation vary between different cells of an organism. In early studies these differences were detected both by bulk and locus-specific DNA methylation measurements [Gruenbaum *et al.*, 1981; Mandel and Chambon, 1979; Shen and Maniatis, 1980]. Later genome-scale profiling in multiple tissues, cell types and lineages confirmed and extended this knowledge, in particular in human [Eckhardt *et al.*, 2006; Varley *et al.*, 2013; Zhang *et al.*, 2009; Ziller *et al.*, 2013]. The current view goes as far as stating that each individual human cell has a potentially unique pattern of methylation at approximately 27 million CpGs present in the haploid genome, collectively denoted as the *DNA methylome* [Pelizzola and Ecker, 2011]. It is stunning

that all the observed methylome diversity originates from a single methylation pattern of the zygote.

Global DNA methylation changes in early embryonic development

Valuable insights about the origins of the intra-organismal diversity of the cell methylomes were provided by embryological studies in mammals. It was shown that shortly after fertilization the paternal and maternal genomes undergo fast and dramatic DNA methylation changes [Feng *et al.*, 2010; Mayer *et al.*, 2000; Oswald *et al.*, 2000; Rougier *et al.*, 1998]. First, a fast TET-mediated demethylation of the paternal genome starts with much of the 5-methylcytosine being substituted by 5-hydroxymethylcytosine prior to the first DNA replication [Iqbal *et al.*, 2011; Wossidlo *et al.*, 2011]. It is followed by a slower demethylation of the maternal genome, taking place over the first several replication cycles [Feng *et al.*, 2010; Geiman and Muegge, 2010]. Both pronuclei lose almost a half of the methylcytosine by the time of implantation [Allis *et al.*, 2007]. The lowest level of genome-wide methylation is observed in the preimplantation blastocyst where only 25% of CpGs are methylated [Lee *et al.*, 2014]. The set of genomic loci maintaining methylation includes, among other, the repetitive elements, such as Intracisternal A-Particle (IAP), and the imprinted loci.

Embryonic stem cells provide a model of the methylome "ground state"

From the bottom levels observed in the blastocyst prior to the implantation, re-establishment of DNA methylation begins, in which a decrease of the TET activity and an increase of *de novo* methylation DNMT3A/3B play a decisive role [Lee *et al.*, 2014]. Since studying DNA methylation in a developing embryo is difficult many insights were obtained using embryonic stem (ES) cells. Depending on the culturing conditions ES cells come in two flavours [Ficz *et al.*, 2013; Habibi *et al.*, 2013]. The "naive" ES cells have a very low global DNA methylation level and directly correspond to the inner cell mass (ICM) of the pre-implantation mammalian embryo. The hypermethylated "primed" ES cells correspond to the ICM after a large re-methylation wave taking place around the implantation time [Lee *et al.*, 2014]. They turned out to be even slightly hypermethylated compared to fully differentiated tissues, which implies that the establishment of tissue- and cell type-specific methylomes is also linked to a loss of methylation [Ziller *et al.*, 2013]

High-resolution profiling of the human ES cells showed that DNA methylation at a large portion of CpGs is dynamic and changes throughout the development [Hodges *et al.*, 2011; Laurent *et al.*, 2010; Lister *et al.*, 2009; Smith *et al.*, 2012; Ziller *et al.*, 2013]. Computational analysis revealed that ES cells seem to be constantly renewing their methylome to maintain a "clean" epigenetic ground state [Landan *et al.*, 2012]. A high methylome turnover rate at this transition stage is confirmed by the analysis of hydroxymethylation in "primed" ES cells [Booth *et al.*, 2012; Ficz *et al.*, 2013]. A stochastic and non-uniform setting of DNA methylation patterns during the transition between the "primed" and "naive" states most probably underlies the ever increasing diversion of the cell methylomes in the subsequent embryonic and postnatal development [Lee *et al.*, 2014]. Thereby emerging heterogeneity can have several possible molecular mechanisms, including differential expression or targeting of TETs and DNMTs, strand-specific effects of *de novo* methylation and oxidation, as well as inefficient maintenance [Lee *et al.*, 2014].

Lineage-commitment leads to a cell type-specific methylome

Diversion of methylation patterns continues during the later developmental stages. A large-scale comparative study of tissues from three different germ layers discovered numerous CpG positions specifically methylated in endoderm, mesoderm and ectoderm [Irizarry *et al.*, 2009]. Lineage determination in hematopoietic cells was shown to cause well defined changes all across the methylome [Ji *et al.*, 2010]. Similar processes were reported to happen during neuronal differentiation [Kim *et al.*, 2014]. As a result, virtually every tissue, cell type and cell population in an adult mammalian organism is characterized by a specific DNA methylation signature [Lister *et al.*, 2013; Varley *et al.*, 2013; Ziller *et al.*, 2013].

1.1.4 Individual differences

The next dimension of DNA methylation diversity was discovered when studying differences between individuals [Bock *et al.*, 2008; Eckhardt *et al.*, 2006]. The major drivers of these differences in a population of healthy human subjects include genetic variation, gender and age [Pirazzini *et al.*, 2012].

Interplay with the genotype

Large-scale studies showed that genetic variation explains a significant portion of the observed DNA methylation differences [Gutierrez-Arcelus *et al.*, 2013; van Dongen *et al.*, 2016]. This statistical association can have several aspects. First, given that the majority of somatic cells are diploid, the methylation state of CpGs can differ between the two homologous chromosomes, the phenomenon known as *allele-specific methylation* (ASM). ASM was shown to be abundant in the human genome and mostly driven by polymorphisms directly affecting the CpG cytosines [Shoemaker *et al.*, 2010]. Second, cis-acting genetic variants can be quantitatively associated with the bulk methylation level at neighboring CpG positions. This type of genetically influenced regions were termed *methylation quantitative trait loci* (methQTLs) [Rakyan *et al.*, 2011].

Gender

Chromosomal basis of sex determination in mammals implies that males and females have different DNA methylation landscapes already due to a different number of genomic CpGs. The hemizygous state of chromosome X in human male organisms limits the potential number of possible methylation states at almost 2.5 million CpG positions. On the other hand chromosome Y carries CpG positions which are absent in female organisms. Except for such trivial differences sex-specific methylation on the autosomal loci was observed in numerous studies [El-Maarri *et al.*, 2007; Sarter *et al.*, 2005]. This was later confirmed by several genome-scale screens [Boks *et al.*, 2009; Liu *et al.*, 2010; van Dongen *et al.*, 2016].

Age

Maintenance errors can accumulate over the lifespan of an individual, resulting in stochastic or directed changes of the methylome. Decrease of 5-methylcytosine abundance over the lifespan was observed very early in multiple vertebrate species [Berdyshev *et al.*, 1967; Vanyushin *et al.*, 1973; Wilson *et al.*, 1987]. The age-related global hypomethylation was subsequently confirmed in human [Fuke *et al.*, 2004]. In addition to this pan-genomic effect, it was noticed that selected, CGI-overlapping loci gain methylation with age [Issa, 2003; Nakagawa *et al.*,

2001; Shen *et al.*, 2003] later associated with the bivalent domains [Rakyan *et al.*, 2010]. In humans these changes seem to occur with a very similar rate across all tissues [Jones *et al.*, 2015], which allowed the creation of a surrogate DNA methylation age calculator [Bocklandt *et al.*, 2011; Horvath, 2013; Weidner *et al.*, 2014].

On the conceptual level the current view is that age-related changes are a result of two independent phenomena, the *epigenetic drift* which is a, supposedly, stochastic divergence of epigenomes from the common origin over time, and the *epigenetic clock* implying directed changes of methylation at certain sites [Jones *et al.*, 2015; Teschendorff *et al.*, 2013]. It was, however, also shown that due to cell type specificity of DNA methylation age-related changes can at least in some cases be a result of changed cellular composition [Jaffe and Irizarry, 2014; Weng *et al.*, 2009].

1.1.5 Biological function

Highly non-random distribution of methylation and CpG dinucleotides over the genomes, the presence of a specialized enzymatic machinery, differences across tissues and cell types as well as variation between individuals suggests that DNA methylation is playing a significant functional role. Furthermore, the importance of DNA methylation was confirmed by the loss-of-function experiments, which showed that the DNMTs are essential for the mammalian embryonic development [Li *et al.*, 1992; Okano *et al.*, 1999].

Association with global repression: X chromosome inactivation, imprinted genes and silencing of transposons

The idea that DNA methylation is directly linked to silencing of the underlying genomic regions was expressed already as a part of the epigenetic memory hypothesis [Holliday and Pugh, 1975; Riggs, 1975] and supported by gene-specific [Bird, 1978; Christman *et al.*, 1977; Desrosiers *et al.*, 1979; McGhee and Ginder, 1979] and transfection-based experiments [Stein *et al.*, 1982b; Vardimon *et al.*, 1982]. Later research revealed the essential role of the DNA methylation in key repressive epigenetic phenomena. First, it was shown that the methylation level of the genes on the inactivated X chromosome is substantially higher compared to the active one [Wolf *et al.*, 1984] and that they can be derepressed by the use of DNMT-inactivating nucleotide analogue 5-azacytidine [Mohandas *et al.*, 1981; Venolia *et al.*, 1982]. Second, the imprinted genes with parent-of-origin-dependent expression were related to clusters of allele-specific differentially methylated regions marking the suppressed allele from the corresponding gamete down to the somatic cells [Bartolomei *et al.*, 1993; Ferguson-Smith *et al.*, 1993; Liu *et al.*, 2000; Shemer *et al.*, 1997; Takada *et al.*, 2002]. Third, since a large portion of the hypermethylated mammalian genome consists of transposable elements, such as L1 and Alu elements in human, which endanger genomic stability, it was suggested that DNA methylation might be a host defence mechanism of suppressing their activity [Yoder *et al.*, 1997].

Methylation at gene promoters

More than a half of the genes in mammalian genomes overlap with CpG islands [Jones, 2012]. A direct silencing-by-methylation model is apparently not applicable here, since the overwhelming majority of them are unmethylated in most of the somatic cells [Jones, 1999]. Nonetheless, the transcription at the overlapping start sites can be blocked irrespectively of their methylation status [Bestor *et al.*, 2015a], apparently involving other mechanisms such as histone modifications, binding of the Polycomb complex and alike. Some genes do have methy-

lated CpG islands overlapping with their TSS, however, most of them are the already described imprinted, X-chromosome and germline-specific genes [Jones, 2012].

The methylation at non-CGI promoters was reported to have a more direct relation to DNA methylation [Jones, 2012], although this has been put in question [Bestor *et al.*, 2015a]. Genome-wide analysis showed that genes with low CpG density at promoters show inverse correlation of their expression and methylation levels [Gal-Yam *et al.*, 2008]. This is of particular importance for certain tissue-specific genes, promoters of which are losing methylation only in defined cell types [Han *et al.*, 2011]. One speculated view that can explain the observed statistical relations between methylation and transcriptional activity is an epigenetic "lock" model, where DNA methylation is acting as a stabilizer of the inactive state established through other mechanisms [Jones, 2012].

Methylation of gene bodies and regulatory elements

Gene bodies, as regions generally having low CpG density, are methylated [Jones, 2012]. This might be necessary to suppress numerous repetitive elements. Moreover, it was also observed that methylation of the gene bodies is positively correlated to the transcriptional level of respective genes [Jones, 2012]. This might be linked to the necessity of suppressing alternative transcription initiation sites [Maunakea *et al.*, 2010] and is speculated to play a role in splicing [Laurent *et al.*, 2010; Maunakea *et al.*, 2010].

Methylation was found to have a particular pattern at genomic regulatory regions. Enhancers, which are key to the fine cell type-specific control of gene activity, were associated with low methylated regions (LMRs) [Stadler *et al.*, 2011]. The LMRs either indicate a very dynamic methylation of enhancers or the presence of several cell subpopulations. Finally, DNA methylation was implicated with altering the functional state of insulators which control the action of enhancers [Jones, 2012].

Summary: DNA methylation may not be the major regulator, but it is a reliable marker

It is now apparent that the direct silencing of genes by *de novo* methylation, suggested a part of the epigenetic memory hypothesis, is not a universal mechanism of gene regulation [Jones, 2012]. Although the association with repression is strong in phenomena as X chromosome inactivation and imprinting, subsequent research has revealed numerous exceptions and counterexamples to the simple model of mechanistic gene deactivation [Schübeler, 2015]. The very question whether DNA methylation plays any causal role in regulation or is merely an indicator mark which faithfully follows other driving regulatory mechanisms remains a matter of a fierce scientific debate [Bestor *et al.*, 2015a,b; Ngo and Sheppard, 2015; Wilkinson, 2015]. Nonetheless, even if the latter is true this does not diminish the importance of DNA methylation mapping. DNA methylation may not be instructive in the differentiation process, but it provides a reliable record of the current functional state of a cell reflecting cell type, subtype or population which is of utter importance for many applications [Schübeler, 2015].

1.1.6 Association with human disease

Regardless of its causality, the strong association with gene activity and regulation automatically implies that DNA methylation can be changed when the gene function is distorted as a result of certain environmental influences as well as pathological conditions. Below the major results of DNA methylation research in the context of diseases and environmental exposures

are reviewed. The concept of DNA methylation association studies, which are the primary instrument of this research field, is introduced at the end.

Imprinting disorders and other "monogenic" diseases

Overwhelming evidence about epigenetic diseases was obtained in single-gene disorders affecting the imprinted genes. Beckwith-Wiedemann syndrome characterized by growth abnormalities at birth is the most well-known example of such disorder. It was shown that the affected individuals have DNA methylation defects in the imprinted cluster which includes *H19/IGF2*, *SLC22A1*, *LIT1* and several other genes [Feinberg, 2007]. Several other similar disorders were described (Prader-Willy, Angelman, PHPIA syndromes) which are associated with lesions at the *SNURF-SNRPN/UBE3A* imprinted locus [Horsthemke and Buiting, 2006]. In addition to the imprinting disorders, several diseases are known that are directly linked to mutations in epigenetic machinery proteins. For instance, mutations in *MeCP2* gene are associated with Rett syndrome, while ICF (immunodeficiency, centromeric instability, facial anomalies) is apparently caused by mutations of *DNMT3B* [Feinberg, 2007].

Cancer

Cancer cells have widespread perturbations of their epigenomes, involving drastic changes of the DNA methylation landscape [Feinberg, 2007]. A widely accepted model associates the malignant transformation with *global hypomethylation* and *locus-specific hypermethylation* [Esteller, 2007].

Cancer-related hypomethylation [Feinberg and Vogelstein, 1983] was shown to play a role in erroneous activation of tissue-specific CGI-promoters, normally expressed only in defined tissues [Feinberg and Tycko, 2004]. Famous examples include the testicle-specific *MAGE* and *CAGE*, hypomethylated and expressed in melanoma and digestive tract malignancies, respectively. Furthermore, since DNA methylation is a known mechanism of transposone silencing, widespread hypomethylation is a hallmark of a global or focused genomic instability associated with severe chromosomal abnormalities in several cancers [Feinberg, 2007].

More recently discovered focal hypermethylation is known to occur at specific promoters, many of which are of the tumor-suppressor genes [Feinberg and Tycko, 2004]. Retinoblastoma gene (*RB*) was the first well-described and proved example of a tumor-suppressor which is hypermethylated in the corresponding cancer type [Greger *et al.*, 1989; Sakai *et al.*, 1991]. Since then the cancer-attributed hypermethylation was described at many other loci, such as *p16(INK4)*, *CDKN2A*, *VHL*, *MLH1* etc.

Subsequent research revealed that the involvement of epigenetic phenomena might be fundamentally linked to the malignant transformation. It was discovered that many types of cancer are associated with mutations in one or more key players [Plass *et al.*, 2013]. For instance, *TET2* mutations were detected in multiple blood cancers, while *DNMT3A* is frequently affected in acute myeloid leukemia [Schübeler, 2015].

Despite the stunning aberrations cancer cell methylomes maintain the signature of the cell type they originated from [Chen *et al.*, 2016]. In the future this should enable the tissue-of-origin detection based on the profiling of a tumor sample particularly useful in te analysis of metastasis [Heyn and Esteller, 2012].

The initial success in finding associations between DNA methylation and cancer provoked large scientific efforts aimed at comprehensive characterization of the cancer methylome. DNA methylation is one of the focuses in such large research consortia as The Cancer Genome

Atlas (TCGA) [Noushmehr *et al.*, 2010] and the International Cancer Genome Consortium (ICGC) [Hudson *et al.*, 2010].

Complex diseases and environmental influences, EWAS

DNA methylation changes were a matter of research in other common diseases, such as cardiovascular disorders, metabolic syndrome, autoimmunity, neurodegeneration etc [Michels *et al.*, 2013]. Most of the investigations are performed in a form of *epigenome-wide association studies* (EWAS) analogously to the genome-wide ones (GWAS) [Rakyan *et al.*, 2011]. As a rule, EWAS use one of the available genome-scale technologies to profile methylation in affected individuals and unaffected control subjects. EWAS aim to detect single CpGs or complete loci (commonly referred to as *biomarkers*) that are statistically associated with the phenotype or exposure. In order to guarantee the statistical soundness and biological reproducibility EWAS are required to fulfill a number of technical, design-related and reporting standards, thoroughly reviewed elsewhere [Heijmans and Mill, 2012; Michels *et al.*, 2013; Rakyan *et al.*, 2011].

So far, the results of several hundred DNA methylation EWAS were published, which vary widely in the target phenotype or environmental factor, study design (case-control cohorts, monozygotic twins, family trios), sampled tissue or cell type and the used profiling technology [Michels *et al.*, 2013]. The non-cancer EWAS predominantly use whole blood as a DNA source since the affected tissue is either unknown or difficult to sample.

Among the plethora of EWAS there are a few success stories. Several potentially causal DNA methylation variants were detected in a large study for rheumatoid arthritis [Liu *et al.*, 2013]. Based on an EWAS in the post-mortem brain tissue *ANK1* and several other genes were found to be significantly associated with the Alzheimer's disease [Lunnon *et al.*, 2014]. Furthermore, EWAS approach was successful in identifying the influence of environmental effects such as smoking, diet or exposure to potentially dangerous substances and physical factors. For instance, in a well-designed EWAS the smoking status was credibly associated with DNA methylation changes at *F2RL* locus [Breitling *et al.*, 2011].

One particular environmental exposure has direct relation to the present thesis. In the context of the metabolic syndrome it was hypothesized that the so called fetal programming, i.e. predetermination of the metabolic patterns *in utero*, may play a role. DNA methylation is one of the speculated mechanism behind such programming [Heijmans *et al.*, 2008]. An EWAS study presented in Chapter 2 tests this hypothesis in a cohort of monozygotic and, hence, genetically identical twins severely discordant for birth weight.

1.1.7 Dimensions of the DNA methylome variability

A brief review of DNA methylation given above can be summarized as a diagram in Figure 1.3. One can stratify three major directions of DNA methylation variability in a fixed species. The first dimension is the characteristic non-random distribution of DNA methylation across the genome, which is usually understood as the dynamic DNA methylation landscape. The second dimension is the variability between cells of the same organism, determined by cell linage, type and population. The third important coordinate is represented by differences between organisms in a population, driven by genetic factors, gender, age and disease-related environmental factors. The temporal coordinate reflecting DNA methylation changes with time is convoluted into each of the thee dimensions above. The biological function of methylation can be seen as a resulting vector in these three variability dimensions. A typical DNA methylation study cannot capture the complete variability space and can be compared with a lower-dimensional

plane providing for a cross-sectional view. The task of the subsequent computational analysis is largely aimed at reconstructing the complete picture based on an achievable projection. The experimental and computational methodology behind the DNA methylation analysis is introduced below.

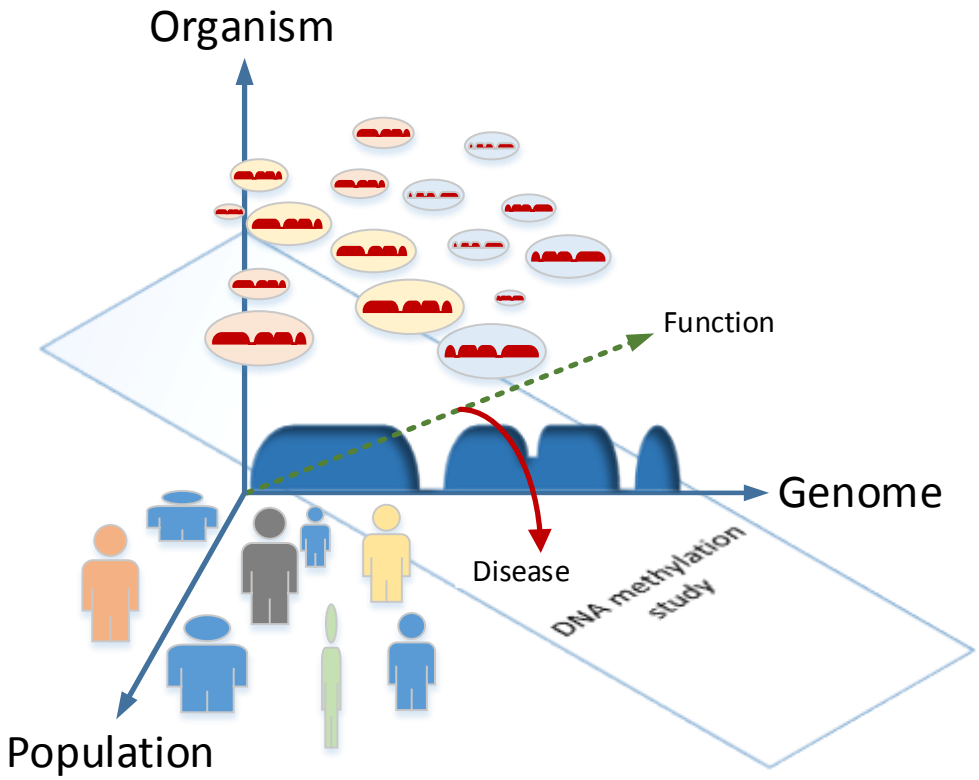


Figure 1.3: Dimensions of the DNA methylation variability. The genomic, intra-organismal and populational components are represented by coordinates of a three-dimensional variability space. The basis is orthonormal for the sake of demonstration, as in reality the above components are clearly not independent. The profile along the genomic coordinate represents a typical DNA methylation landscape with an overall high methylation level but featuring focally absent or decreased methylation at CpG islands and regulatory elements. Cells of different types (color coded) carrying type-specific variants of this landscape are representing the intra-organismal variability dimension. Finally, the individual variability of the global and cell type-specific landscapes, supplemented by genetic, gender-specific and age-attributed effects is visualized by human figures along the "Population" coordinate. The resulting cellular function of a cell type- and individual-specific methylome can be seen as a (dotted green) vector in this three-dimensional space representing a combination of all three variability components. The two-dimensional plane represents a cross-sectional approach behind a typical DNA methylation study that tries to stratify certain variability sources of interest and diminish the irrelevant ones (allegorized by the angle of the plane).

1.2 Profiling DNA methylation

The current view of DNA methylation, briefly outlined above, is a result of an immense progress in the methodology over the last several decades [Harrison and Parle-McDermott, 2011]. In this period the throughput and resolution of the profiling methods grew from bulk measurements of total 5-methylcytosine in a biological specimen to the single-CpG and single-molecule resolution maps available at the moment. This section first gives a brief retrospective overview of early approaches, then describes the bisulfite method that topped the early development efforts, and finally introduces two classes of high-throughput bisulfite-based profiling strategies of great relevance for the present work. The section is concluded by an overview of the technology-specific and more general data analysis aspects.

1.2.1 Overview of early approaches

First methods for DNA methylation analysis did not produce profiles of any kind, but allowed bulk measurements of relative 5-methylcytosine content using various types of chromatography [Bestor *et al.*, 1984; Kuo *et al.*, 1980], radiolabeling [Wu *et al.*, 1993] and immunolabeling [Wu *et al.*, 1993]. Later the gene-specific techniques came of age, based on the action of methylation-sensitive restriction enzymes combined with radiolabeling and subsequent thin-layer smears [Cedar *et al.*, 1979] or Southern blotting. This method was later extended to a genome-wide setting and became known as restriction landmark genomic scanning (RLGS) [Hatada *et al.*, 1991; Hayashizaki *et al.*, 1993; Kawai *et al.*, 1993]. The antibody-based methods enjoyed intensive development in subsequent years. Immunolabeling combined with fluorescent microscopy has become a workhorse method in developmental biology [Mayer *et al.*, 2000; Oakeley *et al.*, 1997; Santos *et al.*, 2002] facilitating the studying of DNA methylation on the cell-to-cell basis, and is widely used till today. With the boost of microarray and next-generation sequencing technologies, immunoprecipitation of methylated DNA formed a basis of several high-throughput methods [Keshet *et al.*, 2006; Weber *et al.*, 2005]. Despite the significant progress, all the above approaches had limited capacity for gene-specific and especially genome-wide studies and were superseded by the methods based on the bisulfite conversion.

1.2.2 The bisulfite method

Sulfonation of certain pyrimidines by sodium bisulfite was known long before it was applied to DNA methylation profiling [Hayatsu *et al.*, 1970]. It was later shown that this reaction had different kinetics for cytosine and 5-methylcytosine [Wang *et al.*, 1980]. Finally, Frommer *et al.* demonstrated that these differences could be used to study DNA methylation patterns using sequencing [Frommer *et al.*, 1992].

The idea of the bisulfite method is to transform the initial methylation mark into a base-change detectable by a variety of existing methods. The bisulfite conversion is a three-step procedure schematically illustrated in Figure 1.4. Its final result is a conversion of an unmethylated cytosine to a uracil. Effectively, when applied to a pool of DNA molecules the bisulfite conversion leads to a substitution of the unmethylated cytosines with uracils, while the methylated ones remain cytosines.

Bisulfite-induced base changes can be read out using a number of methods. Early non-sequencing approaches were predominantly adaptations of well established genetic analysis methods, and included methylation-specific (MS-) variants of PCR [Herman *et al.*, 1996], single

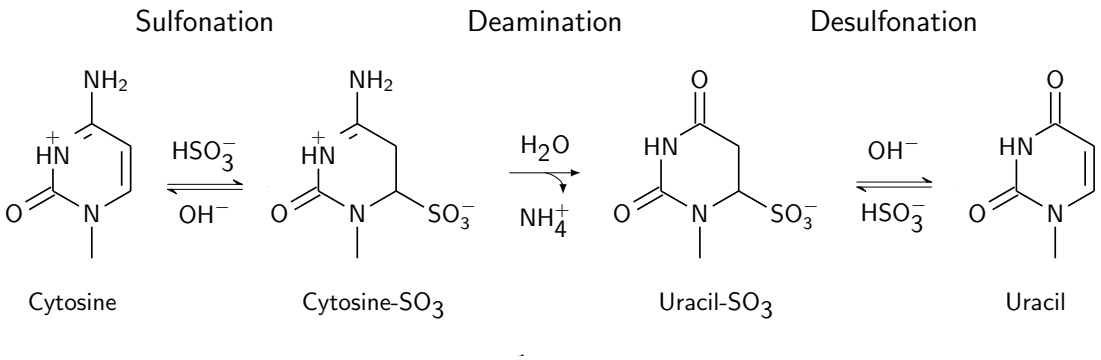


Figure 1.4: Bisulfite conversion of cytosine to uracil.

nucleotide primer extension (MS-SNuPE) [Gonzalgo and Jones, 1997], single-strand conformation analysis (MS-SSCA) [Bianco *et al.*, 1999]. Such methods as, for instance, the combined bisulfite restriction analysis (COBRA) [Xiong and Laird, 1997] make use of the emergence or loss of restriction cutting sites to determine the methylation state of a CpG position. Finally, a number of more exotic methods was developed, for instance, the high-resolution melting curve analysis [Wojdacz and Dobrovic, 2007]. Nevertheless, sequencing methods reviewed below have been proven as the most efficient way to read out the bisulfite-based methylation information.

A modified treatment detects oxidative methylcytosine derivatives

Oxidative forms have different reactivity in the ordinary bisulfite conversion. 5-hydroxymethylcytosine stays predominantly unconverted, while 5-formyl- and 5-carboxylcytosine are deaminated to uracil [Huang *et al.*, 2010; Nestor *et al.*, 2010]. In other terms, the ordinary bisulfite readout corresponds to a bulk measurement of 5-methyl- and 5-hydroxymethylcytosines. Newly invented techniques enable detection of such modifications by combining ordinary bisulfite with a modified treatment which has a different outcome. For instance, oxidative bisulfite protocol applies a soft oxidizing agent $KRuO_4$ to convert 5-hydroxy- to 5-formylcytosine [Booth *et al.*, 2012]. The bisulfite readout after the oxidative step results in a 5-methylcytosine profile with a single-basepair resolution. By comparing the readouts with and without the oxidative step one can estimate a bulk amount of 5-hydroxymethylcytosine at each CpG. Similar techniques were developed to enable estimation of 5-hydroxymethylcytosine as well as other oxidative modifications [Ito *et al.*, 2011; Song *et al.*, 2013]. More detailed background about these methods, their capacity and limitations is given in Chapter 5.

Problems of the bisulfite method

It is important to understand that bisulfite conversion, just like any other chemical reaction, can never guarantee a 100% yield. In the context of methylation state calling, both errors are possible, – *underconversion*, when unmethylated cytosines are not converted to uracils, and *overconversion*, when methylated cytosines are deaminated to uracils [Genereux *et al.*, 2008]. Proper consideration of the inevitable conversion errors is an important part of the bisulfite-based data analysis.

Another important problem is that bisulfite treatment is causing strand breaks in DNA [Mun-

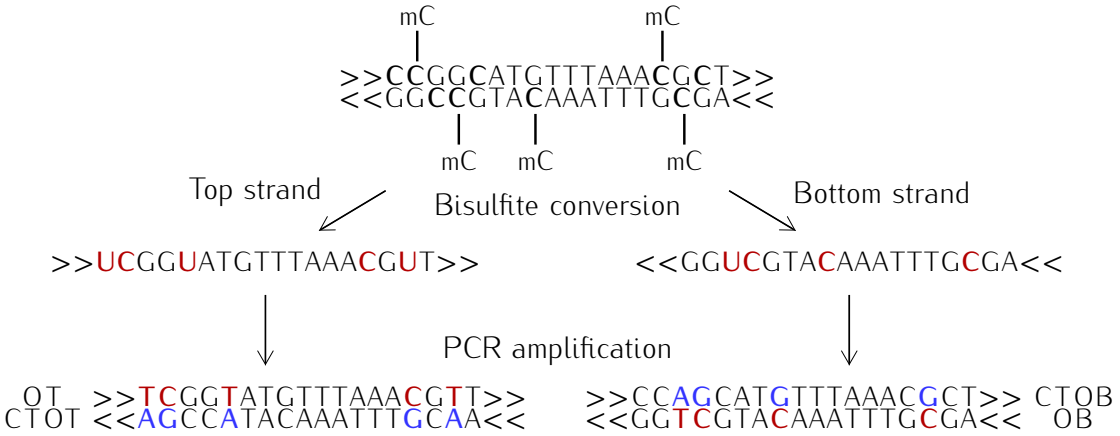


Figure 1.5: General scheme of bisulfite sequencing. Each of the four strands at the bottom can give rise to a sequence read. OT, original top strand; OB, original bottom strand; CTOT, complement of the original top strand; CTOB, complement of the original bottom strand. The figure was adapted from [Krueger *et al.*, 2012]

son *et al.*, 2007]. This decreases the applicability of bisulfite-based methods to low-input samples and in certain protocols requiring that the DNA fragments remain intact after the treatment (e.g. if they were preliminary ligated to adaptors).

1.2.3 Bisulfite sequencing

Sequencing was originally the first method to read out the methylation signal from the bisulfite converted DNA. Along with the bisulfite treatment a subsequent PCR amplification was introduced [Clark *et al.*, 1994; Frommer *et al.*, 1992] which is currently an integrative step of all bisulfite sequencing protocols. In accordance with the base-pairing rules the uracils, substituting the previously unmethylated cytosines, are replaced by thymines in the PCR (Figure 1.5).

Later, bisulfite converted DNA was sequenced using Sanger machines in such assays as direct sequencing [Eckhardt *et al.*, 2006; Rakyan *et al.*, 2004], giving a quantitative readout per CpG, and clonal bisulfite sequencing, generating full methylomes for a representative sample of several dozen DNA molecules. First bioinformatic solutions were developed, e.g. ESME, optimizing the signal processing of electropherograms [Lewin *et al.*, 2004], and BiQ Analyzer, automating the laborious and time-consuming analysis of the clonal sequences [Bock *et al.*, 2005].

Next-generation sequencing (NGS) revolutionized the field enabling DNA methylation profiling on an earlier unseen scale and resolution [Laird, 2010]. The line of method development bifurcated at the trade-off between the genome *coverage*, i.e. the number of sequenced base-pairs of the genomic reference and the sequencing *depth*, i.e. the average number of times a given genomic base-pair is covered by sequence reads. Locus-specific approaches utilize the large amounts of reads to achieve very high sequencing depth at selected regions of interest. Genome-scale methods have similar depth to the pre-NGS approaches, but reach coverage at a large portion of the genome, up to covering its complete mappable part (whole-genome methods).

High-throughput locus-specific bisulfite sequencing

There are several ways to prepare a bisulfite sequencing library for a locus of interest. The most common one is to design a pair of PCR primers hybridizing to the bisulfite converted DNA and amplify a short fragment (amplicon) containing the target CpG sites. Primer design should be performed with certain precautions, e.g. the primer sequences should not contain any underlying CpGs, and can be aided by specialized tools [Tusnády *et al.*, 2005]. The amplicon-based approach is single-stranded by design, and when information from both strands is desired they can be amplified after adding a so called hairpin linker to one of the double stranded fragment ends [Laird *et al.*, 2004].

Massively parallel pyrosequencing [Margulies *et al.*, 2005], commercially available as Roche 454 platform, was applied to sequence amplicon-based libraries. A pilot study in normal blood cells and several types of leukemia [Taylor *et al.*, 2007] demonstrated the power of the method, allowing for simultaneous analysis of 25 CpG-rich regions in over 40 samples. On average over 1600 sequence reads were generated for each case which is two orders of magnitude more compared to the clonal bisulfite sequencing. This supplied the researchers with a detailed picture of methylation pattern heterogeneity in multicellular samples. This approach was further validated and improved in several follow up studies [Gries *et al.*, 2013; Korshunova *et al.*, 2008; Varley *et al.*, 2009]. Later, Illumina introduced MiSeq thereby increasing the throughput of the locus-specific studies by an order of magnitude. The numbers of reads are now reaching tens of thousands per sample-locus pair, challenging the currently available specialized software packages.

Genome-scale bisulfite sequencing

An umbrella of genome-scale methods covers the protocols in which a sequencing library covers a significant portion of the genome, and the included CpGs are from multiple distant loci. Whole-genome bisulfite sequencing (WGBS) is the most generic approach, when a prepared library is not biased towards any underlying region type. Several other methods decrease the sequencing burden by enriching the library for certain kind of regions. Reduced-representation bisulfite sequencing (RRBS) creates preference for the CpG-dense regions, predominantly CpG islands, by the use of restriction enzymes with a CpG dinucleotide in their cutting motif, e.g. MspI [Meissner *et al.*, 2005]. More recent capture-based approaches [Li *et al.*, 2015] enrich the library for target regions using the pools of immobilized oligonucleotides, e.g. bound to magnetic beads. The hybridized target DNA is pooled out and sequenced.

First full methylome was obtained by WGBS of a plant organism. In 2008 a complete genome-wide methylation pattern of *Arabidopsis thaliana* was obtained [Cokus *et al.*, 2008]. As a first genome-scale effort in mammals, a mouse RRBS library was sequenced on a Genome Analyzer machine [Meissner *et al.*, 2008]. Shortly thereafter the first complete and single-base resolution human methylomes were published by Lister and colleagues [Lister *et al.*, 2009]. This pilot study was followed by numerous other reports delivering the first complete methylomes for various human cell lines and tissues [Laurent *et al.*, 2010; Ziller *et al.*, 2013]. Modified bisulfite conversion protocols, such as OxBis-Seq and TAB-Seq, were also used in combination with high-throughput sequencing to map oxidative modifications on a genome-wide scale [Booth *et al.*, 2012; Ito *et al.*, 2011; Song *et al.*, 2013].

Limitations and further progress

Bisulfite sequencing methods made an enormous contribution to the current knowledge of DNA methylation. It is responsible for many breakthroughs in understanding of DNA methylation landscapes and dynamics described in Section 1.1. However, a number of serious drawbacks limits their wider application keeping the alternative approaches in the feasibility range.

One serious limitation is the cost, with the price of a complete WGBS experiment is still at the mark of 3000 Euro. This is the main reason why to date WGBS was mainly used in global epigenomic mapping efforts aiming at reference methylomes such as ENCODE [Bernstein *et al.*, 2010], Roadmap [Roadmap Epigenomics Consortium *et al.*, 2015], BLUEPRINT [Abbott, 2011] and DEEP (`http://www.deutsches-epigenom-programm.de`). The price seriously limits the application of WGBS in EWAS using large cohorts of hundreds to thousands individuals. Due to its high resolution and low error rates, the comparatively cheaper deep locus-specific sequencing is a gold-standard method for the candidate-gene studies and EWAS verification. However, even here many studies decide for the obsolete, yet less costly methods.

The second limitation is the minimal amount of input material. Conventional WGBS protocols still require 10^5 cells to construct a reasonable library. This underlies the methylome heterogeneity problem, introduced in the Section 1.3 below. It is also tightly linked to an issue of over-amplification, inherent to all bisulfite sequencing protocols featuring a PCR step. The fewer cells are submitted to bisulfite treatment, the higher should be the rate of post-bisulfite PCR amplification to deliver enough material for the library preparation. As a result, the majority of generated sequence reads are in fact stemming from the very same original DNA fragments and are known as PCR *duplicates*.

Current progress in the development of genome-scale bisulfite sequencing methods is aimed at improving single-molecule resolution and decreasing the amount of the input material. Several steps have already been made in this direction. For instance, unique molecular identifiers (UMI), attached to the source DNA fragments prior to bisulfite treatment was suggested as a solution for the problem of PCR duplicates. Furthermore, a PCR-free protocol was suggested that involves post-bisulfite adapter tagging (PBAT) [Miura *et al.*, 2012]. The future development will most probably concentrate upon single-cell methods discussed below (Section 1.3).

1.2.4 DNA methylation microarrays

DNA methylation microarrays were developed from the earlier low-throughput methods based on restriction by endonucleases, immunoprecipitation and bisulfite treatment [Harrison and Parle-McDermott, 2011]. The most well-known restriction-based microarrays included HpaII tiny fragment enrichment by ligation-mediated PCR (HELP) [Khulan *et al.*, 2006] and comprehensive high-throughput arrays for relative methylation (CHARM) [Irizarry *et al.*, 2008]. Immunological methods gave rise to methylated DNA immunoprecipitation (MeDIP) [Weber *et al.*, 2005]. Several related affinity-purification methods based on the use of methyl-binding domain proteins (MBDs) were also suggested [Gebhard *et al.*, 2006; Schilling and Rehli, 2007]. Despite of several successful applications, the affinity-based microarrays were outperformed by the bisulfite-based ones due to the limited throughput, the lack of single-CpG resolution, low sensitivity and a strong CpG density-associated bias of the former [Down *et al.*, 2008; Rakyan *et al.*, 2008].

Bisulfite-based bead arrays from Illumina

Illumina's DNA methylation bead arrays are an adaptation of the previously existing genomic platforms used for the high-throughput SNP genotyping. In essence, the assays tracking genomic SNPs were redesigned to identify bisulfite-induced SNPs at CpG positions.

GoldenGate was the first platform of such kind [Bibikova *et al.*, 2006]. Just as its genotyping counterpart it was based on measuring the amount of the product after an allele-specific PCR and enabled simultaneous quantitative profiling of methylation state at 1536 CpG sites mapping to 371 genes. GoldenGate was succeeded by a more progressive Infinium technology which became the basis for three DNA methylation microarrays, HumanMethylation27 [Bibikova *et al.*, 2009], HumanMethylation450 [Bibikova *et al.*, 2011] and MethylationEPIC [Moran *et al.*, 2015]. The Infinium technology is a highly parallelized primer-extension assay, in which each extension reaction is targeting a single genomic CpG (Figure 1.6). The bisulfite converted genomic DNA is hybridized to 50-bp long oligonucleotides attached to nano-sized bead. The oligonucleotides, complementary to the fragments upstream or downstream of the target CpGs, serve as primers in the primer-extension reaction. Free nucleotides in the solution carry two different fluorescent labels (C and G are Cy3-labeled while A and T carry Cy5) allowing to register the incorporation. Currently two variants of the assay exist, denoted as type I and type II, which differ in the way the methylated and unmethylated state intensities are obtained.

HumanMethylation27 (Infinium 27k) platform contained only type I probes tagging 27,578 CpG sites across the human genome. HumanMethylation450, commonly known as Infinium 450k, was a significant move forward with 482,421 CpGs and 3,091 supposed non-CG methylation sites. The array genome-wise coverage reached 0.5% of the total CpG number. Approximately two third of the probes were of type II. The higher density of the array comprised a significant challenge for the oligonucleotide design and placement. Finally, MethylationEPIC platform succeeds 450k microarray bringing the total number of tagged CpGs to over 850,000 which is an astonishing 3% of the genomic total.

Application scope and approaching obsolescence

Combination of several qualities determined the success of the bead arrays. The cost and labor-optimized standard procedure favorably differed them from the genome-scale sequencing approaches. The lack of single-molecule resolution was compensated by a much higher, molecular support for each intensity measurement. While a methylation call in genome-scale sequencing methods is often based on a dozen reads, each intensity measurement relies upon thousands of template molecules hybridizing to a bead, which guarantees a smaller measurement error. Furthermore, unlike many cost-optimized sequencing protocols (RRBS/CapSEQ), microarrays consistently return high-quality calls for the majority of covered CpGs in most of the samples. Owing to these benefits bead arrays have become the most popular profiling method of EWAS [Michels *et al.*, 2013]. Only the Gene Expression Omnibus (GEO) database currently contains data series for almost 1,000 studies using Infinium 27k, 450k and EPIC combined.

Nevertheless, the remaining shortcomings of the bead arrays, along with the constant improvement of the competing approaches, set an applicability horizon for the microarray methods. The highly optimized Infinium procedure has limited potential for further cost reduction, and the only possible upgrade is the increase of the genomic coverage. Due to the probe-based design this becomes progressively difficult. Furthermore, the bead array protocol does not offer an easy way to decrease the amount of the input material, conserving the cell type

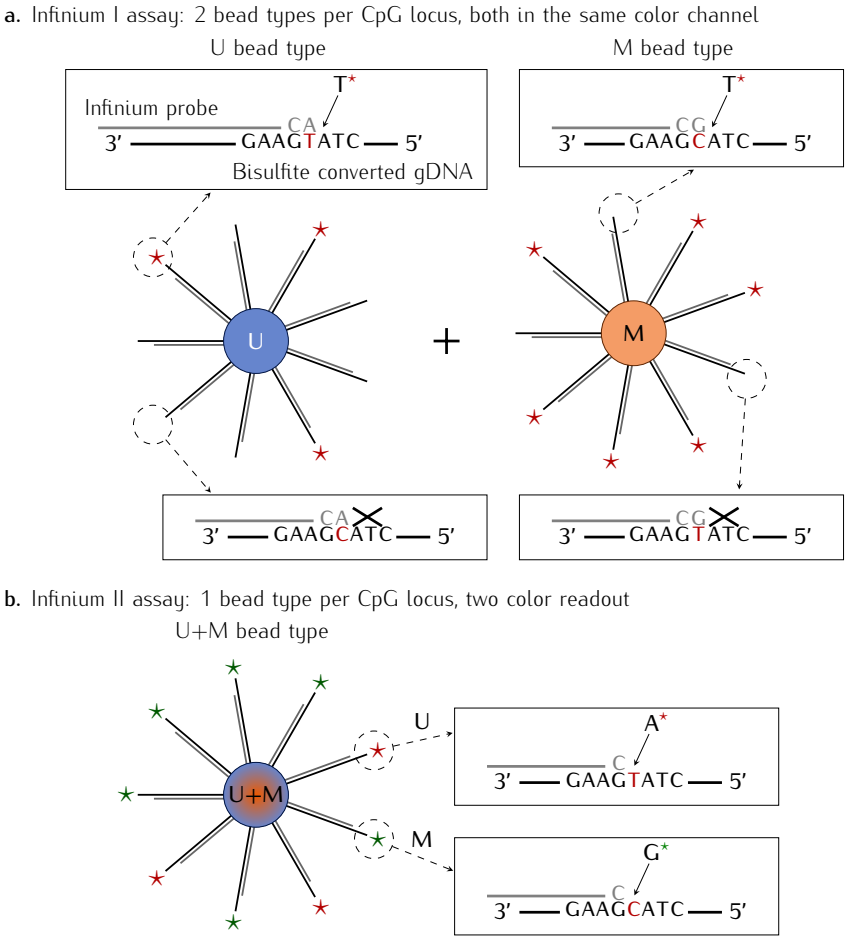


Figure 1.6: Principles of the Infinium methylation assays. **a**. The older type I assay (probe) features two different oligonucleotides per probed CpG, physically separated on two different beads. One of the oligonucleotides is fully complementary to the bisulfite converted sequence terminating with the target CpG guanine under the assumption that the target cytosine is methylated. Analogously, the other oligonucleotide is complementary to the bisulfite converted sequence assuming the target CpG is unmethylated. During the microarray procedure, the oligonucleotides are extended to incorporate the base directly upstream – for the top (Watson) strand – or downstream – for the bottom (Crick) strand – of the target CpG site. Since both oligonucleotides of the type I probe are physically separated and reside on different beads, the rate of the primer extension reaction can be registered in the same color channel. **b**. In the case of type II probes primer-extension reaction is incorporating the base right in the position of the target cytosine. Here the methylated and unmethylated fragments hybridize to the same physical bead and methylation state of the target CpG determines which base gets incorporated in a concrete extension reaction. A type II bead, thus, produces signal in both color channels which is registered as methylated (M) and unmethylated (U) signal intensities. The figure was adapted from [Dedeurwaerder *et al.*, 2011].

heterogeneity problem introduced later in this text. As a consequence, while the current text was written the manufacturer announced the upcoming obsolescence of the complete HiScan platform within the next two years, which concerns all the DNA methylation microarrays (an e-mail circulation). Their place will probably be occupied by more dynamic and adaptable sequencing methods.

1.2.5 Challenges of the DNA methylation data analysis

The wet-lab procedures reviewed above make up only a part of respective profiling methods. The other essential component includes the computational operations, from retrieval of the raw data down to an advanced DNA methylation statistics. This section describes specific problems associated with processing of data from bisulfite sequencing and DNA methylation microarrays. A general outline of the downstream analysis common to all types of DNA methylation data is sketched at the end.

Bisulfite sequencing data

Primary sequencing data is platform-specific and the initial data processing steps are performed by the software of an NGS instrument itself. The current consensus output format for the initially processed sequence reads is FASTQ [Cock *et al.*, 2010] that harbours per-base quality scores in addition to the sequence information. A typical data processing pipeline includes demultiplexing of the raw reads, read-level quality control (QC) and filtering, quality-based and adaptor-aimed trimming, alignment, post-alignment QC and methylation calling (Figure 1.7).

An important problem of practical applications is demultiplexing. Since the capacity of the sequencing runs is most often shared between different and unrelated sequencing libraries, the resulting bulk of sequences contains reads from all users, applications and sequenced samples. The possibility of assigning reads to a specific library is implemented either by utilizing the lanes which separate the space of a sequencing machine, or by incorporating sample-specific sequence tags, also known as multiplex identifiers (MIDs). Several publicly available tools enable read assignment based on short sequence matching algorithms [Blankenberg *et al.*, 2010].

The initial data processing steps are common to all NGS-based bisulfite sequencing methods and involve quality control of the sequencing experiment. FastQC from the Babraham Institute (`http://www.bioinformatics.babraham.ac.uk/projects/fastqc`) has proved itself as an extremely useful tool for visualization and primary analysis of quality scores. Base quality usually decreases in terminal parts of the reads. As a rule such low-quality ends are trimmed [Krueger *et al.*, 2012]. Trimming is also necessary to deal with spurious incorporation of cytosines, which appear as methylated in subsequent analysis, due to end repair, sequenced parts of the adaptor etc [Bock, 2012].

High-quality reads are aligned to the existing genomic reference. The alignment tasks are principally different between locus-specific and genome-scale methods. In the former case most of the reads are originating from a very small part of the genome. A typical problem of aligning up to 10^5 reads to a short reference sequence makes the use of globally optimal dynamic programming algorithms [Needleman and Wunsch, 1970; Waterman *et al.*, 1992] computationally feasible. In the case of genome-scale data, the size of the reference reaches the orders of 10^9 base-pairs, which is intractable by the memory-demanding globally optimal alignment algorithms. This is why heuristic approaches became very popular, the most successful of which are based on the fast searches in hashed data structures [Langmead *et al.*,

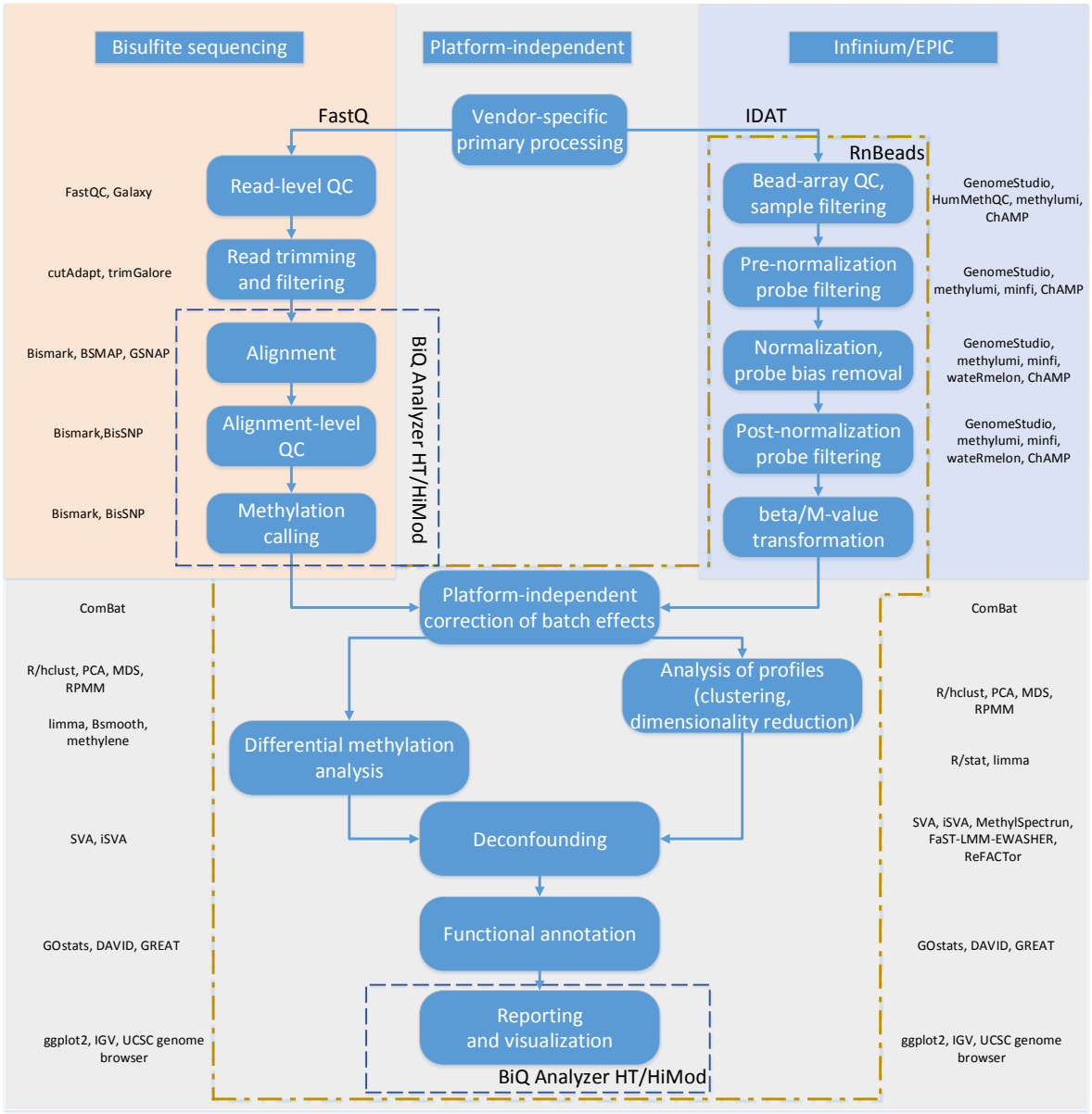


Figure 1.7: General scheme of DNA methylation data processing and analysis. Non-exhaustive lists of popular methods and bioinformatic tools are given for each case. The red dashed-dotted line is enclosing data analysis steps performed by RnBeads (Chapter 3). The blue dashed line demarcates the approximate domains of BiQ Analyzer HT (Chapter 4) and BiQ Analyzer HiMod (Chapter 5).

2009]. In both cases, special preparation of the reference sequence is necessary to account for strand diversion and the alphabet changes due to bisulfite conversion [Krueger *et al.*, 2012].

Some of the quality control operations can only be performed on the aligned reads. This concerns bisulfite conversion rate, which is usually estimated at non-CpG cytosines. Furthermore, alignment allows efficient detection of reads originating from the very same original DNA fragment, commonly known as PCR duplicates. In the paired-end sequencing of the genome scale libraries, out of two or more read pairs having identical starting and ending positions all except one are discarded.

The generated alignment is further on used for calling methylation states at each CpG position in each read. The accurate methylation calling is particularly important at low sequencing depth, and, therefore, the most critical in genome-scale approaches. To tackle this problem specialized callers were developed, which account for most of the possible problems [Liu *et al.*, 2012].

Due to a different computational load the development of bioinformatic tools took different pathways for locus-specific and genome-scale data. Generally smaller amounts of data in locus-specific experiments stimulated the development of complete bioinformatic solutions. These tools are, as a rule, interactive and GUI-based, supporting the complete data processing pipeline from raw reads to the final tables and diagrams with methylation data [Kumaki *et al.*, 2008; Rohde *et al.*, 2010]. Available third-party tools, as well as two software packages that are a contribution of the present thesis are introduced in more detail in Chapters 4 and 5. For the genome-scale methods overwhelming amounts of generated read data impedes the creation of such "turnkey" software. Here custom bioinformatic pipelines are common and consist of tools supporting one or several related steps [Chen *et al.*, 2010; Krueger and Andrews, 2011; Liu *et al.*, 2012]. The most reliable pipelines have now largely taken shape [Krueger *et al.*, 2012]. In due of this, primary processing of the genome-scale sequencing data is not covered by the bioinformatic solutions presented below and is out the scope of the present thesis.

Data of DNA methylation microarrays

Principles of data processing for DNA methylation microarrays were reviewed in detail multiple times elsewhere [Dedeurwaerder *et al.*, 2014; Marabita *et al.*, 2013; Siegmund, 2011; Wilhelm-Benartzi *et al.*, 2013]. A typical low-level pipeline includes data loading, quality control of samples and probes, intensity normalization and background correction, adjustment of the probe type bias and calculation of the methylation statistics. The main purpose of the short overview below is to highlight the challenges and unresolved issues in the context of the solutions that are part of the present thesis.

The primary raw data of the bead arrays are high-resolution images of the microarray plate made in two color channels. The image processing and analysis part is performed by the microarray HiScan platform itself, and the summarized intensities per bead type are stored in a specialized IDAT format [Bock, 2012]. There are several software packages which enable vendor-independent parsing of IDAT files and their import into popular statistical environments such as R [Smith *et al.*, 2013].

As any other microarray platform, Infinium/EPIC features a complex multi-step wet-lab protocol and is inevitably affected by a number of technical biases and noise sources [Dedeurwaerder *et al.*, 2011]. Quality control is facilitated by a panel of specialized control assays which are present on the bead array platform [Illumina Inc., 2014]. The control assays monitor the crucial steps of the procedure and report the intensities in an expected range. The biological samples which show low quality in the control assays should be removed. Although the

task may seem trivial, there are no published methods so far to automate the detection and removal of low-quality samples.

Normalization aims at reduction of technical biases by computational adjustment of the intensity data. First, beads at which no primer extension took place still produce a certain level of background intensity. The manufacturer suggested a simple background correction method making use of the specialized null probes on the bead array [Illumina Inc., 2014]. Later a more advanced method was developed which used the "out-of-band" intensities of type I probes for a more reliable estimation of background in each color channel [Triche Jr. *et al.*, 2013]. Although the bead array technology solved the issue of spatial biases within one array (sample), it was shown that the overall intensity strongly varies between the arrays on the same microarray plate, as well as between the plates [van Eijk *et al.*, 2012; Wilhelm-Benartzi *et al.*, 2013]. Due to large variability this can lead to spurious methylation differences between samples. The platform contains specialized normalization probes, which are used in an intensity scaling normalization developed by the manufacturer [Illumina Inc., 2014]. Subsequently, multiple methods were developed to adjust intensities both within and between the arrays [Fortin *et al.*, 2014; Pidsley *et al.*, 2013; Touleimat and Tost, 2012]. Finally, it was noticed that methylation calls from type I and II assays have a substantially different distribution [Dedeurwaerder *et al.*, 2011]. Although this may stem from the differences in genomic distribution of the two probe types, a number of approaches was developed to adjust these differences as well [Dedeurwaerder *et al.*, 2011; Makismovic *et al.*, 2012; Teschendorff *et al.*, 2013].

Technical confounding usually arises from a non-uniform processing of the samples and is commonly known as batch effects [Leek *et al.*, 2010]. The most extreme form of a batch effect occurs, for instance, when case and control samples in a simple EWAS study are processed independently in two separate batches. In this case any bias in methylation calls between the two batches will be interpreted as effects of interest in the subsequent association analysis and will render the detection of any true differences impossible. More subtle forms of technical confounding arise, for example, in the bead array-based studies. Bead arrays are known to be affected by an intensity bias systematically spreading across a microarray plate [van Eijk *et al.*, 2012]. Some of such effects can be solved by normalization methods. The remaining confounding can be corrected using the specialized methods, such as ComBat [Sun *et al.*, 2011], or the linear mixed models [Barfield *et al.*, 2012].

The normalized and corrected intensities can be used to generate the methylation calls. The bead array manufacturer recommended calculating methylation ratios, also known as a β-value by relating the methylated probe intensity to the total intensity of the methylated and unmethylated probes [Bibikova *et al.*, 2006]. Alternatively, a log-ratio of the methylated and unmethylated probe intensities can be used as a quantitative readout, known as M-value [Du *et al.*, 2010].

In summary, numerous bioinformatic tools exist, enabling one of the above data handling steps, as well as the complete streamlined analysis of the Infinium/EPIC data (reviewed in more detail in Chapter 3). One of the biggest challenges in primary processing of the bead array data is the absence of consensus about the preferred method or combination of methods [Dedeurwaerder *et al.*, 2014]. Multiple comparison meta-studies [Dedeurwaerder *et al.*, 2014; Marabita *et al.*, 2013] apply various benchmarking criteria which results in divided final conclusions. Furthermore, most of the meta-studies are published by one of the method developing groups which inevitably leads to a biased reporting. In this situation a pipeline allowing for a seamless and rapid comparison of the existing methods would greatly simplify the real life applications and facilitate an unbiased benchmarking analysis. This was the main motivation behind the

modules for processing of bead array data in RnBeads presented in Chapter 3.

Challenges of the platform-independent downstream analysis

Regardless of any specific profiling method, high-level DNA methylation data are at certain point represented as data matrices of methylation calls for each CpG in each profiled sample. Downstream high-level analysis of such data is generally similar to that of other genomic data types. The multitude of the analysis strategies can be roughly split into a hypothesis-driven association analysis, hypothesis-free exploratory analysis of the profiles and functional annotation [Wilhelm-Benartzi *et al.*, 2013].

Association analysis, also known as analysis of differential methylation, is central for the comparative studies such as EWAS. The goal is to test a family of hypotheses whether a certain fixed or observed factor is significantly associated with DNA methylation changes at particular CpGs or loci. The analysis is usually performed by fitting linear models to the data of each CpG [Bock, 2012]. The model coefficients of the target phenotype variable are tested for significance which results in a set of P-values. The details of the association analysis were surveyed in numerous publications [Siegmund, 2011; Wilhelm-Benartzi *et al.*, 2013].

Another group of analyses aims predominantly at investigating the observed methylation profiles in their entirety and includes such popular techniques as clustering, dimensionality reduction and factor analysis. The most popular clustering approach is hierarchical clustering, although more advanced methods exist specifically tailored to DNA methylation data, e.g. recursively partitioning mixture model [Houseman *et al.*, 2008]. Dimensionality reduction techniques allow for improved data visualization and discovery of variability components. Principal Component Analysis (PCA) is an exceptionally powerful generic method extensively applied to DNA methylation profiles. Alternative approaches for lower dimensional representation include, among other, Multidimensional Scaling (MDS), and a more recent t-Distributed Stochastic Neighbor Embedding (t-SNE) [van der Maaten and Hinton, 2008].

Both kinds of analysis intersect at the issue of confounding. Confounding occurs when strong factors that affect DNA methylation at a large portion of CpGs are biasing the association analysis and lead to inflation of significance testing. When information about these factors is known in advance, e.g. for an EWAS cohort age and gender of the individuals, microarray plate, processing batch, they can be included into linear modeling. When the confounding factors are not known in full, they can be discovered using the profile analysis methods such as PCA or clustering. Furthermore, Surrogate Variable Analysis (SVA) [Leek and Storey, 2007] and its extension independent SVA (iSVA) [Teschendorff *et al.*, 2011] are two specialized statistical methods specifically developed for deconfounding an association analysis. Finally, one specific confounding problem stemming from the cell type-specificity of DNA methylation profiles is in focus of the present thesis. This problem as well as the experimental and computational methods used to address it are introduced in the concluding Section 1.3.

The high-level analysis of DNA methylation data is remarkably versatile, since it heavily depends on such factors as study goal and design, presence or absence of specific batch effects and confounders, availability of additional data e.g. expression profiles etc. Consequently, it is often performed in a general-purpose statistical or spreadsheet environment such as R, SAS, SPSS or Excel in a highly customized way. Nevertheless, there are a number of computational tools which implement the complete analysis pipeline. One such bioinformatic solution, the R package RnBeads is in part contributed by the current thesis and described in Chapter 3.

1.3 Tackling heterogeneity of the DNA methylomes

1.3.1 Problem definition

The common feature of the mainstream DNA methylation profiling methods reviewed in Section 1.2 is that they are applied to macroscopic samples containing thousands to millions cells. Given the possibility of the same genomic CpG being methylated differently between the profiled cells, the resulting readout will be a superposition of the underlying single-cell CpG states yielding a quantitative signal.

As outlined in Section 1.1 cell methylomes can be different due to multiple reasons. In a healthy organism the largest determinant of the methylation landscape is cell type, affecting the methylation level at numerous CpG positions across the genome [Lam *et al.*, 2012]. In case a sample of a heterogeneous tissue is submitted to any of the analysis methods above, the resulting profile will strongly depend on the proportions of different cell types present in this tissue sample. This notion can be propagated further down to smaller levels. Even if a relatively homogeneous sample is profiled, believed to contain cells of only one well defined cell type, the abundance of stable cell populations will have a major influence upon methylation values of the remaining variable CpGs. More generally, the proportions at which stable methylation signatures, e.g. the cell type methylomes, are present in a profiled cell sample are always convoluted into the average methylome. This is what is usually understood under the cell type (cell population) heterogeneity of DNA methylation profiles. The latter has to be discriminated from other types of heterogeneity. For instance, temporal heterogeneity arises during transition between two stable methylation signatures. If the cells in profiled samples are caught at different stages of this transition, this will result in a quantitative signal at the affected CpGs.

The question whether to define the cell type heterogeneity as a problem depends on the goal of the DNA methylation profiling. Indeed, when the primary objective is to detect cell populations carrying a characteristic methylome, the cell population-attributed heterogeneity is precisely the cause of the sought DNA methylation differences and cannot be seen as a purely deteriorating effect. In this context one can rather speak about characterizing and understanding the methylome heterogeneity. This point of view motivates the methylome deconvolution results presented in Chapter 6 and will be introduced in detail at the end of the current section.

The widespread negative connotation was introduced by the association studies where the cell type composition is usually not the primary analysis target. On the contrary, due to its uncontrolled differences in the compared samples, it comes with a risk of losing DNA methylation effects of interest. Such confounding can be particularly harmful in EWAS, where the statistically justified increase of the cohort sizes imposes limits on sampling unification [Adalsteinsson *et al.*, 2012; Heijmans and Mill, 2012; Liang and Cookson, 2014].

A comprehensive theory of the heterogeneity effects in EWAS was proposed by Houseman *et al.* [Houseman *et al.*, 2012, 2015] (thoroughly reviewed in Ref. [Houseman, 2015]). It was suggested that the association between phenotypic and technical covariates and the observed methylation measurements could be modeled as a sum of "direct" and cell type-mediated effects. The "direct" effects are assumed to influence the methylome in cell type-independent manner, causing changes in all cell types. In contrast, the cell type-mediated effects are considered to be primarily associated with the changes in cell type composition. Due to large differences between the cell type-specific methylomes the changes in composition are translating into the phenotype-associated variability of the measured mixture methylomes. Delineating the direct and cell-type mediated effects is one of the primary goals in DNA methylation

EWAS.

There are several fundamentally different strategies for dealing with heterogeneity in DNA methylation data. First, when standard profiling methods have to be used, one can decrease heterogeneity by cell enrichment or isolation. Second, heterogeneity can be translated into differences between methylomes by profiling very small cell samples or even single cells. Finally, one can try to delineate heterogeneity effects computationally. The remainder of the current section outlines the current state-of-the-art for each of the three approaches.

1.3.2 Cell separation

Most of the cell types are well defined by their morphological or biochemical properties. The cellular dimensions, volume, shape, nuclear morphology are tightly linked to the cell function, and are, therefore, intrinsic characteristics of cell types which can be used for their identification [Gautam and Bhadauria, 2014]. Moreover, cell types are characterized by unique constellation of surface marker proteins, such as the cluster of differentiation (CD) antigens [Engel *et al.*, 2015]. These properties can be used to enrich the samples for certain cell type prior to a DNA methylation analysis, and thereby increase the homogeneity of the resulting methylomes. Cell separation and isolation methods based upon adherence, density and antibody labeling are designed to fulfill this aim [Tomlinson *et al.*, 2013]. While the adherence-based methods are very rough and generally not aimed at the isolation of pure cell subpopulations, the density- and, especially, antibody-based ones are of particular importance for DNA methylation profiling.

Density-based methods have developed from earlier sedimentation approaches, separating cells by difference of their deposition speed in a solution. The most common instance is centrifugation in a density gradient widely applied to separate whole blood into fraction of granulocytes and peripheral blood mononuclear cells (PBMCs) including monocytes, macrophages and lymphocytes. This can be exploited by DNA methylation studies using blood which are not interested in one of the two fractions [Lam *et al.*, 2012]. In most of the cases, however, density centrifugation is used as a pre-enrichment step for the downstream methods for cell isolation [Tomlinson *et al.*, 2013].

In case surface proteins are well defined for the target cell type or population, antibodies can be used to selectively label such cells. Antibodies usually carry additional functional groups enabling subsequent isolation steps. In fluorescence-activated cell sorting (FACS) [Bonner *et al.*, 1972], the antibodies carry a fluorescent dye. The labeled cells are passing through a thin stream and get electrically charged depending on their fluorescence. An electrostatic deflection system is then used to deviate cell-containing droplets into one or another direction and this way collects the labeled cells. In magnetic-assisted cell sorting (MACS) the antibodies are attached to magnetic beads which enable the pullout of bound cells with the magnetic field [Miltenyi *et al.*, 1990; Rembaum *et al.*, 1982].

Other methods exist for cell separation such as laser-capture micro-dissection [Emmert-Buck *et al.*, 1996]. The most modern lab-on-chip approaches utilize microfluidic devices for fine cell sorting based on additional physical properties, such as volume and shape. They are, however, still in early development stages [Tomlinson *et al.*, 2013], and are hardly applicable for large-scale DNA methylation profiling.

The cell separation methods have been successfully used in combination with DNA methylation profiling. It was a method of choice for large epigenome consortia, allowing to profile methylation in a wide range of cell types [Abbott, 2011; Bernstein *et al.*, 2012; Roadmap Epigenomics Consortium *et al.*, 2015]. MACS separation allowed generation of high quality ref-

erence methylomes of various blood cell types [Calvanese *et al.*, 2012; Reinius *et al.*, 2012]. FACS-based sorting of brain cell nuclei antibody-labeled for RBFOX3 (NeuN) nuclear surface protein was used to separate neuronal and glial fractions, significantly increasing the resolution of the DNA methylation studies in brain [Guintivano *et al.*, 2013; Lister *et al.*, 2013].

The potential of cell separation for the future epigenome studies is limited by several pitfalls [Tomlinson *et al.*, 2013]. The immunolabeling relies upon the existence of a well defined cellular markers, which may not exist or, what is more often, may not be exclusive. In due of this difficulty the purity achievable by the labelling-based methods is always a relative notion being limited to the characteristic of carrying a specific surface marker. The purity is further undermined by formation of cell clusters containing both labeled and non-labeled cells, and non-specific antibody binding. Furthermore, the amounts of cells recovered by FACS and MACS might not be sufficient to perform a DNA methylation analysis. Although, DNA methylation is a comparatively stable epigenetic mark, one cannot completely exclude the possibility that the prolonged times out of the natural environment, sample handling as well as interactions during the isolation procedure, for instance, the antibody binding to an receptor, might cause certain methylome alterations. Finally, the methods are often still too laborious and costly to be used in the context of large-scale studies such as EWAS.

1.3.3 Single-cell methods

One possibility to increase the homogeneity of a cell sample is to decrease the sample itself. This involves the downscaling of the DNA methylation profiling method to much smaller amounts of input materials down to a single cell. Comparative stability of the DNA methylation mark is very favorable and makes it easier to profile methylomes in low amounts of cells [Schwartzman and Tanay, 2015].

In certain cases the standard profiling methods can be applied to low amounts of input without major adaptation. This is, for example, the case when targeted bisulfite sequencing is performed on repetitive elements, which was successfully applied to study DNA methylation dynamics in early development stages, such as zygote or even pro-nuclei prior to the fusion [Arand *et al.*, 2012]. Furthermore, any bisulfite sequencing protocol is in essence a single-molecule procedure and a picture for a selected locus reflects the distribution in the analyzed pool of cells. Computational methods were used to draw more insights about methylome dynamics based on this information [Landan *et al.*, 2012; Siegmund *et al.*, 2009].

The truly single-cell methods extend this approach to the complete genome. So far, three single-cell DNA methylation assays were published, all being adaptations of the existing bisulfite sequencing protocols [Schwartzman and Tanay, 2015]. First, a single-tube RRBS protocol was described that allowed profiling of early mouse embryos and ES cells at single-cell resolution [Guo *et al.*, 2013]. Next, PBAT protocol was adapted to single cells [Smallwood *et al.*, 2014]. As the latest development, another bisulfite sequencing protocol was used for low-depth methylome profiling of developmental and drug-induced effects [Farlik *et al.*, 2015].

While having achieved a significant progress in recent years, low-input and single-cell methods are in early development stage. A major drawback of the maps obtained by all the above methods is their sparsity. In the best cases total genomic coverage reaches at most 20% of all CpGs with mean being around 10% [Schwartzman and Tanay, 2015]. This seriously complicates a comparative analysis since, depending on the number of sequenced cells, there might be only a few ones having a call for a particular CpG. Scaling of the protocols to more cells is also difficult in this case, as 20 million reads are typically required to achieve the above genomic coverage, meaning that only around 80 cells can be profiled in one typical two-flowcell

Table 1.1: Computational methods for the correction of cell type heterogeneity

Name	Source	Predecessor	Implementation
Reference-based			
Constrained projection	[Houseman *et al.*, 2012]		R scripts
CETS	[Guintivano *et al.*, 2013]		R package
Reference-free			
RefFreeEWAS	[Houseman *et al.*, 2014]	SVA	R package
FaST-LMM EWASHER	[Zou *et al.*, 2014]	FaST-LMM	R package, python
ReFACTor	[Rahmani *et al.*, 2016]	PCA	R package, python

HiSeq 2500 run. At the moment this is prohibitively expensive to be used for large studies such as EWAS. The future development will, most likely, come from the novel (third-generation) sequencing technologies. The latter will enable profiling of large, potentially chromosome-sized DNA fragments and will natively discriminate the modified bases [Munroe and Harris, 2010].

1.3.4 Computational inference and deconvolution

The advantages of cell separation-based and single-cell methods for decreasing and delineating heterogeneity of the methylomes are doubtless. They are, however, outweighed by inherent drawbacks limiting their utility in many application scenarios. In the same time, an average methylome of a cell sample with a decent depth of support carries information about all cell populations which contributed to it. As a convenient, cost-efficient and time-saving alternative, it was suggested that computational methods can be used to detect and measure tissue heterogeneity of complex samples [Baron *et al.*, 2006; Sehouli *et al.*, 2011]. Earlier this has been successfully demonstrated for the gene expression data [Kuhn *et al.*, 2011; Shen-Orr *et al.*, 2010].

The later developed computational methods for methylome heterogeneity analysis are usually classified in two large groups depending on whether or not they are using any prior information. Reference-based methods utilize the existing, usually genome-scale profiles of purified cell populations. Reference-free methods attempt to infer the heterogeneity effects without any prior information. Existing reference-free and reference-based approaches are summarized in Table 1.1.

Methods for reference-based estimation and adjustment

Houseman *et al.* suggested a regression calibration-like approach to estimate the contribution of each of the two effects for every tested covariate [Houseman *et al.*, 2012]. Their model relies upon reference methylome measurements which were directly used to find cell type-specific quantitative markers and to model the target data using the values for the markers observed in the reference data set. In addition, they devised a systematic quadratic optimization-based procedure, under the name of constrained projection, that allowed estimation of leukocyte proportions in genome-scale DNA methylation profiles of whole blood samples. The estimated proportions could then be considered as covariates in a subsequent association analysis [Houseman *et al.*, 2012]. The approach by Houseman *et al.* was validated in several follow-up studies [Accomando *et al.*, 2014; Koestler *et al.*, 2013].

Similar reference-based approaches were developed focusing at heterogeneity in the brain tissue. Cell epigenotype-specific (CETS) model is using DNA methylation profiles of neuronal-

enriched and depleted fractions for quantifying the proportion of neuronal and non-neuronal (glial) cells in bulk samples of various regions in brain cortex [Guintivano *et al.*, 2013]. CETS mixes measured profiles of $NeuN^+$ and $NeuN^-$ enrichment fractions *in silico* with an incremental range of possible mixing proportions to simulate brain tissue methylomes for each neurons to glia ratio. The observed brain profile is then compared to all simulated mixtures, and the mixing proportions of the best fitting one are considered to represent the proportions of the observed profile. Another study extended the blood-based constrained projection to support the brain tissue data and showed how accurate estimates can be obtained using data from different brain regions instead of the NeuN-sorted references [Montaño *et al.*, 2013].

An important aspect of the reference-based methods is the selection of cell type-specific quantitative marker CpGs. The pilot study in blood applied per-CpG linear modeling of reference data with cell type as an independent variable and selected the top 500 CpGs showing the best F-statistic [Houseman *et al.*, 2012]. Later implementation of the constrained projection in the `minfi` R-package used a slightly different approach, selecting 100 best markers for each cell type [Aryee *et al.*, 2014]. Finally, a recent paper presents an in-depth systematic method IDOL for addressing this problem, and demonstrates its superiority [Koestler *et al.*, 2016].

At the moment, most of the available reference-based methods use the reference data also in their estimation steps. This hinges upon two significant assumptions. First, it is expected that technical differences between the target and the reference data sets are not affecting the analysis substantially. Second, they assume that the cell type methylomes of the reference individuals are, on average, equivalent to those of the target cohort. While the first requirement can, at least in part, be solved by joint preprocessing of both data sets, the second assumption is very hard to verify in practice. This is why it appears more reliable to use reference data set to select ctDMRs, and use the methylation values observed at these loci or CpGs as a basis for estimation or adjustment. An *ad hoc* implementation of such approach for the correction of cell type heterogeneity in saliva samples [Souren *et al.*, 2013] is presented as a part of Chapter 2.

Reference-free correction methods

Reference-based methods showed good performance in multiple data sets [Koestler *et al.*, 2013; Liu *et al.*, 2013]. However, they largely rely upon the availability of reference DNA methylation profiles of purified cell types which are not always at hand. In addition, the complete set of contributing epigenome varieties is not known *a priori*. This stimulated the search for methods that would be independent of such reference measurements.

First two such methods appeared simultaneously [Houseman *et al.*, 2014; Zou *et al.*, 2014]. RefFreeEWAS [Houseman *et al.*, 2014], from the authors of the constrained projection method [Houseman *et al.*, 2012], is essentially an adaptation of the more general Surrogate Variable Analysis (SVA) [Leek and Storey, 2007]. The SVA-inspired assumption behind RefFreeEWAS is that the cell type methylomes and proportions are convoluted into both coefficients and residuals of the phenotype-based model of the observed data. Although not obtainable explicitly, due to orthogonality with other covariates, one can recover the cell type methylomes and proportions in a convoluted form using joint Singular Value Decomposition of the coefficients and residuals. This also allows for respective correction of the phenotype model coefficients to capture only the "direct" methylation effects.

The second method, FaST-LMM-EWASHER [Zou *et al.*, 2014] is a descendant of the FaST-LMM algorithm for the significance analysis in genomic association studies. In a thorough mathematical analysis the authors expose limitations of the standard linear mixed models (LMMs) accounting for covariance structure in capturing the confounding caused by the cell

type variability, as well as of a simple principal component-based correction. The EWASHER model is based on augmenting the full covariance-aware LMM with correction for top principal components of the CpG marker matrix. The fitting algorithm proceeds in iterative fashion re-selecting the features for the covariance matrix after addition of each new principal component. The authors demonstrate that EWASHER successfully eliminates the significance inflation in the statistical analysis of DNA methylation data severely affected by the cell type heterogeneity.

Finally, the most recent reference-free approach ReFACTor [Rahmani *et al.*, 2016] is applying sparse PCA for the same purpose of correcting an association analysis for the phenotype of interest. The idea behind ReFACTor is to find a small number t of cell type-specific markers as CpGs which have the lowest distance between their actual observed data and a low-rank approximation of the observed data based on k principal components, where k is the assumed number of underlying cell types. The PCA of the data subset on t CpGs is then supposed to yield scores correlated to the cell type proportions. The PCA scores can subsequently be used as covariates for the adjustment. In the original publication the authors demonstrate the slight superiority of ReFACTor as compared to the earlier published methods RefFreeEWAS and EWASHER.

The available reference-free methods are provably helpful for the analysis of EWAS data affected by strong cell type heterogeneity effects. Purely from the end user perspective the methods return heterogeneity-corrected P-values for significance of association with the target phenotypic variable of interest at each CpG position. Although, this is often a desired analysis outcome, DNA methylation data sets tend to be increasingly more complex and, as a rule, do not meet all prior expectations about the tested hypothesis and the confounding factors. The internal structure of the sampled methylomes often provides for more interesting findings than the starting hypothesis itself. Routine application of such correction methods may impede an unconstrained data exploration and important insights about the analysed data set might be missed. Furthermore, from the technical perspective most of the reference-free methods are explicitly or implicitly based on PCA. PCA is a general-purpose method and the variability components it captures are not guaranteed to exclusively reflect the variability cell type composition. As a result the above methods are at risk of over-correcting the data for other unknown, yet potentially interesting factors.

Methylome deconvolution

The reference-based and reference-free methods above, apart from their specific advantages and disadvantages, were developed with the primary goal of correcting the statistical analysis in DNA methylation association studies. In general, computational methods for heterogeneity analysis significantly improve tractability of the data obtained on multi-cellular specimens. They require no additional costs and minimal labor overhead while providing for very useful inference results. However, if one abstracts oneself from the specific issue of confounding, and looks at methylome heterogeneity from a more general perspective, one can define a related deconvolution problem. The latter can be formulated as follows: given a set of cell sample average methylomes find all unique underlying methylation signatures as well as their mixing proportions.

This deconvolution problem is closely related to the goals pursued by the methods for DNA methylation profile analysis, i.e. the clustering and dimensionality reduction methods briefly reviewed at the end of Section 1.2. An ideal method for solving it should be able to delineate the observed variability into “biological”, i.e. corresponding to the real DNA methylation changes, and “technical”, i.e. caused by other effects of non-biological nature, such as batch effects in

sample handling and calling procedure, measurement noise etc.

Full deconvolution was earlier attempted on gene expression data [Gaujoux and Seoighe, 2012, 2013; Repsilber *et al.*, 2010]. The promising results obtained with linear methods, such as standard non-negative matrix factorization (NMF), in gene expression are, at the very least, surprising. Unlike DNA methylation, the expression level is a quantitative signal already at the level of a single cell meaning that the mixing of the expression signals in a multi-cellular sample does not have to obey a linear (proportional) mode. An analogous problem in DNA methylation should be much easier to solve due to its discreteness at the single-cell level and consecutive mixing linearity.

In Chapter 6 we present a computational method which attempts to solve this problem in an unsupervised, reference-free manner. We show that, although this problem is computational hard and resides on the verge of mathematical tractability, the main deconvolution goals can nevertheless be achieved.

1.4 Outline

The sections above give a rough overview of DNA methylation, the major dimensions of its variability, the most powerful experimental methods for mapping the latter, and introduce some of the data analysis problems, specifically the problem of heterogeneous methylome measurements. There are deep mutual relations between each of these fundamental and practical aspects. The high non-uniformity of the DNA methylation landscape is characterized by drastic changes of methylation states between neighboring CpG at transitions from an island-overlapping promoter into a gene body, through the CpG-sparse intergenic regions to lowly methylated regulatory elements. This non-uniformity makes it necessary to study DNA methylation at single-CpG resolution. The variability of the landscapes across different cell types within one organism compromises the bulk DNA methylation measurements in large cell samples, especially those coming from heterogeneous tissues, and calls either for cell isolation-based and low-input profiling strategies or for proper statistical correction and interpretation of average methylomes. Variability of the methylomes across human populations is an access point for studying the involvement of DNA methylation into pathological conditions and fixation of various environmental influences, but it also confounds such studies due to the large influence of such factors as genetic information, gender and ageing. This confounding together with technical issues of particular profiling technologies necessitate the creation of powerful yet comprehensive data processing and analysis tools. The contributions presented in the remainder of the present thesis aim to satisfy these needs with bioinformatic tools and novel computational methods.

Chapter 2 sets the stage for the subsequent bioinformatic results, and provides an example of a typical genome-scale DNA methylation study. The chapter presents an EWAS in monozygotic twins discordant for intra-uterine growth. Out of performance and cost considerations the study embarked upon the Infinium 450k arrays to generate high quality average methylomes for 17 female twin pairs. The EWAS demonstrates that even highly controlled settings, such as monozygotic twin-based study design which the genetic, age, gender and many other types of confounding reduces to an absolute minimum, is not insured against other types of undesired variability effects. In this case a seemingly homogeneous DNA source, the neutrophil-containing saliva, was nevertheless contaminated with uncontrolled amounts of epithelium. A simple yet efficient method we developed for the adjustment of the consequent cell type heterogeneity effects is an important result for the current thesis also presented in Chapter 2. The negative result of the study was confirmed by a locus-specific analysis of the potential

candidates and the cell type-specific markers. Taken together, the experience collected while working on the EWAS formed a basis for bioinformatic and computational results of the subsequent chapters.

Chapter 3 presents RnBeads, a pipeline for processing and analysis of genome-scale DNA methylation profiles similar to those of the twin-based EWAS. RnBeads supports both the bead array and WGBS/RRBS data and leads the user from minimally preprocessed input files to a holistic representation of the study data. RnBeads integrates and simplifies application of the state-of-the-art computational methods for dealing with the data analysis problems including but not limited to those presented in Chapter 2.

Chapters 4 and Chapter 5 is devoted to locus-specific analysis of 5-methylcytosine using high-throughput bisulfite sequencing. These approaches sacrifice genomic coverage to obtain very good representation of the variability between cells, and are usually applied in candidate gene studies as well as for the verification analyses of the genome-scale studies, the way it was done in the twin-based EWAS described in Chapter 2. Chapter 4 describes an interactive software package BiQ Analyzer HT specifically developed to handle such data. As pointed out in the introduction, one limitation of the bisulfite method is its inability to discriminate the oxidative methylation varieties. In Chapter 5 the software is extended to support the 5-methylcytosine derivatives: 5-hydroxy-, 5-formyl- and 5-carboxylcytosine.

After facing the cell type heterogeneity problem for the first time, a lot of thinking was invested to understand it better from many different angles. Chapter 6 presents the core result for the DNA methylation heterogeneity problem: a methylome deconvolution method MeDeCom. The concept behind MeDeCom extends beyond the level of statistical correction and tries to approach the heterogeneity from a constructive data-driven perspective. The pilot results of MeDeCom in synthetic and real DNA methylation data sets presented in Chapter 6 are demonstrating the power of such an approach and provide clues for overcoming the current limitations.

The thesis is concluded by an overarching discussion and outlook in Chapter 7. The discussion first puts all the bioinformatic results in a common context of a large-scale DNA methylation study, and then gives in-depth considerations about the methylome deconvolution problem. The outlook sketches a broader picture of the anticipated developments in the field and the author's opinion about the respective role of bioinformatics researchers in making them possible.

References

Abbott, A. Europe to map the human epigenome. *Nature*, 477(7366):518, 2011.

Accomando, W. P., Wiencke, J. K., Houseman, E. A., Nelson, H. H., and Kelsey, K. T. Quantitative reconstruction of leukocyte subsets using DNA methylation. *Genome Biology*, 15(3):R50, 2014.

Adalsteinsson, B. T., Gudnason, H., Aspelund, T., Harris, T. B., Launer, L. J., Eiriksdottir, G. *et al.* Heterogeneity in white blood cells has potential to confound DNA methylation measurements. *PloS one*, 7(10):e46705, 2012.

Allis, C. D., Jenuwein, T., and Reinberg, D. *Epigenetics*. CSHL Press, 2007.

Arand, J., Spieler, D., Karius, T., Branco, M. R., Meilinger, D., Meissner, A. *et al.* In vivo control of CpG and non-CpG DNA methylation by DNA methyltransferases. *PLoS Genetics*, 8(6), 2012.

Arber, W. and Dussoix, D. Host specificity of DNA produced by Escherichia coli. *Journal of Molecular Biology*, 5(1):18–36, 1962.

Aryee, M. J., Jaffe, A. E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A. P., Hansen, K. D. *et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*, 2014.

Barfield, R. T., Kilaru, V., Smith, A. K., and Conneely, K. N. CpGassoc: an R function for analysis of DNA methylation microarray data. *Bioinformatics*, 28(9):1280–1281, 2012.

Baron, U., Türbachova, I., Hellwag, A., Eckhardt, F., Berlin, K., Hoffmuller, U. *et al.* DNA methylation analysis as a tool for cell typing. *Epigenetics*, 1(1):55–60, 2006.

Bartolomei, M. S., Webber, A. L., Brunkow, M. E., and Tilghman, S. M. Epigenetic mechanisms underlying the imprinting of the mouse H19 gene. *Genes & development*, 7(9):1663–73, 1993.

Berdyshev, G. D., Korotaev, G. K., Boiarskikh, G. V., and Vaniushin, B. F. [Nucleotide composition of DNA and RNA from somatic tissues of humpback and its changes during spawning]. *Biokhimiia (Moscow, Russia)*, 32(5):988–93, 1967.

Bernstein, B. E., Birney, E., Dunham, I., Green, E. D., Gunter, C., and Snyder, M. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, 2012.

Bernstein, B. E., Stamatoyannopoulos, J. a., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nature Biotechnology*, 28(10):1045–1048, 2010.

Bestor, T., Laudano, A., Mattaliano, R., and Ingram, V. Cloning and sequencing of a cDNA encoding DNA methyltransferase of mouse cells. *Journal of Molecular Biology*, 203(4):971–983, 1988.

Bestor, T. H., Edwards, J. R., and Boulard, M. Notes on the role of dynamic DNA methylation in mammalian development. *Proceedings of the National Academy of Sciences of the United States of America*, 112(22):6796–9, 2015a.

Bestor, T. H., Edwards, J. R., and Boulard, M. Reply to Wilkinson: Minor role of programmed methylation and demethylation in mammalian development. *Proceedings of the National Academy of Sciences of the United States of America*, 112(17):E2117, 2015b.

Bestor, T. H., Hellewell, S. B., and Ingram, V. M. Differentiation of two mouse cell lines is associated with hypomethylation of their genomes. *Molecular and cellular biology*, 4(9):1800–6, 1984.

Bestor, T. H. and Ingram, V. M. Two DNA methyltransferases from murine erythroleukemia cells: purification, sequence specificity, and mode of interaction with DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 80(18):5559–63, 1983.

Bianco, T., Hussey, D., and Dobrovic, A. Methylation-sensitive, single-strand conformation analysis (MS-SSCA): A rapid method to screen for and analyze methylation. *Human mutation*, 14(4):289–93, 1999.

Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J. M. *et al.* High density DNA methylation array with single CpG site resolution. *Genomics*, 98(4):288–295, 2011.

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

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

Bird, A., Taggart, M., Frommer, M., Miller, O. J., and Macleod, D. A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. *Cell*, 40(1):91–9, 1985.

Bird, A. P. Use of restriction enzymes to study eukaryotic DNA methylation: II. The symmetry of methylated sites supports semi-conservative copying of the methylation pattern. *Journal of molecular biology*, 118(1):49–60, 1978.

Bird, A. P. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Research*, 8(7):1499–1504, 1980.

Bird, A. P. CpG-rich islands and the function of DNA methylation. *Nature*, 321(6067):209–13, 1986.

Bird, A. P. and Taggart, M. H. Variable patterns of total DNA and rDNA methylation in animals. *Nucleic Acids Research*, 8(7):1485–1497, 1980.

Blankenberg, D., Kuster, G. V., Coraor, N., Ananda, G., Lazarus, R., Mangan, M. *et al.* Galaxy: A web-based genome analysis tool for experimentalists. *Current Protocols in Molecular Biology*, Chapter 19(SUPPL. 89):Unit 19 10 1–21, 2010.

Bock, C. Analysing and interpreting DNA methylation data. *Nature Reviews Genetics*, 13(10):705–719, 2012.

Bock, C., Reither, S., Mikeska, T., Paulsen, M., Walter, J., and Lengauer, T. BiQ Analyzer: Visualization and quality control for DNA methylation data from bisulfite sequencing. *Bioinformatics*, 21(21):4067–4068, 2005.

Bock, C., Walter, J., Paulsen, M., and Lengauer, T. Inter-individual variation of DNA methylation and its implications for large-scale epigenome mapping. *Nucleic acids research*, 36(10):e55, 2008.

Bocklandt, S., Lin, W., Sehl, M. E., Sánchez, F. J., Sinsheimer, J. S., Horvath, S. *et al.* Epigenetic predictor of age. *PloS one*, 6(6):e14821, 2011.

Boks, M. P., Derks, E. M., Weisenberger, D. J., Strengman, E., Janson, E., Sommer, I. E. *et al.* The relationship of DNA methylation with age, gender and genotype in twins and healthy controls. *PloS one*, 4(8):e6767, 2009.

Bonner, W. a., Hulett, H. R., Sweet, R. G., and Herzenberg, L. a. Fluorescence activated cell sorting. *Review of Scientific Instruments*, 43(3):404–409, 1972.

Booth, M. J., Branco, M. R., Ficz, G., Oxley, D., Krueger, F., Reik, W. *et al.* Quantitative Sequencing of 5-Methylcytosine and 5-Hydroxymethylcytosine at Single-Base Resolution. *Science*, 336(6083):934–937, 2012.

Breitling, L. P., Yang, R., Korn, B., Burwinkel, B., and Brenner, H. Tobacco-smoking-related differential DNA methylation: 27K discovery and replication. *American journal of human genetics*, 88(4):450–7, 2011.

Calvanese, V., Fernández, A. F., Urdinguio, R. G., Suárez-Alvarez, B., Mangas, C., Pérez-García, V. *et al.* A promoter DNA demethylation landscape of human hematopoietic differentiation. *Nucleic Acids Research*, 40(1):116–131, 2012.

Capuano, F., Mülleder, M., Kok, R., Blom, H. J., and Ralser, M. Cytosine DNA methylation is found in Drosophila melanogaster but absent in Saccharomyces cerevisiae, Schizosaccharomyces pombe, and other yeast species. *Analytical chemistry*, 86(8):3697–702, 2014.

Cedar, H., Solage, A., Glaser, G., and Razin, A. Direct detection of methylated cytosine in DNA by use of the restriction enzyme MspI. *Nucleic acids research*, 6(6):2125–32, 1979.

Chen, P.-Y., Cokus, S. J., and Pellegrini, M. BS Seeker: precise mapping for bisulfite sequencing. *BMC bioinformatics*, 11(1):203, 2010.

Chen, Y., Breeze, C. E., Zhen, S., Beck, S., and Teschendorff, A. E. Tissue-independent and tissue-specific patterns of DNA methylation alteration in cancer. *Epigenetics & chromatin*, 9(1):10, 2016.

Christman, J. K., Price, P., Pedrinan, L., and Acs, G. Correlation between hypomethylation of DNA and expression of globin genes in Friend erythroleukemia cells. *European journal of biochemistry / FEBS*, 81(1):53–61, 1977.

Clark, S. J., Harrison, J., Paul, C. L., and Frommer, M. High sensitivity mapping of methylated cytosines. *Nucleic acids research*, 22(15):2990–7, 1994.

Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L., and Rice, P. M. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic acids research*, 38(6):1767–71, 2010.

Cokus, S. J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C. D. *et al.* Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature*, 452(7184):215–9, 2008.

Cooper, D. N., Taggart, M. H., and Bird, A. P. Unmethylated domains in vertebrate DNA. *Nucleic acids research*, 11(3):647–58, 1983.

Coulondre, C., Miller, J. H., Farabaugh, P. J., and Gilbert, W. Molecular basis of base substitution hotspots in Escherichia coli. *Nature*, 274(5673):775–80, 1978.

Dedeurwaerder, S., Defrance, M., Bizet, M., Calonne, E., Bontempi, G., and Fuks, F. A comprehensive overview of Infinium HumanMethylation450 data processing. *Briefings in bioinformatics*, 15(6):929–41, 2014.

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

Delatte, B., Deplus, R., and Fuks, F. Playing TETris with DNA modifications. *The EMBO journal*, 33(11):1198–211, 2014.

Desrosiers, R. C., Mulder, C., and Fleckenstein, B. Methylation of Herpesvirus saimiri DNA in lymphoid tumor cell lines. *Proceedings of the National Academy of Sciences of the United States of America*, 76(8):3839–43, 1979.

Doscocil, J., Sorm, F., Doskočil, J., and Šorm, F. Distribution of 5-methylcytosine in pyrimidine sequences of deoxyribonucleic acids. *Biochimica et Biophysica Acta (BBA) - Specialized Section on Nucleic Acids and Related Subjects*, 55(6):953–959, 1962.

Down, T. A., Rakyan, V. K., Turner, D. J., Flicek, P., Li, H., Kulesha, E. *et al.* A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nat Biotechnol*, 26(7):779–785, 2008.

Du, P., Zhang, X., Huang, C.-C., Jafari, N., Kibbe, W. a., Hou, L. *et al.* Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC bioinformatics*, 11:587, 2010.

Dunn, D. B. and Smith, J. D. Occurrence of a New Base in the Deoxyribonucleic Acid of a Strain of Bacterium Coli. *Nature*, 175(4451):336–337, 1955.

Eckhardt, F., Lewin, J., Cortese, R., Rakyan, V. K., Attwood, J., Burger, M. *et al.* DNA methylation profiling of human chromosomes 6, 20 and 22. *Nature genetics*, 38(12):1378–1385, 2006.

Ehrlich, M., Gama-Sosa, M. A., Huang, L.-H., Midgett, R. M., Kuo, K. C., McCune, R. A. *et al.* Amount and distribution of 5-methylcytosine in human DNA from different types of tissues or cells. *Nucleic Acids Research*, 10(8):2709–2721, 1982.

El-Maarri, O., Becker, T., Junen, J., Manzoor, S. S., Diaz-Lacava, A., Schwaab, R. *et al.* Gender specific differences in levels of DNA methylation at selected loci from human total blood: a tendency toward higher methylation levels in males. *Human genetics*, 122(5):505–14, 2007.

Emmert-Buck, M. R., Bonner, R. F., Smith, P. D., Chuaqui, R. F., Zhuang, Z., Goldstein, S. R. *et al.* Laser capture microdissection. *Science (New York, N.Y.)*, 274(5289):998–1001, 1996.

Engel, P., Boumsell, L., Balderas, R., Bensussan, A., Gattei, V., Horejsi, V. *et al.* CD Nomenclature 2015: Human Leukocyte Differentiation Antigen Workshops as a Driving Force in Immunology. *Journal of immunology (Baltimore, Md. : 1950)*, 195(10):4555–63, 2015.

Esteller, M. Cancer epigenomics: DNA methylomes and histone-modification maps. *Nature reviews. Genetics*, 8(4):286–298, 2007.

Farlik, M., Sheffield, N., Nuzzo, A., Datlinger, P., Schönegger, A., Klughammer, J. *et al.* Single-Cell DNA Methylome Sequencing and Bioinformatic Inference of Epigenomic Cell-State Dynamics. *Cell Reports*, 10(8):1386–97, 2015.

Feinberg, A. P. Phenotypic plasticity and the epigenetics of human disease. *Nature*, 447(7143):433–440, 2007.

Feinberg, A. P. and Tycko, B. The history of cancer epigenetics. *Nature Reviews. Cancer*, 4(2):143–53, 2004.

Feinberg, A. P. and Vogelstein, B. Hypomethylation distinguishes genes of some human cancers from their normal counterparts. *Nature*, 301(5895):89–92, 1983.

Feng, S., Jacobsen, S. E., and Reik, W. Epigenetic reprogramming in plant and animal development. *Science (New York, N.Y.)*, 330(6004):622–7, 2010.

Ferguson-Smith, A. C., Sasaki, H., Cattanach, B. M., and Surani, M. A. Parental-origin-specific epigenetic modification of the mouse H19 gene. *Nature*, 362(6422):751–5, 1993.

Ficz, G., Hore, T. a., Santos, F., Lee, H. J., Dean, W., Arand, J. *et al.* FGF signaling inhibition in ESCs drives rapid genome-wide demethylation to the epigenetic ground state of pluripotency. *Cell Stem Cell*, 13(3):351–359, 2013.

Fortin, J.-P., Labbe, A., Lemire, M., Zanke, B. W., Hudson, T. J., Fertig, E. J. *et al.* Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome Biology*, 15(12):503, 2014.

Frommer, M., McDonald, L. E., Millar, D. S., Collis, C. M., Watt, F., Grigg, G. W. *et al.* A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proceedings of the National Academy of Sciences of the United States of America*, 89(5):1827–1831, 1992.

Fuke, C., Shimabukuro, M., Petronis, A., Sugimoto, J., Oda, T., Miura, K. *et al.* Age related changes in 5-methylcytosine content in human peripheral leukocytes and placentas: an HPLC-based study. *Annals of human genetics*, 68(Pt 3):196–204, 2004.

Gal-Yam, E. N., Egger, G., Iniguez, L., Holster, H., Einarsson, S., Zhang, X. *et al.* Frequent switching of Polycomb repressive marks and DNA hypermethylation in the PC3 prostate cancer cell line. *Proceedings of the National Academy of Sciences of the United States of America*, 105(35):12979–12984, 2008.

Gardiner-Garden, M. and Frommer, M. CpG islands in vertebrate genomes. *Journal of molecular biology*, 196(2):261–82, 1987.

Gaujoux, R. and Seoighe, C. Semi-supervised Nonnegative Matrix Factorization for gene expression deconvolution: a case study. *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases*, 12(5):913–21, 2012.

Gaujoux, R. and Seoighe, C. CellMix: a comprehensive toolbox for gene expression deconvolution. *Bioinformatics (Oxford, England)*, 29(17):2211–2, 2013.

Gautam, A. and Bhadauria, H. Classification of white blood cells based on morphological features. In *Advances in Computing, Communications and Informatics (ICACCI, 2014 International Conference on*, pages 2363–2368. 2014.

Gebhard, C., Schwarzfischer, L., Pham, T. H., Andreesen, R., Mackensen, A., and Rehli, M. Rapid and sensitive detection of CpG-methylation using methyl-binding (MB)-PCR. *Nucleic Acids Research*, 34(11):e82, 2006.

Geiman, T. M. and Muegge, K. DNA methylation in early development. *Molecular reproduction and development*, 77(2):105–13, 2010.

Genereux, D. P., Johnson, W. C., Burden, A. F., Stöger, R., and Laird, C. D. Errors in the bisulfite conversion of DNA: modulating inappropriate- and failed-conversion frequencies. *Nucleic acids research*, 36(22):e150, 2008.

Gonzalgo, M. L. and Jones, P. A. Rapid quantitation of methylation differences at specific sites using methylation-sensitive single nucleotide primer extension (Ms-SNuPE). *Nucleic acids research*, 25(12):2529–31, 1997.

Greger, V., Passarge, E., H\"opping, W., Messmer, E., and Horsthemke, B. Epigenetic changes may contribute to the formation and spontaneous regression of retinoblastoma. *Human Genetics*, 83(2):155–158, 1989.

Gries, J., Schumacher, D., Arand, J., Lutsik, P., Markelova, M. R., Fichtner, I. *et al.* Bi-PROF: Bisulfite profiling of target regions using 454 GS FLX Titanium technology. *Epigenetics*, 8(7):765–771, 2013.

Gruenbaum, Y., Cedar, H., and Razin, A. Restriction enzyme digestion of hemimethylated DNA. *Nucleic acids research*, 9(11):2509–15, 1981.

Gruenbaum, Y., Cedar, H., and Razin, A. Substrate and sequence specificity of a eukaryotic DNA methylase. *Nature*, 295(5850):620–2, 1982.

Guintivano, J., Aryee, M. J., and Kaminsky, Z. a. A cell epigenotype specific model for the correction of brain cellular heterogeneity bias and its application to age, brain region and major depression. *Epigenetics*, 8(3):290–302, 2013.

Guo, H., Zhu, P., Wu, X., Li, X., Wen, L., and Tang, F. Single-Cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. *Genome Research*, 23(12):2126–2135, 2013.

Guo, J. U., Su, Y., Zhong, C., Ming, G.-l., and Song, H. Hydroxylation of 5-methylcytosine by TET1 promotes active DNA demethylation in the adult brain. *Cell*, 145(3):423–34, 2011.

Gutierrez-Arcelus, M., Lappalainen, T., Montgomery, S. B., Buil, A., Ongen, H., Yurovsky, A. *et al.* Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *eLife*, 2:e00523, 2013.

Habibi, E., Brinkman, A. B., Arand, J., Kroeze, L. I., Kerstens, H. H. D., Matarese, F. *et al.* Whole-genome bisulfite sequencing of two distinct interconvertible DNA methylomes of mouse embryonic stem cells. *Cell stem cell*, 13(3):360–9, 2013.

Hackenberg, M., Previti, C., Luque-Escamilla, P. L., Carpena, P., Martínez-Aroza, J., and Oliver, J. L. CpGcluster: a distance-based algorithm for CpG-island detection. *BMC bioinformatics*, 7(1):446, 2006.

Han, H., Cortez, C. C., Yang, X., Nichols, P. W., Jones, P. A., and Liang, G. DNA methylation directly silences genes with non-CpG island promoters and establishes a nucleosome occupied promoter. *Human molecular genetics*, 20(22):4299–310, 2011.

Harrison, A. and Parle-McDermott, A. DNA methylation: a timeline of methods and applications. *Frontiers in genetics*, 2:74, 2011.

Hatada, I., Hayashizaki, Y., Hirotsune, S., Komatsubara, H., and Mukai, T. A genomic scanning method for higher organisms using restriction sites as landmarks. *Proceedings of the National Academy of Sciences of the United States of America*, 88(21):9523–7, 1991.

Hayashizaki, Y., Hirotsune, S., Okazaki, Y., Hatada, I., Shibata, H., Kawai, J. *et al.* Restriction landmark genomic scanning method and its various applications. *Electrophoresis*, 14(4):251–8, 1993.

Hayatsu, H., Wataya, Y., Kai, K., and Iida, S. Reaction of sodium bisulfite with uracil, cytosine, and their derivatives. *Biochemistry*, 9(14):2858–65, 1970.

Heijmans, B. T. and Mill, J. Commentary: The seven plagues of epigenetic epidemiology. *International Journal of Epidemiology*, 41(1):74–78, 2012.

Heijmans, B. T., Tobi, E. W., Stein, A. D., Putter, H., Blauw, G. J., Susser, E. S. *et al.* Persistent epigenetic differences associated with prenatal exposure to famine in humans. *Proceedings of the National Academy of Sciences of the United States of America*, 105(44):17046–9, 2008.

Herman, J. G., Graff, J. R., Myöhänen, S., Nelkin, B. D., and Baylin, S. B. Methylation-specific PCR: a novel PCR assay for methylation status of CpG islands. *Proceedings of the National Academy of Sciences of the United States of America*, 93(18):9821–6, 1996.

Heyn, H. and Esteller, M. DNA methylation profiling in the clinic: applications and challenges. *Nature Reviews. Genetics*, 13(10):679–92, 2012.

Hodges, E., Molaro, A., Dos Santos, C. O., Thekkat, P., Song, Q., Uren, P. J. *et al.* Directional DNA methylation changes and complex intermediate states accompany lineage specificity in the adult hematopoietic compartment. *Molecular cell*, 44(1):17–28, 2011.

Holliday, R. and Pugh, J. E. DNA modification mechanisms and the evolving role of DNA methylation in animals. *Science*, 187:226–232, 1975.

Horsthemke, B. and Buiting, K. Imprinting defects on human chromosome 15. *Cytogenetic and genome research*, 113(1-4):292–9, 2006.

Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biology*, 14(10):R115, 2013.

Hotchkiss, R. D. The quantitative separation of purines, pyrimidines, and nucleosides by paper chromatography. *The Journal of biological chemistry*, 175(1):315–32, 1948.

Houseman, E. A. *Computational and Statistical Epigenomics*, chapter DNA Methyl, pages 35–50. Springer Netherlands, Dordrecht, 2015. ISBN 978-94-017-9927-0.

Houseman, E. A., Accomando, W. P., Koestler, D. C., Christensen, B. C., Marsit, C. J., Nelson, H. H. *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*, 13(1):86, 2012.

Houseman, E. A., Christensen, B. C., Yeh, R.-F., Marsit, C. J., Karagas, M. R., Wrensch, M. *et al.* Model-based clustering of DNA methylation array data: a recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distributions. *BMC bioinformatics*, 9(1):365, 2008.

Houseman, E. A., Kelsey, K. T., Wiencke, J. K., and Marsit, C. J. Cell-composition effects in the analysis of DNA methylation array data: a mathematical perspective. *BMC Bioinformatics*, 16(1):95, 2015.

Houseman, E. A., Molitor, J., and Marsit, C. J. Reference-free cell mixture adjustments in analysis of DNA methylation data. *Bioinformatics*, 30(10):1431–1439, 2014.

Huang, Y., Pastor, W. a., Shen, Y., Tahiliani, M., Liu, D. R., and Rao, A. The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing. *PLoS ONE*, 5(1):e8888, 2010.

Hudson, T. J., Anderson, W., Artez, A., Barker, A. D., Bell, C., Bernabé, R. R. *et al.* International network of cancer genome projects. *Nature*, 464(7291):993–8, 2010.

Illumina Inc. GenomeStudio software data sheet. 2014.

Iqbal, K., Jin, S.-G., Pfeifer, G. P., and Szabó, P. E. Reprogramming of the paternal genome upon fertilization involves genome-wide oxidation of 5-methylcytosine. *Proceedings of the National Academy of Sciences of the United States of America*, 108(9):3642–7, 2011.

Irizarry, R. a., Ladd-Acosta, C., Carvalho, B., Wu, H., Brandenburg, S. a., Jeddeloh, J. a. *et al.* Comprehensive high-throughput arrays for relative methylation (CHARM). *Genome Research*, 18(5):780–790, 2008.

Irizarry, R. A., Ladd-Acosta, C., Wen, B., Wu, Z., Montano, C., Onyango, P. *et al.* The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nature genetics*, 41(2):178–86, 2009.

Issa, J.-P. Age-related epigenetic changes and the immune system. *Clinical immunology (Orlando, Fla.)*, 109(1):103–8, 2003.

Ito, S., Shen, L., Dai, Q., Wu, S. C., Collins, L. B., Swenberg, J. A. *et al.* Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science (New York, N.Y.)*, 333(6047):1300–3, 2011.

Jaffe, A. E. and Irizarry, R. a. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biology*, 15(2):R31, 2014.

Ji, H., Ehrlich, L. I. R., Seita, J., Murakami, P., Doi, A., Lindau, P. *et al.* Comprehensive methylome map of lineage commitment from haematopoietic progenitors. *Nature*, 467(7313):338–42, 2010.

Jones, M. J., Goodman, S. J., and Kobor, M. S. DNA methylation and healthy human aging. *Aging Cell*, 14(6):n/a–n/a, 2015.

Jones, P. A. The DNA methylation paradox. *Trends in genetics : TIG*, 15(1):34–7, 1999.

Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews. Genetics*, 13(7):484–92, 2012.

Jones, P. A. and Liang, G. Rethinking how DNA methylation patterns are maintained. *Nature Reviews. Genetics*, 10(11):805–11, 2009.

Kawai, J., Hirotsune, S., Hirose, K., Fushiki, S., Watanabe, S., and Hayashizaki, Y. Methylation profiles of genomic DNA of mouse developmental brain detected by restriction landmark genomic scanning (RLGS) method. *Nucleic acids research*, 21(24):5604–8, 1993.

Keshet, I., Schlesinger, Y., Farkash, S., Rand, E., Hecht, M., Segal, E. *et al.* Evidence for an instructive mechanism of de novo methylation in cancer cells. *Nature genetics*, 38(2):149–53, 2006.

Khulan, B., Thompson, R. F., Ye, K., Fazzari, M. J., Suzuki, M., Stasiek, E. *et al.* Comparative isoschizomer profiling of cytosine methylation: the HELP assay. *Genome research*, 16(8):1046–55, 2006.

Kim, K., Ban, H.-J., Seo, J., Lee, K., Yavartanoo, M., Kim, S. C. *et al.* Genetic factors underlying discordance in chromatin accessibility between monozygotic twins. *Genome Biology*, 15(5):R72, 2014.

Koestler, D. C., Christensen, B. C., Karagas, M. R., Marsit, C. J., Langevin, S. M., Kelsey, K. T. *et al.* Blood-based profiles of DNA methylation predict the underlying distribution of cell types: A validation analysis. *Epigenetics*, 8(8):816–826, 2013.

Koestler, D. C., Jones, M. J., Usset, J., Christensen, B. C., Butler, R. A., Kobor, M. S. *et al.* Improving cell mixture deconvolution by identifying optimal DNA methylation libraries (IDOL). *BMC bioinformatics*, 17(1):120, 2016.

Korshunova, Y., Maloney, R. K., Lakey, N., Citek, R. W., Bacher, B., Budiman, A. *et al.* Massively parallel bisulphite pyrosequencing reveals the molecular complexity of breast cancer-associated cytosine-methylation patterns obtained from tissue and serum DNA. *Genome Research*, 18(1):19–29, 2008.

Kriaucionis, S. and Heintz, N. The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science (New York, N.Y.)*, 324(5929):929–30, 2009.

Krueger, F. and Andrews, S. R. Bismark: a flexible aligner and methylation caller for bisulfite-seq applications. *Bioinformatics*, 27(11):1571–1572, 2011.

Krueger, F., Kreck, B., Franke, A., and Andrews, S. R. DNA methylome analysis using short bisulfite sequencing data. *Nature Methods*, 9(2):145–51, 2012.

Kuhn, a., Thu, D., Waldvogel, H., Faull, R., and Luthi-Carter, R. Population-specific expression analysis (PSEA) reveals molecular changes in diseased brain. *Nat Methods*, 8(11):945–947, 2011.

Kumaki, Y., Oda, M., and Okano, M. QUMA: quantification tool for methylation analysis. *Nucleic acids research*, 36(Web Server issue):W170–5, 2008.

Kuo, K. C., McCune, R. A., Gehrke, C. W., Midgett, R., and Ehrlich, M. Quantitative reversed-phase high performance liquid chromatographic determination of major and modified deoxyribonucleosides in DNA. *Nucleic acids research*, 8(20):4763–76, 1980.

Laird, C. D., Pleasant, N. D., Clark, A. D., Sneeden, J. L., Hassan, K. M. A., Manley, N. C. *et al.* Hairpin-bisulfite PCR: assessing epigenetic methylation patterns on complementary strands of individual DNA molecules. *Proceedings of the National Academy of Sciences of the United States of America*, 101(1):204–9, 2004.

Laird, P. W. Principles and challenges of genome-wide DNA methylation analysis. *Nat Rev Genet*, 11(3):191–203, 2010.

Lam, L. L., Emberly, E., Fraser, H. B., Neumann, S. M., Chen, E., Miller, G. E. *et al.* Factors underlying variable DNA methylation in a human community cohort. *Proceedings of the National Academy of Sciences*, 109(Supplement_2):17253–17260, 2012.

Landan, G., Cohen, N. M., Mukamel, Z., Bar, A., Molchadsky, A., Brosh, R. *et al.* Epigenetic polymorphism and the stochastic formation of differentially methylated regions in normal and cancerous tissues. *Nature genetics*, 44(11):1207–14, 2012.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):R25, 2009.

Laurent, L., Wong, E., Li, G., Huynh, T., Tsirigos, A., Ong, C. T. *et al.* Dynamic changes in the human methylome during differentiation. *Genome research*, 20(3):320–31, 2010.

Lee, Y., Ghosh, D., and Zhang, Y. Regression hidden Markov modeling reveals heterogeneous gene expression regulation: a case study in mouse embryonic stem cells. *BMC Genomics*, 15:360, 2014.

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E. *et al.* Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews. Genetics*, 11(10):733–9, 2010.

Leek, J. T. and Storey, J. D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS genetics*, 3(9):1724–35, 2007.

Lewin, J., Schmitt, A. O., Adorján, P., Hildmann, T., and Piepenbrock, C. Quantitative DNA methylation analysis based on four-dye trace data from direct sequencing of PCR amplificates. *Bioinformatics (Oxford, England)*, 20(17):3005–12, 2004.

Li, E., Bestor, T. H., and Jaenisch, R. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell*, 69(6):915–26, 1992.

Li, Q., Suzuki, M., Wendt, J., Patterson, N., Eichten, S. R., Hermanson, P. J. *et al.* Post-conversion targeted capture of modified cytosines in mammalian and plant genomes. *Nucleic acids research*, 43(12):e81, 2015.

Liang, L. and Cookson, W. O. C. Grasping nettles: cellular heterogeneity and other confounders in epigenome-wide association studies. *Human molecular genetics*, 23(1):83–88, 2014.

Lister, R., Mukamel, E. A., Nery, J. R., Urich, M., Puddifoot, C. A., Johnson, N. D. *et al.* Global epigenomic reconfiguration during mammalian brain development. *Science (New York, N.Y.)*, 341(6146):1237905, 2013.

Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 462(7271):315–322, 2009.

Liu, J., Litman, D., Rosenberg, M. J., Yu, S., Biesecker, L. G., and Weinstein, L. S. A GNAS1 imprinting defect in pseudohypoparathyroidism type IB. *The Journal of clinical investigation*, 106(9):1167–74, 2000.

Liu, J., Morgan, M., Hutchison, K., and Calhoun, V. D. A study of the influence of sex on genome wide methylation. *PloS one*, 5(4):e10028, 2010.

Liu, Y., Aryee, M. J., Padyukov, L., Fallin, M. D., Hesselberg, E., Runarsson, A. *et al.* Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nature Biotechnology*, 31(2):142–7, 2013.

Liu, Y., Siegmund, K. D., Laird, P. W., and Berman, B. P. Bis-SNP: combined DNA methylation and SNP calling for Bisulfite-seq data. *Genome Biology*, 13(7):R61, 2012.

Lunnon, K., Smith, R., Hannon, E., De Jager, P. L., Srivastava, G., Volta, M. *et al.* Methylomic profiling implicates cortical deregulation of ANK1 in Alzheimer's disease. *Nature neuroscience*, 17(9):1164–70, 2014.

Makismovic, J., Gordon, L., and Oshlack, A. SWAN: Subset-quantile Within Array Normalization for Illumina Infinium HumanMethylation450 BeadChips. *Genome Biol*, 13(6):R44, 2012.

Mandel, J. and Chambon, P. DNA methylation: organ specific variations in the methylation pattern within and around ovalbumin and other chicken genes. *Nucleic Acids Research*, 7(8):2081–2103, 1979.

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

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. a. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380, 2005.

Maunakea, A. K., Nagarajan, R. P., Bilenky, M., Ballinger, T. J., D'Souza, C., Fouse, S. D. *et al.* Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature*, 466(7303):253–7, 2010.

Mayer, W., Niveleau, A., Walter, J., Fundele, R., and Haaf, T. Embryogenesis: Demethylation of the zygotic paternal genome. *Nature*, 403(6769):501–502, 2000.

McGhee, J. D. and Ginder, G. D. Specific DNA methylation sites in the vicinity of the chicken beta-globin genes. *Nature*, 280(5721):419–20, 1979.

Meissner, A., Gnirke, A., Bell, G. W., Ramsahoye, B., Lander, E. S., and Jaenisch, R. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Research*, 33(18):5868–5877, 2005.

Meissner, A., Mikkelsen, T. S., Gu, H., Wernig, M., Hanna, J., Sivachenko, A. *et al.* Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*, 454(7205):766–770, 2008.

Michels, K. B., Binder, A. M., Dedeurwaerder, S., Epstein, C. B., Greally, J. M., Gut, I. *et al.* Recommendations for the design and analysis of epigenome-wide association studies. *Nature Methods*, 10(10):949–55, 2013.

Miltenyi, S., Müller, W., Weichel, W., and Radbruch, A. High gradient magnetic cell separation with MACS. *Cytometry*, 11(2):231–8, 1990.

Miura, F., Enomoto, Y., Dairiki, R., and Ito, T. Amplification-free whole-genome bisulfite sequencing by post-bisulfite adaptor tagging. *Nucleic acids research*, 40(17):e136, 2012.

Mohandas, T., Sparkes, R. S., and Shapiro, L. J. Reactivation of an inactive human X chromosome: evidence for X inactivation by DNA methylation. *Science (New York, N.Y.)*, 211(4480):393–396, 1981.

Montaño, C. M., Irizarry, R. a., Kaufmann, W. E., Talbot, K., Gur, R. E., Feinberg, A. P. *et al.* Measuring cell-type specific differential methylation in human brain tissue. *Genome Biology*, 14(8):R94, 2013.

Moran, S., Arribas, C., and Esteller, M. Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. *Epigenomics*, 2015.

Munroe, D. J. and Harris, T. J. R. Third-generation sequencing fireworks at Marco Island. *Nature Biotechnology*, 28(5):426–8, 2010.

Munson, K., Clark, J., Lamparska-Kupsik, K., and Smith, S. S. Recovery of bisulfite-converted genomic sequences in the methylation-sensitive QPCR. *Nucleic acids research*, 35(9):2893–903, 2007.

Nakagawa, H., Nuovo, G. J., Zervos, E. E., Martin, E. W., Salovaara, R., Aaltonen, L. A. *et al.* Age-related hypermethylation of the 5' region of MLH1 in normal colonic mucosa is associated with microsatellite-unstable colorectal cancer development. *Cancer research*, 61(19):6991–5, 2001.

Needleman, S. B. and Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453, 1970.

Nestor, C., Ruzov, A., Meehan, R., and Dunican, D. Enzymatic approaches and bisulfite sequencing cannot distinguish between 5-methylcytosine and 5-hydroxymethylcytosine in DNA. *BioTechniques*, 48(4):317–9, 2010.

Ngo, S. and Sheppard, A. The role of DNA methylation: a challenge for the DOHaD paradigm in going beyond the historical debate. *Journal of developmental origins of health and disease*, 6(1):2–4, 2015.

Noushmehr, H., Weisenberger, D. J., Diefes, K., Phillips, H. S., Pujara, K., Berman, B. P. *et al.* Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. *Cancer Cell*, 17(5):510–522, 2010.

Oakeley, E. J., Podestà, A., and Jost, J. P. Developmental changes in DNA methylation of the two tobacco pollen nuclei during maturation. *Proceedings of the National Academy of Sciences of the United States of America*, 94(21):11721–5, 1997.

Okano, M., Bell, D. W., Haber, D. a., and Li, E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell*, 99(3):247–257, 1999.

Oswald, J., Engemann, S., Lane, N., Mayer, W., Olek, A., Fundele, R. *et al.* Active demethylation of the paternal genome in the mouse zygote. *Current biology : CB*, 10(8):475–8, 2000.

Pelizzola, M. and Ecker, J. R. The DNA methylome. *FEBS Letters*, 585(13):1994–2000, 2011.

Pidsley, R., Y Wong, C. C., Volta, M., Lunnon, K., Mill, J., Schalkwyk, L. C. *et al.* A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics*, 14(1):293, 2013.

Pirazzini, C., Giuliani, C., Bacalini, M. G., Boattini, A., Capri, M., Fontanesi, E. *et al.* Space/population and time/age in DNA methylation variability in humans: a study on IGF2/H19 locus in different Italian populations and in mono- and di-zygotic twins of different age. *Aging*, 4(7):509–20, 2012.

Plass, C., Pfister, S. M., Lindroth, A. M., Bogatyrova, O., Claus, R., and Lichter, P. Mutations in regulators of the epigenome and their connections to global chromatin patterns in cancer. *Nature reviews. Genetics*, 14(11):765–80, 2013.

Pollack, Y., Stein, R., Razin, A., and Cedar, H. Methylation of foreign DNA sequences in eukaryotic cells. *Proceedings of the National Academy of Sciences of the United States of America*, 77(11):6463–7, 1980.

Rahmani, E., Zaitlen, N., Baran, Y., Eng, C., Hu, D., Galanter, J. *et al.* Sparse PCA corrects for cell type heterogeneity in epigenome-wide association studies. *Nature Methods*, 2016.

Rakyan, V. K., Down, T. a., Balding, D. J., and Beck, S. Epigenome-wide association studies for common human diseases. *Nature Reviews. Genetics*, 12(8):529–541, 2011.

Rakyan, V. K., Down, T. A., Maslau, S., Andrew, T., Yang, T.-P., Beyan, H. *et al.* Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains. *Genome research*, 20(4):434–9, 2010.

Rakyan, V. K., Down, T. a., Thorne, N. P., Flicek, P., Kulesha, E., Gräf, S. *et al.* An integrated resource for genome-wide identification and analysis of human tissue-specific differentially methylated regions (tDMRs). *Genome Research*, 18(9):1518–1529, 2008.

Rakyan, V. K., Hildmann, T., Novik, K. L., Lewin, J., Tost, J., Cox, A. V. *et al.* DNA methylation profiling of the human major histocompatibility complex: a pilot study for the human epigenome project. *PLoS biology*, 2(12):e405, 2004.

Reinius, L. E., Acevedo, N., Joerink, M., Pershagen, G., Dahlén, S.-E., Greco, D. *et al.* Differential DNA Methylation in Purified Human Blood Cells: Implications for Cell Lineage and Studies on Disease Susceptibility. *PLoS ONE*, 7(7):e41361, 2012.

Rembaum, A., Yen, R. C., Kempner, D. H., and Ugelstad, J. Cell labeling and magnetic separation by means of immunoreagents based on polyacrolein microspheres. *Journal of immunological methods*, 52(3):341–51, 1982.

Repsilber, D., Kern, S., Telaar, A., Walzl, G., Black, G. F., Selbig, J. *et al.* Biomarker discovery in heterogeneous tissue samples -taking the in-silico deconfounding approach. *BMC bioinformatics*, 11:27, 2010.

Riggs, A. D. X inactivation, differentiation, and DNA methylation. *Cytogenet Cell Genet*, 14(1):9–25, 1975.

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–30, 2015.

Rohde, C., Zhang, Y., Reinhardt, R., and Jeltsch, A. BISMA–fast and accurate bisulfite sequencing data analysis of individual clones from unique and repetitive sequences. *BMC bioinformatics*, 11:230, 2010.

Rougier, N., Bourc'his, D., Gomes, D. M., Niveleau, A., Plachot, M., Paldi, A. *et al.* Chromosome methylation patterns during mammalian preimplantation development. *Genes & Development*, 12(14):2108–2113, 1998.

Ruppel, G. W. Zur chemie der tuberkelbacillen. *Z Physiol Chem*, 26:218–232, 1898.

Sakai, T., Toguchida, J., Ohtani, N., Yandell, D. W., Rapaport, J. M., and Dryja, T. P. Allele-specific hypermethylation of the retinoblastoma tumor-suppressor gene. *American journal of human genetics*, 48(5):880–8, 1991.

Santos, F., Hendrich, B., Reik, W., and Dean, W. Dynamic reprogramming of DNA methylation in the early mouse embryo. *Developmental biology*, 241(1):172–82, 2002.

Sarter, B., Long, T. I., Tsong, W. H., Koh, W.-P., Yu, M. C., and Laird, P. W. Sex differential in methylation patterns of selected genes in Singapore Chinese. *Human genetics*, 117(4):402–3, 2005.

Schilling, E. and Rehli, M. Global, comparative analysis of tissue-specific promoter CpG methylation. *Genomics*, 90(3):314–23, 2007.

Schübeler, D. Function and information content of DNA methylation. *Nature*, 517(7534):321–326, 2015.

Schwartzman, O. and Tanay, A. Single-cell epigenomics: techniques and emerging applications. *Nature Reviews Genetics*, 16(12):716–726, 2015.

Sehouli, J., Loddenkemper, C., Cornu, T., Schwachula, T., Hoffmüller, U., Grützkau, A. *et al.* Epigenetic quantification of tumor-infiltrating T-lymphocytes. *Epigenetics*, 6(2):236–246, 2011.

Shemer, R., Birger, Y., Riggs, A. D., and Razin, A. Structure of the imprinted mouse Snrpn gene and establishment of its parental-specific methylation pattern. *Proceedings of the National Academy of Sciences of the United States of America*, 94(19):10267–72, 1997.

Shen, C. K. and Maniatis, T. Tissue-specific DNA methylation in a cluster of rabbit beta-like globin genes. *Proceedings of the National Academy of Sciences*, 77(11):6634–6638, 1980.

Shen, L., Kondo, Y., Hamilton, S. R., Rashid, A., and Issa, J.-P. J. P14 methylation in human colon cancer is associated with microsatellite instability and wild-type p53. *Gastroenterology*, 124(3):626–33, 2003.

Shen-Orr, S. S., Tibshirani, R., Khatri, P., Bodian, D. L., Staedtler, F., Perry, N. M. *et al.* Cell type-specific gene expression differences in complex tissues. *Nature Methods*, 7(4):287–289, 2010.

Shoemaker, R., Deng, J., Wang, W., and Zhang, K. Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. *Genome Research*, 20(7):883–889, 2010.

Siegmund, K. D. Statistical approaches for the analysis of DNA methylation microarray data. *Human genetics*, 129(6):585–95, 2011.

Siegmund, K. D., Marjoram, P., Woo, Y.-J., Tavaré, S., and Shibata, D. Inferring clonal expansion and cancer stem cell dynamics from DNA methylation patterns in colorectal cancers. *Proceedings of the National Academy of Sciences of the United States of America*, 106(12):4828–33, 2009.

Smallwood, S. a., Lee, H. J., Angermueller, C., Krueger, F., Saadeh, H., Peat, J. *et al.* Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nature Methods*, 11(8):817–20, 2014.

Smith, J. D., Arber, W., and Kühnlein, U. Host specificity of DNA produced by Escherichia coli. XIV. The role of nucleotide methylation in in vivo B-specific modification. *Journal of molecular biology*, 63(1):1–8, 1972.

Smith, M. L., Baggerly, K. A., Bengtsson, H., Ritchie, M. E., and Hansen, K. D. illuminaio: An open source IDAT parsing tool for Illumina microarrays. *F1000Research*, 2:264, 2013.

Smith, Z. D., Chan, M. M., Mikkelsen, T. S., Gu, H., Gnirke, A., Regev, A. *et al.* A unique regulatory phase of DNA methylation in the early mammalian embryo. *Nature*, 484(7394):339–44, 2012.

Song, C. X., Szulwach, K. E., Dai, Q., Fu, Y., Mao, S. Q., Lin, L. *et al.* Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming. *Cell*, 153(3):678–691, 2013.

Souren, N. Y., Lutsik, P., Gasparoni, G., Tierling, S., Gries, J., Riemenschneider, M. *et al.* Adult monozygotic twins discordant for intra-uterine growth have indistinguishable genome-wide DNA methylation profiles. *Genome Biology*, 14(5):R44, 2013.

Srinivasan, P. R. and Borek, E. Enzymatic Alteration of Nucleic Acid Structure: Enzymes put finishing touches characteristic of each species on RNA and DNA by insertion of methyl groups. *Science*, 145(3632):548–553, 1964.

Stadler, M. B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Schöler, A. *et al.* DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature*, 480(7378):490–5, 2011.

Stein, R., Gruenbaum, Y., Pollack, Y., Razin, A., and Cedar, H. Clonal inheritance of the pattern of DNA methylation in mouse cells. *Proceedings of the National Academy of Sciences of the United States of America*, 79(1):61–5, 1982a.

Stein, R., Razin, A., and Cedar, H. In vitro methylation of the hamster adenine phosphoribosyltransferase gene inhibits its expression in mouse L cells. *Proceedings of the National Academy of Sciences*, 79(11):3418–3422, 1982b.

Sun, Z., Chai, H. S., Wu, Y., White, W. M., Donkena, K. V., Klein, C. J. *et al.* Batch effect correction for genome-wide methylation data with Illumina Infinium platform. *BMC medical genomics*, 4(1):84, 2011.

Swartz, M. N., Trautner, T. A., and Kornberg, A. Enzymatic synthesis of deoxyribonucleic acid. XI. Further studies on nearest neighbor base sequences in deoxyribonucleic acids. *The Journal of biological chemistry*, 237:1961–7, 1962.

Szathmáry, E. Why are there four letters in the genetic alphabet? *Nature Reviews. Genetics*, 4(12):995–1001, 2003.

Tahiliani, M., Koh, K. P., Shen, Y., Pastor, W. a., Bandukwala, H., Brudno, Y. *et al.* Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science (New York, N.Y.)*, 324(5929):930–935, 2009.

Takada, S., Paulsen, M., Tevendale, M., Tsai, C.-E., Kelsey, G., Cattanach, B. M. *et al.* Epigenetic analysis of the Dlk1-Gtl2 imprinted domain on mouse chromosome 12: implications for imprinting control from comparison with Igf2-H19. *Human molecular genetics*, 11(1):77–86, 2002.

Takai, D. and Jones, P. A. The CpG island searcher: a new WWW resource. *In silico biology*, 3(3):235–40, 2003.

Taylor, K. H., Kramer, R. S., Davis, J. W., Guo, J., Duff, D. J., Xu, D. *et al.* Ultradeep bisulfite sequencing analysis of DNA methylation patterns in multiple gene promoters by 454 sequencing. *Cancer Research*, 67(18):8511–8518, 2007.

Teschendorff, A. E., Marabita, F., Lechner, M., Bartlett, T., Tegner, J., Gomez-Cabrero, D. *et al.* A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics*, 29(2):189–196, 2013.

Teschendorff, A. E., Zhuang, J., and Widschwendter, M. Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies. *Bioinformatics (Oxford, England)*, 27(11):1496–505, 2011.

Tomlinson, M. J., Tomlinson, S., Yang, X. B., and Kirkham, J. Cell separation: Terminology and practical considerations. *Journal of tissue engineering*, 4:2041731412472690, 2013.

Touleimat, N. and Tost, J. Complete pipeline for Infinium(®) Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. *Epigenomics*, 4(3):325–41, 2012.

Townsend, L. B. *Chemistry of nucleosides and nucleotides*, volume 3. Springer Science & Business Media, 2013.

Triche Jr., T. J., Weisenberger, D. J., Van Den Berg, D., Laird, P. W., Siegmund, K. D., Triche, T. J. *et al.* Low-level processing of Illumina Infinium DNA Methylation BeadArrays. *Nucleic Acids Research*, 41(7):e90–e90, 2013.

Tusnády, G. E., Simon, I., Váradi, A., and Arányi, T. BiSearch: primer-design and search tool for PCR on bisulfite-treated genomes. *Nucleic acids research*, 33(1):e9, 2005.

van der Maaten, L. and Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

van Dongen, J., Nivard, M. G., Willemsen, G., Hottenga, J.-J., Helmer, Q., Dolan, C. V. *et al.* Genetic and environmental influences interact with age and sex in shaping the human methylome. *Nature communications*, 7:11115, 2016.

van Eijk, K. R., de Jong, S., Boks, M. P. M., Langeveld, T., Colas, F., Veldink, J. H. *et al.* Genetic analysis of DNA methylation and gene expression levels in whole blood of healthy human subjects. *BMC genomics*, 13(1):636, 2012.

Vanyushin, B. F., Nemirovsky, L. E., Klimenko, V. V., Vasiliev, V. K., and Belozersky, A. N. The 5-methylcytosine in DNA of rats. Tissue and age specificity and the changes induced by hydrocortisone and other agents. *Gerontologia*, 19(3):138–52, 1973.

Vardimon, L., Kressmann, A., Cedar, H., Maechler, M., and Doerfler, W. Expression of a cloned adenovirus gene is inhibited by in vitro methylation. *Proceedings of the National Academy of Sciences of the United States of America*, 79(4):1073–1077, 1982.

Varley, K. E., Gertz, J., Bowling, K. M., Parker, S. L., Reddy, T. E., Pauli-Behn, F. *et al.* Dynamic DNA methylation across diverse human cell lines and tissues. *Genome research*, 23(3):555–67, 2013.

Varley, K. E., Mutch, D. G., Edmonston, T. B., Goodfellow, P. J., and Mitra, R. D. Intra-tumor heterogeneity of MLH1 promoter methylation revealed by deep single molecule bisulfite sequencing. *Nucleic Acids Research*, 37(14):4603–4612, 2009.

Venolia, L., Gartler, S. M., Wassman, E. R., Yen, P., Mohandas, T., and Shapiro, L. J. Transformation with DNA from 5-azacytidine-reactivated X chromosomes. *Proceedings of the National Academy of Sciences of the United States of America*, 79(7):2352–4, 1982.

Wang, R. Y., Gehrke, C. W., and Ehrlich, M. Comparison of bisulfite modification of 5-methyldeoxycytidine and deoxycytidine residues. *Nucleic acids research*, 8(20):4777–90, 1980.

Waterman, M. S., Eggert, M., and Lander, E. Parametric sequence comparisons. *Proceedings of the National Academy of Sciences of the United States of America*, 89(13):6090–6093, 1992.

Weber, M., Davies, J. J., Wittig, D., Oakeley, E. J., Haase, M., Lam, W. L. *et al.* Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nature genetics*, 37(8):853–862, 2005.

Weidner, C. I., Lin, Q., Koch, C. M., Eisele, L., Beier, F., Ziegler, P. *et al.* Aging of blood can be tracked by DNA methylation changes at just three CpG sites. *Genome Biology*, 15(2):R24, 2014.

Weng, N.-P., Akbar, A. N., and Goronzy, J. CD28(-) T cells: their role in the age-associated decline of immune function. *Trends in immunology*, 30(7):306–12, 2009.

Wilhelm-Benartzi, C. S., Koestler, D. C., Karagas, M. R., Flanagan, J. M., Christensen, B. C., Kelsey, K. T. *et al.* Review of processing and analysis methods for DNA methylation array data. *British journal of cancer*, 109(6):1394–402, 2013.

Wilkinson, M. F. Evidence that DNA methylation engenders dynamic gene regulation. *Proceedings of the National Academy of Sciences of the United States of America*, 112(17):E2116, 2015.

Wilson, V. L., Smith, R. A., Ma, S., and Cutler, R. G. Genomic 5-methyldeoxycytidine decreases with age. *J. Biol. Chem.*, 262(21):9948–9951, 1987.

Wojdacz, T. K. and Dobrovic, A. Methylation-sensitive high resolution melting (MS-HRM): a new approach for sensitive and high-throughput assessment of methylation. *Nucleic acids research*, 35(6):e41, 2007.

Wolf, S. F., Jolly, D. J., Lunnen, K. D., Friedmann, T., and Migeon, B. R. Methylation of the hypoxanthine phosphoribosyltransferase locus on the human X chromosome: implications for X-chromosome inactivation. *Proceedings of the National Academy of Sciences of the United States of America*, 81(9):2806–10, 1984.

Wossidlo, M., Nakamura, T., Lepikhov, K., Marques, C. J., Zakhartchenko, V., Boiani, M. *et al.* 5-Hydroxymethylcytosine in the mammalian zygote is linked with epigenetic reprogramming. *Nature communications*, 2:241, 2011.

Wu, J., Issa, J. P., Herman, J., Bassett, D. E., Nelkin, B. D., and Baylin, S. B. Expression of an exogenous eukaryotic DNA methyltransferase gene induces transformation of NIH 3T3 cells. *Proceedings of the National Academy of Sciences of the United States of America*, 90(19):8891–5, 1993.

Wyatt, G. R. Recognition and estimation of 5-methylcytosine in nucleic acids. *The Biochemical journal*, 48(5):581–584, 1951.

Xiong, Z. and Laird, P. W. COBRA: a sensitive and quantitative DNA methylation assay. *Nucleic acids research*, 25(12):2532–4, 1997.

Yen, R. W., Vertino, P. M., Nelkin, B. D., Yu, J. J., El-Deiry, W., Cumaraswamy, A. *et al.* Isolation and characterization of the cDNA encoding human DNA methyltransferase. *Nucleic acids research*, 20(9):2287–91, 1992.

Yoder, J. A., Walsh, C. P., and Bestor, T. H. Cytosine methylation and the ecology of intragenomic parasites. *Trends in Genetics*, 13(8):335–340, 1997.

Zhang, Y., Rohde, C., Tierling, S., Jurkowski, T. P., Bock, C., Santacruz, D. *et al.* DNA methylation analysis of chromosome 21 gene promoters at single base pair and single allele resolution. *PLoS genetics*, 5(3):e1000438, 2009.

Ziller, M. J., Gu, H., Müller, F., Donaghey, J., Tsai, L. T.-Y., Kohlbacher, O. *et al.* Charting a dynamic DNA methylation landscape of the human genome. *Nature*, 500(7463):477–81, 2013.

Zou, J., Lippert, C., Heckerman, D., Aryee, M., and Listgarten, J. Epigenome-wide association studies without the need for cell-type composition. *Nature Methods*, 11(3):309–11, 2014.

Chapter 2

Adult monozygotic twins discordant for intra-uterine growth have indistinguishable genome-wide DNA methylation profiles

The full text of this chapter has been earlier published as:

Nicole Y.P. Souren[1-3,#], Pavlo Lutsik[1], Gilles Gasparoni[1], Sascha Tierling[1], Jasmin Gries[1], Matthias Riemenschneider[4], Jean-Pierre Fryns[5], Catherine Derom[5], Maurice P. Zeegers[2,3,6], Jörn Walter[1,#] (2011) *Genome Biology* **15**(5), R44.

The author of the present thesis processed and analyzed the bead array data, co-designed (with N.S.) and implemented the marker-based adjustment for the correction of the cell type heterogeneity, visualized and deposited the processed data in GEO. He generated most of the display items and participated in writing of the manuscript (with N.S. and J.W.).

[1] Laboratory of EpiGenetics, FR 8.3 Life Sciences, Saarland University, Saarbrücken, 66123, Saarland, Germany
[2] Department of Genetics and Cell Biology
[3] Nutrition and Toxicology Research Institute Maastricht (NUTRIM), Maastricht University, Maastricht, 6200 MD, Limburg, The Netherlands
[4]Department of Psychiatry and Psychotherapy, Saarland University Hospital, Homburg, 66424, Saarland, Germany
[5] Department of Human Genetics, University Hospital Gasthuisberg, Katholieke Universiteit Leuven, Leuven, B-3000, Vlaams-Brabant, Belgium
[6] Unit of Urologic and Genetic Epidemiology, Department of Public Health and Epidemiology, University of Birmingham, Birmingham, B15 2TT, West Midlands, United Kingdom.
[#] To whom correspondence should be addressed.

Abstract

Background: Low birth weight is associated with an increased adult metabolic disease risk. It is widely discussed that poor intra-uterine conditions could induce long-lasting epigenetic modifications, leading to systemic changes in regulation of metabolic genes. To address this, we acquire genome-wide DNA methylation profiles from saliva DNA in a unique cohort of 17 monozygotic monochorionic female twins very discordant for birth weight. We examine if adverse prenatal growth conditions experienced by the smaller co-twins lead to long-lasting DNA methylation changes.
Results: Overall,co-twins show very similar genome-wide DNA methylation profiles. Since observed differences are almost exclusively caused by variable cellular composition, an original marker-based adjustment strategy was developed to eliminate such variation at affected CpGs. Among adjusted and unchanged CpGs 3,153 are differentially methylated between the heavy and light co-twins at nominal significance, of which 45 show sensible absolute mean β-value differences. Deep bisulfite sequencing of eight such loci reveals that differences remain in the range of technical variation, arguing against a reproducible biological effect. Analysis of methylation in repetitive elements using methylation-dependent primer extension assays also indicates no significant intra-pair differences.
Conclusions: Severe intra-uterine growth differences observed within these monozygotic twins are not associated with long-lasting DNA methylation differences in cells composing saliva, detectable with up-to-date technologies. Additionally, our results indicate that uneven cell type composition can lead to spurious results and should be addressed in epigenomic studies.

Keywords: DNA methylation, birth weight, monozygotic monochorionic twins, saliva, intra-uterine growth restriction, Infinium HumanMethylation450 BeadChip

Competing interests

The authors declare that they have no competing interests associated with this manuscript.

Authors' contributions

NS designed the study, conducted the experimental work, participated in the statistical data analysis and wrote the manuscript. PL performed bioinformatical and statistical analyses and participated in writing the manuscript. ST and JG assisted in the DBS and methylation-dependent primer extension experiments. GG and MR generated the HumanMethylation450 profiles. CD, JF and MZ were involved in the study design, collecting the twin data and samples. JW supervised the study, participated in the study design, writing the manuscript and provided technical and material support. All authors read and approved the final manuscript.

Acknowledgements

This project has been supported by the Foundation "De Drie Lichten" in The Netherlands. Nicole YP Souren was supported by an Alexander von Humboldt research fellowship. The EFPTS has been partly supported by grants from the Fund for Scientific Research Flanders and by Twins, Association for Scientific Research in Multiple Births Belgium. We are grateful to all the twins participating in this study. We thank Lut De Zeure for excellent fieldwork and we thank Sandra Rubil and Şebnem Akbaş for excellent laboratory assistance.

2.1 Background

Both observational human and experimental animal studies have confirmed that low birth weight is associated with an increased risk of metabolic diseases, like Type 2 Diabetes (T2D) [Barker, 2006; McMillen and Robinson, 2005; Newsome *et al.*, 2003]. Although genetic factors are likely to contribute [Lindsay *et al.*, 2000; Wannamethee *et al.*, 2004], studies assessing the association between low birth weight and T2D precursors in monozygotic (MZ) twins showed that the twin who was lighter at birth had a more adverse metabolic profile in adulthood compared to its genetically identical co-twin, who was heavier at birth [Bo *et al.*, 2000; Grunnet *et al.*, 2007; Iliadou *et al.*, 2004; Monrad *et al.*, 2009; Poulsen *et al.*, 1997]. This suggests that the association between low birth weight and increased T2D risk is at least partly independent of genetic factors.

One of the possible molecular mechanisms explaining this non-genetic association suggests that poor prenatal conditions induce epigenetic modifications [Gluckman *et al.*, 2008]. These epigenetic modifications are believed to cause a "thrifty" metabolic state, which is beneficial for survival under circumstances of insufficient nutrient supply, but unfavorable when nutrient supply is abundant in postnatal life. An important epigenetic phenomenon is DNA methylation that almost exclusively occurs at cytosines within CpG dinucleotides and correlates with transcriptional repression, while loss of methylation can result in transcriptional activation [Bird, 2002].

The notion that poor intra-uterine conditions cause epigenetic modifications during prenatal development is supported by data from animal studies, where dietary restriction or surgical interventions are used to induce fetal growth restriction, resulting in epigenetic modifications on metabolic disease-related genes (reviewed by [Seki *et al.*, 2012]). The number of studies assessing the relation between an adverse fetal environment and epigenetic alterations in humans is gradually growing as well. For instance, humans who were periconceptionally exposed to famine during the Dutch Hunger Winter (1944-1945) were reported to show significant methylation differences at several imprinted and non-imprinted genes in comparison to their unexposed siblings in peripheral blood cells [Heijmans *et al.*, 2008; Tobi *et al.*, 2009]. A genome-wide DNA methylation study performed on CD34+ hematopoietic stem cells from cord blood of five intra-uterine growth restricted (IUGR) neonates and five gestational age and gender-matched controls [Einstein *et al.*, 2010], identified among others significant methylation differences at the *HNF4A* gene, which is involved in monogenic diabetes.

However, epigenetic association studies using population or family-based designs suffer from confounding caused by DNA sequence variation. Specifically, since birth weight is partly controlled by genetic factors [Yaghootkar and Freathy, 2012], a study with genetically unmatched cases and controls cannot dissect whether a small size at birth is due to a poor prenatal environment or genetic predisposition. On the other hand, epigenetic variation is often a result of genetic variation, e.g. allele-specific methylation where the methylation pattern of a DNA molecule is determined by a *cis*- or *trans*-acting genetic variant [Meaburn *et al.*, 2010]. Since MZ twins originate from one zygote, they are almost absolutely genetically identical, which makes them ideal to search for epigenetic phenomena associated with phenotypic discordancy. In addition, MZ twins are matched for gender, (gestational) age, maternal factors (e.g. parity, age) and a broad range of environmental factors.

Depending on whether the embryo splits during an early or later developmental stage, MZ twins can be dichorionic (DC) or monochorionic (MC), respectively. MZ DC twins have two separate placentas, while MZ MC twins share a single placenta [Derom *et al.*, 2006]. It has been shown that the degree of DNA methylation dissimilarity varies between MZ MC and MZ DC

twins [Kaminsky *et al.*, 2009], indicating that for epigenetic purposes one should either study MZ MC or MZ DC twins. Due to placental blood vessel connections and unequal sharing of the placenta, imbalanced blood and nutrient supply is more common in MZ MC compared to MZ DC twins [Lewi *et al.*, 2010]. Poor prenatal conditions experienced by only one co-twin often result in large intra-pair birth weight differences within MZ MC pairs [Loos *et al.*, 2001], turning them into a "natural experiment" to study the fetal programming origins of late onset human diseases.

We hypothesized that if poor prenatal conditions induce changes in DNA methylation patterns that remain throughout life, these changes should be visible in MZ MC twins discordant for birth weight, irrespective of their health status (degree of insulin resistance, obesity etc) in adulthood. To identify loci that are differentially methylated due to poor prenatal conditions, we performed an epigenome-wide association study (EWAS) in 17 adult female MZ MC twin pairs with a relative birth weight difference greater than 20%. The twins were recruited from the East Flanders Prospective Twin Survey (EFPTS), which started in 1964 and is unique due to its long term extensive collection of perinatal (e.g. birth weight, gestational age, parity) and placental data (e.g. chorionicity) of nearly 8800 twin pairs [Derom *et al.*, 2006]. DNA was isolated from saliva, a bio-fluid that is easily accessible via a totally non-invasive method and therefore widely used in large cohort studies and perfect for diagnostic purposes. Genome-wide DNA methylation profiles were determined using the Infinium HumanMethylation450 BeadChip and validated using targeted deep coverage bisulfite sequencing. Additionally, repetitive element methylation levels were determined using methylation-dependent primer extension assays. Our thorough DNA methylation analyses in saliva of birth weight discordant MZ MC twins show that the adverse prenatal growth conditions experienced by the smaller co-twins do not lead to long-lasting DNA methylation changes in cells composing saliva (i.e. buccal epithelium and leukocytes), detectable with up-to-date technologies. In addition, we observe that EWAS studies can be hampered by variation in cellular composition, which can lead to spurious results. We present an adjustment method to normalize the DNA methylation data with respect to variable cell-type content.

2.2 Material and methods

2.2.1 Participants

For this study, 17 spontaneously conceived MZ MC female twin pairs discordant for birth weight were recruited from the EFPTS [Derom *et al.*, 2006], which is a population-based twin register that started in 1964 and recorded all multiple births in the Belgian Province of East Flanders until the present. Discordancy was defined as relative birth weight difference $\geq$20% ([highest birth weight – lowest birth weight]/highest birth weight), with the lightest twin having a birth weight below the 10^{th} percentile and the heavier twin having a birth weight between the 10^{th} and 90^{th} percentile for that gestational age, gender, parity and chorion type (based on twin-specific growth charts [Gielen *et al.*, 2008]). To minimize variation due to gender-specific methylation differences [Liu *et al.*, 2010], only female twins were included. In addition, to assure that the DNA methylation changes remain throughout life, only adult ($\geq$18 years) discordant MZ MC female twins were included (in total 61 pairs satisfy these selection criteria). Of the 17 twin pairs, 15 pairs were newly recruited for this study, while two pairs were previously recruited for another study [Souren *et al.*, 2011]. None of the participants suffered from severe postnatal complications. The Ethics Committee of the Faculty of Medicine of the Katholieke Universiteit Leuven approved the project and all participants gave written

informed consent. The study was conducted according to the principles of the Declaration of Helsinki.

2.2.2 Phenotypes

Information on birth weight and parity was obtained from obstetric records within 24 hours after delivery. Gestational age was reported by the obstetrician and was calculated as the number of completed weeks of pregnancy based on the last menstrual period. The obstetricians and the pediatricians answered a structured questionnaire that provided among other items information on the mode of conception, abnormalities of the children and the health status of the children for the period they stayed in the neonatal unit. A trained midwife examined the placentas within 24 hours of delivery and assessed chorionicity macroscopically following a standardized protocol [Derom *et al.*, 1995]. Adult phenotypic data were retrieved from a mailed questionnaire, which included self-reported items on current body weight, body height, physical activity level, medical history, smoking behavior and alcohol consumption. Body mass index was calculated as self-reported body weight divided by the square of height (kg/m^2).

2.2.3 Genomic DNA extraction

Saliva samples were collected using the Oragene DNA Self-collection Kit (DNA Genotek, Ottawa, Canada). Genomic DNA was extracted from saliva using the GenElute Mammalian Genomic DNA Miniprep Kit (Sigma-Aldrich, St. Louis, MO, USA) according to the manufacturer's instructions. DNA was quantified using the Qubit fluorometer (Invitrogen GmbH, Karlsruhe, Germany) and qualified using the Nanodrop 2000C spectrophotometer (Thermo Scientific, Wilmington, DE, USA). Per DNA extraction batch, only one member per twin pair was processed.

2.2.4 Zygosity confirmation

Although all twins were monochorionic and thus assumed to be monozygotic, zygosity was confirmed by genotyping 17 highly polymorphic microsatellite markers using the PowerPlex ESI 17 system (Promega Corporation, Madison, WI, USA), with an average certainty exceeding 99.99%.

2.2.5 Genome-wide DNA methylation analysis

To avoid differences in methylation levels within twins due to bisulfite treatment or PCR bias, both members of a twin pair were always processed in the same batch.

Bisulfite treatment. For the genome-wide study, per sample 1 ug (2x 500 ng) of genomic DNA (OD260/280>1.8) extracted from saliva was treated with bisulfite using the EZ DNA Methylation-Gold Kit (D5005, Zymo Research, Orange, CA, USA). In brief, 700 μl water, 300 μl M-Dilution Buffer and 50 μl M-Dissolving Buffer was added to the CT Conversion Reagent tube. After mixing, 110 μl of the CT Conversion Reagent was added to 40 μl of DNA, which was then incubated for 10 min at 98°C followed by 3 hours on 64°C. Afterwards, bisulfite-converted DNA samples were purified by loading, desulfonating and washing on the provided Zymo-Spin™ IC columns (following the manufacturer's instructions), eluted in 12 μl M-Elution Buffer and stored at-20°C prior to processing.

Infinium HumanMethylation450 BeadChip. Genome-wide DNA methylation profiles were generated using Illumina's Infinium HumanMethylation450 Beadchip assay (Illumina, San Diego, CA, USA) at the Department of Psychiatry and Psychotherapy of the Saarland University Hospital. The assay allows determination of DNA methylation levels at >450,000 CpG sites covering all designable RefSeq genes, including promoter, 5', and 3' regions; it captures CpG islands and shores, non-CpG methylated sites, miRNA promoter regions and disease-associated regions identified through GWAS [Bibikova *et al.*, 2011]. The Infinium Methylation Assay was performed according to manufacturer's instructions. In brief, 4 μl of denatured bisulfite-treated DNA was isothermally amplified overnight at 37°C, followed by an enzymatic fragmentation step using end-point fragmentation. The fragmented DNA was then precipitated, resuspended and loaded on the 12-sample BeadChip (see for the distribution of the samples across the beadchips Additonal file 1, Table 2.S4). The chips were incubated overnight at 48°C, allowing the fragmented DNA to hybridize to the locus-specific 50-mers on the chip. Unhybridized and non-specifically hybridized DNA was washed away, followed by a single-base extension reaction using DNP- and Biotin-labeled ddNTPs.

Subsequently, the hybridized DNA was washed away and a multi-layer staining process was carried out to attach fluorescent dyes to the labeled extended primers. The fluorescently stained chips were imaged using an Illumina HiScanSQ scanner and the Illumina's GenomeStudio software (Methylation Module v1.8) was used to extract the data, subtract the background and to normalize the data using internal controls present on the chip (see for details Supplemental methods). The overall performance of the normalization procedure is illustrated in Figure 2.S13. Subsequently, for each CpG site a β-value was calculated, which represents the fraction of methylated cytosines at that particular CpG site (0=unmethylated, 1=fully methylated). Only CpGs with a detection *P*-value <0.001 in all samples were included (see Table 2.S4). In total, 4325 out of 482,421 CpGs were excluded, of which 351 were located on the Y-chromosome.

2.2.6 Deep bisulfite sequencing (DBS) analysis

Selected MVPs were validated using DBS. As long as the stocks lasted, the bisulfite DNA used for the genome-wide scan was used for validation analysis or a new bisulfite treatment was performed using standard protocols. In brief, 2 M sodium bisulfite and 0.6 M NaOH was added to 300 ng genomic DNA, which was then incubated for 15 min at 99°C and 30 min at 50°C, followed by 2 cycles of 5 min at 99°C and 90 min at 50°C. Afterwards, bisulfite-treated DNA was sequentially desulfonated (with 0.3 M NaOH), washed with 1xTE and recovered in 50 µl 0.5xTE using centrifugal filter units YM-30 (Millipore, Schwalbach, Germany).

Amplicons were generated using region-specific primers having on their 5-ends the recommended GS-FLX A and B adaptors sequences (Lib-L) and multiplex identifiers (MID) (Roche, Mannheim, Germany). Bisulfite PCRs were carried out in 30 µl mixes, including 1-3 µl bisulfite-treated DNA, 0.2 mM of each dNTP, 3 U HOT FIREPol DNA polymerase (Solis BioDyne, Tartu, Estonia), 1x reaction buffer B (Solis BioDyne), 2.5 mM MgCl2, or 1.5 U HotStarTaq DNA polymerase (Qiagen, Hilden, Germany) and 1x PCR buffer (Qiagen). Primer sequences, concentrations and PCR conditions are summarized in Table 2.S5.

PCR products were visualized on 1.2% agarose gels, purified using the Gel/PCR DNA Fragments extraction kit (AVEGENE, Taipei, Taiwan) and measured by intercalating fluorescence dye using the Qubit Fluorometer (Qubit HS-Kit, Invitrogen, Darmstadt, Germany). After equimolar amplicon pooling, emulsion PCR was performed using Lib-L emPCR protocols. DNA containing beads were recovered, enriched and sequenced on a XLR70 Titanium Pi-

coTiterPlate (Roche) separated into 8 regions, according to the manufacturer's protocols. Reads were extracted from primary sff-files and assigned to the reference sequence. Afterwards, the reads were imported into BiQ Analyzer HT [Lutsik *et al.*, 2011], to filter out low quality reads and analyze the methylation levels and patterns. In total, 340 amplicons were sequenced and 331,768 high quality sequences were obtained with an average conversion rate >99%.

2.2.7 DNA methylation analysis of repetitive elements

DNA methylation levels in the repetitive DNA elements *HERVK* and *LINE1* (not covered by the Illumina Beadchip) were determined using methylation-dependent primer extension assays (SIRPH). The bisulfite PCRs were performed as described in the previous section (see Table 2.S5, for primer sequences, concentrations and PCR conditions). The degree of methylation was determined using single-nucleotide primer extension (SNuPE) assays in combination with ion-pair reversed-phase high-performance liquid chromatography (IP-RP-HPLC) separation techniques (SIRPH) as previously described [El-Maarri *et al.*, 2002]. In brief, after amplification, unincorporated dNTPs and primers were removed by treating 5 µl PCR product with Exonuclease I/SAP mix (1U each, USB) for 30 min at 37°C, followed by an inactivation step of 15 min at 80°C. Afterwards, 14 µl primer extension mastermix (2.0 mM MgCl2, 0.05 mM ddCTP, 0.05 mM ddTTP, 3.6 µM SNuPE primer, 5 U TERMIPol DNA Polymerase (Solis BioDyne, Tartu, Estonia) and 1x reaction buffer C (Solis BioDyne)) were added to the Exonuclease 1/SAP treated PCR product. A primer extension reaction was performed with a primer annealing next to a CpG (in the context of the original genomic sequence) being extended by either a ddCTP or a ddTTP, depending on whether the site was methylated prior the bisulfite treatment and PCR or not. SNuPE reaction conditions and primer sequences are described in Table 2.S6. Obtained SNuPE products were loaded directly on a DNASepTM (Transgenomic, Omaha, USA) column and separated on the WAVE^{TM} system (Transgenomic) using acetonitril gradient elution. The elution gradient parameters were adjusted specifically for each SNuPE primer. Methylation indices (MIs) were obtained by calculating the ratio AC/(AC+AT), where AC and AT is the area under the peak corresponding to the ddCTP and the ddTTP-extended primer, respectively, as calculated by Wave Maker v4.1 (Transgenomic). Two CpG sites per amplicon were analyzed. The assays were validated using DBS of 8 DNA samples (4 chorion and 4 decidua). Correlation coefficients (r) between the SIRPH and the DBS data were for CpG_1 and CpG_2 of *HERVK* and LINE1, 0.98, 0.96, 0.73 and 0.82, respectively.

2.2.8 Whole blood and buccal genome-wide reference methylation data

To generate reference datasets for whole blood and buccal DNA, genome-wide DNA methylation data were obtained from Gene Expression Omnibus (GEO, `http://www.ncbi.nlm.nih.gov/geo/`) (Table 2.S1). As there are currently no directly comparable 450k datasets available, we used data obtained with the Illumina HumanMethylation27 BeadChip (shares 25978 CpGs with the 450k chip). Whole blood DNA methylation data were obtained from 274 postmenopausal female controls (GEO accession number: [GSE19711]) [Teschendorff *et al.*, 2010]. Buccal DNA methylation data were obtained from 60 female samples (mean age ± SD, 15.1 ± 0.5) (GEO accession number: [GSE25892]) [Essex *et al.*, 2013]. Subsequently, genome-wide DNA methylation reference datasets for whole-blood and buccal were generated by averaging the DNA methylation data.

2.2.9 Data analysis

To determine whether the phenotypic characteristics differed significantly between the heavier and lighter co-twins, a paired t-test for continuous data and Fisher's exact test for categorical data were carried out using the statistical package SAS (version 9.2, SAS Institute Inc., Cary, NC, USA). Using the Bioconductor *methylumi* library, the genome-wide methylation data were loaded into R statistical environment for analysis [Gentleman *et al.*, 2004]. Quality control of the Infinium methylation data were assessed using the HumMeth27QCReport library. For all different types of methylation data (i.e. Infinium, DBS and SIRPH), DNA methylation differences between the heavy and light co-twins were tested using the non-parametric Wilcoxon signed-rank test, which tests the null hypothesis that the mean DNA methylation differences are equal to zero (see for details Supplemental methods).

2.2.10 Power calculation

Power analysis of the Wilcoxon signed rank test is complicated, because its power function is difficult to express [Shieh *et al.*, 2007]. Therefore to estimate the power of our analysis we did the calculation for its closest parametric equivalent (paired T-test). With a sample size of 16 discordant twin pairs, 99% power is achieved to detect a mean β-value difference of 0.05 using a two-sided paired T-Test assuming a standard deviation of 0.025 (which is the true standard deviation observed in our data) and a significance threshold of 0.01. The details of this power calculation and calculations using lower significance thresholds are presented in Table 2.S7.

2.3 Results

2.3.1 Phenotypic characteristics of the discordant MZ MC twins

Perinatal, maternal and adult phenotypic characteristics of the 17 spontaneously conceived MZ MC female twins discordant for birth weight are presented in Table 2.1. Compared to the heavier co-twins, the birth weight of the lighter co-twins was on average 698 gram (26.7%) lower (P<0.0001), with absolute and relative intra-pair birth weight differences ranging from 500 to 1000 gram and from 21.3 to 35.7%, respectively. In addition, the frequency of a (para)central umbilical cord insertion was significantly higher in the heavier co-twins (P=0.008). The mean age of the twins when the saliva samples were taken was 34.4 years, the youngest twin pair was 22 years old and the oldest 45 years. The adult phenotypic characteristics body height, body weight and body mass index did not differ between the discordant twins (P>0.05). In the questionnaires, none of the twins reported that they experienced diabetes, cancer, cardiovascular or cerebrovascular disease events.

2.3.2 Exploratory analysis of the Infinium methylation profiles

Genome-wide DNA methylation profiles of the 17 MZ MC twins were established using the Infinium HumanMethylation450 BeadChip assay. After quality control and filtering, methylation data of 478,096 CpG sites were available. In order to identify global DNA methylation changes across the samples, the Infinium methylation data were used to calculate pair-wise array-wide Pearson correlations coefficients for each pair of samples. As depicted in Figure 2.1 and in Figure 2.S1, twin pair 1 showed severe genome-wide DNA methylation changes compared to all other samples, resulting in a relatively low correlation to other samples (r=0.846-0.930) while the intra-pair correlation of this twin pair was high (r=0.996). In addition, the

Table 2.1: Perinatal, maternal and adult phenotypic characteristics of the female MZ twins discordant for birth weight.

Characteristic	Heavier co-twins	Lighter co-twins	Range	P^a
N	17	17		
Perinatal				
Gestational age (wks)[b]	37.9 ± 2.4	37.9 ± 2.4	(34 - 42)	
Birth weight (g)	2619 ± 319	1921 ± 278	(1440 - 3100)	<0.0001
Umbilical cord insertion[b]				
(Para)central	12 (75%)	4 (25%)		
(Para)marginal	4 (25%)	9 (56%)		
Velamentous	0 (0%)	3 (19%)		0.008
Maternal				
Maternal age (yrs)	26.9 ± 5.4	26.9 ± 5.4	(18 - 43)	
Parity	1.8 ± 1.0	1.8 ± 1.0	(1 - 4)	
Adult				
Age (yrs)	34.4 ± 7.1	34.4 ± 7.1	(22 - 45)	
Body height (cm)	166.9 ± 6.1	165.5 ± 6.7	(155 - 177)	0.13
Body weight (kg)	62.7 ± 12.3	61.2 ± 14.2	(47.5 - 102)	0.18
Body mass index (kg/m^2)	22.5 ± 3.8	22.3 ± 4.5	(16.7 - 33.7)	0.60

Continuous data expressed as mean ± SD. Categorical data expressed as: number of observations (%).
[a]Heavy vs. light calculated using a paired T-test for continuous data and Fisher's exact test for categorical data.
[b]Missing for one twin pair.

overall methylation profiles of samples 6_H, 10_H and 12_H deviated such that intra-pair correlation coefficients (r=0.975-0.989) for the pairs 6, 10 and 12 were low compared to otherwise constantly high intra-pair correlation coefficients for all other twin pairs (r=0.992-0.997). Careful analysis of the sample independent and sample dependent Infinium methylation control probes present on the BeadChip (see Figure 2.S2 and 2.S3) revealed that the aberrant methylation profiles observed for some of the samples are unlikely the result of technical failure.

2.3.3 Cellular composition of saliva as a cause of aberrant methylation profiles

Since saliva DNA is derived from leukocytes and epithelial cells [Aps *et al.*, 2002; Thiede *et al.*, 2000], we hypothesized that the deviating DNA methylation profiles observed for some samples were at least partially attributed to inter-sample differences in cell type proportions. We therefore compared our data to genome-wide DNA methylation reference datasets obtained from whole blood and buccal epithelial cells (HumanMethylation27 BeadChip), respectively. Indeed, in contrast to all other samples, the five samples with the most deviating profiles (1_H, 1_L, 6_H, 10_H and 12_H) showed lower array-wide correlation coefficients to the reference dataset for whole blood (r_{norm}=0.943-0.951 *vs.* $r_{deviant}$=0.812-0.939), than to the buccal epithelium reference dataset (r_{norm}=0.875-0.915 *vs.* $r_{deviant}$=0.926-0.975) (see Figure 2.1 and Figure 2.S4). This suggests that the deviating methylation profiles observed for these samples were a consequence of cellular composition differences, i.e. a higher amount of buccal epithelial cells in the respective saliva samples.

Based on this finding we assumed that individual DNA methylation markers specific for buccal epithelium (or conversely for whole blood) can be used to determine the relative amount of buccal epithelium-derived DNA present in the samples. To illustrate that this is a valid assumption we performed an experiment with *in vitro* generated series of two cell type mixtures profiled on the Infinium HumanMethylation450 BeadChip. The results showed that the methylation values of CpGs that are differentially methylated between the two cell types provided a good estimate of the mixed proportions (see Figure 2.S5).

Subsequently, we screened for CpGs that were highly discriminatively methylated between blood and buccal (see for details Supplemental methods) and selected the top 10 most discriminatively methylated CpGs. One of them was cg18384097 in *PTPN7* (protein tyrosine phosphatase non-receptor type 7; β-$value_{Buccal}$ =0.82 and β-$value_{Blood}$ = 0.05), a gene preferentially expressed in hematopoietic cells (Entrez Gene ID: 5778). When correlating the methylation values of every CpG with the methylation levels of the *PTPN7* CpG, we observed that 58,987 CpGs were strongly correlated with the *PTPN7* CpG (|r|>0.8) and 134,265 CpGs were moderately correlated (|r|=0.4-0.8) (see Figure 2.S6). This indicates that for a large number of CpGs a great amount of the observed variation in DNA methylation levels could be explained by variation in cellular composition of the saliva samples.

2.3.4 Adjustment for cell type heterogeneity

Following this observation, we decided to diminish the confounding effect of the saliva cell type composition using linear regression. The detailed procedure is described in Additional file 1. In brief, first a model fitting approach was used to select, out of the top ten most discriminatively methylated CpGs, a marker CpG of which the methylation measurements behave the most linear with respect to the changing cell proportions. In our data the *PTPN7*-associated CpG (cg18384097) gave the best linear fit to the biggest number of CpG positions, and was

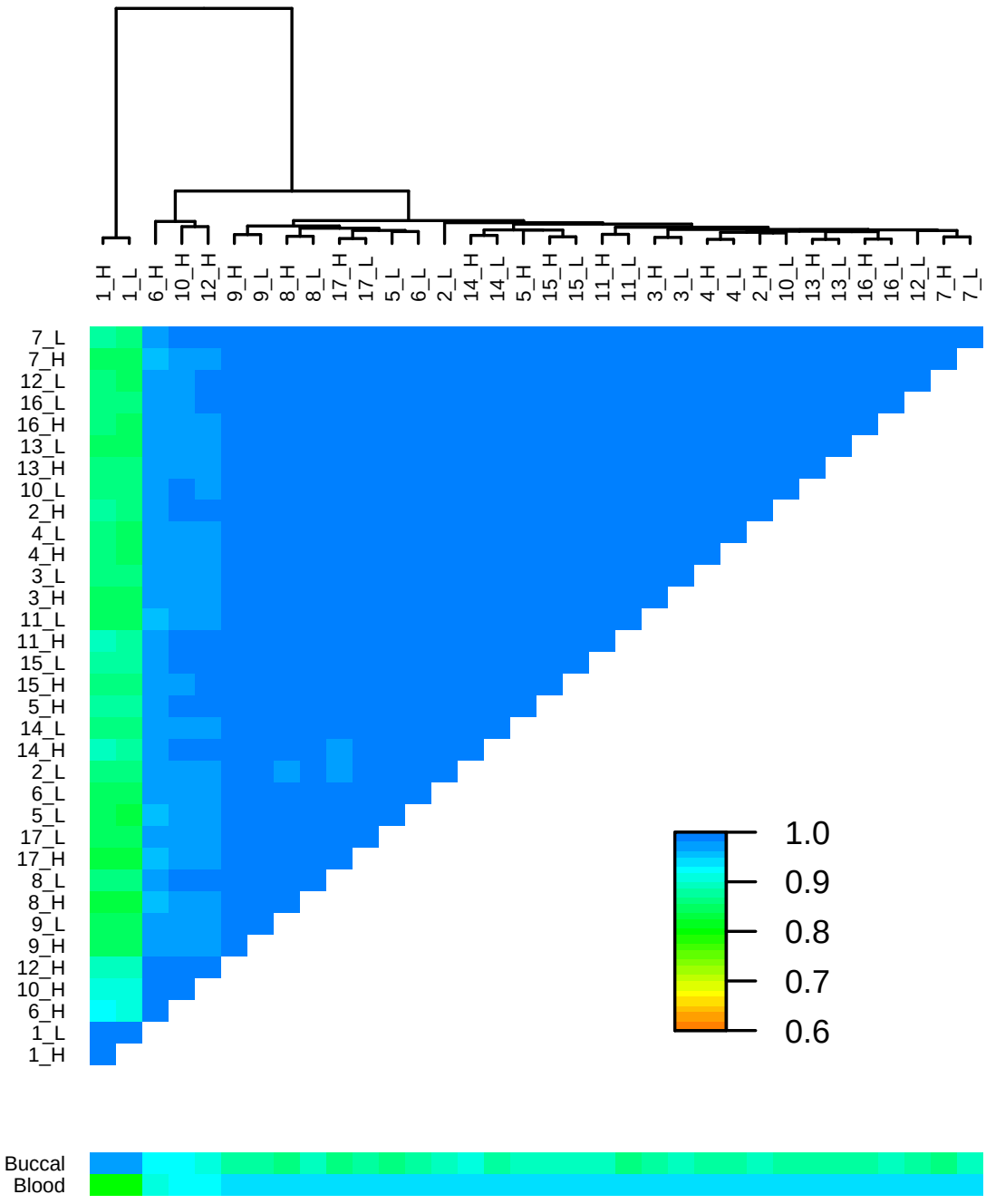


Figure 2.1: Heatmap representing the pair-wise correlations for each pair of samples plus the reference data set for whole-blood and buccal (27K), calculated from ≈25,978 CpGs. H = high birth weight, L = low birth weight

therefore selected as the quantitative marker measuring the proportion of buccal epithelium. Subsequently, this CpG was used to adjust the Infinium data so that the methylation level at each probe becomes linearly independent of the cellular composition in the studied sample. Noteworthy, this procedure adjusted only those CpGs that were significantly affected by the saliva composition (based on the marker model fit), while the methylation levels of the other CpGs (45%) remained unchanged. Due to the extremely high buccal epithelium content in the saliva samples of pair 1, some CpGs showed extreme values which could not be approximated by the linear model and greatly affected the regression slopes. Therefore we excluded pair 1 from the analysis. Interestingly, pair 1 were the only current smokers in the sample and the only individuals with an intensive and long smoking history (>10 cigarettes/day for 25 years). This most probably caused the highly different cell composition in their saliva. The methylation data of the remaining 32 samples got adjusted for the buccal epithelium content and in Figure 2.S7, the pair-wise correlations for each pair of samples of the adjusted data are graphically shown, confirming the robustness of our approach.

2.3.5 Birth weight associated methylation variable positions

Next we tested the hypothesis that poor prenatal conditions lead to significant DNA methylation differences between the heavy and light co-twins for each CpG site independently using the non-parametric Wilcoxon signed-rank test. Figure 2.2 gives the volcano plots, that is, distributions of the resulting *P*-values versus the corresponding mean β-value difference for each CpG position, for both the unadjusted data (a) and the data adjusted for buccal epithelium content (b). The plot documents that large significant DNA methylation differences could not be detected between the heavy and light co-twins (upper corners of both plots are void of data points). In addition, one can also notice that the uneven cellular composition resulted in a considerable amount of CpGs that had relatively big effect sizes but high (non-significant) *P*-values. This confirms that the vast majority of the changes were not associated with birth weight, but were the result of within-pair variation in cellular composition. We therefore proceeded with the results of the data where the affected CpG positions were adjusted for buccal epithelium content variation.

Given the small number of significant changes, we selected a non-stringent significance threshold (uncorrected *P*-value <0.01). In addition, to identify CpG sites likely to be validatable using other methods [Rakyan *et al.*, 2011], we focused on CpG positions that showed an absolute mean β-value difference >0.05. 7859 CpGs out of 478,096 sites had a *P*-value <0.01 with absolute mean β-value differences ranging from 0.0012 to 0.1049 and *P*-values ranging from 0.0092 to 3.05×10^{-5}. Only 131 of these CpGs showed an absolute mean β-value difference >0.05. To exclude the potential influence of other blood-derived cells present in saliva [Vidovic *et al.*, 2012], we extended the adjustment model to include markers of leukocyte subtypes (i.e. neutrophils, B-lymphocytes, CD4+ T-lymphocytes, CD8+ T-lymphocytes and natural killers) (see for details Supplemental methods, Table 2.S1 and 2.S2). In total, 3153 CpGs remained significant (*P*<0.01) in both analyses of which only 45 CpGs showed an absolute mean β-value difference >0.05 (ranging from 0.05-0.08). We treated this set of 45 CpGs, further denoted as "birth weight associated methylation variable positions" (BW-MVPs), as being most likely differentially methylated between the discordant MZ twins, regardless of the cellular composition (see for details Table 2.S3). Subsequently, we tested these BW-MVPs using state-of-the-art technical validation to prove or disprove them being true biological effects.

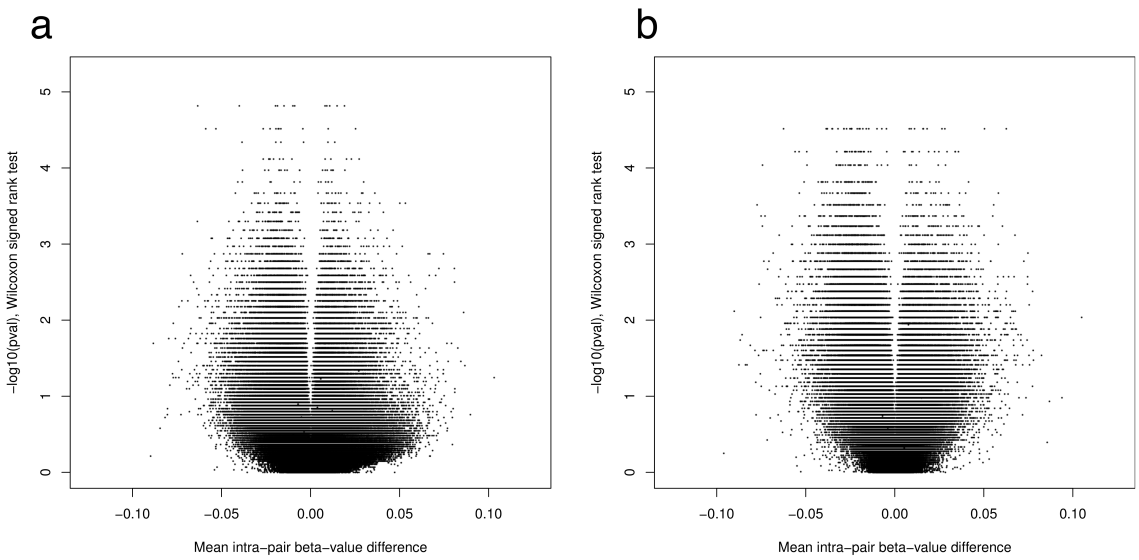


Figure 2.2: Volcano plots of the distributions of the *P*-values resulting from the Wilcoxon signed-rank test versus the corresponding mean β-value difference for each CpG position. **a**. Unadjusted data (72 CpG sites had a *P*-value <0.01 and mean β-value difference >0.05). **b**. Data adjusted for buccal epithelium content using the *PTPN7* CpG (131 CpG sites had a *P*-value <0.01 and mean β-value difference >0.05) (pair 1 was excluded).

2.3.6 BW-MVP validation using deep bisulfite sequencing (DBS)

The 45 BW-MVPs were prioritized for independent DNA methylation validation either based on their biological significance, regulatory relevance and/or whether neighboring probes were also differentially methylated. In total, eight BW-MVPs were validated using DBS and their functional characteristics are summarized in Table 2.2. Some of the prioritized BW-MVPs were situated in genes strongly involved in glucose and/or lipid metabolism or have been implicated in T2D or obesity risk (e.g. *APPL2*, *IGF2BP2*, *PHKG2* and *PPARGC1B*), which is in line with the observation that low birth weight is associated with increased metabolic disease risk. Of the eight selected loci only three were affected (*IGF2BP2*, *PAPOLA* and *PPARGC1B*) by buccal content, so the methylation values of the other five remained unchanged. Moreover, in the unadjusted data the *IGF2BP2* and *PPARGC1B* CpGs were also significantly differentially methylated between the discordant twins. To be able to adjust the DBS data of the *IGF2BP2*, *PAPOLA* and *PPARGC1B* CpGs for buccal epithelium content, the *PTPN7* CpG (buccal epithelium marker) was also analyzed using DBS and simultaneously served as a positive control. High quality DBS data were obtained for all amplicons, with an average sequencing coverage of 988 reads (i.e. individual chromosomal patterns) per amplicon per sample. Examples of the methylation profiles obtained by DBS are given in Figure 2.S8.

Technical performance of both methods was first analyzed by comparing the (unadjusted) DBS data of the validated CpGs with the (unadjusted) Infinium data for every analyzed sample separately. The individual correlation plots are presented in Figure 2.S9, and the correlation coefficients calculated for each sample are summarized in a box plot in Figure 2.S10. Overall, the Infinium data correlates very well with the DBS data, with a median correlation coefficient of 0.97. From both figures it becomes also clear that sample 12_H is an outlier with a correlation coefficient of only 0.70. When we correlate the (unadjusted) Infinium data with the (unadjusted) DBS data for each validated CpG site separately (see Table 2.3 and Additional data file 1, Figure 2.S11 and 2.S12), then the DBS data of the CpG sites in *APBA1*, *APPL2*, *PHKG2*,

Table 2.2: Characteristics of the eight BW-MVPs that were prioritized for further validation using DBS.

CpG Number	Region	Gene/ Element	Full name	Function
cg14123607	Intron 1/insulator	*APBA1*	Amyloid beta A4 precursor protein-binding family A member 1	Associated with reduced production of amyloid-β peptide, which is considered a key player in Alzheimer's disease [Chai *et al.*, 2012].
cg12170649	Intron 2/enhancer	*APPL2*	Adaptor protein, phosphotyrosine interaction, PH domain and leucine zipper containing 2	Involved in cell proliferation and embryonic development. Acts as negative controller of adiponectin signalling and SNPs in *APPL2* have been associated with obesity [Jiang *et al.*, 2012; Miaczynska *et al.*, 2004; Schenck *et al.*, 2008].
cg26404226	Enhancer	Chr10q23.3		
cg15487251[a]	Promoter	*IGF2BP2*	Insulin-like growth factor 2 mRNA-binding protein 2	mRNA binding protein involved in RNA localization, stability and translation. Intronic SNP in *IGF2BP2* has been identified as T2D risk factor by several GWAS [Christiansen *et al.*, 2009].
cg10362113[a]	Intron 1/weak enhancer	*PAPOLA*	Poly(A) polymerase alpha	Plays a predominant role in addition of the 3'-poly(A) tail to mRNA precursors, which is important for mRNA stability, transport and translation [Rapti *et al.*, 2010].
cg02409150	Intron 4	*PHKG2*	Phosphorylase kinase, gamma 2 (testis)	Encodes the gamma subunit of the liver-specific phosphorylase kinase, which activates the enzyme glycogen phosphorylase resulting in glycogen breakdown [Burwinkel *et al.*, 2003].
cg15049370[a]	Intron 2/enhancer	*PPARGC1B*	Peroxisome proliferator-activated receptor gamma coactivator 1-beta	Multifunctional transcriptional co-regulator involved in many metabolic processes, including mitochondrial oxidative metabolism and hepatic lipogenesis [Liu and Lin, 2011].
cg22768222	Intron 1/weak enhancer	*RUNX2*	Runt-related transcription factor 2	Transcription factor essential for osteoblast and chondrocyte differentiation [Komori, 2011].

GWAS = genome-wide association studies, SNP = single nucleotide polymorphism, T2D = type 2 diabetes. [a]Affected by buccal content.

Table 2.3: Validation of eight BW-MVPs and the *PTPN7* CpG using DBS in the 17 discordant MZ twins.

CpG Number	Gene	Infinium		DBS			$r_{Pearson}$	*p*
		Mean[a] ± SD	Range	Mean[b] ± SD	Range	Mean # reads (range)		
cg14123607	*APBA1*	0.19 ± 0.08	0.06 – 0.43	0.09 ± 0.05	0.04 – 0.24	950 (440 – 1670)	0.61	0.0001
cg12170649	*APPL2*	0.83 ± 0.05	0.71 – 0.91	0.97 ± 0.02	0.93 – 0.99	1120 (785 – 1725)	-0.23	0.18
cg26404226	NA	0.52 ± 0.07	0.38 – 0.72	0.35 ± 0.08	0.21 – 0.54	1201 (716 – 2345)	0.73	<0.0001
cg15487251	*IGF2BP2*	0.57 ± 0.08	0.30 – 0.69	0.55 ± 0.10	0.24 – 0.68	1156 (531 – 1824)	0.80	<0.0001
cg10362113	*PAPOLA*	0.83 ± 0.12	0.41 – 0.98	0.84 ± 0.16	0.25 – 0.95	997 (188 – 1689)	0.87	<0.0001
cg02409150	*PHKG2*	0.91 ± 0.05	0.80 – 0.98	0.93 ± 0.02	0.87 – 0.97	749 (294 – 1301)	0.37	0.03
cg15049370	*PPARGC1B*	0.78 ± 0.07	0.62 – 0.98	0.96 ± 0.02	0.93 – 1.00	1019 (522 – 1792)	0.32	0.06
cg18384097[c]	*PTPN7*[c]	0.40 ± 0.17	0.16 – 0.89	0.28 ± 0.20	0.09 – 0.98	848 (383 – 1334)	0.91	<0.0001
cg22768222	*RUNX2*	0.32 ± 0.09	0.05 – 0.55	0.20 ± 0.05	0.02 – 0.26	853 (302 – 1347)	0.55	0.0008

Pearson's correlation coefficients ($r_{Pearson}$) between the (unadjusted) Infinium and the (unadjusted) DBS data are given as well as the overall mean methylation, standard deviation (SD) and the overall methylation range observed for the Infinium and the DBS data.

[a]Infinium data (unadjusted) are expressed as mean β-value ± SD. [b]DBS data (unadjusted) are expressed as mean methylation level ± SD, where the methylation level is calculated by dividing the number of reads in which the particular CpG is methylated by the total number of sequenced reads. [c]Buccal marker. Mean # reads (range) = average number of sequenced bisulfite reads per amplicon (min – max number of sequenced bisulfite reads per amplicon). *P* = *P*-value under H_0: $r_{Pearson} = 0$.

PPARGC1B and *RUNX2* correlate poorly with the Infinium data ($r \leq 0.60$). For the other four CpG sites (Chr10q23.3, *IGF2BP2*, *PAPOLA* and *PTPN7*) the correlation coefficients were higher (r=0.73-0.91) and, as expected, the highest correlation was observed for the buccal epithelium marker (*PTPN7*).

Subsequently, we tested whether the significant DNA methylation differences between the heavy and light co-twins at these BW-MVPs obtained using the Infinium data, could be confirmed with the DBS data. We again performed a Wilcoxon signed-rank test and the results of both technologies are presented in Table 2.4. Unfortunately, in the DBS data no significant DNA methylation differences were observed between the heavy and light co-twins (p>0.01). Moreover, in the DBS data we can also include CpGs neighboring the selected MVPs and also for those CpGs no significant methylation differences between the heavy and light co-twins were observed (data not shown). Since the correlation analysis revealed that sample 12_H suffers from technical problems (Figure 2.S9), we repeated the analysis without pair 12. Still, in the DBS data no significant differences between the heavy and light co-twins were observed, while the differences in the Infinium data remained significant.

2.3.7 HNF4A methylation

In addition we analyzed the hepatocyte nuclear factor 4 alpha (*HNF4A*) promoter, which has previously been identified as differentially methylated between IUGR neonates and controls in a genome-wide scan [Einstein *et al.*, 2010]. Since the significant region was not covered by the Infinium chip, we analyzed it using DBS with a mean sequence coverage of 865 reads ranging from 176 to 1324 reads per sample. The *HNF4A* methylation levels correlated significantly with the buccal epithelium marker *PTPN7*, and therefore the *HNF4A* methylation data were adjusted for *PTPN7* methylation as described earlier. Following such an adjustment none of the 10 CpG sites within the *HNF4A* amplicon retained significant methylation differences between the discordant twins (p>0.01, data not shown).

2.3.8 Global DNA methylation analysis on repetitive elements

Finally, the DNA methylation levels of the genome dispersed repetitive elements human endogenous retrovirus type K (*HERVK*) and long interspersed nuclear element-1 (LINE1) were evaluated using methylation-dependent primer extension assays. For every CpG analyzed, mean methylation indices (MIs, similar to Illumina's β-values) were very similar among the heavy and light co-twins and no significant differences were observed (p>0.05) (Table 2.5). Some CpGs strongly correlated with *PTPN7* methylation (*HERVK* CpG1 r=-0.89, LINE1 CpG1 r=-0.49), indicating that global methylation levels are lower in buccal epithelium. Accordingly, when we repeated the analysis following *PTPN7* methylation adjustment, no significant associations could be detected (data not shown).

2.4 Discussion

We aimed to identify loci that remain differentially methylated in adult body fluid cells as a consequence of a poor prenatal environment. Our hypothesis was that DNA methylation changes induced by adverse intra-uterine conditions are detectable in adult MZ MC twins with large intra-pair weight differences at birth, irrespective of their health status in adulthood. We used Infinium HumanMethylation450 BeadChip to profile DNA methylation changes genome-wide, and applied 454 GSFLX-based single-molecule deep bisulfite sequencing (DBS) to vali-

Table 2.4: Differential methylation analysis of the eight selected BW-MVPs using the Infinium and DBS data.

CpG Number	Gene	Infinium				DBS			
		Mean Heavy Co-twins[a]	Mean Light Co-twins[a]	Mean Difference	p[b]	Mean Heavy Co-twins[c]	Mean Light Co-twins[c]	Mean Difference	p[b]
cg14123607	*APBA1*	0.22 ± 0.07	0.15 ± 0.07	0.07 ± 0.05	0.0008	0.09 ± 0.05	0.09 ± 0.05	0.006 ± 0.04	0.30
cg12170649	*APPL2*	0.81 ± 0.04	0.86 ± 0.02	-0.06 ± 0.05	<0.0001	0.97 ± 0.02	0.97 ± 0.02	-0.002 ± 0.02	0.86
cg26404226	NA	0.50 ± 0.05	0.56 ± 0.07	-0.05 ± 0.04	<0.0001	0.35 ± 0.08	0.36 ± 0.06	-0.01 ± 0.05	0.46
cg15487251	*IGF2BP2*[d]	0.53 ± 0.05	0.58 ± 0.04	-0.05 ± 0.05	0.002[e]	0.65 ± 0.05	0.64 ± 0.06	0.01 ± 0.03	0.04
cg10362113	*PAPOLA*[d]	0.83 ± 0.08	0.77 ± 0.06	0.06 ± 0.07	0.008	0.98 ± 0.04	0.99 ± 0.03	-0.01 ± 0.05	0.50
cg02409150	*PHKG2*	0.88 ± 0.05	0.94 ± 0.03	-0.06 ± 0.05	<0.0001	0.93 ± 0.02	0.93 ± 0.02	-0.006 ± 0.03	0.43
cg15049370	*PPARGC1B*[d]	0.70 ± 0.06	0.77 ± 0.07	-0.07 ± 0.07	0.002[e]	0.96 ± 0.01	0.97 ± 0.02	-0.007 ± 0.02	0.23
cg22768222	*RUNX2*	0.37 ± 0.07	0.31 ± 0.05	0.06 ± 0.07	0.008	0.21 ± 0.03	0.21 ± 0.03	0.001 ± 0.05	1.00

Results of the Wilcoxon signed-rank test of the 8 selected MVPs performed on the Infinium and the DBS data in the 16 MZ twins discordant for birth weight (pair 1 was excluded) are presented. [a]Infinium data are expressed as mean β-value ± SD.[b]Heavy *vs.* light calculated using a Wilcoxon signed-rank test. [c]DBS data are expressed as mean methylation level ± SD, where the methylation level is calculated by dividing the number of reads in which the particular CpG is methylated by the total number of sequenced reads. [d]Infinium and DBS data were adjusted for buccal content (using *PTPN7*) and the adjusted values are presented. [e]Also significant without adjusting for buccal marker. Mean difference = heavy co-twin – light co-twin.

Table 2.5: Methylation analysis of *HERVK* and LINE1 in the 16 discordant MZ twins (pair 1 excluded).

Element	CpG[a]	Mean MI Heavy Co-twins	Mean MI Light Co-twins	Mean MI difference	p[b]
HERVK	1	0.61 ± 0.05	0.63 ± 0.02	-0.02 ± 0.06	0.38
	2	0.36 ± 0.01	0.36 ± 0.01	0.0008 ± 0.006	0.86
LINE1	1	0.58 ± 0.02	0.58 ± 0.02	-0.0007 ± 0.01	0.86
	2	0.37 ± 0.02	0.38 ± 0.02	-0.004 ± 0.02	0.50

All procedures (including bisulfite treatment, PCR and SIRPH assay) were performed in duplicate; the mean values were used for the statistical analysis. Data are expressed as mean MI ± SD. MI = methylation index, MI difference = MI heavy co-twin − MI light co-twin. [a]CpG 1 and 2 corresponds to the CpG tagged by SNuPE primer 1 and 2, respectively. [b]Heavy *vs.* light calculated using a Wilcoxon signed-rank test.

date potential methylation variable positions (MVPs). To assess possible changes in repetitive element methylation, we applied bisulfite-based primer extension HPLC assays (SIRPH). Despite cellular composition differences, our thorough DNA methylation analyses show that the methylomes in saliva of birth weight discordant MZ MC twins are very similar.

All analyses were performed on DNA isolated from saliva, a bio-fluid that contains adequate amounts of DNA and is easy accessible via a totally non-invasive method. Assuming that MVPs are maintained in a systemic way, saliva DNA should be suitable for diagnostic and prognostic purposes, like any other accessible body fluid such as blood. However, we observed that the composition of saliva can be highly variable, possibly causing confounding effects since DNA methylation signatures are cell type-specific. Such hardly controllable effects should be accounted for, when studying the association of DNA methylation to the phenotype of interest. This can either be done by separating cells prior to the methylation analysis, which is usually a challenging experimental task, or by using methods that allow a post-sampling adjustment for cellular composition. Indeed, here we show that cell type-specific epigenetic signatures of cells can be used for such post-sampling adjustment.

When (young) MZ twins are used for epigenetic studies, tissues other than blood are preferred (often buccal epithelium). This is because MZ twins often have a shared blood supply during intra-uterine development, and therefore epigenetically discordant MZ twins can display the same epigenetic defect in blood while for instance in fibroblasts the epigenetic defect is restricted to the affected twin only [Tierling *et al.*, 2011; Weksberg *et al.*, 2002]. Interestingly, Kaminsky et al. [Kaminsky *et al.*, 2009] observed that methylation profiles of buccal swab DNA are significantly more variable within MZ MC twins compared to MZ DC twins. They suggested that this epigenetic dissimilarity may reflect differences in epigenetic divergence among embryonic cells at the time of splitting. Since buccal swabs also contain saliva [Thiede *et al.*, 2000], our results indicate that the previously reported epigenetic differences within MZ MC twins observed by Kaminsky et al. [Kaminsky *et al.*, 2009] might be caused by sample composition-attributed variation of e.g. leukocytes and epithelial cells [Thiede *et al.*, 2000], rather than a real developmental difference.

Our results show that MZ twins have very similar genome-wide DNA methylation profiles. After controlling for sample composition-attributable variation, we obtained 3153 CpGs that were differentially methylated between the heavy and light co-twins with nominal significance (p<0.01) of which only 45 CpGs showed an absolute mean β-value difference >0.05. To verify whether these loci were true BW-MVPs, eight of these 45 loci were validated using state-of-the-art targeted DBS. When correlating the Infinium with the DBS data for each

individual separately, the two technologies gave consistent results and the correlations were high. However, the DBS data did not replicate the DNA methylation differences between the heavy and light co-twins. Nevertheless, when correlating the Infinium with the DBS data for each validated CpG site separately, we observed a wide range of Pearson correlation coefficient values. The highest correlation was observed for the CpG in the buccal epithelium marker *PTPN7* (r=0.91), which served as a positive control, and the buccal content-affected *PAPOLA* and *IGF2BP2* CpGs (r=0.87 and r=0.80, respectively) (Table 2.3). This indicates that true biological variation, linked to the variation of the cell type proportions in saliva samples, is confirmed by DBS. On the other hand, the decreasingly significant correlation values observed for the remaining CpGs indicates low or absent true biologically-meaningful variation in the measured DNA methylation levels. The fact that DBS did not replicate the differences between heavy and light co-twins, might thus indicate that the few BW-MVPs identified using the HumanMethylation450 assay are the result of technical noise, that is, false positives. This is coherent with the fact that in case a more stringent significance criteria would have been used to correct for multiple testing, none of the BW-MVPs would have been called significant. In addition, DAVID tool did not identify enrichments in any of the numerous functional annotation categories for the genes underlying the BW-MVPs [Huang *et al.*, 2009], indicating that there was no evidence of coordinated DNA methylation changes at BW-MVPs that would reflect potential regulation events in groups of loci.

Compared to other widely used genome-wide methylation profiling technologies that are based on methylation-sensitive restriction digestion (HELP, CHARM) or affinity-based enrichment (MeDIP, MethylCap), the Infinium assay has ahigher resolution (single base pair) and therefore expected to have a higher sensitivity [Rakyan *et al.*, 2011]. However, whether the Infinium assay is sensitive enough to distinguish between an absolute mean β-value difference 0.05-0.07 is unclear. Bibikova et al. [Bibikova *et al.*, 2009] estimated that with the Infinium HumanMethylation27 BeadChip on average β-value differences of 0.14 or larger can be detected, with a higher sensitivity at unmethylated and fully methylated sites (e.g. at unmethylated promoters on average β-value changes of 0.07 were detectable). We made an attempt to estimate the technical noise level by examining the 64 SNP probes that are present on the chip. For the twin samples that were heterozygous, the SNP probes showed an absolute intra-pair mean β-value difference of 0.00-0.03. Hence, trying to replicate absolute mean β-value differences of 0.05-0.07 can be a realistic goal, assuming that all probes on the chip perform as good as these SNP probes. Nonetheless, this is a strong assumption and a number of technical issues are likely to undermine the performance of the HumanMethylation450 BeadChip, e.g. differences between the Infinium I and II technologies, multiple CpGs in the probe sequences, cross-hybridization of (repetitive) sequences (e.g. *PPARGC1B* contains *Alu* element) [Chen *et al.*, 2013; Dedeurwaerder *et al.*, 2011]. On the other hand, whether DBS, which is currently considered as the gold standard, is sensitive enough to replicate a 5% methylation difference is also questionable. On average we obtained 988 high quality reads per sample per amplicon, thus the lack of replication is unlikely to be the result of low quality DBS data. Still, in three of the sequenced amplicons informative SNPs were present and for the heterozygous twins a mean absolute intra-pair allele frequency difference of 0.05-0.08 was observed. This indicates that DBS also suffers from technical variation, which is probably the result of random bias induced by PCR amplification. Since all currently used diagnostic methods are PCR-based, focusing on small methylation differences might currently not be worthwhile. These aspects should be more carefully considered in EWAS. Taken together, we cannot exclude the possibility that the BW-MVPs identified in this study are false positives. The fact that the detected differences are on the border of technical variation, makes it unlikely that they can be regarded

as biologically significant.

The first EWAS for birth weight was performed on CD34+ hematopoietic stem cells from cord blood of five IUGR neonates and five controls using the HELP assay and identified moderate changes at 56 loci [Einstein *et al.*, 2010]. The authors validated only one locus, the *HNF4A* promoter, using another technology (bisulfite MassArray). We also analyzed this locus using DBS, but observed no significant differences between the discordant twins. Since the authors studied CD34+ hematopoietic stem cells from cord blood, our negative results might indicate that the changes they observed are not maintained until adulthood or that they are specific for CD34+ hematopoietic stem cells. On the other hand, they only observed a 6% methylation difference between IUGR neonates and controls [Einstein *et al.*, 2010], which remains according to our technical observations difficult to replicate.

Currently, six genome-wide DNA methylation studies for birth weight have been published of which the details are presented in Table 2.6 [Adkins *et al.*, 2012; Banister *et al.*, 2011; Einstein *et al.*, 2010; Fryer *et al.*, 2011; Gordon *et al.*, 2012; Turan *et al.*, 2012]. All of them studied fetal tissues i.e. umbilical cord blood, umbilical vascular endothelial cells and/or placenta. One study used the HELP assay [Einstein *et al.*, 2010], while the other five studies used the HumanMethylation27 BeadChip. In addition, one of the studies [Gordon *et al.*, 2012] also used a twin design (18 MZ and 10 DZ), however the authors did not select the twins based on birth weight discordancy and therefore the mean relative intra-pair birth weight difference of the MZ twins included in their study was only 10.1%. All these EWAS studies for birth weight reported a number of differentially methylated loci. However, only one gene (*PRSS3*) was reported by more than one study [Einstein *et al.*, 2010; Fryer *et al.*, 2011] and none of the loci identified in these studies was significant in our analysis. Moreover, none of these studies performed an intense technical validation comparable to the one made in our study or controlled for sample composition-attributed variation. In summary, while the number of potential candidate loci that become differentially methylated due to an adverse prenatal environment is rapidly growing, the validity of these effects, perhaps with the exception of *PRSS3* that was reported by two studies, is questionable since there is no overlap between the reported loci.

Some studies also examined the relation between birth weight and global DNA methylation levels by assessing repetitive elements. Fryer et al. [Fryer *et al.*, 2011] observed in 12 cord blood samples that LINE1 methylation was higher among the heavier newborns. However, Michels et al. [Michels *et al.*, 2011] observed in cord blood of 319 newborns a significant correlation between low birth weight, high birth weight and preterm birth with reduced LINE1 methylation, while in placental tissue they observed that low birth weight individuals had higher LINE1 methylation compared to normal birth weight individuals. In addition, Wilhelm-Benartzi et al. [Wilhelm-Benartzi *et al.*, 2012] observed in 184 placenta samples a positive association between LINE1 and AluYb8 methylation and birth weight. We did not observe any significant differences in LINE1 and *HERVK* methylation between the heavy and light co-twins, but we observed differences in LINE1 and *HERVK* methylation between leukocytes and epithelial cells. Further studies should consider such sample composition-attributed variation as it might be responsible for the inconsistent reports concerning global DNA methylation and intra-uterine growth.

The majority of the genome-wide methylation studies for birth weight published thus far used a population based design [Adkins *et al.*, 2012; Banister *et al.*, 2011; Einstein *et al.*, 2010; Fryer *et al.*, 2011; Turan *et al.*, 2012], thus their outcome variable birth weight suffers from variation induced by gestational age, gender, maternal factors (e.g. maternal weight, age, parity) and, most importantly, genetic differences. For all these factors our twin study is controlled

Table 2.6: Genome-wide DNA methylation studies for birth weight.

Study	Design	Sample	Tissue	Method	Significant loci	Remark
Einstein et al. [Einstein *et al.*, 2010]	Population based	5 IUGR & 5 AGA	CD34+ hematopoietic stem cells (cord blood)	HELP	56 loci (p<0.00001)	
Banister et al. [Banister *et al.*, 2011]	Population based	89 IUGR & 117 AGA	Placenta	HM27	22 loci (number predetermined)	
Fryer et al. [Fryer *et al.*, 2011]	Population based	12 newborns	Cord blood	HM27	304 loci (p<0.05)	Samples were selected to give a range in LINE1 methylation values.
Adkins et al. [Adkins *et al.*, 2012]	Population based	201 newborns	Cord blood	HM27	10 loci (p<0.001)	
Turan et al. [Turan *et al.*, 2012]	Population based	48 newborns	Cord blood & placenta	HM27	23 loci	Regularized regression model fit was used (R^2>0.80).
Gordon et al. [Gordon *et al.*, 2012]	Twin design	18 MZ & 10 DZ[a]	CBMCs, placenta & UVECs	HM27	7 loci in DZ CBMCs 1 loci in MZ UVECs (FDR<0.1)	Twins not selected for birth weight discordancy.

AGA = appropriate for gestational age, CBMCs = cord blood mononuclear cells, DZ = dizygotic twins, FDR = false discovery rate, HELP = HpaII tiny fragment Enrichment by Ligation-mediated PCR assay, HM27 = HumanMethylation27 BeadChip, IUGR = intra-uterine growth retarded, MZ = monozygotic twins, UVECs = umbilical vascular endothelial cells. [a]Maximal 18 MZ and 10 DZ twins per tissue.

and hence methylation variations associated with them should be eliminated. Since we fail to identify any major methylation change, the contribution of such variable factors in data interpretation should be considered more carefully. One might argue that our negative outcome is the result of the shared intra-uterine blood supply of MZ MC twins, which "diluted" any differential methylation signals. Nevertheless we are aware of this problem and focused on adult twins, since we earlier observed that epigenetic discordance in MZ MC twins, that even suffered from twin-to-twin transfusion syndrome (TTTS), becomes measurable in saliva when they grow older [Tierling *et al.*, 2011]. Moreover, for twin pair 2 it was recorded that they suffered from TTTS *in utero*. In case the severely unbalanced intra-uterine blood flow would still have an impact on their leukocyte populations, then for this pair one would expect to see a higher intra-pair correlation, which was not higher than expected on average (see Figure 2.S7).

Since none of the twins reported in the questionnaire to suffer from acute diabetes, cancer, cardiovascular or cerebrovascular diseases, one might reason that our approach enriched for healthy individuals. However, the twins were only recruited based on being very discordant for birth weight and adult health status was never used as an inclusion criteria. In addition, the EFPTS is a prospective and population-based twin registry [Derom *et al.*, 2006]. Therefore neither our selection strategy nor the EFPTS ever enriched for healthy individuals. In addition, a medical examination was not conducted for this study, therefore the actual adult health status of the twins is unknown. Our twin sample is also relatively young and clear symptoms are expected to appear at later age. Moreover, manifestation of metabolic disorders is strongly related to lifestyle factors.

Finally, through the post-hoc power calculation presented in Table 2.S7, we demonstrate that our approach has sufficient power to detect an absolute mean β-value difference of at least 0.05. To exclude that our negative study outcome is the result of a high false negative rate, we applied a nominal significance threshold of 0.01 (gives approximately 99% power). Note that the statistical analyses show that our design can also easily detect smaller methylation differences, since 3108 out of the 3153 CpGs having a p<0.01 in the final analysis showed an absolute mean β-value difference below 0.05. However, these small differences are not reproducible using DBS and thus remain in the range of technical inaccuracy.

2.5 Conclusions

Our study is based on the assumption that methylation changes caused by a poor prenatal environment remain throughout life in many cell types (systemic). The fact that we used saliva instead of whole blood, is in this respect an advantage since saliva contains ectoderm- and mesoderm derived cells, while blood only contains mesoderm-derived cells. Nevertheless, our negative results might indicate that the methylation differences are restricted to biologically relevant metabolic tissues (e.g. pancreas, liver, muscle, adipose tissue) and thus absent in cells composing saliva. It is also possible that the methylation differences are temporary and do not maintain until adulthood. Due to placental blood vessel connections, blood of young MZ MC birth weight discordant twins is not suitable for epigenetic studies. Studying young MZ DC twins would be an alternative as they do not experience intra-uterine vascular connections, but they are more rare (33% of all MZ twins) and have smaller intra-pair birth weight differences. On the other hand, the HumanMethylation450 BeadChip covers just approximately 2% of all CpGs in the genome and gene bodies and regulatory intergenic regions are underrepresented on the chip. In addition, birth weight discordancy in MZ MC twins can arise from different pathologies and in this respect our group is certainly not homogeneous (e.g. differ-

ent locations of umbilical cord insertion). Despite of these limitations, we can conclude that genome-wide and locus specific DNA methylation perturbations are small and not abundant in cells composing saliva (i.e. epithelium and leukocytes) of individuals that experienced severe intra-uterine growth restriction.

2.6 Accession codes

HumanMethylation450 profiles of the twin samples are available in GEO under accession no. [GSE39560]. The DBS data is available in the Sequence Read Archive under accession no. [SRA075928].

2.7 Supplementary Data

Supplemental Methods

Infinium HumanMethylation450 data pre-processing

The raw data from the Illumina scanner after the internal image processing stage were loaded for initial analysis into Genome Studio software (Illumina, San Diego, CA, USA). The background level – defined as the 5^{th} percentile of the negative control probe signal distribution – was subtracted from the intensities at each probe in each channel separately, setting the negative values to 0. The intensities were then normalized by multiplying the intensity at each probe by the scaling factor. The latter was defined for each sample as the ratio of the mean intensity of (positive) normalization control probes in this sample to same mean intensity in an arbitrary-selected reference sample ("normalization to internal controls"). The data was exported as Genome Studio analysis report, and loaded into R using the *methylumi* library (`http://bioconductor.org/packages/2.5/bioc/html/methylumi.html`). The sample-dependent and sample-independent quality control probes were visualized using the *HumMethQC27* library (see Figure 2.S2 and 2.S3). Low-quality probes were defined as those having a detection *p*-value (one minus quantile of the negative probe intensity distribution into which the intensity of the given probe falls) greater than or equal to 0.001 in one of the samples, and were excluded. The methylation level for each probe was estimated as β-value and M-value [Du *et al.*, 2010] using *methylumi* routines. We set the offset α, added to the intensities of methylated and unmethylated probes in order to decrease the influence of the extremely low intensities on the resulting methylation levels, to be equal 25.

Adjustment for cell type heterogeneity

Since the uneven cellular composition could undermine the downstream statistical analysis, we decided to normalize the Infinium data so that the methylation level at each probe becomes linearly independent of the cellular composition in the studied sample. In order to do this, we first selected quantitative cell type-specific markers using reference data sets of purified cell types, supposedly present in saliva [Vidovic *et al.*, 2012]. We composed the reference data set of publicly available DNA methylation profiles obtained with the Illumina Infinium HumanMethylation27 BeadChip (**Table S1**). The buccal epithelium data was obtained from 60 female samples generated by Essex et al. [Essex *et al.*, 2013]. Leukocyte data was obtained from Calvanese et al. [Calvanese *et al.*, 2012], who generated HumanMethylation27 profiles from

purified leukocyte-subtypes (e.g. neutrophils, B-lymphocytes, CD4+ T-lymphocytes, CD8+ T-lymphocytes and natural killers) that were isolated from the same blood pool. Selection of cell type-specific markers was performed on the set of probes, which is shared between the HumanMethylation27 and HumanMethylation450 BeadChip.

Our method is based on the assumption that the methylation level of CpGs that are highly discriminatively methylated between buccal epithelium and leukocytes provide an estimate of the amount of buccal epithelial cell derived DNA present in the saliva samples. As a consequence, such marker CpGs can be used to adjust the methylation data for the varying cell type composition of the saliva samples. We are aware that alternative approaches can be used to estimate the proportion of buccal derived DNA present in the samples (e.g Principal Component Analysis). However, we prefer to adjust using a "physical" marker (a real CpG). which has the advantage that in the validation phase the marker can be measured using another technology (e.g. deep bisulfite sequencing) and thus also in samples of which no array data (so no PCA estimated cell type proportion) is available.

Although in theory, the relation between the β-value of the marker CpG and the proportion of buccal derived DNA should be linear, in reality due to technical reasons the precise behavior of the Infinium methylation data with changing cell type composition can be quite different. This is illustrated by a mixing experiment with KG1a and K562 cells, of which the mixed proportions were profiled on the Infinium HumanMethylation450 BeadChip. In Figure 2.S12, the β-values of marker CpGs that were hypermethylated in K562 (and thus hypomethylated in KG1a) are plotted against the corresponding mixing proportions. Figure 2.S12 indicates that the methylation values of the marker CpGs provide a good estimate of the mixed proportions, but it also shows that many probes do not behave linear. Besides, we observed that marker CpGs often have a different range of variation, e.g. β-value Buccal/ β-value Blood Marker_1 = 0.80/0.20, Marker_2 = 0.90/0.05. Accordingly, an intuitive approach like averaging methylation data of different marker CpGs will not reduce noise and therefore not result in a good cell composition surrogate for adjustment. Instead we used a model fitting approach to identify that single marker CpG of which the methylation measurements behave the most linear with respect to the changing cell proportions.

The identification of the marker CpGs was performed as follows:

1. For each cell type c from the reference data set, all marker CpGs were ranked by the absolute difference between the methylation level in c (M_c) and the average of methylation levels in all other contributing cell types ($\bar{M}_a$):

$$\bar{M}_a = \frac{1}{m-1} \sum_i M_i \tag{2.1}$$

   where M_i is the methylation value of the CpG in the contributing cell type i of all cell types except c, and m is the number of contributing cell types in the sample.

2. For every considered cell type the top 10 highest ranking CpGs were selected as quantitative marker candidates.

3. The optimal quantitative marker was determined as the one giving the best linear fit to the majority of measured CpGs in the saliva data set. For all the 450k probes, 11 one-parameter linear models were fitted, one for each candidate marker and the intercept-only ("zero") model. Then the fit of the 11 models, measured by Akaike information criterion (AIC), was compared and the model with the lowest AIC was considered to

give the best linear fit. The candidate marker CpG which was most often generating the best fitting model, was considered as the most linear quantitative marker.

We realize that the marker CpGs could potentially be affected by genetic differences between the samples of Essex et al. [Essex *et al.*, 2013] and the pooled sample of Calvanese et al. [Calvanese *et al.*, 2012], i.e. due to SNPs in the probes or allele specific methylation (meth-QTLs). In order to exclude this, we carefully examined the methylation levels of the final marker CpGs in the twin data, in which such effects can easily be detected through the small intra-pair and large inter-pair beta-value differences that are observed in case of a SNP or meth-QTL. In addition, we also examined large whole blood DNA methylation (n=274) and buccal DNA methylation (n=60) data sets of Teschendorff et al. [Teschendorff *et al.*, 2010] and Essex et al. [Essex *et al.*, 2013], respectively. Moreover, we did an extensive annotation search for potential presence of polymorphisms and/or other genetic variability features in the HumanMethylation450 probe of the marker CpG. The results allowed to be confident that the selected marker CpGs are not affected by genetic variation.

In case the methylation levels of the marker CpGs are correlated (confounded) with the phenotype (low birth weight vs. high birth weight), the adjustment procedure will also remove the DNA methylation changes induced by the intra-uterine growth restriction. However, we verified this and none of the marker CpGs is associated with the phenotype ($p_{\text{paired t-test}}$>0.10, $p_{\text{Wilcoxon signed rank}}$>0.10).

Finally, the genome-wide DNA methylation data, expressed as M-values, was adjusted for variation introduced by cellular composition differences using standard linear regression. The selected cell type-specific markers (Table 2.S2) were used as explanatory variables in the linear regression model, with the methylation levels measured at each probe as the response variable. First the adjustment was performed using the buccal epithelium marker only. For a given probe the levels were only adjusted if the AIC of the one-parameter model was lower than the AIC of the intercept-only, "zero" model (i.e. CpGs not affected by saliva composition do not get adjusted by this procedure). The residuals of the fitted adjustment model incremented with the intercept term were treated as the adjusted methylation levels. The second adjustment was done using all selected quantitative markers (Table 2.S2) as predictors and the best adjustment model for each probe was selected from the set of all possible one- and two-parameter models plus the "zero" model. The adjusted methylation levels were transformed into β-values via the logit-transformation.

Association analysis and candidate selection

The hypothesis that poor prenatal conditions lead to significant DNA methylation differences between the heavy and light co-twins was tested using the non-parametric Wilcoxon signed rank test on the intra-pair β-value differences obtained for each CpG site independently. The intra-pair difference was defined as β-value observed for the heavier co-twin minus the β-value of the lighter co-twin, and the two-sided null was tested for the symmetry of the sample distribution around zero. The calculations were performed using the standard R function *wilcox.test*. First the Wilcoxon signed rank test was performed on the dataset that was only adjusted for the buccal marker. Subsequently, the analysis was repeated on a dataset that was adjusted for additional blood derived cellular subtypes. CpGs that were significant in both analysis (p<0.01) and showed an absolute mean β-value difference >0.05 in both analyses were considered as being stably differentially methylated between the discordant MZ twins. The non-stringent nominal significance threshold of 0.01 was chosen to limit the number of false

negatives. To compensate for the substantially increased false positive rate due to multiple testing, we carried out an extensive technical validation.

Supplemental Figures

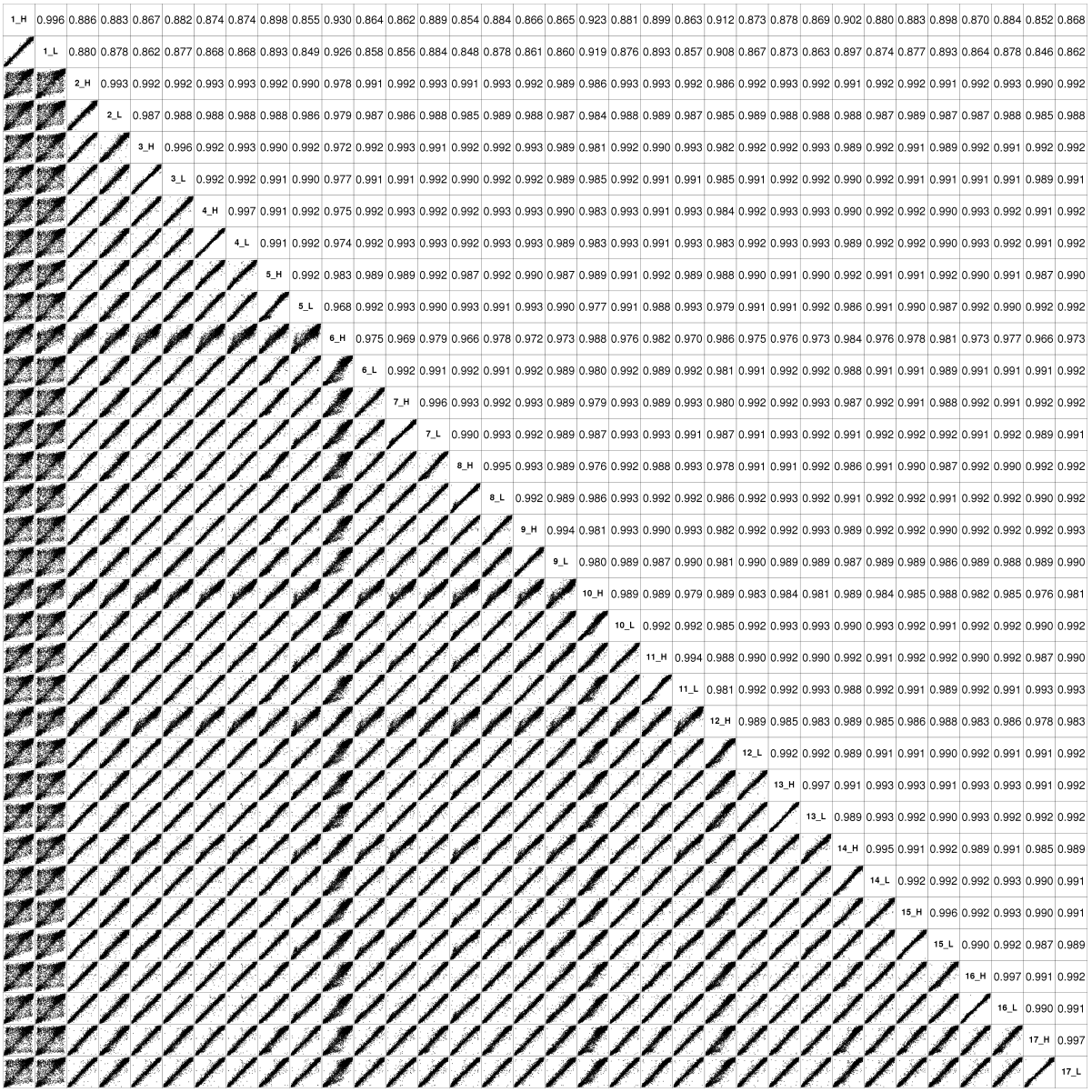


Figure 2.S1: Pair-wise correlations for each pair of samples, calculated from ≈480,000 CpGs. Sample labels are shown on the diagonal. Pearson correlation coefficients are shown in the upper part of the figure and the dotplots under the diagonal illustrate a visual representation of the similarity between two samples (H = high birth weight, L = low birth weight)

.

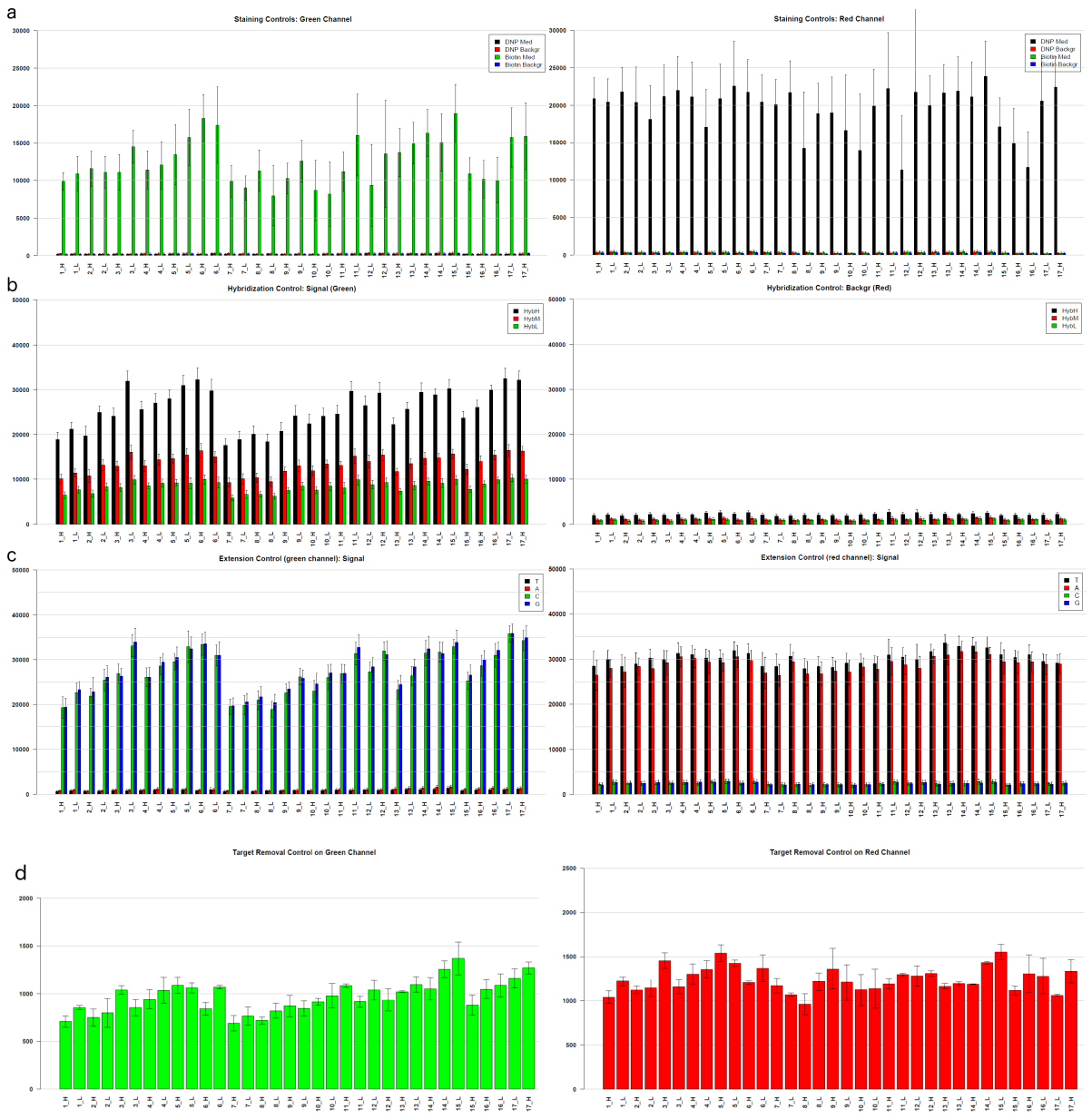


Figure 2.S2: Sample-independent Infinium methylation controls. Staining controls are used to examine the efficiency of the staining step in both the red and green channels, and are independent of the hybridization and extension step. **b**. Hybridization controls test the overall performance of the Infinium assay using synthetic targets that are present in the hybridization buffer at three levels (high (5 pM), medium (1 pM) and low concentration (0.2 pM)) and complement the sequence on the array perfectly, which allows the probe to extend on the synthetic target as a template. The performance of the hybridization controls should be monitored only in the green channel. **c**. Extension controls test the extension efficiency of A, T, C, and G nucleotides from a hairpin probe, and their performance should be monitored in the red (A,T) and green (C,G) channels. **d**. Target removal controls test the efficiency of the stripping step after the extension reaction. Target removal controls are present in the hybridization buffer RA1 and these oligos are extended using the probe sequence as a template. This process generates labelled targets and extension from the probe does not occur. All target removal controls should result in low signal compared to the hybridization controls, indicating that the targets were removed efficiently after extension. H = high birth weight, L = low birth weight.

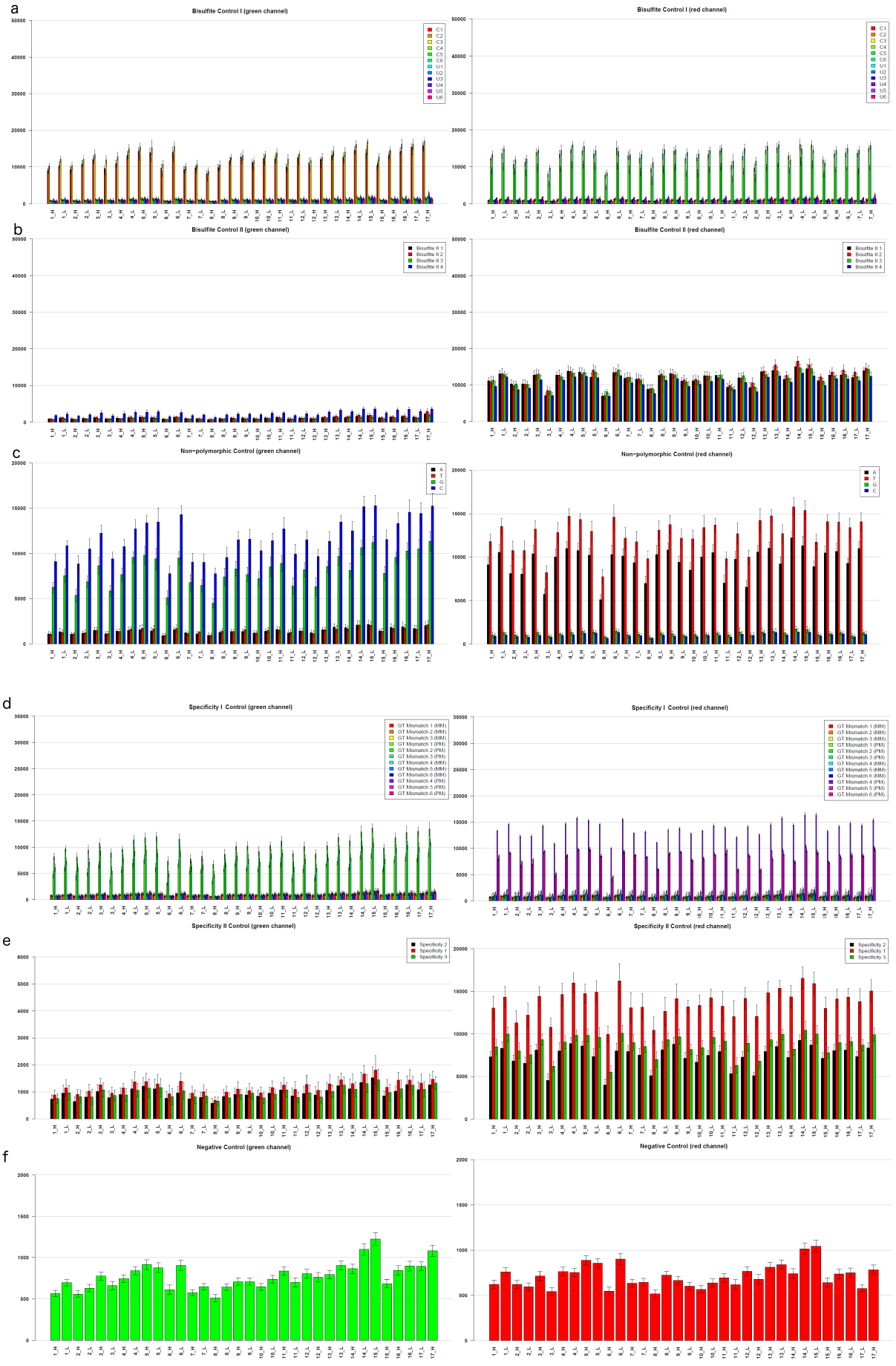

a
Bisulfite Control I (green channel)
Bisulfite Control I (red channel)
b
Bisulfite Control II (green channel)
Bisulfite Control II (red channel)
c
Non−polymorphic Control (green channel)
Non−polymorphic Control (red channel)
d
Specificity I Control (green channel)
Specificity I Control (red channel)
e
Specificity II Control (green channel)
Specificity II Control (red channel)
f
Negative Control (green channel)
Negative Control (red channel)

Figure 2.S3: (on the previous page) Sample-dependent Infinium methylation controls. a. Bisulfite controls I use the Infinium I probe design to monitor the efficiency of the bisulfite conversion. If the bisulfite conversion was successful, the "C" (converted) probes will match the converted sequence and get extended. If the sample has unconverted DNA, the "U" (unconverted) probes will get extended. Performance of bisulfite controls C1, C2 and C3 should be monitored in the green channel, and controls C4, C5 and C6 should be monitored in red channel. b. Bisulfite controls II use the Infinium II probe design to monitor efficiency of bisulfite conversion. If the bisulfite conversion reaction was successful, the "A" base will get incorporated and the probe will have intensity in the red channel. If the sample has unconverted DNA, the "G" base will get incorporated across the unconverted cytosine, and the probe will have an elevated signal in the green channel. **c**. Non-polymorphic controls test the overall performance of the assay, from amplification to detection, by querying a particular base in a non-polymorphic region of the genome. They allow comparing assay performance across different samples. One non-polymorphic control has been designed for each of the four nucleotides (A, T, C, and G). **d**. Specificity I controls are designed to monitor extension specificity for the Infinium I probes. G/T mismatch controls check for non-specific detection of methylation signal over unmethylated background. PM controls correspond to A/T perfect match and should give a high signal. MM controls correspond to G/T mismatch and should give a low signal. Performance of GT mismatch controls should be monitored in both green and red channels. **e**. Specificity II controls are designed to monitor extension specificity for the Infinium II probes. Specificity II probes should incorporate the "A" base across the non-polymorphic T and have intensity in the red channel. In case of nonspecific incorporation of the "G" base, the probe will have elevated signal in the green channel. **f**. Negative controls target bisulfite-converted sequences that do not contain CpG dinucleotides. Assay probes are randomly permuted and should not hybridize to the DNA template. The mean signal of these probes defines the system background. The performance of the negative controls should be monitored in both the green and red channel. H = high birth weight, L = low birth weight.

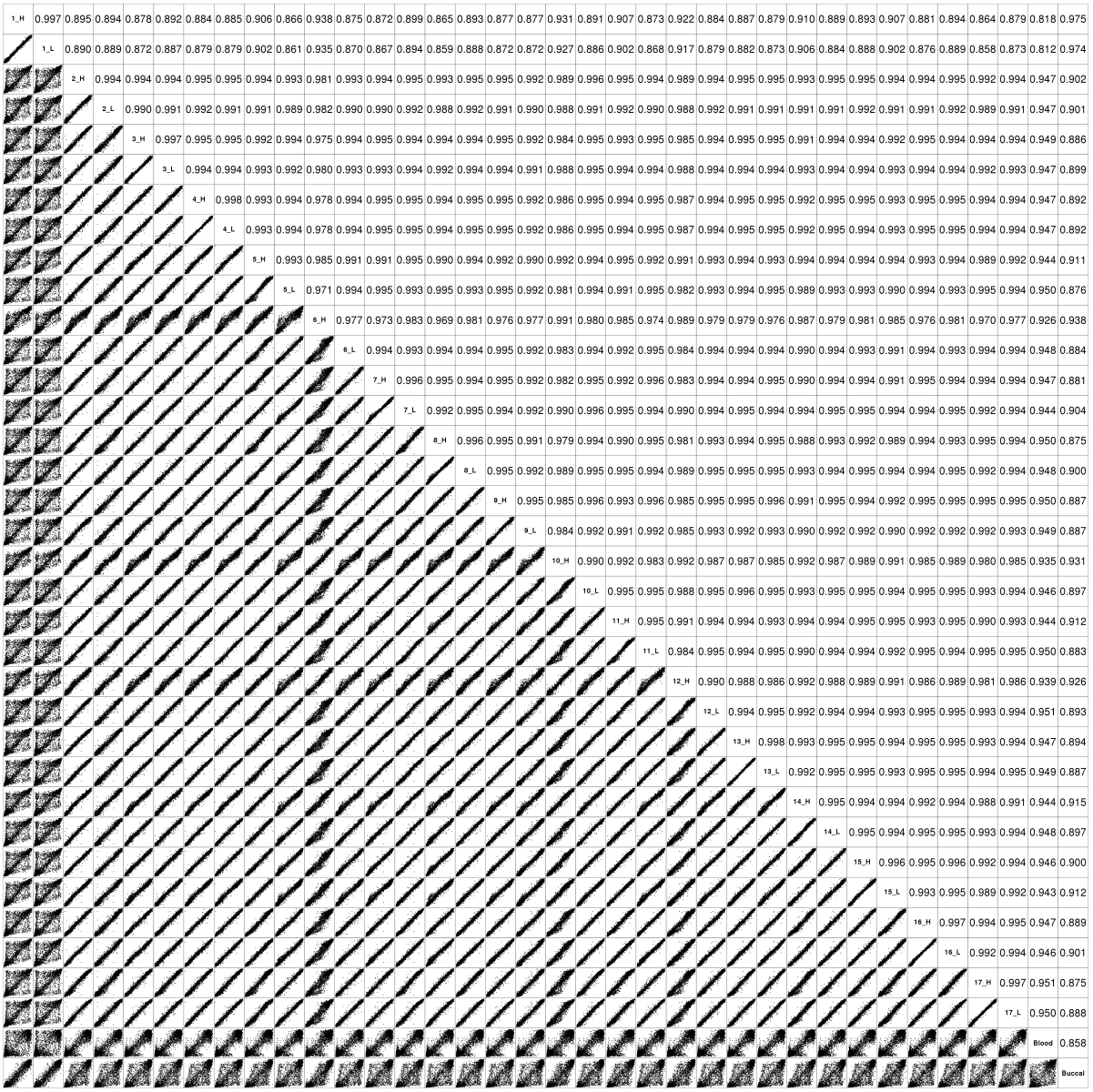


Figure 2.S4: Pair-wise correlations for each pair of samples, including the reference dataset for whole-blood and buccal (27k), calculated from ≈25,978 CpGs. Sample labels are shown on the diagonal. Pearson correlation coefficients are shown in the upper part of the figure and the dotplots under the diagonal illustrate a visual representation of the similarity between two samples. H = high birth weight, L = low birth weight.

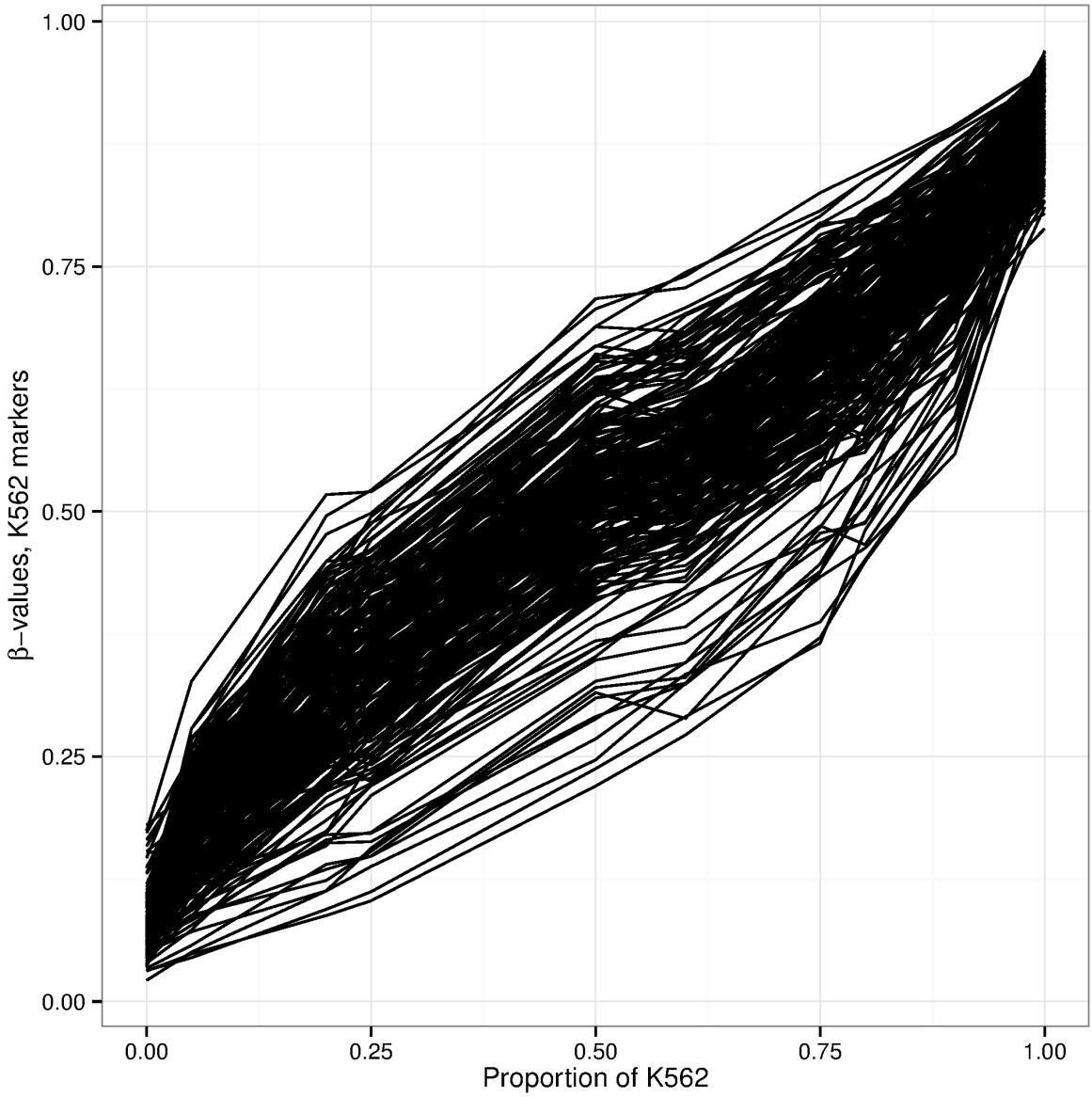


Figure 2.S5: Mixing experiment with KG1a and K562 cells profiled on the Infinium HumanMethylation450 BeadChip. β-values of marker CpGs hypermethylated in K562 cells (and thus hypomethylated in KG1a cells) plotted against the corresponding mixing proportions.

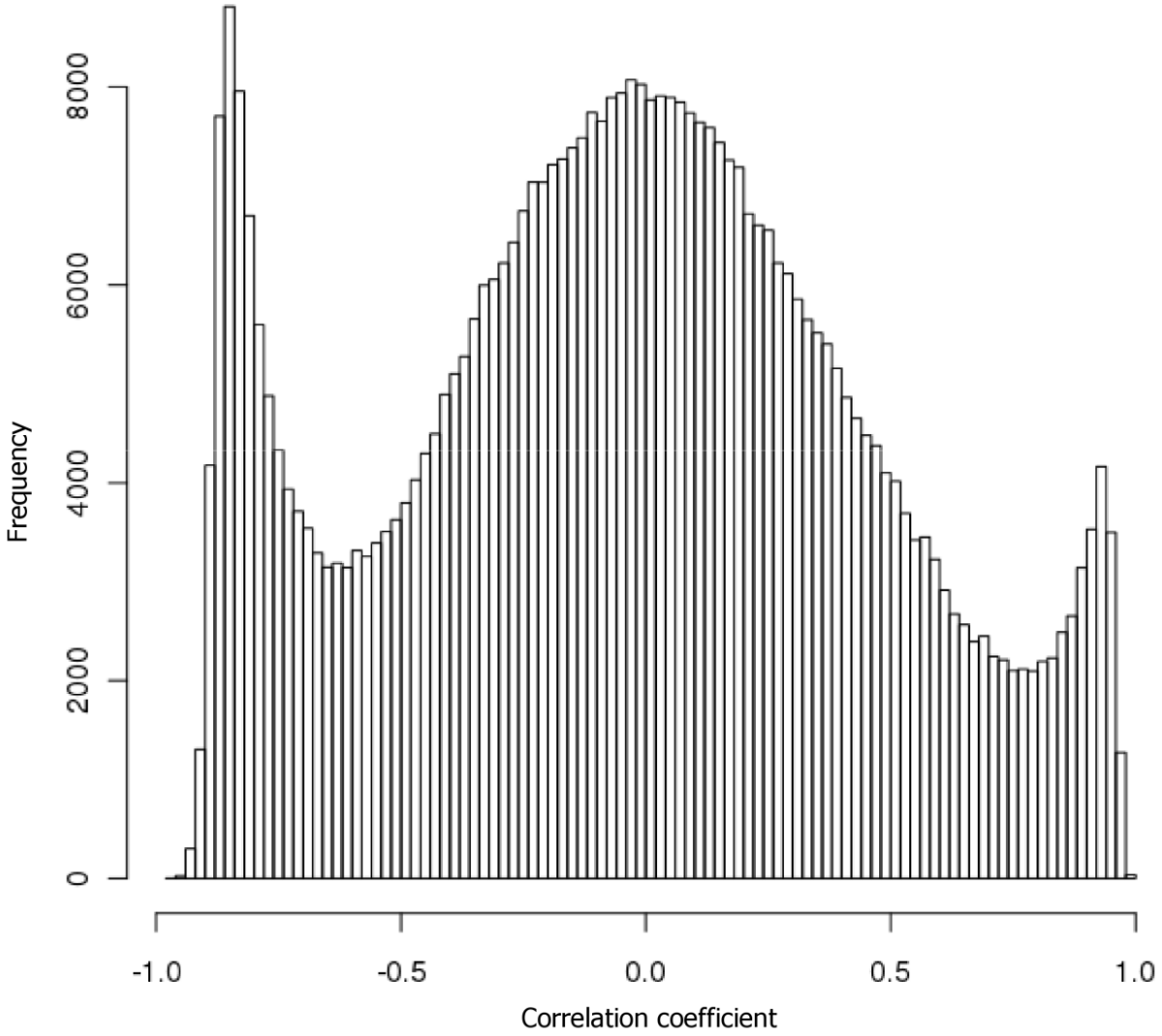


Figure 2.S6: Distribution of the correlation coefficients of the methylation values of the ≈480,000 CpGs to the methylation values of the *PTPN7* CpG (cg18384097).

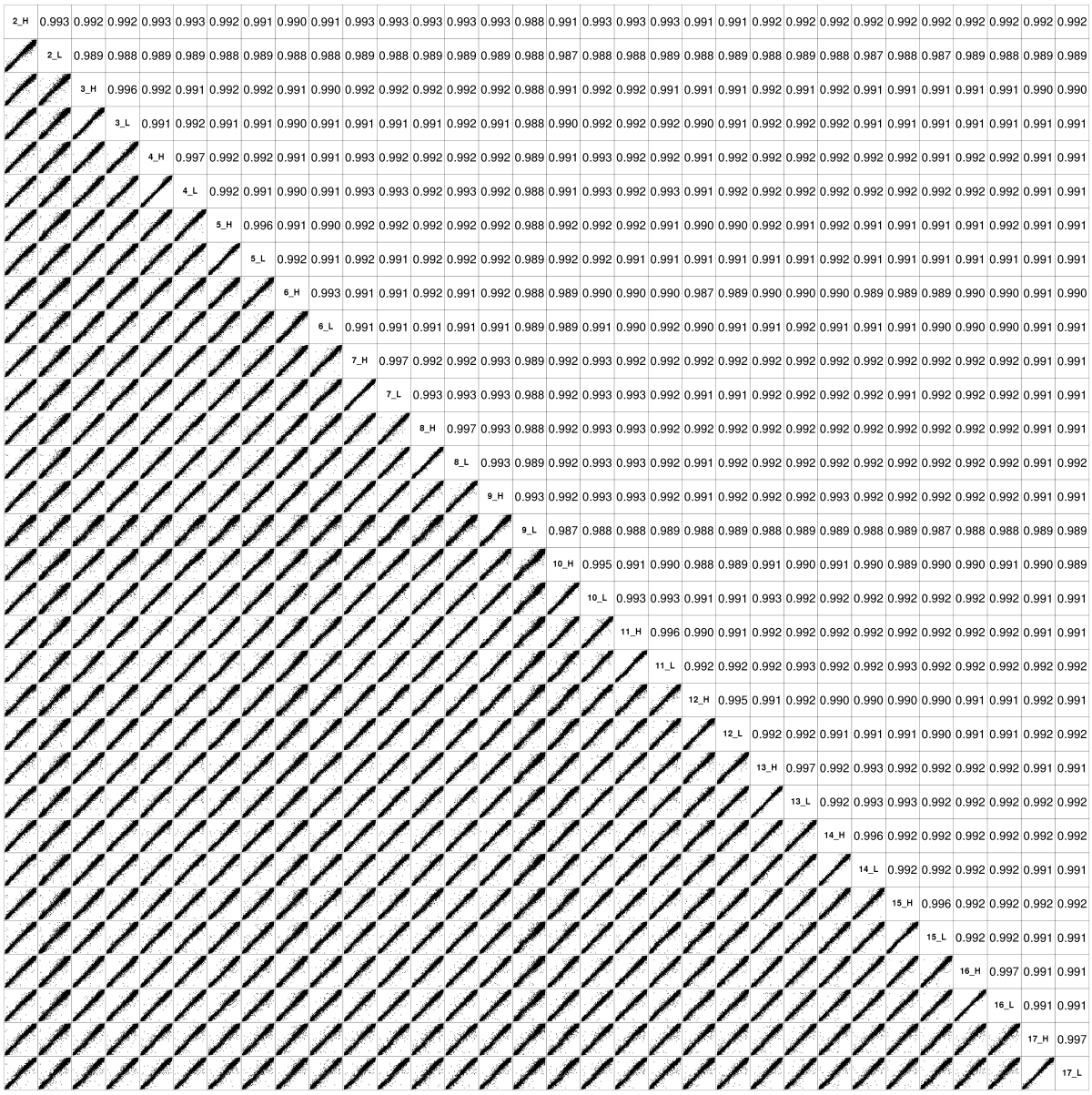


Figure 2.S7: Pair-wise correlations for each pair of samples after adjusting for cell type composition using the *PTPN7* CpG (cg18384097), calculated from ≈480,000 CpGs. Sample labels are shown on the diagonal. Pearson correlation coefficients are shown in the upper part of the figure and the dotplots under the diagonal illustrate a visual representation of the similarity between two samples. H = high birth weight, L = low birth weight.

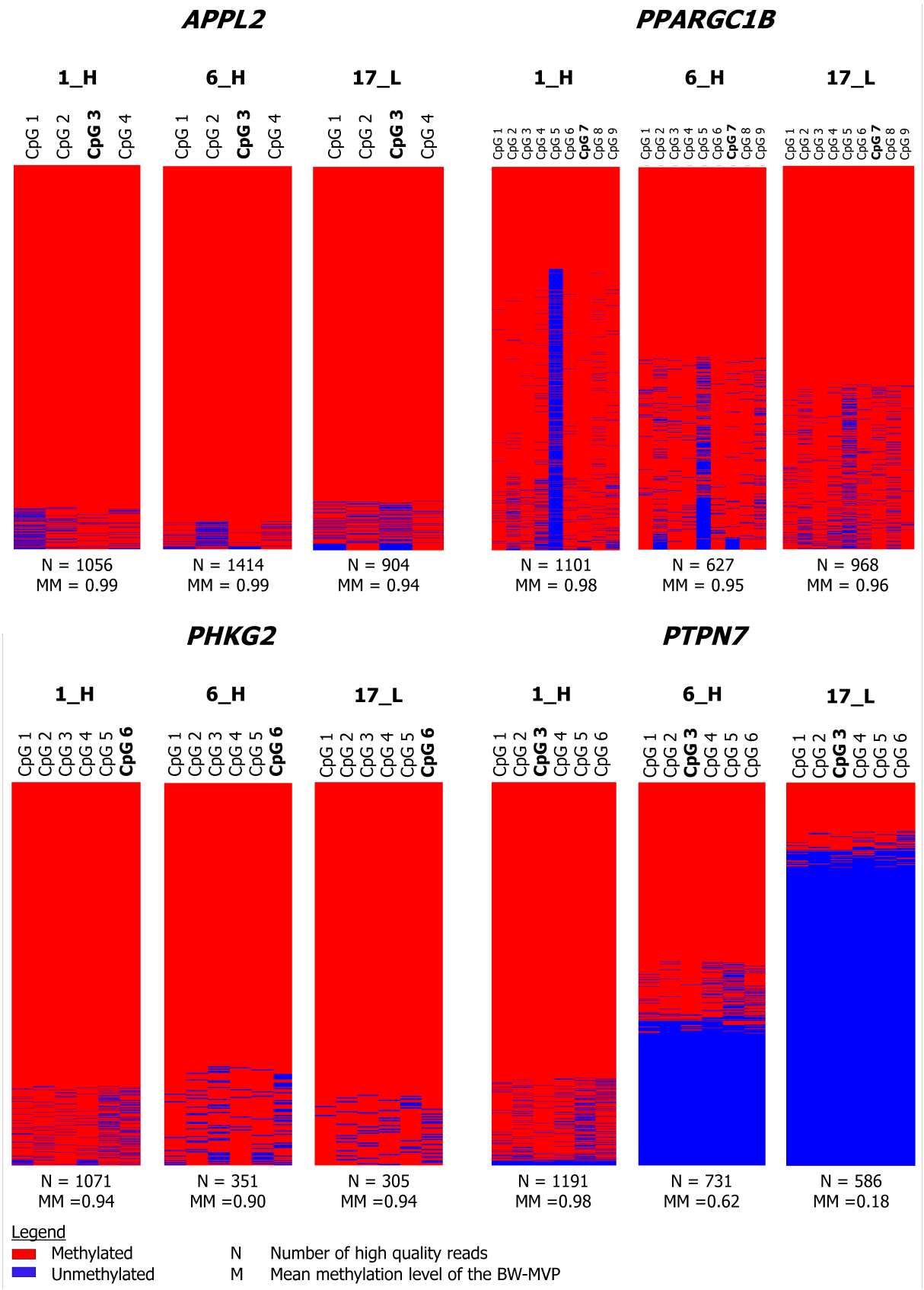


Figure 2.S8: Examples of methylation profiles generated using the deep bisulfite sequencing data of the *APPL2*, *PPARGC1B*, *PHKG2* and *PTPN7* amplicons. The bold CpGs correspond to the BW-MVPs identified using the Infinium HumanMethylation450 BeadChip.

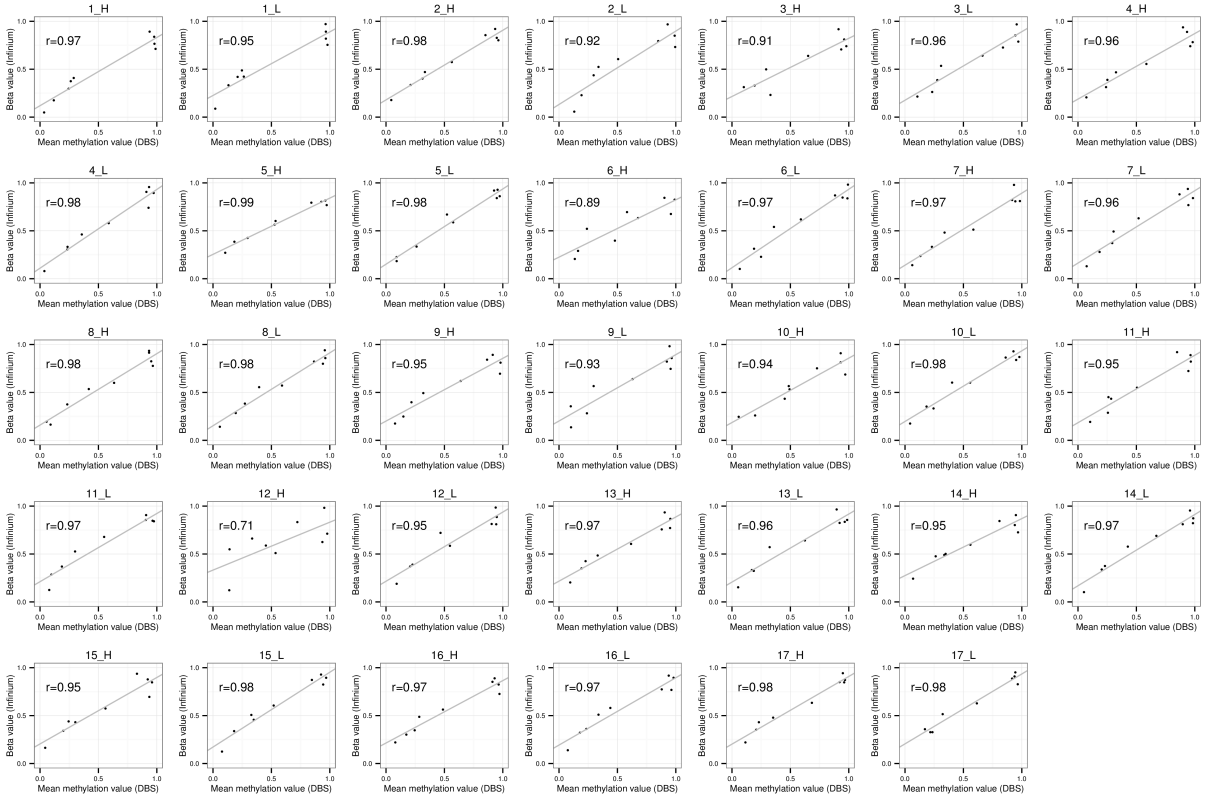


Figure 2.S9: Correlation plots in which the (unadjusted) Infinium 450K data of the validated CpGs is plotted against the (unadjusted) deep bisulfite sequencing (DBS) data for every sample separately. Infinium data are expressed as β-value. DBS data are expressed as mean methylation level, in which the methylation level is calculated by dividing the number of reads in which the particular CpG is methylated by the total number of sequenced reads. H = high birth weight, L = low birth weight.

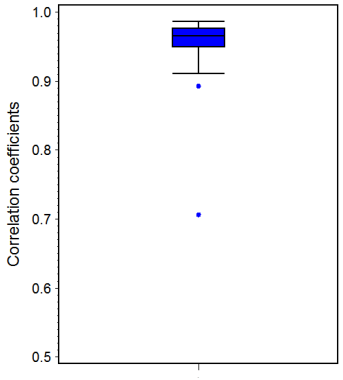


Figure 2.S10: Box plot of the correlation coefficients calculated between the Infinium 450K data and the deep bisulfite sequencing (DBS) data of the validated CpGs for every individual sample.

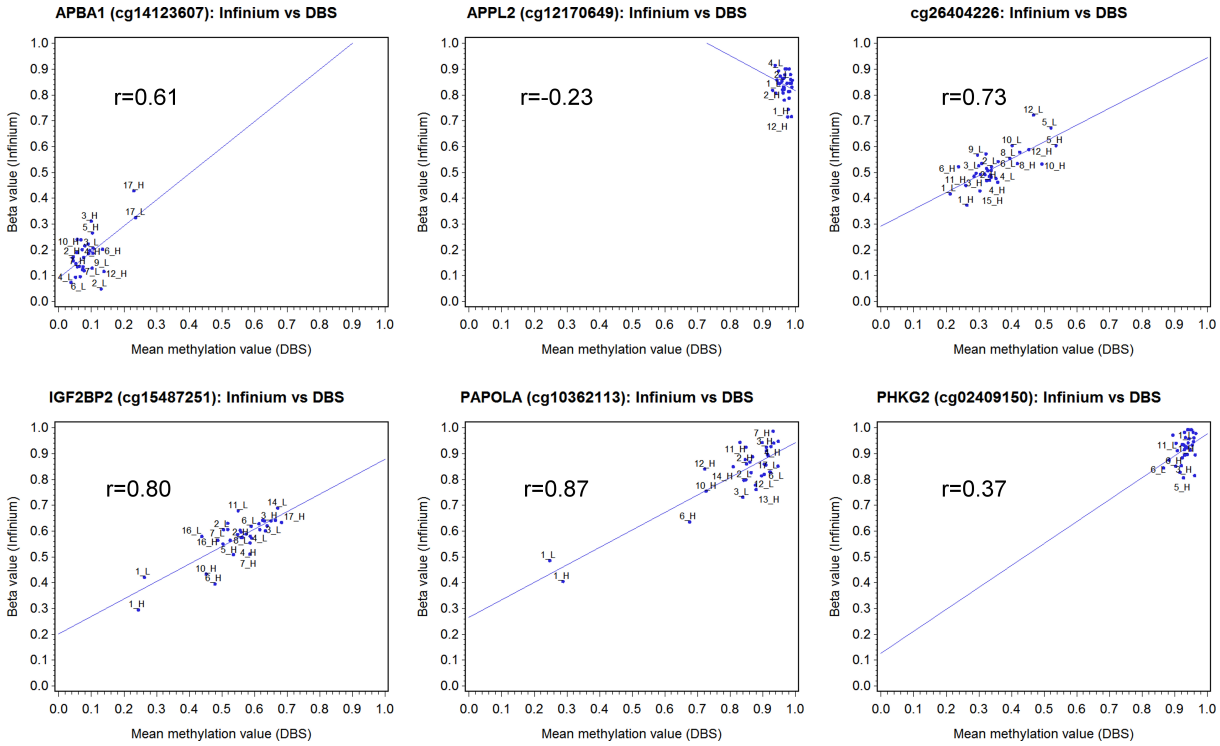


Figure 2.S11: Comparison of the (unadjusted) Infinium data with the (unadjusted) deep bisulfite sequencing (DBS) data of the validated BW-MVPs and the buccal marker (*PTPN7*) of the 17 discordant MZ twin pairs. Infinium data are expressed as β-value. DBS data are expressed as mean methylation level, where the methylation level is calculated by dividing the number of reads in which the particular CpG is methylated by the total number of sequenced reads. H = high birth weight, L = low birth weight.

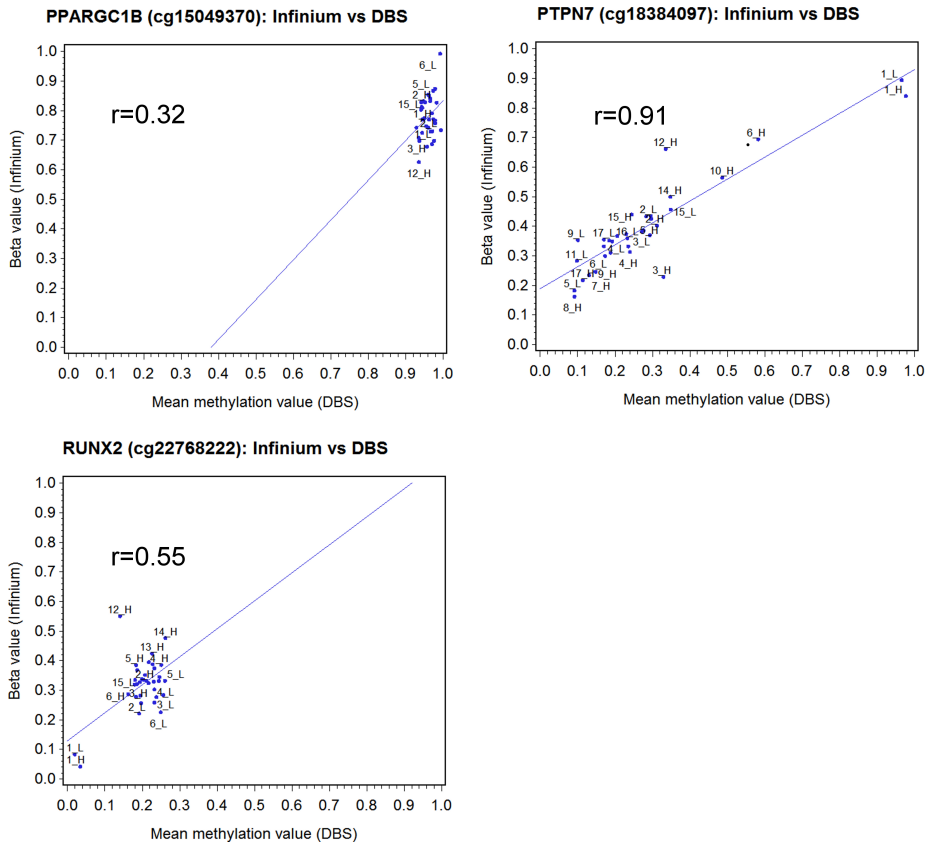


Figure 2.S12: Continuation of Figure 2.S11.

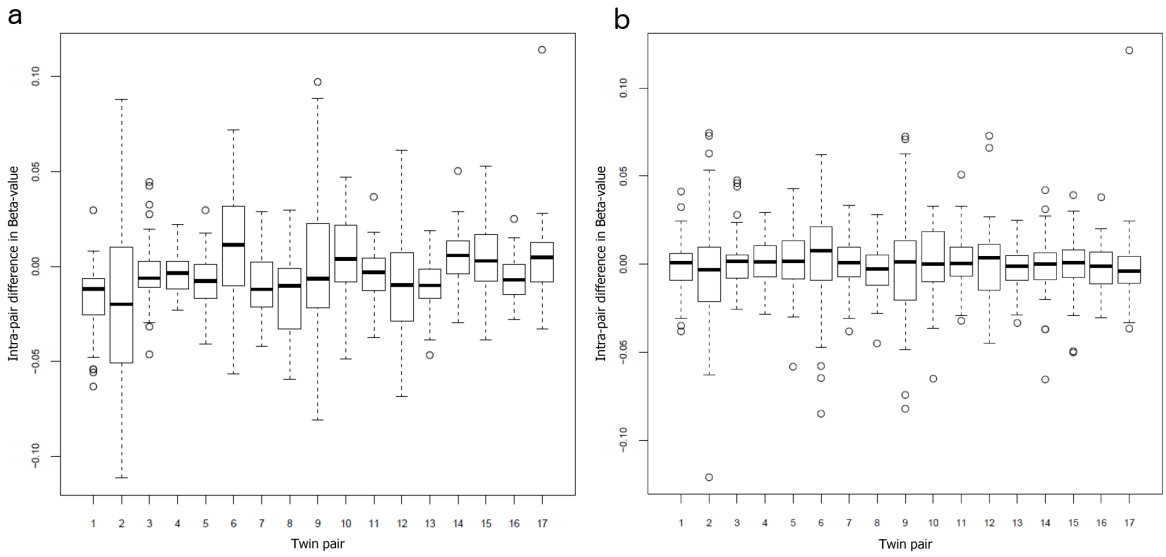


Figure 2.S13: Box-plot of the intra-pair differences in Beta-values of the 64 SNPs present on the Infinium HumanMethylation450 BeadChip before (a) and after (b) normalisation using internal controls and background subtraction by the GenomeStudio software (Methylation Module v1.8).

Supplemental Tables

Table 2.S1: DNA methylation profiles used to create the cell-type reference data set.

Cell type/Tissue	Accession	#samples	Reference	Comments
Tissues				
Whole blood	GEO: GSE19711	274[a]	[Teschendorff *et al.*, 2010]	
Buccal epithelium	GEO: GSE25892	66[a]	[Essex *et al.*, 2013]	Mouth scrabs
Purified cell types				
Neutrophils (granulocytes)	GEO: GSE30090	1[b]	[Calvanese *et al.*, 2012]	Centrifugation, negative selection while sorting for other cell types
B-lymphocytes	GEO: GSE30090	1[b]	[Calvanese *et al.*, 2012]	Centrifugation, CD19+ staining and sorting
CD4+ T-lymphocytes	GEO: GSE30090	1[b]	[Calvanese *et al.*, 2012]	Magnetic bead isolation
CD8+ T-lymphocytes	GEO: GSE30090	1[b]	[Calvanese *et al.*, 2012]	Magnetic bead isolation
NK cells	GEO: GSE30090	1[b]	[Calvanese *et al.*, 2012]	Centrifugation, CD56+ staining and sorting

[a]The reference profiles were obtained by averaging of the profiles of individual samples. [b]The reference profiles were obtained by methylation profiling of the pooled samples (5 individuals). GEO = Gene Expression Omnibus.

Table 2.S2: Cell type-specific quantitative markers used as explanatory variables in heterogeneity adjustment.

Probe ID[a]	Cell-type	Buccal epithelium	Neutrophils	B-cells	CD4+ T-cells	CD8+ T-cells	NK-cells	Chr.	Position	Gene
cg18384097	Buccal	**0.82**	0.04	0.06	0.04	0.04	0.09	1	202129566	*PTPN7*
cg20748065	Neutrophils	0.79	**0.02**	0.78	0.92	0.85	0.69	7	75583421	*POR*
cg00226923	B-cells	0.82	0.96	**0.01**	0.85	0.78	0.80	6	36972027	*FGD2*
cg22858308	CD4+ T-cells	0.93	0.94	0.95	**0.19**	0.68	0.75	6	143095613	*HIVEP2*
cg10163825	CD8+ T-cells	0.07	0.10	0.15	0.18	**0.74**	0.15	16	776685	-
cg06900776	NK-cells	0.28	0.02	0.01	0.10	0.12	**0.78**	X	100878107	*ARMCX3*

[a]Official Illumina Infinium probe identifier, valid both for Illumina HumanMethylation27 and HumanMethylation450 BeadChips.

Table 2.S3: Characteristics of the 45 CpG sites that are significantly differentially methylated between the heavy and light co-twins (BW-MVPs) identified using the Infinium HumanMethylation450 BeadChip.

	CpG Number	Mean β-value difference	*P*-value	Chromosomal position	Gene name	Gene region	Relation to CpG island	Enh.	AbB
1	cg16826055	0.06	0.00003	Chr 7: 41087207				Yes	Yes
2	**cg02409150**	-0.06	0.00003	Chr 16: 30764007	*PHKG2*	Body	S_Shelf		
3	**cg12170649**	-0.06	0.00006	Chr 12: 105622107	*APPL2*	Body		Yes	
4	**cg26404226**	-0.05	0.00006	Chr 10: 90686467				Yes	
5	cg18699337	-0.07	0.00009	Chr 17: 44159144	*KANSL1*	Body		Yes	Yes
6	cg05518778	-0.08	0.00031	Chr 7: 148730143			S_Shelf		Yes
7	cg07574216	-0.08	0.00043	Chr 3: 156394398		5'UTR, TSS1500	S_Shore		Yes
8	cg21234955	-0.07	0.00043	Chr 9: 97713896	*C9orf3*	Body		Yes	Yes
9	cg12149795	-0.05	0.00043	Chr 21: 47882121	*DIP2A*	Body	S_Shelf		Yes
10	cg14696311	-0.07	0.00058	Chr 13: 114855198	*RASA3*	Body	S_Shelf	Yes	Yes
11	cg09683440	0.06	0.00058	Chr 16: 83869927				Yes	Yes
12	cg02350090	-0.06	0.00058	Chr 5: 1951736			S_Shore		
13	**cg14123607**	0.07	0.00076	Chr 9: 72164709	*APBA1*	5'UTR			
14	cg05680237	0.06	0.00101	Chr 6: 27103185			N_Shelf		Yes
15	cg13071869	-0.06	0.00101	Chr 3: 112948716	*BOC*	5'UTR		Yes	Yes
16	cg03839714	-0.07	0.00131	Chr 10: 31361534				Yes	Yes
17	cg22979546	-0.05	0.00131	Chr 15: 28389947	*HERC2*	Body			Yes
18	**cg15049370**	-0.07	0.00168	Chr 5: 149186389	*PPARGC1B*	Body		Yes	Yes
19	cg04416414	0.07	0.00168	Chr 19: 14260587	*LPHN1*	3'UTR	N_Shore	Yes	
20	cg10984962	-0.07	0.00168	Chr 2: 236462202	*AGAP1*	Body		Yes	Yes
21	cg14868128	-0.06	0.00168	Chr 6: 22367352				Yes	Yes
22	cg08846459	-0.06	0.00168	Chr 5: 2176047					
23	**cg15487251**	-0.05	0.00168	Chr 3: 185544216	*IGF2BP2*	TSS1500	S_Shore		Yes
24	cg25064052	-0.05	0.00168	Chr 4: 166216151	*KLHL2*	Body		Yes	Yes
25	cg21450228	-0.05	0.00214	Chr 3: 23727711				Yes	Yes
26	cg10773972	-0.07	0.00269	Chr 15: 81035670	*FAM108C1*	Body		Yes	Yes
27	cg20438460	-0.05	0.00269	Chr 2: 146580994				Yes	Yes
28	cg15940337	-0.05	0.00269	Chr 8: 142456489	*FLJ43860*	Body	S_Shelf	Yes	Yes
29	cg08561071	0.06	0.00336	Chr 19: 41627236	*CYP2F1*	Body	N_Shelf		Yes
30	cg10156499	-0.05	0.00336	Chr 11: 112161291			S_Shore	Yes	Yes
31	cg02832477	-0.06	0.00418	Chr 2: 121501895			S_Shelf		
32	cg13181022	0.05	0.00418	Chr 12: 69247976	*CPM*	3'UTR			
33	cg00587523	0.06	0.00516	Chr 17: 46212998	*SKAP1*	3'UTR			
34	cg24994002	0.06	0.00516	Chr 5: 14118611				Yes	
35	cg04416247	-0.05	0.00516	Chr 3: 85213671	*CADM2*	Body		Yes	Yes
36	cg24049629	0.07	0.00763	Chr 3: 50376475	*RASSF1*	TSS1500, Body	N_Shore		Yes
37	cg25828093	0.07	0.00763	Chr 8: 26041003				Yes	Yes
38	**cg22768222**	0.06	0.00763	Chr 6: 45383690	*RUNX2*	Body	N_Shelf		
39	cg12394706	0.06	0.00763	Chr 8: 99988676			S_Shore		Yes
40	**cg10362113**	0.06	0.00763	Chr 14: 96978537	*PAPOLA*	Body		Yes	Yes
41	cg20540235	-0.08	0.00919	Chr 8: 91683243				Yes	Yes
42	cg13035743	0.06	0.00919	Chr 6: 32119685	*PRRT1*	1st Exon, 5'UTR	S_Shore	Yes	Yes
43	cg26544458	-0.06	0.00919	Chr 12: 109045006	*CORO1C*	Body		Yes	
44	cg12222588	0.05	0.00919	Chr 18: 34823808	*CELF4*	3'UTR	Island		Yes
45	cg07965300	0.05	0.00919	Chr 12: 56368138	*RAB5B*	5'UTR	Island		Yes

Bold and underlined CpG sites were validated using deep bisulfite sequencing. β-value difference = β-value heavy co-twin – β-value light co-twin, Chr = chromosome, NA = not applicable, Shelves = 2-4 kb from CpG island, Shores = 0-2 kb from CpG island, TSS200 = within 200 bp from transcription start site, TSS1500 = within 1500 bases from transcription start site, UTR = untranslated region, Enh.= Enhancer, AbB = Affected by buccal.

Table 2.S4: Distribution of the samples across the beadchips, detected CpGs (detection *p*-value<0.001) and the corresponding call rate per sample.

Pair	Twin	Array	Row	Column	Detected CpGs[a]	Call rate[b]
1	H	1	1	1	481967	99.9
	L	1	2	1	481792	99.9
2	H	1	3	1	481942	99.9
	L	1	4	1	481839	99.9
3	H	1	5	1	481902	99.9
	L	1	6	1	481826	99.9
4	H	1	1	2	481937	99.9
	L	1	2	2	481923	99.9
5	H	1	3	2	481876	99.9
	L	1	4	2	481968	99.9
6	H	1	5	2	480062	99.5
	L	1	6	2	481727	99.8
7	H	2	1	1	482018	99.9
	L	2	2	1	481996	99.9
8	H	2	3	1	481949	99.9
	L	2	4	1	481913	99.9
9	H	2	5	1	481880	99.9
	L	2	6	1	481048	99.7
10	H	2	1	2	481644	99.8
	L	2	2	2	481941	99.9
11	H	2	3	2	481934	99.9
	L	2	4	2	481941	99.9
12	L	2	5	2	481800	99.9
	H	2	6	2	481775	99.9
13	H	3	2	2	481873	99.9
	L	3	3	2	481843	99.9
14	H	3	4	2	481300	99.8
	L	3	5	2	481767	99.9
15	L	3	6	2	481663	99.8
	H	4	1	1	481806	99.9
16	H	4	2	1	481839	99.9
	L	4	3	1	481845	99.9
17	L	5	5	2	481888	99.9
	H	5	6	2	481772	99.9

[a]Number of CpGs with a detection *p*-value<0.001. [b]Call rate (%), based on a detection *p*-value <0.001. Total number of CpGs are 482421. H = high birth weight, L = low birth weight.

Table 2.S5: Reaction conditions and primer sequences of the bisulfite-PCRs.

Method	CpG number	Gene/ element	Forward primer sequence (5'→3')	C	Reverse primer sequence (5'→3')	C	T	Cyc	Product size (bp)	#CpGs
DBS[a]	cg14123607	*APBA1*	ATGATATGGTTTATGAATTTTAATTTTT	83	AAAAAATTATATCTCCTCTATCCCAATA	83	55	45	306	9
	cg12170649	*APPL2*	TATTGAATGTGGTATAGGTTAAGTATTT	83	AATTAAACATCCCAAATTTAAAAA	83	58	45[c]	228	4
	cg26404226	Chr10q23.3	GTTTATGGAAAGAATGATTTTTGTTT	83	CCCTACCAAAACCTAAATCTCAA	83	56	45[c]	228	2
	cg15487251	*IGF2BP2*	TTAGTTTTAAAGTTAGGGTGGTGG	83	AATTTCATTTCCTAACTAAAACCAAA	83	54	42[c]	280	14
		HNF4A	GGGAAGTTATTGAATTAGGGGATT	83	ACCCTCTCTACCTTCCTTTCAAAC	83	59	42[c]	394	9
	cg10362113	*PAPOLA*	TTATTATAGGGGGTTATTTTAGTTTTT	83	CAAACTAATCTAAACTCTTAAACTCAAA	83	55	42[c]	308	3
	cg02409150	*PHKG2*	GAGAGTGTTTAGTAGTGGATTGATTAT	83	TCCTCAAAATAACATTATCAAATACTA	83	53	45[b]	369	6
	cg15049370	*PPARGC1B*	TTTTAGATGGAGTGTGTGGGATAT	83	CACAATAACTCAAACCTATAATTCCAA	83	62	45[c]	339	11
	cg18384097	*PTPN7*	TAAGAGGGTATTTTTAGGAGGGAGT	83	AAAATATTAAATCACCAAAACAACAAAA	83	54	40[c]	253	6
	cg22768222	*RUNX2*	AATTTTTTTTTTTTGTAATGTTTTTTT	83	CCATCTAATCATAATCAATTTCTAAAAA	83	53	45[b]	312	4
SIRPH		*HERVK*	TATTTTTTAATTTTAAGTATTTAGGGAT	200	ATACCTTCCTCTTATCTCAACTACA	200	55	30	237	6
		LINE1[d]	TTATTAGGGAGTGTTAGATAGTGGG	200	CCTCTAAACCAAATATAAAATATAATCT	200	55	30	246	18

[a]Amplicons were generated using region-specific primers having on their 5-ends the recommended GS-FLX A and B adaptors (Lib-L) that included multiplex identifiers (MID). [b]HotStarTaq DNA polymerase (Qiagen) was used. [c]2 µg Hotstart-IT Binding Protein (USB, Cleveland, USA) added. [d]Primers from [Marques *et al.*, 2008]. Bisulfite-treated DNA was denatured for 15 min at 95°C, followed by *n* cycles of 1 min at 95°C, 1 min at T°C and 1 min at 72°C, and a final extension step of 5 min at 72°C. C = concentration (nM), Cyc = number of cycles, DBS = deep bisulfite sequencing, LTR = long terminal repeat, SIRPH = single-nucleotide primer extension (SNuPE) assays in combination with ion pair reverse phase high performance liquid chromatography (IP-RP-HPLC) separation techniques, T = annealing temperature (°C), #CpGs = number of CpGs present in the amplicon.

Table 2.S6: Reaction conditions and primer sequences of the SIRPH analysis.

Gene/ element	Region	SNuPE primer 1	SNuPE primer 2	Acetonitril Gradient	Oven Temperature
HERVK	LTR	5'-TAGGGATATAAAAATTG-3'	5'-GGAAAGATTTGAT-3'	15 min: 5%-7.5%	50°C
LINE1[a]	5' UTR CpG island	5'-CCTAACTCCTTAC-3'	5'-CCCCTTTCTTTAACTC-3'	13 min: 4.75%-7.25%	50°C

SNuPE reactions were performed starting with 2 min denaturation at 96°C followed by 50 cycles of 96°C for 20 s, 50°C for 30 sec and 60°C for 30 sec. Products are loaded directly onto the DNASepTM column (Transgenomic) and separated applying the respective acetonitril gradient. [a]Extended with ddATP and ddGTP. LTR = long terminal repeat.

Table 2.S7: Statistical power of the twin study.

Magnitude of the correlation	Power at alpha = 0.01	Power at alpha = $1 \cdot 10^{-4}$	Power at alpha = $1 \cdot 10^{-6}$
0.0	0.99	0.65	0.11
0.2	1.00	0.80	0.21
0.4	1.00	0.94	0.40
0.6	1.00	1.00	0.75
0.8	1.00	1.00	1.00
1.0	1.00	1.00	1.00

The table shows the statistical power to detect a mean β-value difference of 0.05 with a sample size of 16 twin pairs based on a standard deviation of 0.025 (which is the true median standard deviation observed in the data) using a two-sided paired T-test.

Abbreviations

APBA1 = Amyloid beta A4 precursor protein-binding family A member 1
APPL2 = Adaptor protein, phosphotyrosine interaction, PH domain and leucine zipper containing 2
BW-MVP = Birth weight associated methylation variable positions
DBS = Deep bisulfite sequencing
DC = Dichorionic
DZ = Dizygotic
EFPTS = East Flanders Prospective Twin Survey
EWAS = Epigenome-wide association study
GEO = Gene Expression Omnibus
GWAS = Genome-wide association study
HELP = HpaII tiny fragment enrichment by ligation-mediated PCR
HERVK = Human endogenous retrovirus type K

HNF4α = Hepatocyte nuclear factor 4 alpha
IGF2BP2 = Insulin-like growth factor 2 mRNA-binding protein 2
IP-RP-HPLC = Ion-pair reversed-phase high-performance liquid chromatography
IUGR = Intra-uterine growth restriction
LINE1 = Long interspersed nuclear element-1
MC = Monochorionic
MI = Methylation index
MVP = Methylation variable positions
MZ = Monozygotic
PAPOLA = Poly(A) polymerase alpha
PHKG2 = Phosphorylase kinase, gamma 2 (testis)
PPARGC1B = Peroxisome proliferator-activated receptor gamma coactivator 1-beta
PTPN7 = Protein tyrosine phosphatase non-receptor type 7
RUNX2 = Runt-related transcription factor 2
SIRPH = Single-nucleotide primer extension assays in combination with ion-pair reversed-phase high-performance liquid chromatography separation techniques
SNP = Single nucleotide polymorphism
SNuPE = Single nucleotide primer extension
TTTS = Twin-to-twin transfusion syndrome
T2D = Type 2 Diabetes

References

Adkins, R. M., Tylavsky, F. A., and Krushkal, J. Newborn umbilical cord blood DNA methylation and gene expression levels exhibit limited association with birth weight. *Chem Biodivers*, 9(5):888–899, 2012.

Aps, J. K. M., Van Den Maagdenberg, K., Delanghe, J. R., and Martens, L. C. Flow cytometry as a new method to quantify the cellular content of human saliva and its relation to gingivitis. *Clinica Chimica Acta*, 321(1-2):35–41, 2002.

Banister, C. E., Koestler, D. C., Maccani, M. A., Padbury, J. F., Houseman, E. A., and Marsit, C. J. Infant growth restriction is associated with distinct patterns of DNA methylation in human placentas. *Epigenetics*, 6(7):920–927, 2011.

Barker, D. J. Adult consequences of fetal growth restriction. *Clin Obstet Gynecol*, 49(2):270–283, 2006.

Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J. M. *et al.* High density DNA methylation array with single CpG site resolution. *Genomics*, 98(4):288–295, 2011.

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

Bird, A. DNA methylation patterns and epigenetic memory. *Genes and Development*, 16(1):6–21, 2002.

Bo, S., Cavallo-Perin, P., Scaglione, L., Ciccone, G., and Pagano, G. Low birthweight and metabolic abnormalities in twins with increased susceptibility to Type 2 diabetes mellitus. *Diabet Med*, 17(5):365–370, 2000.

Burwinkel, B., Rootwelt, T., Kvittingen, E. A., Chakraborty, P. K., and Kilimann, M. W. Severe phenotype of phosphorylase kinase-deficient liver glycogenosis with mutations in the PHKG2 gene. *Pediatr Res*, 54(6):834–839, 2003.

Calvanese, V., Fernández, A. F., Urdinguio, R. G., Suárez-Alvarez, B., Mangas, C., Pérez-García, V. *et al.* A promoter DNA demethylation landscape of human hematopoietic differentiation. *Nucleic Acids Research*, 40(1):116–131, 2012.

Chai, K. H., McLoughlin, D. M., Chan, T. F., Chan, H. Y., and Lau, K. F. Genomic Organization and Promoter Cloning of the Human X11alpha Gene APBA1. *DNA Cell Biol*, 31(5):651–659, 2012.

Chen, C. C., Xiao, S., Xie, D., Cao, X., Song, C. X., Wang, T. *et al.* Understanding Variation in Transcription Factor Binding by Modeling Transcription Factor Genome-Epigenome Interactions. *PLoS Computational Biology*, 9(12):e1003367, 2013.

Christiansen, J., Kolte, A. M., Hansen, T. O., and Nielsen, F. C. IGF2 mRNA-binding protein 2: biological function and putative role in type 2 diabetes. *J Mol Endocrinol*, 43(5):187–195, 2009.

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

Derom, C. A., Vlietinck, R. F., Thiery, E. W., Leroy, F. O., Fryns, J. P., and Derom, R. M. The East Flanders Prospective Twin Survey (EFPTS). *Twin Res Hum Genet*, 9(6):733–738, 2006.

Derom, R., Derom, C., and Vlietinck, R. Placentation. In L. G. Keith, E. Papiernik, D. M. Keith, and B. Luke, editors, *Multiple pregnancy: Epidemiology, gestation & perinatal outcome.*, pages 113–128. The Parthenon Publishing Group, New York, 1995.

Du, P., Zhang, X., Huang, C.-C., Jafari, N., Kibbe, W. a., Hou, L. *et al.* Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC bioinformatics*, 11:587, 2010.

Einstein, F., Thompson, R. F., Bhagat, T. D., Fazzari, M. J., Verma, A., Barzilai, N. *et al.* Cytosine methylation dysregulation in neonates following intrauterine growth restriction. *PLoS One*, 5(1):e8887, 2010.

El-Maarri, O., Herbiniaux, U., Walter, J., and Oldenburg, J. A rapid, quantitative, non-radioactive bisulfite-SNuPE-IP RP HPLC assay for methylation analysis at specific CpG sites. *Nucleic Acids Res*, 30(6):e25, 2002.

Essex, M. J., Thomas Boyce, W., Hertzman, C., Lam, L. L., Armstrong, J. M., Neumann, S. M. a. *et al.* Epigenetic Vestiges of Early Developmental Adversity: Childhood Stress Exposure and DNA Methylation in Adolescence. *Child Development*, 84(1):58–75, 2013.

Fryer, A. A., Emes, R. D., Ismail, K. M. K., Haworth, K. E., Mein, C., Carroll, W. D. *et al.* Quantitative, high-resolution epigenetic profiling of CpG loci identifies associations with cord blood plasma homocysteine and birth weight in humans. *Epigenetics*, 6(1):86–94, 2011.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5(10):R80, 2004.

Gielen, M., Lindsey, P. J., Derom, C., Loos, R. J., Souren, N. Y., Paulussen, A. D. *et al.* Twin-specific intrauterine 'growth' charts based on cross-sectional birthweight data. *Twin Res Hum Genet*, 11(2):224–235, 2008.

Gluckman, P. D., Hanson, M. A., Cooper, C., and Thornburg, K. L. Effect of in utero and early-life conditions on adult health and disease. *N Engl J Med*, 359(1):61–73, 2008.

Gordon, L., Joo, J. E., Powell, J. E., Ollikainen, M., Novakovic, B., Li, X. *et al.* Neonatal DNA methylation profile in human twins is specified by a complex interplay between intrauterine environmental and genetic factors, subject to tissue-specific influence. *Genome Res*, 22(8):1395–1406, 2012.

Grunnet, L., Vielwerth, S., Vaag, A., and Poulsen, P. Birth weight is nongenetically associated with glucose intolerance in elderly twins, independent of adult obesity. *J Intern Med*, 262(1):96–103, 2007.

Heijmans, B. T., Tobi, E. W., Stein, A. D., Putter, H., Blauw, G. J., Susser, E. S. *et al.* Persistent epigenetic differences associated with prenatal exposure to famine in humans. *Proceedings of the National Academy of Sciences of the United States of America*, 105(44):17046–9, 2008.

Huang, D. W., Sherman, B. T., Lempicki, R. a., Huang da, W., Sherman, B. T., and Lempicki, R. a. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*, 4(1):44–57, 2009.

Iliadou, A., Cnattingius, S., and Lichtenstein, P. Low birthweight and Type 2 diabetes: a study on 11 162 Swedish twins. *Int J Epidemiol*, 33(5):948–953, 2004.

Jiang, S., Fang, Q., Yu, W., Zhang, R., Hu, C., Dong, K. *et al.* Genetic variations in APPL2 are associated with overweight and obesity in a Chinese population with normal glucose tolerance. *BMC Med Genet*, 13(1):22, 2012.

Kaminsky, Z. A., Tang, T., Wang, S. C., Ptak, C., Oh, G. H., Wong, A. H. *et al.* DNA methylation profiles in monozygotic and dizygotic twins. *Nat Genet*, 41(2):240–245, 2009.

Komori, T. Signaling networks in RUNX2-dependent bone development. *J Cell Biochem*, 112(3):750–755, 2011.

Lewi, L., Gucciardo, L., Van Mieghem, T., de Koninck, P., Beck, V., Medek, H. *et al.* Monochorionic diamniotic twin pregnancies: natural history and risk stratification. *Fetal Diagn Ther*, 27(3):121–133, 2010.

Lindsay, R. S., Dabelea, D., Roumain, J., Hanson, R. L., Bennett, P. H., and Knowler, W. C. Type 2 diabetes and low birth weight: the role of paternal inheritance in the association of low birth weight and diabetes. *Diabetes*, 49(3):445–449, 2000.

Liu, C. and Lin, J. D. PGC-1 coactivators in the control of energy metabolism. *Acta Biochim Biophys Sin (Shanghai)*, 43(4):248–257, 2011.

Liu, J., Morgan, M., Hutchison, K., and Calhoun, V. D. A study of the influence of sex on genome wide methylation. *PloS one*, 5(4):e10028, 2010.

Loos, R. J., Beunen, G., Fagard, R., Derom, C., and Vlietinck, R. The influence of zygosity and chorion type on fat distribution in young adult twins consequences for twin studies. *Twin Res*, 4(5):356–364, 2001.

Lutsik, P., Feuerbach, L., Arand, J., Lengauer, T., Walter, J., and Bock, C. BiQ Analyzer HT: Locus-specific analysis of DNA methylation by high-throughput bisulfite sequencing. *Nucleic Acids Research*, 39(S2):W551–6, 2011.

Marques, C. J., Costa, P., Vaz, B., Carvalho, F., Fernandes, S., Barros, A. *et al.* Abnormal methylation of imprinted genes in human sperm is associated with oligozoospermia. *Molecular human reproduction*, 14(2):67–74, 2008.

McMillen, I. C. and Robinson, J. S. Developmental origins of the metabolic syndrome: prediction, plasticity, and programming. *Physiol Rev*, 85(2):571–633, 2005.

Meaburn, E. L., Schalkwyk, L. C., and Mill, J. Allele-specific methylation in the human genome: implications for genetic studies of complex disease. *Epigenetics*, 5(7):578–582, 2010.

Miaczynska, M., Christoforidis, S., Giner, A., Shevchenko, A., Uttenweiler-Joseph, S., Habermann, B. *et al.* APPL proteins link Rab5 to nuclear signal transduction via an endosomal compartment. *Cell*, 116(3):445–456, 2004.

Michels, K. B., Harris, H. R., and Barault, L. Birthweight, maternal weight trajectories and global DNA methylation of LINE-1 repetitive elements. *PLoS One*, 6(9):e25254, 2011.

Monrad, R. N., Grunnet, L. G., Rasmussen, E. L., Malis, C., Vaag, A., and Poulsen, P. Age-dependent nongenetic influences of birth weight and adult body fat on insulin sensitivity in twins. *J Clin Endocrinol Metab*, 94(7):2394–2399, 2009.

Newsome, C. A., Shiell, A. W., Fall, C. H., Phillips, D. I., Shier, R., and Law, C. M. Is birth weight related to later glucose and insulin metabolism?–A systematic review. *Diabet Med*, 20(5):339–348, 2003.

Poulsen, P., Vaag, A. A., Kyvik, K. O., Moller Jensen, D., and Beck-Nielsen, H. Low birth weight is associated with NIDDM in discordant monozygotic and dizygotic twin pairs. *Diabetologia*, 40(4):439–446, 1997.

Rakyan, V. K., Beyan, H., Down, T. a., Hawa, M. I., Maslau, S., Aden, D. *et al.* Identification of type 1 Diabetes-associated DNA methylation variable positions that precede disease diagnosis. *PLoS Genetics*, 7(9):e1002300, 2011.

Rapti, A., Trangas, T., Samiotaki, M., Ioannidis, P., Dimitriadis, E., Meristoudis, C. *et al.* The structure of the 5'-untranslated region of mammalian poly(A) polymerase-alpha mRNA suggests a mechanism of translational regulation. *Mol Cell Biochem*, 340(1-2):91–96, 2010.

Schenck, A., Goto-Silva, L., Collinet, C., Rhinn, M., Giner, A., Habermann, B. *et al.* The endosomal protein Appl1 mediates Akt substrate specificity and cell survival in vertebrate development. *Cell*, 133(3):486–497, 2008.

Seki, Y., Williams, L., Vuguin, P. M., and Charron, M. J. Minireview: Epigenetic Programming of Diabetes and Obesity: Animal Models. *Endocrinology*, 153(3):1031–1038, 2012.

Shieh, G., Jan, S. L., and Randles, R. H. Power and sample size determinations for the Wilcoxon signed-rank test. *Journal of Statistical Computation and Simulation*, 77(8):717–724, 2007.

Souren, N. Y., Tierling, S., Fryns, J. P., Derom, C., Walter, J., and Zeegers, M. P. DNA Methylation Variability at Growth-Related Imprints Does not Contribute to Overweight in Monozygotic Twins Discordant for BMI. *Obesity (Silver Spring)*, 19(7):1519–1522, 2011.

Teschendorff, A. E., Menon, U., Gentry-Maharaj, A., Ramus, S. J., Weisenberger, D. J., Shen, H. *et al.* Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. *Genome Research*, 20(4):440–446, 2010.

Thiede, C., Prange-Krex, G., Freiberg-Richter, J., Bornhäuser, M., and Ehninger, G. Buccal swabs but not mouthwash samples can be used to obtain pretransplant DNA fingerprints from recipients of allogeneic bone marrow transplants. *Bone marrow transplantation*, 25(5):575–577, 2000.

Tierling, S., Souren, N. Y., Reither, S., Zang, K. D., Meng-Hentschel, J., Leitner, D. *et al.* DNA methylation studies on imprinted loci in a male monozygotic twin pair discordant for Beckwith-Wiedemann syndrome. *Clin Genet*, 79(6):546–553, 2011.

Tobi, E. W., Lumey, L. H., Talens, R. P., Kremer, D., Putter, H., Stein, A. D. *et al.* DNA methylation differences after exposure to prenatal famine are common and timing- and sex-specific. *Hum Mol Genet*, 18(21):4046–4053, 2009.

Turan, N., Ghalwash, M. F., Katari, S., Coutifaris, C., Obradovic, Z., and Sapienza, C. DNA methylation differences at growth related genes correlate with birth weight: a molecular signature linked to developmental origins of adult disease? *BMC Med Genomics*, 5:10, 2012.

Vidovic, A., Vidovic Juras, D., Vucicevic Boras, V., Lukac, J., Grubisic-Ilic, M., Rak, D. *et al.* Determination of leucocyte subsets in human saliva by flow cytometry. *Archives of Oral Biology*, 57(5):577–583, 2012.

Wannamethee, S. G., Lawlor, D. A., Whincup, P. H., Walker, M., Ebrahim, S., and Davey-Smith, G. Birthweight of offspring and paternal insulin resistance and paternal diabetes in late adulthood: cross sectional survey. *Diabetologia*, 47(1):12–18, 2004.

Weksberg, R., Shuman, C., Caluseriu, O., Smith, A. C., Fei, Y. L., Nishikawa, J. *et al.* Discordant KCNQ1OT1 imprinting in sets of monozygotic twins discordant for Beckwith-Wiedemann syndrome. *Hum Mol Genet*, 11(11):1317–1325, 2002.

Wilhelm-Benartzi, C. S., Houseman, E. A., Maccani, M. A., Poage, G. M., Koestler, D. C., Langevin, S. M. *et al.* In utero exposures, infant growth, and DNA methylation of repetitive elements and developmentally related genes in human placenta. *Environ Health Perspect*, 120(2):296–302, 2012.

Yaghootkar, H. and Freathy, R. M. Genetic origins of low birth weight. *Curr Opin Clin Nutr Metab Care*, 15(3):258–264, 2012.

Chapter 3

Comprehensive Analysis of DNA Methylation Data with RnBeads

The full text of this chapter has been earlier published as:

Yassen Assenov[1,*†], Fabian Müller[1,*#], Pavlo Lutsik[2,*], Jörn Walter[2], Thomas Lengauer[1] and Christoph Bock[1,3,4,#] (2014) *Nature Methods* **11**(11), pp. 1138–1140

The author of the present thesis implemented the core infrastructure of the data handling classes, most of the data loading, quality control and normalization modules, as well as functionality for the cell type estimation (the inference module) and reference-free correction of the differential methylation analysis (option "refFreeEWAS"). He also participated in writing of the manuscript.

[1]Max Planck Institute for Informatics, Saarbrücken, Germany
[2]Department of Genetics/Epigenetics, Saarland University, Saarbrücken, Germany
[3]CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, Vienna, Austria
[4]Department of Laboratory Medicine, Medical University of Vienna, Vienna, Austria
[*]These authors contributed equally to this work
[†]Present address: Division of Epigenomics and Cancer Risk Factors, German Cancer Research Center (DKFZ), Heidelberg, Germany
[#]Correspondence: rnbeads@mpi-inf.mpg.de

Abstract

We present RnBeads, a software tool for large-scale analysis and interpretation of DNA methylation data. A user-friendly, automated and customizable analysis workflow gives rise to detailed hypertext reports. Supported assays include whole genome bisulfite sequencing, reduced representation bisulfite sequencing, Infinium microarrays, and any other protocol that yields high-resolution DNA methylation data. RnBeads facilitates reproducible analysis of epigenome-wide association studies, cancer epigenome profiling and many other applications of DNA methylation mapping.

Keywords: DNA methylation analysis, computational epigenetics, whole genome bisulfite sequencing, reduced representation bisulfite sequencing, epigenotyping microarrays, Illumina Infinium HumanMethylation450 assay, bioinformatics software, epigenome-wide association studies, medical epigenomics

Acknowledgements

We would like to thank David Brocks, Hector Hernandez-Vargas, Eberhard Schneider, Andreas Schönegger and all users of RnBeads for their extensive testing and feedback. Furthermore, we would like to thank Georg Friedrich, Joachim Büch and the Information Services and Technology team at the Max Planck Institute for technical support. This work is funded in part by the European Union's Seventh Framework Programme (FP7/2007-2013) grant agreement No. 282510 (BLUEPRINT) and grant agreement No. 267038 (NOTOX) as well as by the German Science Ministry grant No. 01KU1216A (DEEP).

Author contributions

Y.A., F.M. and P.L. developed and maintain RnBeads; J.W., T.L. and C.B. supervised the project; all authors contributed to the writing of the manuscript.

3.1 Main text

DNA methylation is an important epigenetic mark and widely studied in the context of biological processes and diseases. Several assays are now available for mapping DNA methylation genome-wide, at high resolution and in a large number of samples. Whole genome bisulfite sequencing (WGBS) provides comprehensive genome-wide coverage (approximately 28 million CpGs in the human genome) at the cost of resequencing the whole genome [Lister *et al.*, 2009]. Reduced representation bisulfite sequencing (RRBS) focuses the sequencing on a defined subset of DNA fragments that contain at least one CpG each, thereby covering approximately two million individual CpGs in the human genome [Gu *et al.*, 2010]. The Infinium HumanMethylation450 ("450k") assay uses an adapted genotyping microarray to measure DNA methylation at approximately 0.5 million CpGs [Bibikova *et al.*, 2011]. In addition, enrichment-based assays such as MeDIP-seq [Down *et al.*, 2008] and restriction-enzyme based MRE-seq [Harris *et al.*, 2010] can be combined with bioinformatic algorithms to infer high-resolution DNA methylation data for a large proportion of CpGs [Stevens *et al.*, 2013]. The technical accuracy and reproducibility of these assays is generally high [Bock *et al.*, 2010; Harris *et al.*, 2010], but bioinformatic analysis of the resulting datasets remains a complex task with many pitfalls [Bock, 2012].

We developed the RnBeads software with the goal of establishing a user-friendly workflow for the analysis and interpretation of large-scale DNA methylation data. RnBeads builds upon extensive prior research on bioinformatic and statistical methods for DNA methylation analysis. We have reviewed the features of 22 related software tools (Table 3.S1), and based on our assessment of existing algorithms and software we defined the following key elements of RnBeads: (i) Support for all genome-scale and genome-wide DNA methylation assays that provide single-basepair resolution; (ii) extensive functionality for high-level DNA methylation analysis, including data visualization, quality control, exploratory analysis, handling of batch effects, correction for tissue heterogeneity, and differential DNA methylation analysis; (iii) generation of interactive reports that allow users to select results and adjust parameters without having to rerun the analysis; (iv) implementation of a standardized pipeline mode that is essentially self-configuring, with the additional option to adapt the workflow using custom parameter settings and/or custom scripts; (v) flexibility to run RnBeads on a personal computer, on high-performance computing infrastructure, via a web-based service and in a cloud computing environment, depending on the scale of the analysis; (vi) sufficient performance to process – on a suitable scientific computing cluster – the largest DNA methylation datasets that are currently available (10s of WGBS profiles, 100s of RRBS profiles, or 1000s of Infinium 450k profiles); (vii) reproducibility and easy results sharing through automatic documentation of parameters and analysis methods in the RnBeads report.

To be able to support all protocols for large-scale DNA methylation mapping, RnBeads builds upon existing software tools that can convert raw data into high-resolution DNA methylation profiles. Sequencing data should be preprocessed prior to running RnBeads using software tools such as Bismark [Krueger and Andrews, 2011], BSMAP [Xi *et al.*, 2012] and/or Bis-SNP [Liu *et al.*, 2012] (for WGBS/RRBS), MEDIPS [Lienhard *et al.*, 2014], MEDUSA [Wilson *et al.*, 2012] or BayMeth [Riebler *et al.*, 2014] (for MeDIP-seq), or methylCRF [Stevens *et al.*, 2013] (for MRE-seq). Raw IDAT files from Infinium 450k experiments can be imported directly into RnBeads, in which case the preprocessing and normalization are performed by RnBeads using low-level functionality imported from other R/Bioconductor packages (methylumi, minfi and wateRmelon, cf. Table 3.S1). Alternatively, users can preprocess and normalize the data prior to importing them into RnBeads, for example using the Illumina GenomeStudio

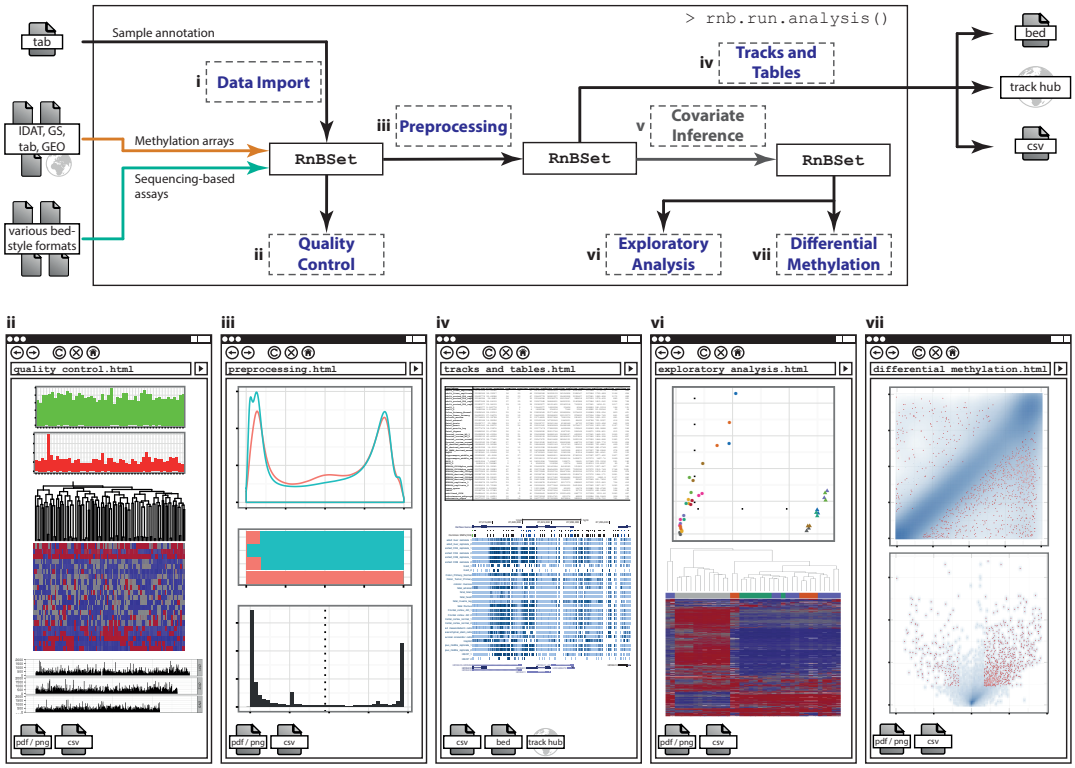


Figure 3.1: Comprehensive analysis of DNA methylation data with RnBeads RnBeads provides a workflow for large-scale DNA methylation analysis. The software is R-based and can be run locally on a personal computer (small analyses), on a scientific computing cluster (large analyses), or in the cloud using the RnBeads web service (small analyses) or a custom instance of Galaxy CloudMan (large analyses). The RnBeads workflow consists of seven modules and is essentially self-configuring based on a sample annotation table provided by the user. Each module generates part of the RnBeads hypertext report, which includes method descriptions, results diagrams and links to data tables. Furthermore, all data and annotations are stored in *RnBSet* objects to facilitate custom analysis workflows in R, and they are exported for visualization using genome browsers and follow-up analyses using other software tools.

software. In addition to data import, the core workflow of RnBeads comprises quality control, preprocessing and filtering, generation of genome browser tracks and data tables, optional inference of confounding covariates (e.g., different cell type compositions), exploratory analysis and differential DNA methylation analysis. Each step of the workflow is performed by a dedicated RnBeads module, as illustrated in Figure 3.1 and described in more detail in the Online Methods.

RnBeads is straightforward to run even for inexperienced users, and the standard pipeline requires an R installation `http://r-project.org` but no prior R programming experience. It can be launched using a single R command: *rnb.run.analysis()*, which takes a user-provided sample annotation table as input and extracts relevant information needed to automatically configure the analysis. For example, annotation columns containing many unique labels are interpreted as sample identifiers, while columns with several different labels are regarded as sample groups that are to be compared against each other. It is also possible to run some or all steps of the RnBeads workflow interactively and to write R scripts that operate directly on the *RnBSet* object containing all DNA methylation data and sample annotations of a given analysis. The main result of the RnBeads pipeline is an interactive hypertext report with publication-quality figures (box plots, bar charts, heatmaps, dendrograms, histograms, density plots, quantile-quantile plots, scatter plots, deviation plots, volcano plots, word clouds, etc.) and tables (DNA methylation profiles, ranked lists of differentially methylated regions, attribute enrichment scores, etc.) covering a broad spectrum of topics and analyses. These reports can be viewed from a local directory or over the Internet, and they facilitate data integration with web-based tools such as the UCSC Genome Browser [Meyer *et al.*, 2013], Ensembl [Flicek *et al.*, 2013], Galaxy [Giardine *et al.*, 2005], the WashU Epigenome Browser [Zhou *et al.*, 2011], and EpiExplorer [Halachev *et al.*, 2012].

To illustrate the practical use of RnBeads, we applied the software to two datasets for which the underlying biology is relatively well understood. The resulting RnBeads reports are available online (`http://rnbeads.mpi-inf.mpg.de/examples.php`), and it is straightforward to rerun these analyses using the *rnb.run.example()* function in RnBeads. The first example is based on Infinium 450k profiles for 124 glioblastoma patients generated by The Cancer Genome Atlas (TCGA) project [Weisenberger, 2014]. We show how RnBeads can identify and characterize samples with glioblastoma CpG island methylator phenotype, an epigenetically defined subtype of brain tumors (Figure 3.S1 and Supplementary Note).

The second example focuses on an RRBS dataset describing the DNA methylation dynamics of blood and skin stem cell differentiation in mice [Bock, 2012]. This dataset comprises 13 blood and 6 skin cell populations with biological replicates and DNA methylation data for slightly more than two million CpGs in each sample. The RnBeads analysis report provides an overview of this dataset (Figure 3.2 and `http://rnbeads.mpi-inf.mpg.de/examples.php`). The global distribution of DNA methylation is characteristically bimodal, and discrete peaks at 33%, 50% and 67% DNA methylation disappear after filtering out CpGs with low sequencing coverage (Figure 3.2, a). Exploratory analysis confirms that the difference between blood and skin cell types dominates the analysis (Figure 3.2, b), and DNA methylation levels are generally higher in blood cells than in skin cells when taking regional averages over all annotated genes (Figure 3.2, c). Hierarchical clustering perfectly discriminates between blood and skin cell types (Figure 3.2, d), confirming that DNA methylation patterns tend to be determined more strongly by cellular lineage than by other properties such as cell proliferation or differentiation status.

RnBeads also identifies differentially methylated regions (DMRs) that are statistically significant and exhibit pronounced DNA methylation differences between the two lineages. This

analysis is performed for single CpGs and also for sets of pre-defined genomic regions such as CpG islands, genes, promoters and genome-wide tiling regions. Such region-of-interest based DMR analyses provide an effective way of increasing the statistical power to detect differential DNA methylation [Bock, 2012]. The priority-ranked list of DMRs between blood and skin cell types comprises many genes with established roles in blood and skin biology, such as members of the homeobox and keratin gene families. Scatterplots provide a convenient way of visualizing the overall frequency of DMRs for a region type of interest (Figure 3.2, e shows data for entire gene loci), and volcano plots illustrate the relationship between effect size and significance of the DMRs (Figure 3.2, f). In Figure 3.2, the *Hoxb3* gene is highlighted as an example of blood-specific DNA methylation, and we can use the *rnb.plot.locus.profile()* function of RnBeads to produce a genomic view of this locus – thus providing an example of custom R scripting on top of the *RnBSet* object calculated by the standard pipeline (Figure 3.2, g).

In addition to these two relatively small examples, we assessed the performance of RnBeads when applied to large-scale datasets from the ENCODE, TCGA, IHEC, Roadmap Epigenomics, and BLUEPRINT consortia. All analyses could be completed within reasonable time on a standard scientific computing cluster (Table 3.S2), and the resulting Methylome Resource (Figure 3.S2 and `http://rnbeads.mpi-inf.mpg.de/methylomes.php`) provides comprehensive analysis reports for some of the largest publicly available DNA methylation datasets. On this website we also provide preconfigured RnBeads analyses for these large-scale epigenome collections, which can be run along as reference maps when analyzing custom DNA methylation datasets. Such reference-based analyses are particularly valuable for researchers who have generated a specialized DNA methylation dataset and who want to assess the data quality and/or biological relevance in context with a broad range of reference methylomes. The concept of preconfigured and re-runnable analyses of reference epigenome data also provides a means of making data from large-scale epigenome mapping projects more useful for smaller-scale and mechanism-centered studies, thereby contributing to the broader relevance of large-scale epigenome mapping projects [Bock, 2014].

In summary, RnBeads combines good practices from prior research into a comprehensive and efficient pipeline, thereby facilitating large-scale analysis and interpretation of DNA methylation data. The software also fosters standardization, reproducibility and data sharing between labs and across collaborative projects. Detailed documentation and examples of RnBeads analyses as well as the Methylome Resource with its pre-configured analyses are available from the RnBeads website `http://rnbeads.mpi-inf.mpg.de/`.

3.2 Online Methods

3.2.1 RnBeads software overview

RnBeads is written in the R programming language (`http://www.r-project.org`). It is available under the GPLv3 open source license and has been submitted for inclusion in Bioconductor [Gentleman *et al.*, 2004]. RnBeads follows a modular design that supports automated pipeline workflows as well as flexible interactive analyses. The default RnBeads workflow is executed by calling the *rnb.run.analysis()* command, either in an interactive R session or via R's support for scripted analyses. Optionally, an XML configuration file can be provided in order to execute RnBeads analyses with standard parameter sets. RnBeads analyses can also be run on the Internet using either the RnBeads web service (`http://rnbeads.mpi-inf.mpg.de/webservice.php`, which is restricted to small datasets, or using the Galaxy integration of Rn-

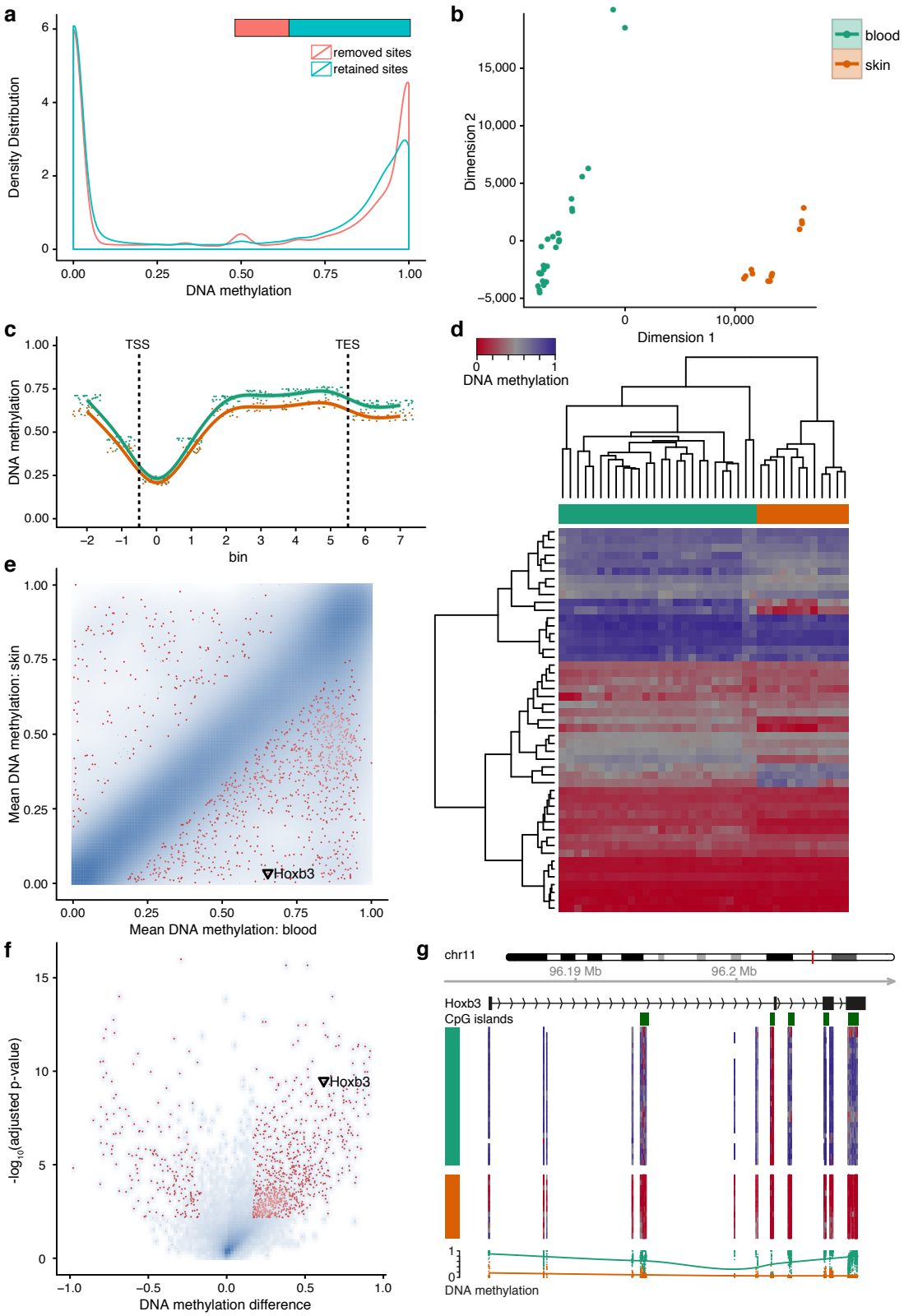

a
Density Distribution
DNA methylation
removed sites
retained sites
b
Dimension 2
Dimension 1
blood
skin
c
DNA methylation
TSS
TES
bin
d
DNA methylation
e
Mean DNA methylation: skin
Mean DNA methylation: blood
Hoxb3
f
-log10(adjusted p-value)
DNA methylation difference
Hoxb3
g
chr11
96.19 Mb
96.2 Mb
Hoxb3
CpG islands
DNA methylation

Figure 3.2: (on the previous page) Analysis of DNA methylation during adult stem cell differentiation based on RRBS data. DNA methylation data for 19 cell types of the blood and skin lineages [Bock, 2012] were reanalyzed with RnBeads. The full analysis report is available online (http://rnbeads.mpi-inf.mpg.de/examples.php). All diagrams in this figure were calculated by RnBeads but have been simplified and reformatted according to journal standards. **a**. Global distribution of DNA methylation levels among retained and removed CpGs after the preprocessing step. **b**. Two-dimensional representation of DNA methylation profiles calculated using multi-dimensional scaling based on average methylation levels in 5kb tiling regions. Samples are color-coded according to tissue type. **c**. Composite plot of DNA methylation levels in blood (green) and skin (orange) cell types averaged across all genes. Each gene was covered by six equally sized bins and by two flanking regions of the same size. Smoothing was done using cubic splines. **d**. Heatmap with hierarchical clustering of DNA methylation levels among lineage marker genes that are specifically expressed in the blood lineage. Clustering used average linkage and Manhattan distance. **e**. Scatterplot of groupwise mean DNA methylation levels across genes, with the 1,000 highest ranking differentially methylated genes highlighted in red. Point density is shown as blue shading. **f**. Volcano plot illustrating effect size and statistical significance across genes, with the 1,000 highest ranking differentially methylated genes highlighted in red. Point density is shown as blue shading. **g**. DNA methylation profile of the *Hoxb3* gene locus. Heatmaps show DNA methylation levels of single CpGs according to the color scheme in panel d. Smoothing of DNA methylation levels (bottom) was done using cubic splines.

Beads available from Galaxy tool shed (http://toolshed.g2.bx.psu.edu). On a sufficiently powerful computing infrastructure, RnBeads can process very large cohorts (Supplementary Table 2). To exploit the parallelization options of RnBeads and to avoid out-of-memory problems, users who want to run large analyses should carefully review the corresponding sections in the RnBeads documentation.

When used with default options on small to medium-scale datasets, RnBeads is essentially self-configuring: It parses a user-provided sample annotation table, configures the analysis accordingly and then executes the RnBeads modules as shown in Figure 1. RnBeads workflows can also be fine-tuned using global configuration parameters, which are specified using *rnb.options()*. During execution of an RnBeads analysis, each step is tracked by extensive logging functionality. Upon successful completion, the modules write their results into an interactive report comprising method descriptions, publication-quality diagrams and links to data tables. The reports generated by RnBeads use client-side scripting and the dynamic features of XHTML to enable interactive data exploration of pre-calculated results. RnBeads can also save the analysis options and data objects in binary RData objects, which makes it straightforward to rerun an analysis with the same parameters and to comply with the paradigm of reproducible research [Gentleman *et al.*, 2004]. Finally, custom workflows can be designed by running the analysis modules individually or by using R functions that operate directly on *RnBSet* objects (these objects are instances of an R S4 class and constitute the RnBeads representation of all DNA methylation and metadata within a given dataset). For instance, Figure 2g was created using the function *rnb.plot.locus.profile()* for plotting genome browser like views of individual genomic loci.

The following paragraphs describe the methodology and functionality behind RnBeads and its modules in more detail. Further information on RnBeads is also available in the package vignette (http://rnbeads.mpi-inf.mpg.de/data/RnBeads.pdf), from the example re-

ports (http://rnbeads.mpi-inf.mpg.de/examples.php), and from the other materials on the RnBeads website (http://rnbeads.mpi-inf.mpg.de). RnBeads was also compared with 22 related software tools for DNA methylation analysis in terms of supported assays, data and analysis types, visualizations and other functionalities (Supplementary Table 1). This comparison includes the following software tools: BEAT [Akman *et al.*, 2014], BiSeq [Hebestreit *et al.*, 2013], Bisulfighter [Saito *et al.*, 2014], BSmooth [Hansen *et al.*, 2012], ChAMP [Morris *et al.*, 2014], COHCAP [Warden *et al.*, 2013], CpGassoc [Barfield *et al.*, 2012], DMEAS [He *et al.*, 2013], EpiDiff (QDMR) [Zhang *et al.*, 2013], Genome Studio (Methylation module)[Illumina Inc., 2014], FastDMA [Wu *et al.*, 2013], HumMeth27QCReport [Mancuso *et al.*, 2011], IMA [Wang *et al.*, 2012], LumiWCluster [Kuan *et al.*, 2010], MethLAB [Kilaru *et al.*, 2012], methyAnalysis [Du *et al.*, 2014], methylkit [Akalin *et al.*, 2012], MethylSig [Park *et al.*, 2014], methylumi [Davis *et al.*, 2014], minfi [Aryee *et al.*, 2014], Shinymethyl [Fortin *et al.*, 2014] and wateRmelon [Pidsley *et al.*, 2013]. For each software tool, the supported features were determined by manual review of the respective publication, software documentation and supplementary material.

3.2.2 Data import

RnBeads supports a broad range of DNA methylation assays, comprising the Illumina Infinium microarray platform (in both its 450k and 27k version), various types of bisulfite sequencing (including WGBS and RRBS) and other sequencing-based methods that can be used to bioinformatically infer DNA methylation measurements at the level of single CpGs (such as MeDIP, MDB-seq and MRE-seq). RnBeads analyses are configured by providing a user-generated sample annotation table that not only identifies the input data files but also includes columns with analysis-relevant information such as tissue types or disease states. RnBeads accepts a broad range of tab-separated or comma-separated text files, and concrete examples of such sample annotation tables are available from the RnBeads website. The data import module of RnBeads parses the annotation table and uses the contained information to configure the analysis, for example identifying the data files that are to be loaded and inferring which specific comparisons may be of interest to the user.

For Infinium microarrays, it is recommended to start the analysis from signal intensity data (IDAT) files and to let RnBeads perform the normalization and DNA methylation calling. Alternatively, RnBeads can load pre-normalized data from Illumina GenomeStudio report files, import Infinium datasets directly from the Gene Expression Omnibus (GEO) database, or read preprocessed data in one of several tabular formats. When IDAT files are loaded into RnBeads, the R/Bioconductor package methylumi [Davis *et al.*, 2014] is internally used for performing the low-level processing. RnBeads offers several alternative options for signal intensity-based normalization, which is an important step to reduce probe biases that could interfere with the analysis. The RnBeads default for Infinium data normalization is SWAN [Makismovic *et al.*, 2012], which is implemented in the minfi package [Aryee *et al.*, 2014] and which – in our experience – provides a good balance of accuracy, robustness, runtime performance and software stability. In addition, RnBeads supports Illumina's standard normalization procedure as implemented in methylumi, the BMIQ normalization method [Teschendorff *et al.*, 2013], and all modular normalization algorithms that are available in the wateRmelon package [Pidsley *et al.*, 2013]. RnBeads also supports the background correction techniques implemented in methylumi [Davis *et al.*, 2014], which can optionally be combined with the normalization algorithms.

For sequencing-based methods, data preparation requires steps that are highly protocol-

dependent, including sequence alignment and DNA methylation calling for single CpGs [Bock, 2012]. These steps need to be completed prior to loading the data into RnBeads; and the RnBeads analysis starts with importing BED files or data tables that provide the number of methylated and unmethylated observations for each covered CpG. For example, bisulfite sequencing data can be preprocessed with the Bismark software [Krueger and Andrews, 2011], whose export format for DNA methylation values is directly supported by RnBeads without the need for file conversion. Furthermore, the combination of BSMAP [Xi *et al.*, 2012; Xi and Li, 2009] and Bis-SNP [Liu *et al.*, 2012] is well-suited for preprocessing reduced representation bisulfite sequencing data, and the output format of Bis-SNP is also a supported input format for RnBeads. Enrichment-based and restriction-enzyme based assays require specialized algorithms for inferring DNA methylation levels at single-basepair resolution. Software tools such as MEDIPS [Chavez *et al.*, 2010], MEDUSA [Wilson *et al.*, 2012] and methylCRF [Stevens *et al.*, 2013] give rise to DNA methylation tables that can be imported into RnBeads as BED files or in one of several other data file formats.

After the DNA methylation data have been loaded from any of the supported input formats, RnBeads combines the data of all samples into a single *RnBSet* object that constitutes the basis for all further analysis steps. This object can become very large when performing genome-wide analyses in large numbers of samples (e.g., up to 100 GB for some of the benchmarking analyses shown in Supplementary Table 2). RnBeads thus provides the option to maintain the *RnBSet* object on hard disk rather than in main memory using the *ff* package [Adler *et al.*, 2014], which is essential for performing large analyses on computers with limited memory. The *RnBSet* object also links the DNA methylation data to genome annotations such as CpG islands, genes and promoters, genome-wide tiling regions and user-defined genomic region sets. RnBeads currently supports the human, mouse and rat genomes with auxiliary data packages named *RnBeads.hg19*, *RnBeads.mm9*, *RnBeads.mm10* and *RnBeads.rn5*. The FAQ section on the RnBeads website describes how users can prepare additional genome assemblies for DNA methylation analysis with RnBeads. The *RnBSet* object primarily stores DNA methylation levels as beta values, which are used by most modules; nevertheless, RnBeads also calculates M-values [Du *et al.*, 2010] and uses them for the *limma* analysis as part of the differential DNA methylation module.

Quality control

RnBeads helps the user identify certain technical and biological biases that are common in large-scale DNA methylation datasets, which includes technical assay failures, sample mix-ups, and batch effects (the latter are addressed by the Exploratory Analysis module and described in the corresponding section below). Quality issues are highlighted in the RnBeads reports, but it is ultimately left to the user to handle them appropriately, e.g. by excluding samples with low technical data quality, by resolving sample mix-ups using genotyping data, or by statistically correcting for batch effects. When RnBeads reports significant quality issues it is typically advisable to consult with an experienced statistician, in order to assess whether or not these issues may be symptoms of more severe problems with the study design or the assay that was used.

The detection of technical failures is assay-specific and differs between sequencing-based and microarray-based analyses. For Infinium data, RnBeads plots the microarray's quality control probes to monitor technical parameters such as bisulfite conversion efficiency and unspecific probe hybridization. For sequencing-based datasets, the quality assessment is largely focused on sequencing coverage, given that bisulfite conversion and clonal read rates are typ-

ically dealt with already during alignment and DNA methylation calling.

RnBeads also addresses the relatively common problem of sample mix-ups [Westra *et al.*, 2011], for example using the genotyping probes that are present on the Infinium microarray to confirm sample identity. As illustrated in Supplementary Figure 1a, clustered heatmaps based on genotype measurements provide a straightforward graphical approach for identifying sample duplications and mix-ups, genetically related individuals, and other types of genetic similarity. RnBeads also calculates inter-sample distances based on these genetic data, which enables users to quantitatively compare sample pairs with respect to their genetic similarity. In addition, RnBeads uses DNA methylation data to predict which samples were derived from male and female donors, based on their X-inactivation status and the presence or absence of measurements on the Y-chromosome. This classifier makes it easy to detect discrepancies between gender information from the sample annotation table and the biological sex of the analyzed samples, which are often indicative of sample mix-ups.

3.2.3 Preprocessing

To minimize the risk of measurement biases affecting the analysis, RnBeads implements a framework for rule-based filtering of samples, CpG sites and DNA methylation measurements. Filtering is performed in two steps, in order to provide flexibility and to avoid biasing the normalization procedure of Infinium analyses with problematic samples. First, RnBeads removes low-quality data that could bias an analysis, discarding samples and CpGs that contain a substantial fraction of measurements with low technical quality (e.g., bad detection p-value for Infinium data or low sequencing coverage in the case of bisulfite sequencing data) as well as CpGs and measurements that may be unreliable for other reasons. For example, RnBeads can remove Infinium probes overlapping SNPs that stand a high chance of influencing DNA methylation measurements; and the default pipeline implements a previously published heuristic for identifying such probes [Nordlund *et al.*, 2013]. Users who wish to apply different criteria can also switch off the default filtering in RnBeads and instead provide a custom list of probes or CpGs that should always be excluded. In a second step, RnBeads discards those samples and CpGs that should be included in the normalization but not in the analysis. Examples are CpGs with too many missing values or with zero variability in their methylation values. Furthermore, users can configure additional filtering rules and define a custom blacklist of CpGs that should always be excluded and/or a whitelist of CpGs that should always be retained. The default filtering criteria of RnBeads were chosen relatively conservatively with the goal of reducing the risk of spurious or misleading results. For datasets with significant quality issues, it can be worthwhile to change the filtering criteria in order to remove problematic probes and samples more aggressively, whereas low-coverage bisulfite sequencing data may require more lenient filtering criteria. All filtering is tracked in the RnBeads report, and before-after plots visualize any changes in the global distribution of DNA methylation levels that may arise from the filtering.

3.2.4 Tracks and Tables

Before proceeding with detailed data analysis, RnBeads exports the preprocessed and quality-controlled data in several formats, thus facilitating data visualization with genome browsers and complementary analyses with other software tools. On the one hand, RnBeads provides track hubs that that can be loaded into various genome browsers, thus providing a common reference point for exploring the bigBed and bigWig data tracks that RnBeads generates. On

the other hand, the software aggregates the preprocessed data in CSV and BED files that can be loaded and analyzed with custom scripts and with web-based tools such as Galaxy [Giardine *et al.*, 2005], the Genomic HyperBrowser [Sandve *et al.*, 2013], EpiGRAPH [Bock, 2009], and EpiExplorer [Halachev *et al.*, 2012]. Furthermore, sample-wise statistics including the number of assayed CpGs and genomic regions, the number of assayed CpGs per region type, and the average read coverage (for sequencing data) are summarized in a dedicated table.

3.2.5 Exploratory Analysis

Global changes in DNA methylation can often be identified by visual inspection of the normalized and quality-controlled DNA methylation data, prior to in-depth analysis of differential DNA methylation. To facilitate this type of exploratory analysis, RnBeads visualizes sample-specific DNA methylation profiles at the single-CpG level and for genomic regions of interest. The global distribution of DNA methylation levels is summarized by density plots, which help identify samples and sample groups that deviate from the characteristic bimodal shape with its clear-cut distinction between highly methylated loci and essentially unmethylated loci (e.g., due to global gain or loss of DNA methylation). RnBeads also provides two types of visualization for DNA methylation variation within and across sample groups, which facilitates the detection of hypervariable samples (e.g., due to technical issues or biological effects such as high tissue heterogeneity). The aforementioned DNA methylation profiles are computed not only based on single CpG measurement values, but also based on methylation levels in predefined regions such as gene promoters or enhancer elements. Furthermore, if the user includes biological or technical replicates in the analysis and identifies them as such (as described in the package vignette), RnBeads calculates pairwise correlations and visualizes them as scatterplots, thereby providing a global assessment of the reproducibility between the experiments.

Hierarchically clustered heatmaps provide a global assessment of sample subtypes in the dataset. This analysis is quantitatively supported by various distance metrics, by the calculation of silhouette statistics to identify the best fitting number of clusters, and by systematic association testing between the obtained clusters and the user-provided sample annotations. Dimension reduction using principal component analysis and multi-dimensional scaling is also available within RnBeads. In combination with interactive sample coloring, this functionality provides a powerful way of visualizing associations between sample annotations and global trends in DNA methylation data. Finally, RnBeads generates composite plots of DNA methylation levels around genes and other genomic regions; and these plots can for example help detect global changes in DNA methylation that affect gene promoters differently compared to intragenic or intergenic regions.

The analysis of global trends and associations is also helpful for detecting batch effects, which can arise from technical confounders such as date and duration of sample processing, the person running the assay, and the sample origin. Batch effects are not uncommon in large-scale DNA methylation datasets, in particular among those generated with microarrays or with enrichment sequencing protocols such as MeDIP and MBD-seq. To systematically detect batch effects, RnBeads runs tests for significant association between user-provided sample annotations (we recommend to include at least the sample collection data, the processing date, and the sample origin) and the directions of largest variance identified in a principal component analysis of the DNA methylation dataset. Statistical testing is also performed to identify significant associations among the sample annotations (e.g., in order to identify problematic confounding between collection date and sample type or disease status) and with quality control indicators such as bisulfite conversion rates and non-specific binding (for Infinium data).

In these comparisons, RnBeads automatically selects the appropriate statistical test (Fisher's exact test, Wilcoxon rank sum test, Kruskal-Wallis one-way analysis of variance, or Pearson correlation coupled with a permutation test) based on the type of annotation data. All results are visualized in the RnBeads report, thus providing a systematic assessment of associations between trends in the DNA methylation data and sample annotations.

3.2.6 Differential DNA methylation

DNA methylation differences can be analyzed not only at the level of individual CpGs, but also by combining measurements across larger genomic regions, which increases statistical power and can result in more interpretable sets of differentially methylated regions [Bock, 2012; Bock and Lengauer, 2008]. In each comparison defined by the sample annotation table, RnBeads initially computes p-values for all covered CpGs. By default, this analysis is performed with hierarchical linear models as implemented in the *limma* package [Smyth, 2004] and using M-values [Du *et al.*, 2010], which exhibit a distribution that is more consistent with *limma*'s statistical model assumptions than the beta values that RnBeads uses in most parts of its analysis. Alternatively, by configuring the *differential.site.test.method* option, p-values can also be calculated using two-sided t-tests or the RefFreeEWAS method [Houseman *et al.*, 2014], which is described in more detail in the section on covariate inference below. In addition to the default unpaired analysis, RnBeads also supports paired-samples analysis, which can substantially increase statistical power when analyzing matched pairs such as tumor versus normal or disease-discordant twins. The CpG-level p-values are corrected for multiple testing using the false discovery rate (FDR) method. Furthermore, to obtain aggregate p-values at the level of predefined genomic regions, the uncorrected, CpG-specific p-values within a given region are combined using an extension of Fisher's method [Makambi, 2003]. This procedure results in a single aggregate p-value for each region, and the aggregate p-values are subjected to multiple testing correction using the FDR method.

In order to address the problem that minimal but consistent differences tend to receive low p-values that do not reflect biological significance, RnBeads ranks the differentially methylated regions according to the combination of statistical significance and effect size. The effect size is estimated in two ways, namely as the absolute difference in DNA methylation and as the relative ratio of mean DNA methylation levels between sample groups. These two measurements differ in their relevance for regions with low versus high DNA methylation levels and thus complement each other. In regions of the genome that exhibit DNA methylation values near 0%, the DNA methylation ratio between sample groups tends to overestimate the effect size, and the absolute DNA methylation difference is a more appropriate measure. The opposite is true for high DNA methylation values near 100%, where the relative ratio is the more stringent and appropriate measure of effect size.

In summary, RnBeads combines statistical testing with a priority ranking scheme that is based on the absolute and relative effect size of the differences between sample groups; and it assigns a combined rank score for differential DNA methylation to each analyzed CpG site and genomic region. This combined rank is defined as the maximum (i.e. worst) of three individual rankings: (i) by absolute difference in mean DNA methylation levels, (ii) by the relative difference in mean DNA methylation levels, which is calculated as the absolute value of the logarithm of the quotient of mean DNA methylation levels, and (iii) by the CpG-based or region-based p-value calculated as described above. The priority-ranked lists can be used directly for downstream analysis, such as manual inspection of the top-ranking regions in a genome browser or for web-based analysis using tools such as Galaxy and EpiExplorer. In

addition to the ranking of differential DNA methylation, RnBeads visualizes the observed differences using scatterplots and volcano plots, and it performs enrichment analysis for Gene Ontology (GO) terms associated with strongly differentially methylated regions.

3.2.7 Covariate inference

Even well-designed studies performed with accurate DNA methylation assays can include confounders and potential sources of batch effects. For example, the samples in an epigenome-wide association study may be collected using different preprocessing steps in different countries or from genetically distinct populations. Furthermore, many large cohort studies are currently being conducted on whole blood, which is characterized by significant cellular heterogeneity. RnBeads implements a number of methods for data correction that can be used to help control such biases.

Batch effects arise from variation in the sample origin or sample handling [Leek *et al.*, 2010], and their effect on the measurements can obscure biologically relevant differences. As long as the batch effects are not too strongly confounded with the biological comparisons of interest, RnBeads together with specialized statistical tools can correct for the resulting biases. To that end, known sources of batch effects (e.g., sample processing date, the microarray slide or the sequencing machine, the origin of clinical samples or the person performing the sample preparation) should be documented by dedicated columns in the sample annotation table, and these columns can then be specified as known confounders when performing the *limma*-based analysis of differential DNA methylation. RnBeads also integrates the surrogate variable analysis method as implemented in the *sva* package [Leek *et al.*, 2012] as an optional step of the standard workflow, which can detect batch effects of unknown origin and annotate them in such a way that they can be controlled for as covariates during *limma* analysis. Furthermore, other methods for batch effect detection such as ComBat [Johnson *et al.*, 2007], ISVA [Teschendorff *et al.*, 2011] and RUV-2 [Gagnon-Bartsch and Speed, 2012] can be applied to *RnBSet* objects as part of custom RnBeads workflows. Any such adjustments should be carefully monitored to avoid introducing additional biases, and it is typically advisable to consult with an experienced statistician when strong batch effects are detected in a dataset.

DNA methylation differences between heterogeneous samples (such as blood, tumor tissue, and most other types of tissue biopsies) can arise not only from cell-intrinsic differences in DNA methylation but also from differences in the cell composition between cases and controls. It is often important to distinguish between these two causes of DNA methylation differences, particularly because they give rise to different biological interpretations [Jaffe and Irizarry, 2014]. RnBeads supports three alternative methods for handling cell type heterogeneity in the context of analyzing differential DNA methylation. First, for certain sample types such as whole blood it is possible to purify reference populations of the most prevalent cell types in the heterogeneous sample and to use their DNA methylation as reference for quantifying differences in cell composition between samples [Houseman *et al.*, 2012]. The estimated cell composition percentages can then be included as covariates in the *limma*-based analysis of differential DNA methylation. This method is most commonly used for epigenome wide association studies performed on patient cohorts for which only whole blood samples are available [Michels *et al.*, 2013]. Suitable reference maps have been generated for the Infinium 450k assay [Reinius *et al.*, 2012]. Any such reference maps must be generated with the same assay and processed in the same RnBeads analysis to minimize bias. It is also important to assess whether there are any strong batch effects between the reference samples and the samples that are to be analyzed, which can be a major issue when using published reference

datasets rather than reestablishing the reference populations in-house. Second, the RefFreeEWAS method has recently been proposed for inferring global trends indicative of cell type heterogeneity directly from the data [Houseman *et al.*, 2014]. RnBeads supports this method as an alternative to *limma* and t-tests in the differential DNA methylation module. Third, the FaST-LMM-EWASher software provides an alternative approach to reference-free analysis of tissue heterogeneity [Zou *et al.*, 2014], and RnBeads can export preprocessed DNA methylation data in a format that can be directly loaded into FaST-LMM-EWASher. However, users should be aware that especially the reference-free methods are still relatively new and susceptible to various biases in the data, such that the results of these analyses – and in fact of any analysis that attempts to correct for tissue heterogeneity – should be carefully checked for statistical as well as biological plausibility.

3.2.8 Implementation details and package design

RnBeads and its companion data packages currently comprise a code base of approximately 32,000 lines of R code, and they export over 200 functions, classes and methods. To structure all functionality in a flexible and easily understandable way, RnBeads utilizes elements of object-oriented programming available in R, and all DNA methylation data are organized in an R S4 class hierarchy. Each analysis module is implemented as an independent unit operating on an *RnBSet* object, and the modules write their results into a hypertext report that employs XHTML and JavaScript to enable self-contained interactivity. The RnBeads reports are organized by figures, which are collections of related plots spanning relevant parts of the parameter space. This setup allows users to dynamically explore each figure without the need to rerun the analysis. The *ggplot2* package [Wickham, 2009] is used to generate publication-grade plots, which are incorporated in the reports as bitmaps for quick visualization and as vector graphics for high-resolution printing and for custom postprocessing using vector graphics software. Heatmaps are visualized using the heatmap.2 functionality of the *gplots* package [Warnes, 2012]. Genome browser like views are created using the *Gviz* package [Hahne *et al.*, 2014].

3.2.9 Scalability and performance

RnBeads has been designed to be scalable to large sample sizes and efficient in its use of computational resources. Parallel computation is implemented using the *foreach* and *doParallel* packages; and large R objects can be maintained directly on hard disk using the *ff* package, which leads to a massive reduction of the memory required for large analyses. Small RnBeads analyses can be completed on a standard personal computer, while large analyses should be run a scientific computing cluster (or on adequately powered cloud computing infrastructure). For users who prefer a web-based workflow, a web server supporting analyses with up to 24 samples is available on the RnBeads website. Furthermore, it is relatively straightforward to run RnBeads in an academic or commercial cloud computing environment using an instance of Galaxy CloudMan [Afgan *et al.*, 2010], as described in the FAQ section on the RnBeads website. RnBeads has been tested successfully on Infinium datasets comprising thousands of samples, on reduced presentation bisulfite sequencing datasets with hundreds of samples and on whole genome bisulfite sequencing datasets with dozens of deeply sequenced methylomes. Nevertheless, analyses of this scale require careful planning and configuration to avoid out-of-memory problems or excessive runtime. Supplementary Table 2 lists runtime measurements of RnBeads for several large datasets, and the RnBeads documentation provides additional instructions on how to set up large analyses.

3.2.10 Methylome resource

The Methylome Resource on the RnBeads website (http://rnbeads.mpi-inf.mpg.de/methylomes.php) was established by applying RnBeads to the largest public datasets that are currently available for whole genome bisulfite sequencing, for reduced representation bisulfite sequencing, and for the Infinium 450k assay. This resource provides a reference of large-scale DNA methylation analyses that can be used in various ways. For example, researchers can browse through the reports online, explore biological hypotheses, and investigate relevant aspects of the data visually or through custom data analysis with R or other software tools. Furthermore, researchers can download the data and configuration files of the Methylome Resource, add their own DNA methylation data, and then run RnBeads in order to analyze their data in the context of high-quality methylome datasets that span a broad set of tissue types.

For whole genome bisulfite sequencing, the Methylome Resource covers DNA methylation profiles of 41 samples with coverage of 28,158,385 CpGs [Ziller *et al.*, 2013]. These methylomes are compiled from several sources, including the activities of the Roadmap Epigenomics Project and the International Human Epigenome Consortium [Satterlee *et al.*, 2010], and they span a broad range of human cell types. For reduced representation bisulfite sequencing, we obtained DNA methylation profiles for 216 samples with coverage of 2,295,083 CpGs from the ENCODE project [ENCODE Project Consortium, 2004], which comprises cell lines and primary samples of various normal and cancerous tissue types [Varley *et al.*, 2013]. Finally, for the Illumina Infinium 450k assay, we downloaded raw intensity files for 4,034 primary tumor and normal control samples with microarray coverage of 485,577 CpGs, which have been collected by the TGCA consortium [Weisenberger, 2014]. All data were processed according to a standardized RnBeads workflow, and these analyses could be completed in no more than a few days on a standard scientific computing cluster (Supplementary Table 2).

3.2.11 Availability and website

Additional materials, including the RnBeads download, the package vignette, the source code, an RnBeads web service, commands and configurations for cloud-based RnBeads analysis, example analysis reports, the methylome resource, documentation and FAQs are available on the RnBeads website (http://rnbeads.mpi-inf.mpg.de/).

3.3 Supplementary Material

Supplementary Note

As an example for RnBeads-based analysis of Infinium 450k data, we performed a reanalysis of a publicly available glioblastoma dataset generated by The Cancer Genome Atlas (TCGA) project [Weisenberger, 2014]. Glioblastoma multiforme is an aggressive type of brain cancer with a median survival time of little more than a year and substantial variation between patients [Wen and Kesari, 2008]. In an attempt to stratify patients according to the molecular characteristics of the tumors, recent research has identified a subtype that is characterized by elevated levels of DNA methylation, prolonged survival and high frequency of mutations in the IDH1 gene [Noushmehr *et al.*, 2010]. The discovery of this "glioblastoma CpG island methylator phenotype positive" (G-CIMP+) subtype was based on Illumina's Infinium 27k assay, prompting us to validate this observation using RnBeads and an extended dataset of Infinium 450k profiles for 124 glioblastoma patients.

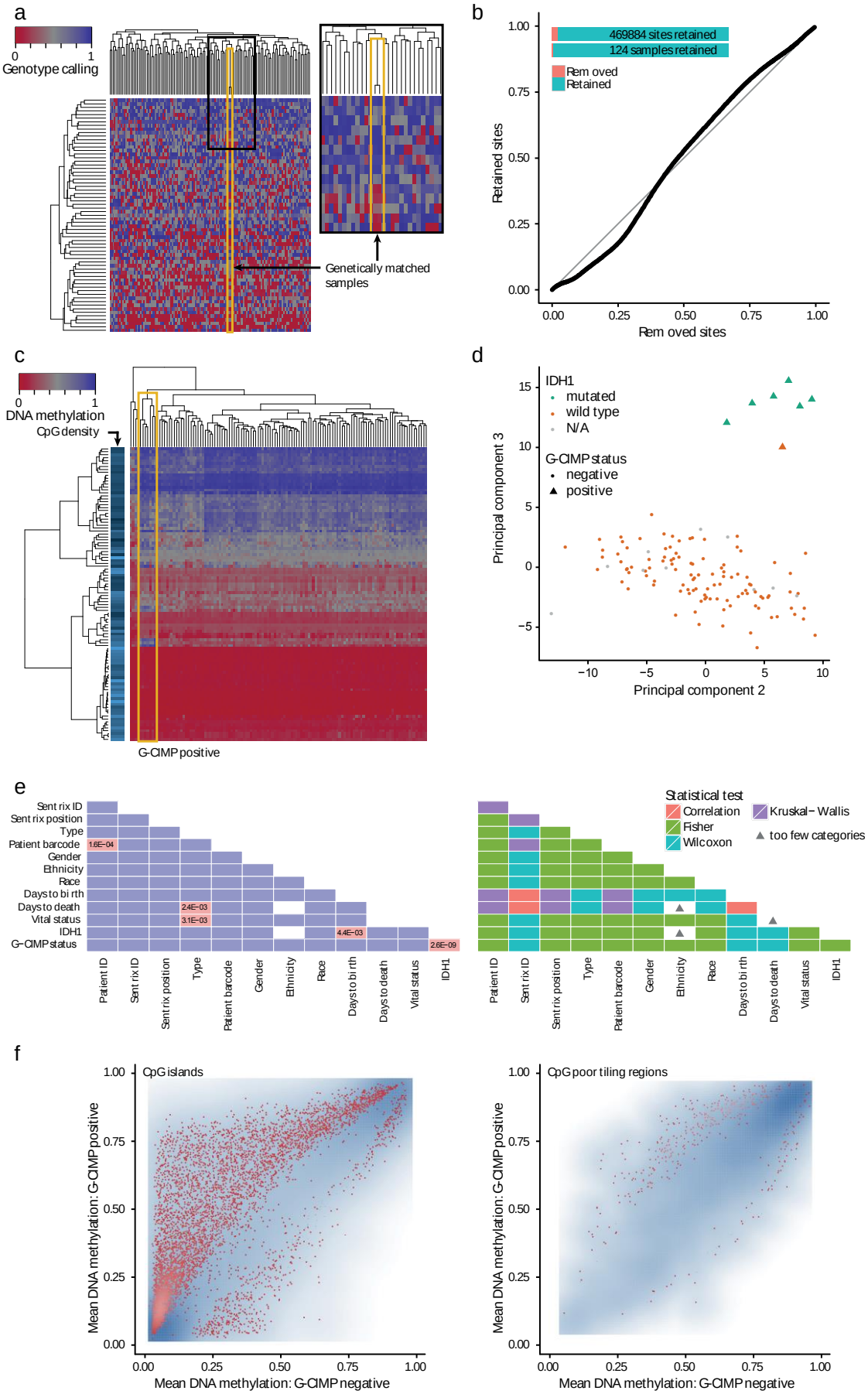
We downloaded the raw microarray signal intensity files (in IDAT format) from the TCGA website (http://tcga-data.nci.nih.gov), created a sample annotation file that contains the available patient data – including IDH1 mutation status – and then launched RnBeads. The software identifies the data directory and input file format from the annotation file and normalizes the raw intensity data using SWAN [Makismovic *et al.*, 2012] (other normalization algorithms are supported as well, as described in the Online Methods). CpG-specific DNA methylation levels are obtained from the normalized data and collected in an *RnBSet* object that provides the basis for all subsequent analyses. During quality control, RnBeads performs clustering of all samples based on genotype fingerprinting probes included on the Infinium microarray (Supplementary Figure 1a), which provides an effective method for identifying sample mix-ups and duplications. Here, we identified two samples with identical SNP patterns, in concordance with their TCGA annotation as primary and recurrent tumors from the same patient. All other samples were taken from genetically unrelated patients. RnBeads provides flexible features for data filtering as part of the preprocessing module (Supplementary Figure 1b), which are useful for excluding measurements that could bias the analysis (e.g., due to low signal quality, overlap with SNPs, or X-chromosome association in case of different sex ratios between cases and controls).

Based on the filtered and quality-controlled dataset, RnBeads performs hierarchical clustering to facilitate data exploration and outlier detection. In the clustered heatmap, we observe a small and distinct group of samples with increased promoter hypermethylation suggestive of the G-CIMP+ subtype (Supplementary Figure 1c). These putative G-CIMP+ samples indeed exhibit the characteristic enrichment of IDH1 mutations and a clear separation with respect to their global DNA methylation levels – patterns that are particularly evident from a low-dimensional projection of the entire dataset that has been annotated with IDH1 mutation status and G-CIMP subtype information (Supplementary Figure 1d). The significance of this association is also confirmed by pairwise statistical tests for associations that RnBeads performs between all sample annotations (Supplementary Figure 1e). Furthermore, RnBeads calculates groupwise comparisons between the mean DNA methylation levels in the G-CIMP positive versus negative samples for CpG islands and for genome-wide tiling regions (Supplementary Figure 1f). The resulting scatterplots show that the gain of DNA methylation among the G-CIMP+ samples is more pronounced in CpG islands than in genomic regions exhibiting low CpG content.

These automated, exploratory analyses provide a starting point for dissecting the patterns and mechanisms of epigenetic deregulation that may affect DNA methylation in G-CIMP+ tumors. Follow-up analyses can be performed directly in R, most conveniently by using the precalculated *RnBSet* data object that RnBeads prepares as part of the initial analysis. Furthermore, RnBeads makes it easy to export the data and results in a variety of formats and to hand them over to stand-alone or web-based bioinformatic tools for further analysis.

Supplementary Figures

Figure 3.S1: (on the next page) Analysis of DNA methylation in a cancer cohort based on Infinium 450k data. RnBeads was used to rediscover a clinically distinct subgroup of glioblastoma patients characterized by increased DNA methylation levels (termed G-CIMP+), and to predict the G-CIMP status for a total of 124 patients using Infinium 450k **a**. Detection of genetic duplicates among the patient samples (columns) using a clustered heatmap of intensity values for the genotyping probes that are present on the Infinium microarray (rows). The inset shows that two samples exhibit a high level of genetic identity, and they are indeed derived from tumors of the same patient. **b**. Quality control plot summarizing the outcome of the data filtering. The bar plots on the top left show that the majority of CpG sites (top) and samples (bottom) are of good quality and can be retained. The relatively straight line in the quantile-quantile plot indicates that the probe filtering does not have a major impact on the distribution of DNA methylation in the dataset. **c**. Identification of a small but clearly distinguished cluster of G-CIMP+ glioblastoma samples with elevated DNA methylation levels especially in CpG-rich genomic regions (dark blue in the leftmost column). In the heatmap, blue colors denote high levels of DNA methylation, red indicates low levels and grey represents intermediate levels. For visualization purposes, only the 100 gene promoters (rows) with the highest levels of inter-sample variation in DNA methylation are shown (columns), but the hierarchical clustering is based on **d**. Global assessment of the similarity between the DNA methylation profiles, plotting all glioblastoma samples according to their second and third principal components. The samples exhibit strong separation according to the G-CIMP status (denoted by point shape) and IDH1 mutation status (denoted by point color). **e**. Analysis of significant associations between all user-provided right triangle (orange: Pearson correlation followed by permutation-based estimation of the p-value; green: Fisher's exact test; blue: Wilcoxon rank sum test; violet: Kruskal-Wallis one-way analysis of variance). **f**. Genome-scale comparison between the DNA methylation levels of G-CIMP positive (y-axis) and G-CIMP negative (x-axis) tumor samples, focusing on CpG islands (left scatterplot) and on 5-kilobase tiling regions with a CpG content in the bottom quartile (right scatterplot), respectively. Genomic regions that are differentially methylated with an FDR below 0.05 are presented as red points. All other regions are displayed in blue, and color brightness denotes point density.


a
Genotype calling
Genetically matched samples
b
469884 sites retained
124 samples retained
Rem oved
Retained
Retained sites
Rem oved sites
c
DNA methylation
CpG density
G-CIMP positive
d
IDH1
mutated
wild type
N/A
G-CIMP status
negative
positive
Principal component 3
Principal component 2
e
Sent rix ID
Sent rix position
Type
Patient barcode
Gender
Ethnicity
Race
Days to bi rth
Days to death
Vital status
IDH1
G-CIMP status
Statistical test
Correlation
Fisher
Wilcoxon
Kruskal− Wallis
too few categories
f
CpG islands
CpG poor tiling regions
Mean DNA methylation: G-CIMP positive
Mean DNA methylation: G-CIMP negative

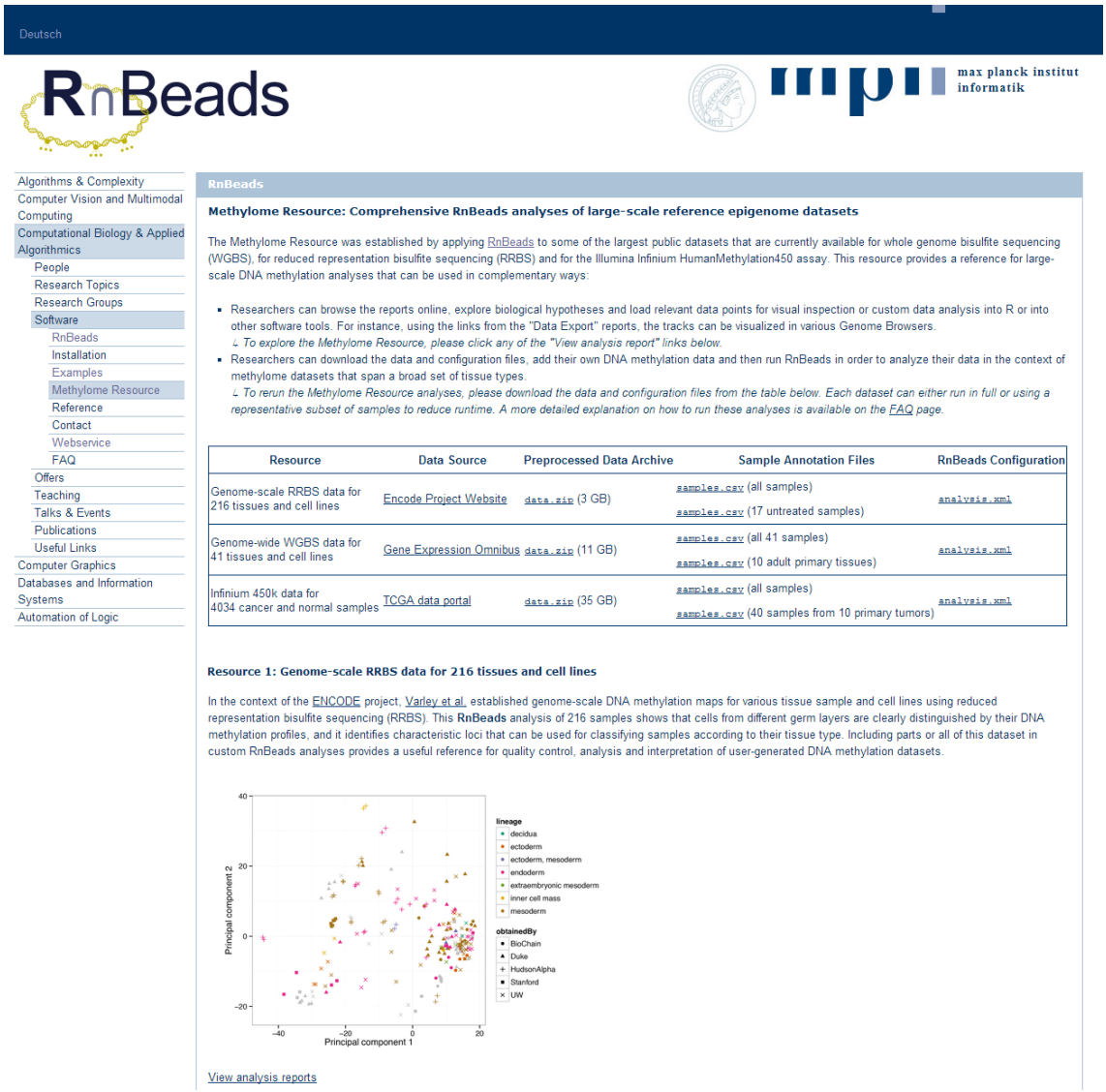


Figure 3.S2: RnBeads-based Methylome Resource of reference epigenome datasets. Screenshot of the Methylome Resource (http://rnbeads.mpi-inf.mpg.de/methylomes.php), which makes large-scale DNA methylation datasets readily available for follow-up research. On the one hand, it provides detailed analysis reports for publicly available methylome datasets that can be explored interactively. On the other hand, the Methylome Resource website lets RnBeads users download all data and configurations that are needed to re-run all or part of the DNA methylation analyses in their local or cloud-based computing environment. These re-runnable analysis configurations make it straightforward for RnBeads users to analyze their own DNA methylation data in the context of publicly available reference epigenome maps.

Supplementary Tables

Table 3.S1: Comparison between software tools for DNA methylation analysis

The table as an Excel (.xlsx) file is available from:
http://www.nature.com/nmeth/journal/v11/n11/extref/nmeth.3115-S2.xlsx

Table 3.S2: Performance benchmark for large DNA methylation analyses with RnBeads

Data type[1]	No. of Samples[2]	No. of CpGs[3]	No. of Annotations[4]	No. of Comparisons[5]	Runtime (node)[6]	Runtime (cluster)[7]
Infinium 450k	100	485,577	2	2	2h 12min	1h 9min
Infinium 450k	500	485,577	6	6	15h 2min	7h 29min
Infinium 450k	1000	485,577	10	10	1d 13h 51min	20h 15min
Infinium 450k	4034*	485,577	5	18	9d 7h 21min	6d 18h 40min
RRBS	10	1,804,103	2	2	1h 56min	49min
RRBS	50	2,169,859	6	6	5h 32min	1h 54min
RRBS	100	2,221,920	10	10	10h 13min	2h 57min
RRBS	216*	2,295,083	7	11	1d 8h 50min	14h 27min
WGBS	5	28,132,494	2	2	20h 43min	8h 23min
WGBS	10	28,150,344	6	6	2d 10h 23min	20h 5min
WGBS	20	28,154,125	10	10	4d 12h 21min	1d 15h 34min
WGBS	41*	28,158,385	5	6	3d 4h 54min	1d 9h 27min

[1] Data from the following sources were included in the analysis: TCGA (Infinium 450k), ENCODE (RRBS), Ziller et al. (WGBS)
[2] Subsets of the full datasets were randomly generated in order to assess the effect of sample size on runtime
[3] Number of CpG sites present in at least one sample. For RRBS/WGBS, low-coverage sites are removed prior to counting
[4] Adding more columns to the sample annotation table increases the complexity and runtime of the analysis
[5] Including more pairwise comparisons in the analysis strongly increases runtime but can be parallelized effectively
[6] Serial runtime measured on a scientific computing cluster (16 nodes), summing up the runtime of all contributing nodes
[7] Parallel runtime / time to completion on a scientific computing cluster (16 nodes) with optimal use of job parallelization
* The analysis results for the full datasets are available as part of the Methylome Resource on the RnBeads website

References

Adler, D., Gläser, C., Nenadic, O., Oehlschlägel, J., and Zucchini, W. ff: Memory-efficient storage of large data on disk and fast access functions. Reference manual. http://cran.r-project.org/web/packages/ff/index.html. 2014.

Afgan, E., Baker, D., Coraor, N., Chapman, B., Nekrutenko, A., and Taylor, J. Galaxy CloudMan: delivering cloud compute clusters. *BMC Bioinformatics*, 11 Suppl 1:S4, 2010.

Akalin, A., Kormaksson, M., Li, S., Garrett-Bakelman, F. E., Figueroa, M. E., Melnick, A. *et al.* methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol*, 13(10):R87, 2012.

Akman, K., Haaf, T., Gravina, S., Vijg, J., and Tresch, A. Genome-wide quantitative analysis of DNA methylation from bisulfite sequencing data. *Bioinformatics*, 30(13):1933–1934, 2014.

Aryee, M. J., Jaffe, A. E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A. P., Hansen, K. D. *et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*, 2014.

Barfield, R. T., Kilaru, V., Smith, A. K., and Conneely, K. N. CpGassoc: an R function for analysis of DNA methylation microarray data. *Bioinformatics*, 28(9):1280–1281, 2012.

Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J. M. *et al.* High density DNA methylation array with single CpG site resolution. *Genomics*, 98(4):288–295, 2011.

Bock, C. Epigenetic biomarker development. *Epigenomics*, 1(1):99–110, 2009.

Bock, C. Analysing and interpreting DNA methylation data. *Nature Reviews Genetics*, 13(10):705–719, 2012.

Bock, C. Synergy and competition between cancer genome sequencing and epigenome mapping projects. *Genome Med*, 6(5):41, 2014.

Bock, C. and Lengauer, T. Computational epigenetics. *Bioinformatics*, 24(1):1–10, 2008.

Bock, C., Tomazou, E. M., Brinkman, A. B., Müller, F., Simmer, F., Gu, H. *et al.* Quantitative comparison of genome-wide DNA methylation mapping technologies. *Nature Biotechnology*, 28(10):1106–1114, 2010.

Chavez, L., Jozefczuk, J., Grimm, C., Dietrich, J., Timmermann, B., Lehrach, H. *et al.* Computational analysis of genome-wide DNA methylation during the differentiation of human embryonic stem cells along the endodermal lineage. *Genome Res*, 20(10):1441–1450, 2010.

Davis, S., Du, P., Bilke, S., Triche Jr., T. J., and Bootwalla, M. methylumi: Handle Illumina methylation data. Reference manual. http://bioconductor.org/packages/release/bioc/html/methylumi.html. 2014.

Down, T. A., Rakyan, V. K., Turner, D. J., Flicek, P., Li, H., Kulesha, E. *et al.* A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nat Biotechnol*, 26(7):779–785, 2008.

Du, P., Zhang, X., Huang, C.-C., Jafari, N., Kibbe, W. a., Hou, L. *et al.* Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC bioinformatics*, 11:587, 2010.

Du, Y., Murani, E., Ponsuksili, S., and Wimmers, K. BiomvRhsmm: Genomic segmentation with hidden semi-Markov model. *BioMed Research International*, 2014:910390, 2014.

ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, 306(5696):636–640, 2004.

Flicek, P., Ahmed, I., Amode, M. R., Barrell, D., Beal, K., Brent, S. *et al.* Ensembl 2013. *Nucleic Acids Res*, 41(D1):D48–55, 2013.

Fortin, J.-P., Fertig, E., and Hansen, K. shinyMethyl: interactive quality control of Illumina 450k DNA methylation arrays in R. *F1000Research*, 3:175, 2014.

Gagnon-Bartsch, J. A. and Speed, T. P. Using control genes to correct for unwanted variation in microarray data. *Biostatistics*, 13(3):539–552, 2012.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5(10):R80, 2004.

Giardine, B., Riemer, C., Hardison, R. C., Burhans, R., Elnitski, L., Shah, P. *et al.* Galaxy: A platform for interactive large-scale genome analysis. *Genome Research*, 15(10):1451–1455, 2005.

Gu, H., Bock, C., Mikkelsen, T. S., Jäger, N., Smith, Z. D., Tomazou, E. *et al.* Genome-scale DNA methylation mapping of clinical samples at single-nucleotide resolution. *Nat Methods*, 7(2):133–136, 2010.

Hahne, F., Durinck, S., Ivanek, R., Mueller, A., Lianoglou, S., Tan, G. *et al.* Gviz: Plotting data and annotation information along genomic coordinates. Reference manual. http://bioconductor.org/packages/release/bioc/html/Gviz.html. 2014.

Halachev, K., Bast, H., Albrecht, F., Lengauer, T., and Bock, C. EpiExplorer: live exploration and global analysis of large epigenomic datasets. *Genome Biol*, 13(10):R96, 2012.

Hansen, K. D., Langmead, B., and Irizarry, R. A. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol*, 13(10):R83, 2012.

Harris, R. A., Wang, T., Coarfa, C., Nagarajan, R. P., Hong, C., Downey, S. L. *et al.* Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nature Biotechnology*, 28(10):1097–1105, 2010.

He, J., Sun, X., Shao, X., Liang, L., and Xie, H. DMEAS: DNA methylation entropy analysis software. *Bioinformatics*, 29(16):2044–2045, 2013.

Hebestreit, K., Dugas, M., and Klein, H. U. Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. *Bioinformatics*, 29(13):1647–1653, 2013.

Houseman, E. A., Accomando, W. P., Koestler, D. C., Christensen, B. C., Marsit, C. J., Nelson, H. H. *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*, 13(1):86, 2012.

Houseman, E. A., Molitor, J., and Marsit, C. J. Reference-free cell mixture adjustments in analysis of DNA methylation data. *Bioinformatics*, 30(10):1431–1439, 2014.

Illumina Inc. GenomeStudio software data sheet. 2014.

Jaffe, A. E. and Irizarry, R. a. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biology*, 15(2):R31, 2014.

Johnson, W. E., Li, C., and Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1):118–127, 2007.

Kilaru, V., Barfield, R. T., Schroeder, J. W., Smith, A. K., and Conneely, K. N. MethLAB: A graphical user interface package for the analysis of array-based DNA methylation data. *Epigenetics*, 7(3):225–229, 2012.

Krueger, F. and Andrews, S. R. Bismark: a flexible aligner and methylation caller for bisulfite-seq applications. *Bioinformatics*, 27(11):1571–1572, 2011.

Kuan, P. F., Wang, S., Zhou, X., and Chu, H. A statistical framework for Illumina DNA methylation arrays. *Bioinformatics*, 26(22):2849–2855, 2010.

Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., and Storey, J. D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, 28(6):882–883, 2012.

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E. *et al.* Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews. Genetics*, 11(10):733–9, 2010.

Lienhard, M., Grimm, C., Morkel, M., Herwig, R., and Chavez, L. MEDIPS: genome-wide differential coverage analysis of sequencing data derived from DNA enrichment experiments. *Bioinformatics*, 30(2):284–286, 2014.

Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 462(7271):315–322, 2009.

Liu, Y., Siegmund, K. D., Laird, P. W., and Berman, B. P. Bis-SNP: combined DNA methylation and SNP calling for Bisulfite-seq data. *Genome Biology*, 13(7):R61, 2012.

Makambi, K. H. Weighted inverse chi-square method for correlated significance tests. *Journal of Applied Statistics*, 30(2):225–234, 2003.

Makismovic, J., Gordon, L., and Oshlack, A. SWAN: Subset-quantile Within Array Normalization for Illumina Infinium HumanMethylation450 BeadChips. *Genome Biol*, 13(6):R44, 2012.

Mancuso, F. M., Montfort, M., Carreras, A., Alibes, A., and Roma, G. HumMeth27QCReport: an R package for quality control and primary analysis of Illumina Infinium methylation data. *BMC Res Notes*, 4(1):546, 2011.

Meyer, L. R., Zweig, A. S., Hinrichs, A. S., Karolchik, D., Kuhn, R. M., Wong, M. *et al.* The UCSC Genome Browser database: extensions and updates 2013. *Nucleic Acids Res*, 41(Database issue):D64–9, 2013.

Michels, K. B., Binder, A. M., Dedeurwaerder, S., Epstein, C. B., Greally, J. M., Gut, I. *et al.* Recommendations for the design and analysis of epigenome-wide association studies. *Nature Methods*, 10(10):949–55, 2013.

Morris, T. J., Butcher, L. M., Feber, A., Teschendorff, A. E., Chakravarthy, A. R., Wojdacz, T. K. *et al.* ChAMP: 450k Chip Analysis Methylation Pipeline. *Bioinformatics*, 30(3):428–430, 2014.

Nordlund, J., Backlin, C. L., Wahlberg, P., Busche, S., Berglund, E. C., Eloranta, M. L. *et al.* Genome-wide signatures of differential DNA methylation in pediatric acute lymphoblastic leukemia. *Genome Biol*, 14(9):r105, 2013.

Noushmehr, H., Weisenberger, D. J., Diefes, K., Phillips, H. S., Pujara, K., Berman, B. P. *et al.* Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. *Cancer Cell*, 17(5):510–522, 2010.

Park, Y., Figueroa, M. E., Rozek, L. S., and Sartor, M. A. MethylSig: A whole genome DNA methylation analysis pipeline. *Bioinformatics*, 30(17):2414–2422, 2014.

Pidsley, R., Y Wong, C. C., Volta, M., Lunnon, K., Mill, J., Schalkwyk, L. C. *et al.* A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics*, 14(1):293, 2013.

Reinius, L. E., Acevedo, N., Joerink, M., Pershagen, G., Dahlén, S.-E., Greco, D. *et al.* Differential DNA Methylation in Purified Human Blood Cells: Implications for Cell Lineage and Studies on Disease Susceptibility. *PLoS ONE*, 7(7):e41361, 2012.

Riebler, A., Menigatti, M., Song, J. Z., Statham, A. L., Stirzaker, C., Mahmud, N. *et al.* BayMeth: improved DNA methylation quantification for affinity capture sequencing data using a flexible Bayesian approach. *Genome Biol*, 15(2):R35, 2014.

Saito, Y., Tsuji, J., and Mituyama, T. Bisulfighter: accurate detection of methylated cytosines and differentially methylated regions. *Nucleic Acids Res*, 42(6):e45, 2014.

Sandve, G. K., Gundersen, S., Johansen, M., Glad, I. K., Gunathasan, K., Holden, L. *et al.* The Genomic HyperBrowser: an analysis web server for genome-scale data. *Nucleic Acids Res*, 41(Web Server issue):W133–41, 2013.

Satterlee, J. S., Schubeler, D., and Ng, H. H. Tackling the epigenome: challenges and opportunities for collaboration. *Nat Biotechnol*, 28(10):1039–1044, 2010.

Smyth, G. K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, 3:Article3, 2004.

Stevens, M., Cheng, J. B., Li, D., Xie, M., Hong, C., Maire, C. L. *et al.* Estimating absolute methylation levels at single-CpG resolution from methylation enrichment and restriction enzyme sequencing methods. *Genome Res*, 23(9):1541–1553, 2013.

Teschendorff, A. E., Marabita, F., Lechner, M., Bartlett, T., Tegner, J., Gomez-Cabrero, D. *et al.* A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics*, 29(2):189–196, 2013.

Teschendorff, A. E., Zhuang, J., and Widschwendter, M. Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies. *Bioinformatics (Oxford, England)*, 27(11):1496–505, 2011.

Varley, K. E., Gertz, J., Bowling, K. M., Parker, S. L., Reddy, T. E., Pauli-Behn, F. *et al.* Dynamic DNA methylation across diverse human cell lines and tissues. *Genome research*, 23(3):555–67, 2013.

Wang, D., Szyf, M., Benkelfat, C., Provençal, N., Turecki, G., Caramaschi, D. *et al.* Peripheral SLC6A4 DNA methylation is associated with in vivo measures of human brain serotonin synthesis and childhood physical aggression. *PLoS ONE*, 7(6), 2012.

Warden, C. D., Lee, H., Tompkins, J. D., Li, X., Wang, C., Riggs, A. D. *et al.* COHCAP: an integrative genomic pipeline for single-nucleotide resolution DNA methylation analysis. *Nucleic Acids Res*, 41(11):e117, 2013.

Warnes, G. R. Includes R source code and/or documentation contributed by Ben Bolker, Lodewijk Bonebakker, Robert Gentleman, Wolfgang Huber Andy Liaw, Thomas Lumley, Martin Maechler, Arni Magnusson, Steffen Moeller, Marc Schwartz and Bill Venables. 2012. gplots: Various. 2012.

Weisenberger, D. J. Characterizing DNA methylation alterations from The Cancer Genome Atlas. *J Clin Invest*, 124(1):17–23, 2014.

Wen, P. Y. and Kesari, S. Malignant gliomas in adults. *N Engl J Med*, 359(5):492–507, 2008.

Westra, H. J., Jansen, R. C., Fehrmann, R. S., te Meerman, G. J., van Heel, D., Wijmenga, C. *et al.* MixupMapper: correcting sample mix-ups in genome-wide datasets increases power to detect small genetic effects. *Bioinformatics*, 27(15):2104–2111, 2011.

Wickham, H. *ggplot2: elegant graphics for data analysis*. Springer Publishing Company, Incorporated, 2009. ISBN 0387981403.

Wilson, G., Dhami, P., Feber, A., Cortazar, D., Suzuki, Y., Schulz, R. *et al.* Resources for methylome analysis suitable for gene knockout studies of potential epigenome modifiers. *GigaScience*, 1(1):3, 2012.

Wu, D., Gu, J., and Zhang, M. Q. FastDMA: An Infinium HumanMethylation450 Beadchip Analyzer. *PLoS One*, 8(9):e74275, 2013.

Xi, Y., Bock, C., Müller, F., Sun, D., Meissner, A., and Li, W. RRBSMAP: a fast, accurate and user-friendly alignment tool for reduced representation bisulfite sequencing. *Bioinformatics*, 28(3):430–432, 2012.

Xi, Y. and Li, W. BSMAP: whole genome bisulfite sequence MAPping program. *BMC Bioinformatics*, 10:232, 2009.

Zhang, Y., Su, J., Yu, D., Wu, Q., and Yan, H. EpiDiff: Entropy-based quantitative identification of differential epigenetic modification regions from epigenomes. *Conf Proc IEEE Eng Med Biol Soc*, 2013:655–658, 2013.

Zhou, X., Maricque, B., Xie, M., Li, D., Sundaram, V., Martin, E. A. *et al.* The Human Epigenome Browser at Washington University. *Nat Methods*, 8(12):989–990, 2011.

Ziller, M. J., Gu, H., Müller, F., Donaghey, J., Tsai, L. T.-Y., Kohlbacher, O. *et al.* Charting a dynamic DNA methylation landscape of the human genome. *Nature*, 500(7463):477–81, 2013.

Zou, J., Lippert, C., Heckerman, D., Aryee, M., and Listgarten, J. Epigenome-wide association studies without the need for cell-type composition. *Nature Methods*, 11(3):309–11, 2014.

Chapter 4

BiQ Analyzer HT: Locus-specific analysis of DNA methylation by high-throughput bisulfite sequencing

The full text of this chapter has been earlier published as:

Pavlo Lutsik[1-2], Lars Feuerbach[2], Julia Arand[1], Thomas Lengauer[2], Jörn Walter[1,#] and Christoph Bock[2-5,#] (2011) *Nucleic Acids Research* **39**(Suppl. 2), W551–6.

The author of the thesis co-designed and coded BiQ Analyzer HT software, created the program web-site, performed the benchmarking analysis, wrote the first draft of the manuscript, prepared figures and tables.

[1]Genetics/Epigenetics, Saarland University, Saarbrücken, Germany
[2]Max Planck Institute for Informatics, Saarbrücken, Germany
[3]Broad Institute, Cambridge, Massachusetts, USA
[4]Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, Massachusetts, USA
[5]Harvard Stem Cell Institute, Cambridge, Massachusetts, USA
[#]To whom correspondence should be addressed.

Abstract

Bisulfite sequencing is a widely used method for measuring DNA methylation in eukaryotic genomes. The assay provides single-basepair resolution and, given sufficient sequencing depth, its quantitative accuracy is excellent. High-throughput sequencing of bisulfite-converted DNA can be applied either genome-wide or targeted to a defined set of genomic loci (e.g. using locus-specific PCR primers or DNA capture probes). Here we describe BiQ Analyzer HT (http://biq-analyzer-ht.bioinf.mpi-inf.mpg.de/), a user-friendly software tool that supports locus-specific analysis and visualization of high-throughput bisulfite sequencing data. The software facilitates the shift from time-consuming clonal bisulfite sequencing to the more quantitative and cost-efficient use of high-throughput sequencing for studying locus-specific DNA methylation patterns. In addition, it is useful for locus-specific visualization of genome-wide bisulfite sequencing data.

Availability

http://biq-analyzer-ht.bioinf.mpi-inf.mpg.de (This website/software is free and open to all users and there is no login requirement).

Funding

CANCERDIP project (HEALTH-F2-2007-200620); ColoNet project (BMBF 0315417-D). Funding for open access charge: Max Planck Institute for Informatics and Saarland University.

Conflict of interest statement. None declared.

Acknowledgements

We would like to thank Dr Sascha Tierling and Dirk Schuemacher for helpful discussions and the provision of test data, Yassen Assenov for advice with Java programming and Fabian Müller for advice on the BAM format.

4.1 Introduction

DNA methylation is a widely studied epigenetic modification. It is present in all vertebrates and many invertebrate animals as well as in plants [Suzuki and Bird, 2008]. In mammals, DNA methylation plays an important role for developmental gene regulation and for germline repression of repetitive elements [Bird, 2002]. Aberrant DNA methylation patterns are frequently observed in cancer [Esteller, 2008] and may also occur in many other human diseases [Feinberg, 2007]. The link between locus-specific DNA methylation alterations and common diseases has created significant interest in using these epigenetic alterations as biomarkers in drug discovery and clinical diagnostics [Laird, 2003].

To investigate the many roles of DNA methylation in development and disease, researchers depend on experimental methods that accurately measure DNA methylation patterns at high accuracy and affordable cost. Many technologies with different advantages and disadvantages have been developed over the last twenty years, but only bisulfite-based methods provide quantitative DNA methylation data at single-basepair resolution [Laird, 2010]. In bisulfite sequencing the DNA is treated with sodium bisulfite, which selectively converts unmethylated cytosines into uracils but leaves methylated cytosines untouched [Frommer *et al.*, 1992]. Hydroxymethylated DNA, which has recently been detected in some mammalian cell types, is also left unconverted and is indistinguishable from methylated DNA using bisulfite-based methods [Huang *et al.*, 2010].

Bisulfite sequencing has recently been used to obtain the first genome-wide, high-resolution maps of DNA methylation in the human genome [Laurent *et al.*, 2010; Lister *et al.*, 2009]. Bisulfite-based methods also performed well in a benchmarking study of DNA methylation mapping technologies [Bock *et al.*, 2010]. Along with technologies for DNA methylation mapping at a genomic scale, locus-specific bisulfite sequencing plays an important role as gold-standard validation method and promises to become a standard technology in clinical diagnostics [Bock, 2009].

Locus-specific bisulfite sequencing has traditionally been performed by Sanger sequencing of a few dozen hand-picked DNA clones, making this method rather time-consuming and costly. To address these limitations, researchers increasingly use high-throughput sequencing instead of Sanger sequencing [Korshunova *et al.*, 2008; Taylor *et al.*, 2007; Varley and Mitra, 2010], which has three major advantages: (i) Due to the increased sequencing throughput it becomes feasible to obtain highly quantitative DNA methylation patterns for the loci of interest. This is particularly relevant for studying heterogeneous tissue samples and for clinical diagnostics. (ii) Due to lower per-base costs and the use of multiplexing to sequence many samples and/or loci in a single machine run, the sequencing costs are substantially reduced. (iii) The cloning step for isolating DNA populations that carry the DNA sequence of a single DNA molecule becomes obsolete because current methods for high-throughput sequencing measure the sequences of individual DNA clones.

A major roadblock for the wider use of high-throughput bisulfite sequencing is the lack of software tools for processing and analyzing the vast number of sequencing reads that are characteristic of this method. Several software tools have been developed for processing small-scale bisulfite sequencing data obtained by conventional Sanger sequencing. The BiQ Analyzer [Bock *et al.*, 2005] software from our group has recently been updated to version 2.0 and continues to be a useful tool for interactive analysis of small-scale bisulfite sequencing data. Alternative tools include the QUMA web service [Kumaki *et al.*, 2008], BISMA [Rohde *et al.*, 2010], and several more specialized programs [Carr *et al.*, 2007; Grunau *et al.*, 2000; Hetzl *et al.*, 2007; Xu *et al.*, 2007]. None of these tools can be scaled to the read numbers that are typically

obtained by high-throughput sequencing. For this reason, recent studies utilized custom data analysis scripts, none of which are publicly available [Korshunova *et al.*, 2008; Taylor *et al.*, 2007; Varley and Mitra, 2010].

Here we describe BiQ Analyzer HT, a comprehensive software tool for locus-specific analysis of high-throughput bisulfite sequencing data. BiQ Analyzer HT builds on concepts that we originally developed for the popular BiQ Analyzer software [Bock *et al.*, 2005], but it was redesigned and rewritten to meet the challenges arising for the analysis of high-throughput bisulfite sequencing data. All functionality of BiQ Analyzer HT is available through a web-startable graphical user interface, which guides the user through the data analysis (Figure 4.1). As an additional option, it is possible to run the computationally intensive parts of the software on a remote high-performance computer while maintaining the user-friendliness of a graphical interface run locally. Finally, BiQ Analyzer HT provides an optional command-line interface to facilitate integration into automatic data analysis pipelines.

4.2 Program Overview

BiQ Analyzer HT facilitates locus-specific analysis, quality control and visualization of high-throughput bisulfite sequencing data. The tool takes sequencing read data as input, and it produces quality-controlled output tables and diagrams of the inferred DNA methylation information for each sample, locus and DNA methylation site.

BiQ Analyzer HT is a Java-based program which can be run on any computer which has a recent version of the Java Virtual Machine installed. The tool is available as a self-installing Java Web Start distribution, and as a downloadable installation package for computers that are not connected to the Internet. BiQ Analyzer HT's project-based user interface supports the interactive analysis of bisulfite sequencing data for multiple target loci in multiple samples. A typical analysis consists of three phases: (i) data import, (ii) sequence alignment and quality control, and (iii) visualization and export of the inferred DNA methylation information (Figure 4.1).

To prepare high-throughput sequencing data for analysis with BiQ Analyzer HT, the user first applies vendor-specific software to perform base-calling, to resolve any sample multiplexing and to convert the data into one of two standard formats, FASTA or BAM. When importing FASTA files obtained by locus-specific bisulfite sequencing, BiQ Analyzer HT expects one file per sample and locus. We currently provide a custom script that automatizes data preparation for the Roche 454 sequencing platform (`http://biq-analyzer-ht.bioinf.mpi-inf.mpg.de/`), and we will add similar scripts for other platforms based on user demand. Alternatively, genome-scale bisulfite sequencing data can be imported as BAM files, which are most conveniently generated with BSMAP [Xi and Li, 2009].

When a new BiQ Analyzer HT project is initialized, an output directory is created into which the software writes its analysis results (Table 4.1). The project structure is defined by the user, adding samples and loading FASTA files containing the single genomic reference sequences that define each locus. The resulting tree structure is shown in BiQ Analyzer HT's main window. Once the data are loaded, this tree can be ordered either by samples or by loci, depending on the biological question of interest.

Read alignment and inference of DNA methylation information are controlled by parameters that the user selects on the setup screen. While the default values often provide good results, it is recommended that a user runs a first analysis with default parameters, inspects the results and then adjusts the parameters as necessary. Dataset-specific choice of quality-control parameters can sometimes compensate for quality issues that may be present in the primary

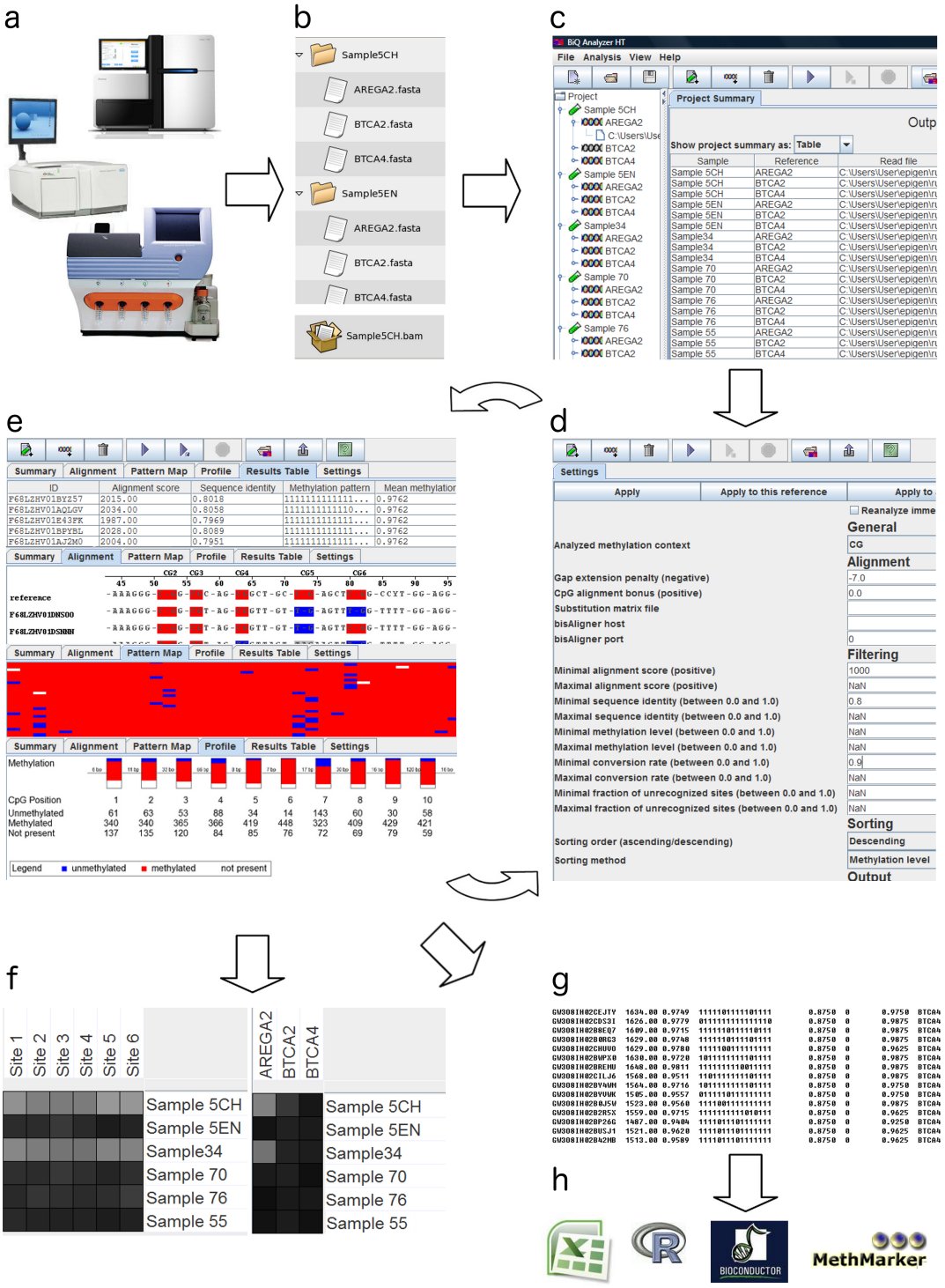


Figure 4.1: BiQ Analyzer HT workflow. Bisulfite sequencing data are generated either for the entire genome or selectively for a defined set of genomic loci using commercially available high-throughput sequencers (**a**). To reduce sequencing cost, bisulfite-converted DNA from several samples and/or loci is typically barcoded and combined into a single sequencing run. The multiplexed read data are separated and converted into FASTA or BAM files using vendor-provided software and/or custom scripts (**b**.), before they are loaded into BiQ Analyzer HT (**c**). Once loaded, the user sets alignment and quality control parameters (**d**), inspects the inferred DNA methylation patterns (**e**) and adjusts the parameters until satisfactory results are obtained. Finally, the DNA methylation measurements can be visualized graphically (**f**) and exported as tab-separated tables (**g**) for in-depth analysis using spreadsheets such as Excel, statistical software such as R/Bioconductor and biomarker development tools such as MethMarker (**h**).

data. For example, a decrease in alignment stringency parameters allows for retaining reads with reduced similarity to the reference, which would be removed by the default filtering criteria. This can be essential to process highly polymorphic sequences such as retrotransposable elements and DNA repeats.

Once satisfactory results are obtained, the inferred DNA methylation data can be exported in several formats, including sequence alignments, data tables and DNA methylation plots. Table 4.1 summarizessummarizes all output items. The sequence alignments provide a detailed account of how the DNA methylation levels were inferred. In addition, they can be used to identify allele-specific single-nucleotide polymorphisms or evidence of structural variation in the sequence data. The data tables facilitate exploratory data analysis using spreadsheets, in-depth statistics using statistical software such as R/Bioconductor [Gentleman *et al.*, 2004] and epigenetic biomarker development using BiQ Analyzer's companion tool MethMarker [Schuffler *et al.*, 2009]. Finally, the DNA methylation plots visualize the results of BiQ Analyzer HT analyses, for example for use in papers and scientific reports.

The visualization module of BiQ Analyzer HT utilizes the publicly available GSEA library [Subramanian *et al.*, 2005] for plotting DNA methylation heatmaps. BAM file handling is implemented using the Picard library (`http://picard.sourceforge.net/`), and parts of the sequence processing code are based on the BioJava framework [Holland *et al.*, 2008].

4.3 Data Processing

BiQ Analyzer HT implements a data processing pipeline that is run for each combination of locus and sample in the project tree. The pipeline aligns all sequencing reads from the corresponding input file to the locus-specific genomic reference sequence, and based on these alignments it infers which cytosines are methylated or unmethylated by comparing the read sequence with the reference sequence. The key steps of the data processing pipeline are outlined in more detail below. All analyses are conveniently accessible via the graphical interface. They can also be run from the command line, which facilitates integration with automatic data processing pipelines.

Read alignment. The analysis of bisulfite sequencing data crucially depends on accurate alignments. This is an inherently difficult task when complex genomic regions with repetitive elements and structural variation are studied and further complicated by the fact that bisulfite-converted DNA has substantially lower information content than genomic DNA. For this reason, speed-optimized seed-based aligners such as BLAT [Kent *et al.*, 2002], MAQ [Li *et al.*, 2008] and BWA [Li and Durbin, 2009] – which are commonly used for aligning high-throughput sequencing data – could undermine the accuracy of BiQ Analyzer HT. After exploring several alternatives, we chose to use the Needleman-Wunsch algorithm [Needleman and Wunsch, 1970], which is guaranteed to find the optimal (although not necessarily the correct) alignment between each sequencing read and the reference sequence. Furthermore, we made several modifications to the algorithm that account for recurrent issues with bisulfite-converted DNA (Supplementary Text S1). To partially compensate for the fact that the Needleman-Wunsch algorithm is substantially slower than current short-read aligners, we use a highly optimized implementation of this algorithm. This implementation provides excellent performance for read numbers in the order of 10^4 per locus on a standard laptop computer (Table 4.2). Furthermore, the read alignment can be outsourced to a remote high-performance computer, which makes it feasible to process in the order of one million reads per locus on a standard laptop computer.

Quality control and read filtering. Based on the pairwise alignment of the sequencing reads

with their corresponding genomic reference sequence, the data quality of the bisulfite sequencing experiment is estimated. Basic quality measures include the alignment score and sequence identity with the bisulfite-converted reference sequence, the estimated bisulfite conversion rate (fraction of unconverted cytosines outside of the analyzed methylation context, e.g. "CG") and the number of DNA methylation sites with missing data. The sequencing read data can be filtered for each of these quality measures, in order to quickly discard low-quality or otherwise unsuitable reads. The threshold values of each quality measure are set to empirically chosen defaults, but users may need to adjust these parameters interactively to account for the characteristics of their specific datasets.

Inference of DNA methylation patterns. BiQ Analyzer HT focuses by default on CpG methylation, which is the most common modification of eukaryotic DNA. The user can also choose to include other symmetric and asymmetric methylation contexts in the analysis, such as CpHpG and CpHpH. A methylation context is defined as a pair of DNA sequence motifs which match the methylated and unmethylated states of a site. The positions of the potential methylation sites are detected by finding matches of the first motif in the aligned reference sequence. The methylation state is then determined by comparing the sequences at the corresponding positions of the aligned sequence read, and the site is characterized as methylated, unmethylated or missing value ("1", "0" and "x", respectively). The collection of DNA methylation states for all sites in a given sequencing read constitute its methylation pattern, and the number of methylated sites divided by the total number of sites that are not missing values defines the mean methylation of a sequencing read.

Data visualization and export. The inferred DNA methylation data and quality control information can be exported for documentation and follow-up analysis using statistical tools (Table 4.1). The resulting tables list the quality measures, DNA methylation patterns and mean methylation levels for each sequencing read that has not been filtered out during quality control. Prior to exporting these tables, they can be sorted by one of the quality measures or by the inferred DNA methylation information.

4.4 Performance Evaluation

To confirm the practical utility of BiQ Analyzer HT for large datasets and to assess its performance relative to existing low-throughput tools, we benchmarked the tools on datasets with up to one million reads (Table 4.2). These data sets were obtained by multiplexed loci-specific bisulfite sequencing on the Roche 454 sequencing platform. Briefly, three classes of repetitive elements (RE1, RE2 and RE3) were amplified from bisulfite-treated mouse DNA, and several thousand reads were sequenced for these repetitive elements. To evaluate BiQ Analyzer HT's performance for one million reads, we further constructed artificial test sets from the actual dataset of region RE3 by reusing sequencing reads multiple times. The results of this benchmarking show that all existing tools have severe limitations in the number of reads that can be processed (Table 2). In contrast, with BiQ Analyzer HT we could successfully analyze a dataset with one million reads mapping to a single locus.

4.5 Conclusions

BiQ Analyzer HT provides comprehensive support for locus-specific analysis, quality control and visualization of high-throughput bisulfite sequencing data. It addresses the bioinformatic challenges of using high-throughput sequencing as a fast and cost-efficient alternative

Table 4.1: Analysis results generated by BiQ Analyzer HT

Category	Title	Access	Format	Description
Tabular				
	Project summary	GUI	TSV	Basic information summarizing the analysis
	Sample summary	GUI	TSV	DNA methylation summary for each locus in each sample
	Results table	OD	TSV	Alignment quality, estimated bisulfite conversion rate and DNA methylation summary for each sequencing read
	Methylation pattern table	GUI	TSV	DNA methylation patterns for each sequencing read. Columns correspond to DNA methylation sites (typically CpG positions)
	Project results table	GUI	TSV	Combined results table for all samples and loci
Graphical				
	Methylation pattern map	OD, GUI	PNG	Heatmap-style representation of DNA methylation patterns for each sequencing read. Columns correspond to DNA methylation sites
	Methylation profile	OD, GUI	PNG, SVG	Diagram visualizing the frequency of methylated, unmethylated and missing-value observations for each DNA methylation site
	Project methylation heatmap	GUI[i]	PNG	Heatmap of mean DNA methylation levels for each locus in each sample
	Methylation profile heatmap	GUI[i]	PNG	Heatmap of mean DNA methylation levels for each DNA methylation site at a specific locus
Sequence				
	Alignment	OD	FASTA	Multiple alignment of sequencing reads for each locus in each sample
	Filtered reads	OD	FASTA	Sequences of all reads that passed quality filtering

OD ("output directory") – the item is written to the project output directory tree; GUI ("graphical user interface") – the item can be exported via "Save as ..." or "Copy to clipboard" in the corresponding context menu; TSV – tab-separated value table; FASTA – sequencing reads in multiple-sequence FASTA format. [i] The data table from which the heatmap is generated can also be exported for follow-up analysis.

Table 4.2: Performance comparison of software packages for locus-specific analysis of bisulfite sequencing data

Region	Read Count	Performance							
		BiQ Analyzer 2.0		QUMA[a]		BiQ Analyzer HT		BiQ Analyzer HT[b]	
		Memory	ET	Memory	ET	Memory	ET	Memory	ET
RE1	400	350	300	NA[c]	10	95	30	1000	6
RE2	1054	500	911	NA[c]	25	200	50	1000	9
RE3	3150	>1000	3455	NA[c]	70	200	95	1000	16
RE3[d]	10^4	NA[f]		NA[c]	323	300	285	1500	50
RE3[d]	10^5	NA[f]		NA[f]		NA[g]		3500	440
RE3[d]	10^6	NA[f]		NA[f]		NA[g]		10 000	1940

All tests, except for the cases noted explicitly, were run on a standard laptop with dual-core processor and 2 GB main memory. The values of peak memory usage are given in MB. ET ("execution time") denotes the total duration of the analysis in seconds.

[a]The QUMA web-server runs on a high-performance machine (8 dual-core processors, 16 GB main memory). [b]BiQ Analyzer HT running on a high-performance machine (8 dual-core processors, 16 GB main memory). [c]Memory usage of the web-servers does not affect performance for the end user. [d]The dataset was obtained by concatenating multiple copies of the initial set of reads obtained for RE3. [e]The test could not be performed because the number of reads in the set exceeded the maximum read threshold of the web server. [f]The calculation could not be finished due to an error. [g]Tests with BiQ Analyzer HT for the last two read sets were performed only on the high-performance computer.

to clonal bisulfite sequencing, and it is fully compatible with multiplex analysis of several loci and samples. The alignment algorithm was specifically optimized for bisulfite-converted sequences, and it supports the analysis of both CpG and non-CpG methylation patterns. In summary, the combination of locus-specific high-throughput sequencing and interactive data analysis with BiQ Analyzer HT provides a highly practical approach for measuring the DNA methylation patterns of 10s to 100s of loci in 100s to 1000s of samples, for example in the context of biomarker validation and clinical diagnostics.

4.6 Supplementary Material

Supplementary Text S1. Bisulfite Sequence Alignment

Wildcard alignment. By selectively converting unmethylated but not methylated cytosines into uracils (the uracils are subsequent converted into thymines during PCR amplification), bisulfite treatment gives rise to significant discrepancies between the sequencing reads and the genomic reference sequence of the corresponding locus. Simply aligning the bisulfite sequencing reads to the genomic reference sequence therefore results in low-quality alignments. Several approaches exist to improve the alignment of the bisulfite sequence reads. Most commonly, all cytosines in both the read sequence and the reference sequence are replaced by thymines. BiQ Analyzer HT uses a related but slightly more complex method that makes use of wildcard sequence characters. Prior to the alignment, in the genomic reference sequence all cytosines within the methylation context (typically "CG") are converted into Zs, and all cytosines outside of the methylation context are converted into Ys (Fig. S1). These wild-

card characters are assigned different substitution scores for cytosines and thymines in the sequencing read (Fig. S2).

Scoring scheme. In the general case, no prior information is available on the length of the reads obtained in a bisulfite sequencing experiment. In other words, the read lengths may differ substantially from the length of the reference sequence. We therefore use local alignment following the Needleman-Wunsch algorithm, which assigns zero gap-extension penalties to all gaps located before the first and after the last non-gap character in both the sequencing read and the genomic reference sequence.

Methylation site alignment bonus. Correct alignment of the methylation context (typically "CG") is optionally rewarded by an increment of the alignment score, which can improve the results when aligning low-quality or highly polymorphic reads. This bonus discourages the alignment algorithm from introducing gaps at the methylation sites that are being analyzed.

Bisulfite conversion rate. Prior information is usually available regarding the expected methylation context of the analyzed DNA sample. For example, in differentiated mammalian cells one expects methylated cytosines to occur almost exclusively at CpG dinucleotides. The cytosine-to-thymine conversion rate for cytosines outside the expected methylation context is used by BiQ Analyzer HT to estimate the bisulfite conversion rate, providing an important indicator of the technical success of the experiment.

Supplementary Figures

a) **AACCTCTTGAACGAGTTCAGA**
b) **AATTTCTTGAACGAGTTCAGA**
c) **AAYYTCTTGAAZGAGTTCAGA**

Figure 4.S1: Modifications of the sequence alignment alphabet for aligning bisulfite sequences. a) Genomic reference sequence. b) *In silico* bisulfite-converted reference sequence. c) The BiQ Analyzer HT conversion scheme introduces different wildcards for cytosines inside and outside of the expected *methylation context* (typically "CG")

.

	A	**G**	**Y**	**Z**	**T**	**N**	**-**
A	MS	MP	MP	MP	MP	MP	MP
G	MS	MS	MP	MP	MP	MP	MP
C	MS	MP	CCR	MS	MP	MP	MP
T	MS	MP	CR	MS	MS	MP	MP
N	MS	MP	MP	MP	MP	MP	MP
-	MS	MP	MP	MP	MP	MP	MP

Figure 4.S2: BiQ Analyzer HT substitution matrix. The columns correspond to the genomic reference sequence, and the rows correspond to the sequencing read. MP – mismatch penalty, MS – match score, CR – a negative term proportional to the conversion rate, CCR – a complement of CR such that CR+CCR=MS+MP.

References

Bird, A. DNA methylation patterns and epigenetic memory. *Genes and Development*, 16(1):6–21, 2002.

Bock, C. Epigenetic biomarker development. *Epigenomics*, 1(1):99–110, 2009.

Bock, C., Reither, S., Mikeska, T., Paulsen, M., Walter, J., and Lengauer, T. BiQ Analyzer: Visualization and quality control for DNA methylation data from bisulfite sequencing. *Bioinformatics*, 21(21):4067–4068, 2005.

Bock, C., Tomazou, E. M., Brinkman, A. B., Müller, F., Simmer, F., Gu, H. *et al.* Quantitative comparison of genome-wide DNA methylation mapping technologies. *Nature Biotechnology*, 28(10):1106–1114, 2010.

Carr, I. M., Valleley, E. M., Cordery, S. F., Markham, A. F., Bonthron, D. T., and Carr IM, Valleley EMA, Cordery SF, Markham AF, B. D. Sequence analysis and editing for bisulphite genomic sequencing projects. *Nucleic Acids Research*, 35(10):e79, 2007.

Esteller, M. Epigenetics in cancer. *N Engl J Med*, 358(11):1148–1159, 2008.

Feinberg, A. P. Phenotypic plasticity and the epigenetics of human disease. *Nature*, 447(7143):433–440, 2007.

Frommer, M., McDonald, L. E., Millar, D. S., Collis, C. M., Watt, F., Grigg, G. W. *et al.* A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proceedings of the National Academy of Sciences of the United States of America*, 89(5):1827–1831, 1992.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5(10):R80, 2004.

Grunau, C., Schattevoy, R., Mache, N., and Rosenthal, a. MethTools–a toolbox to visualize and analyze DNA methylation data. *Nucleic acids research*, 28(5):1053–1058, 2000.

Hetzl, J., Foerster, A. M., Raidl, G. R., and Mittelsten Scheid, O. CyMATE: A New Tool for Methylation Analysis of Plant Genomic DNA after Bisulfite Sequencing. *Plant J*, 51(3):526–536, 2007.

Holland, R. C. G., Down, T. a., Pocock, M., Prlic, a., Huen, D., James, K. *et al.* BioJava: an open-source framework for bioinformatics. *Bioinformatics*, 24(18):2096–2097, 2008.

Huang, Y., Pastor, W. a., Shen, Y., Tahiliani, M., Liu, D. R., and Rao, A. The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing. *PLoS ONE*, 5(1):e8888, 2010.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, a. M. *et al.* BLAT–the BLAST-like alignment tool. *Genome Res*, 12(4):656–664, 2002.

Korshunova, Y., Maloney, R. K., Lakey, N., Citek, R. W., Bacher, B., Budiman, A. *et al.* Massively parallel bisulphite pyrosequencing reveals the molecular complexity of breast cancer-associated cytosine-methylation patterns obtained from tissue and serum DNA. *Genome Research*, 18(1):19–29, 2008.

Kumaki, Y., Oda, M., and Okano, M. QUMA: quantification tool for methylation analysis. *Nucleic acids research*, 36(Web Server issue):W170–5, 2008.

Laird, P. W. The power and the promise of DNA methylation markers. *Nature Reviews. Cancer*, 3(4):253–66, 2003.

Laird, P. W. Principles and challenges of genome-wide DNA methylation analysis. *Nat Rev Genet*, 11(3):191–203, 2010.

Laurent, L., Wong, E., Li, G., Huynh, T., Tsirigos, A., Ong, C. T. *et al.* Dynamic changes in the human methylome during differentiation. *Genome research*, 20(3):320–31, 2010.

Li, H. and Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

Li, H., Ruan, J., and Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res*, 18(11):1851–1858, 2008.

Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 462(7271):315–322, 2009.

Needleman, S. B. and Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453, 1970.

Rohde, C., Zhang, Y., Reinhardt, R., and Jeltsch, A. BISMA–fast and accurate bisulfite sequencing data analysis of individual clones from unique and repetitive sequences. *BMC bioinformatics*, 11:230, 2010.

Schuffler, P., Mikeska, T., Waha, A., Lengauer, T., and Bock, C. MethMarker: user-friendly design and optimization of gene-specific DNA methylation assays. *Genome Biol*, 10(10):R105, 2009.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 102(43):15545–15550, 2005.

Suzuki, M. M. and Bird, A. DNA methylation landscapes: provocative insights from epigenomics. *Nat Rev Genet*, 9(6):465–476, 2008.

Taylor, K. H., Kramer, R. S., Davis, J. W., Guo, J., Duff, D. J., Xu, D. *et al.* Ultradeep bisulfite sequencing analysis of DNA methylation patterns in multiple gene promoters by 454 sequencing. *Cancer Research*, 67(18):8511–8518, 2007.

Varley, K. E. and Mitra, R. D. Bisulfite patch PCR enables multiplexed sequencing of promoter methylation across cancer samples. *Genome Research*, 20(9):1279–1287, 2010.

Xi, Y. and Li, W. BSMAP: whole genome bisulfite sequence MAPping program. *BMC Bioinformatics*, 10:232, 2009.

Xu, Y. H., Manoharan, H. T., and Pitot, H. C. CpG PatternFinder: A Windows-based utility program for easy and rapid identification of the CpG methylation status of DNA. *BioTechniques*, 43(3):334–342, 2007.

Chapter 5

BiQ Analyzer HiMod – an interactive software tool for high-throughput locus-specific analysis of 5-methylcytosine and its oxidized derivatives

The full text of this chapter has been earlier published as:

Daniel Becker[1,2], Pavlo Lutsik[1,#], Peter Ebert[2], Christoph Bock[2-4], Thomas Lengauer[2] and Jörn Walter[1,#] (2014) *Nucleic Acids Research* **32**(W1), W501–7.

The author of the thesis conceived the project (with J.W.), co-designed and supervised the implementation of the software package as well as wrote the major parts of the manuscript.

[1]Department of Genetics, Saarland University, Saarbrücken, 66123, Germany
[2]Department of Computational Biology and Applied Algorithms, Max-Planck Institute for Informatics, Saarbrücken, 66123, Germany
[3]CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, Vienna, 1090, Austria
[4]Department of Laboratory Medicine, Medical University of Vienna, Vienna, 1090, Austria
[#]To whom correspondence should be addressed: Jörn Walter and Pavlo Lutsik

Abstract

Recent data suggest an important biological role for oxidative modifications of methylated cytosines in DNA, specifically hydroxymethylation, formylation and carboxylation. Several assays have been described to profile these modifications genome-wide as well as in targeted, locus-specific settings. So far no comprehensive software has been described to aid the analysis of sequencing data from these assays. Here we present BiQ Analyzer HiMod, an interactive and user-friendly software tool for low-level processing, quality control and initial analysis of the high-throughput locus-specific DNA modification sequencing data. The software supports four different profiling assays, leading the user from raw sequence reads to summarized modification measurements and publication-quality plots. The program combines well-established graphical user interface of its predecessor tool, BiQ Analyzer HT, with new and extended analysis modes. BiQ Analyzer HiMod also includes updates of the analysis workspace, an intuitive interface, a custom vector graphics engine, and support of additional input and output data formats. The tool is freely available as a stand-alone installation package from http://biq-analyzer-himod.bioinf.mpi-inf.mpg.de/.

Availability

The BiQ Analyzer HiMod installation package is freely available from http://biq-analyzer-himod.bioinf.mpi-inf.mpg.de/ .

Funding

This work was supported by the European Union's Seventh Framework Programme (FP7/2007-2013) [grant numbers 267038 (NOTOX), to J.W and P.L, and HEALTH-F5-2011-282510 (BLUEPRINT)], and by the German Science Ministry [grant number 01KU1216A (DEEP Project), to P.E.].

Acknowledgements

The authors would like to thank Julia Arand and Pascal Giehr for providing the oxBS-seq data for the biological example, Georg Friedrich for technical support with the BiQ HiMod web site, and Karl Noerdstrom for valuable discussion on preprocessing of sequencing data.

5.1 Introduction

DNA methylation is widely recognized as a key epigenetic mechanism playing a crucial role in development and disease [Bergman and Cedar, 2013; Smith and Meissner, 2013]. Sequencing of bisulfite-treated DNA is the gold standard for base resolution mapping of DNA-methylation [Bock *et al.*, 2010; Harris *et al.*, 2010]. Recently, oxidized derivatives of 5-methylcytosine (5mC) were identified that are assumed to have important biological function [Branco *et al.*, 2011]. Like 5-methylcytosine the oxidized derivatives occur predominantly at CpG-dinucleotide cytosines [Kohli and Zhang, 2013]. However, conventional bisulfite sequencing does not discriminate between different oxidized forms. Sodium bisulfite treatment converts unmodified cytosine, 5-carboxy (5caC) and 5-formylcytosyne (5fC) into deaminated forms, which are subsequently replaced by thymines during PCR amplification, while methylated and hydroxymethylated (5hmC) cytosines remain unchanged [Huang *et al.*, 2010]. Several new methods have been suggested recently to overcome these limitations and transform the specific oxidative modifications into sequence-based signals. Oxidative bisulfite sequencing (oxBS-seq) involves an oxidation reaction to convert 5hmC to 5fC [Booth *et al.*, 2012]. A subsequent bisulfite reaction followed by the amplification step converts 5fC to thymine whereas the original 5mC is replaced by cytosines in PCR amplicons (Figure 5.1, a). TET-assisted bisulfite sequencing (TAB-seq) utilizes the opposite scheme and applies an enzymatic oxidation by the TET protein to convert all 5mC to 5hmC and further to 5fC and 5caC, while the initially present 5hmC is protected by glycosylation prior to the bisulfite reaction (Figure 5.S1). The subsequent bisulfite treatment and PCR converts the initially present 5mC (as well as 5fC and 5caC) to thymine while the 5hmC is replaced with cytosines [Schüler and Miller, 2012]. Similar methods were developed for mapping 5fC and 5caC, such as formyl-chemically-assisted bisulfite sequencing (fCAB-seq) [Song *et al.*, 2013] (Figure 5.1, b) and chemical-modification-assisted bisulfite sequencing (CAB-seq) [Lu *et al.*, 2013] (Figure 5.S1). Following the modified bisulfite treatments of either oxBS-seq, TAB-Seq, CAB-seq or fCAB-Seq the DNA is amplified and sequenced. The abundance of any modification under inspection is estimated by comparing such sequencing results to data obtained by conventional bisulfite sequencing. Note that all modified bisulfite methods are only partially modification-specific and mostly produce cumulative estimates of several different modifications.

While several tools have been developed for the analysis of conventional locus-specific bisulfite sequencing data of various scale [Bock *et al.*, 2005; Kumaki *et al.*, 2008; Lutsik *et al.*, 2011], no software package currently supports an integrated analysis of data from multiplexed locus-specific high-throughput DNA modification profiling. For the genome-scale data the MLML tool has been designed to produce accurate DNA methylation and hydroxymethylation levels from preprocessed oxBS-seq, TAB-seq and other similar data [Quy *et al.*, 2013].

To fill this gap we developed BiQ Analyzer HiMod (or shortly BiQ HiMod), a specialized interactive software package for the preprocessing, quality control and analysis of various DNA modifications in high-throughput targeted sequencing experiments. BiQ HiMod is an extension of our previously published BiQ Analyzer HT tool [Lutsik *et al.*, 2011]. It provides a user-friendly interface, comprehensive and customizable streamlined analysis pipeline, rich graphical and tabular output. The tool currently supports four experimental assays and can be easily extended to include the upcoming methods.

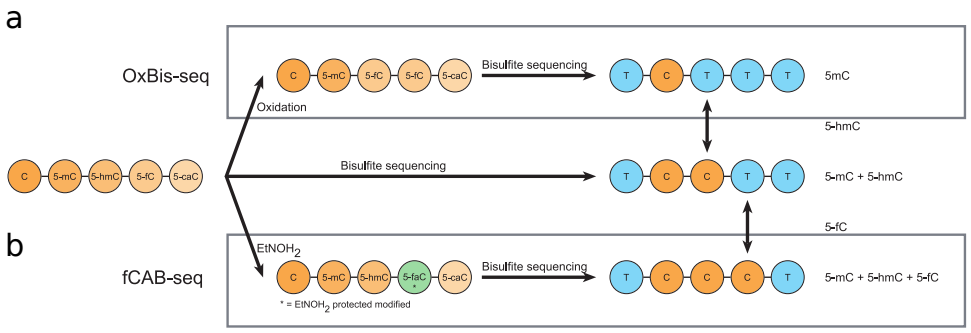


Figure 5.1: Principal scheme of oxBS-seq and fCAB-seq methods. **a**. oxBS-seq. Oxidative bisulfite treatment followed by PCR leads to conversion of the oxidative modifications to thymines while only 5mC appears as cytosine in the sequencing reads. The cumulative level of 5hmC can be established for each CpG site by comparing to ordinary bisulfite sequencing. **b**. In fCAB-seq 5fC is protected from the conversion and together with 5mC and 5hmC appears as cytosine after PCR. The bulk 5fC abundance is calculated by subtracting the cumulative levels of 5mC + 5hmC pair obtained from ordinary bisulfite sequencing.

Table 5.1: Information that can be extracted from each DNA modification profiling method supported by BiQ Analyzer HiMod

Method	**Underlying modification**				**Measurements**
	Bisulfite treatment		**Modified treatment**		
	C	**T**	**C**	**T**	
oxBS-seq	5mC 5hmC	5fC 5caC C	5mC	5hmC 5fC 5caC C	5mC (SM) 5hmC (B) 5fC + 5caC + C (B)
TAB-seq			5hmC	5mC 5fC 5caC C	5hmC (SM) 5mC (B) 5fC + 5caC + C (B)
fCAB-seq			5mC 5hmC 5fC	5caC C	5fC (B) 5mC + 5hmC (B) 5caC + C (B)
CAB-seq			5mC 5hmC 5caC	5fC C	5caC (B) 5mC + 5hmC (B) 5fC + C (B)

SM, single-molecule resolution; B, bulk per-CpG measurement

5.2 BiQ Analyzer HiMod

5.2.1 Overview

BiQ Analyzer HiMod is a cross-platform interactive Java application, which can run on any system with properly installed and configured Java Runtime Environment. The package installation is performed via a simple "click-through" installer. BiQ HiMod facilitates a fully interactive primary processing and analysis of sequencing data from locus-specific DNA modification profiling experiments, and currently supports oxBS-seq, TAB-seq, fCAB-seq and CAB-seq assays. The comprehensive and ergonomic graphical user interface provides means for multi-level overview of the complete multiplexed sequencing project, in which a number of genomic loci of interest are sequenced in several biological samples. The underlying software architecture uses the fact that most of the basic data processing steps are invariant to the type of chemical treatment and thus the reads from ordinary and modified bisulfite procedure can be preprocessed independently. During the quality control cycle the user refines the quality thresholds of the processing pipeline until a satisfactory quality level is reached for each read batch. After low-level processing, cumulative levels of each DNA modification are summarized based on the results of the individual preprocessing pipeline runs. Final results can be exported as tables, figures, alignments, and genome browser tracks (see Supplementary text for a detailed list). Importantly, BiQ HiMod maintains the full functionality of the previous versions by supporting the analysis of conventional bisulfite sequencing data in an independent analysis mode.

5.2.2 Data preparation and project setup

BiQ HiMoD is designed to import multiple data sets of individual sequencing reactions and loci. Today's data sets are often generated by NGS-based multiplexed locus-specific sequencing [Holland *et al.*, 2008; Robinson *et al.*, 2011; Rohde *et al.*, 2010]. In NGS approaches sequences are usually indexed by short sequence tags (using multiplexing library kits), pooled and submitted to a high-throughput sequencing platform, e.g. Illumina MiSeq or Roche 454. In such a setting BiQ HiMod requires the direct sequencing output to be demultiplexed using available third-party tools. We recommend the barcode splitter from Galaxy [Blankenberg *et al.*, 2010; Giardine *et al.*, 2005; Goecks *et al.*, 2010] as an adequate solution which suits researchers with minimal bioinformatics experience. Reads ready to be loaded must be in FASTA or FASTQ format with one file per each sample-locus-treatment combination.

Besides demultiplexing, data preparation usually includes several additional standard steps, e.g. trimming of low quality parts of the sequence reads (often at the 3-end) and – in case of paired-end sequencing – joining of overlapping mate reads. In the FAQ section of the BiQ HiMod web site (`http://biq-analyzer-himod.bioinf.mpi-inf.mpg.de/FAQ.php`) we describe an example of such a data preparation workflow in a step-by-step way using third-party tools integrated into the Galaxy framework [Blankenberg *et al.*, 2010; Giardine *et al.*, 2005; Goecks *et al.*, 2010].

Following data preparation one needs to set up an analysis project and import the data. For a relatively small project this can be done directly using the GUI controls. For larger data sets one should prepare a spreadsheet-like text file, (see our example on the program web site). Samples are defined by deliberate string identifiers, while the genomic loci exhibit one-to-one mapping to their genomic reference sequences, supplied as single- or multi-sequence FASTA files. We recommend extracting sequence information via the Galaxy "Extract Sequence" tool starting from a BED file with genomic coordinates of each amplicon, such that the generated

FASTA records contain genomic location information. This information will be used further on, in order to bring the analysis into the genomic context and generate genome browser tracks.

Finally, sequencing reads can be imported directly from the genome-wide sequencing experiment (e.g. whole-genome or reduced-representation bisulfite sequencing), if the data are available as aligned sequencing reads (BAM files). BiQ HiMod facilitates loading the reads of a corresponding region based on coordinate information and analyzing them in the context of the corresponding genomic reference. This feature of BiQ HiMod can be extremely useful for quickly scanning several locus specific read distributions in regions of interest in large genome-wide WGBS or RRBS datasets provided they are accessible as BAM archives.

5.2.3 Primary processing pipeline

The software backend was subjected to several modifications compared to the predecessor version of the tool. The FASTQ files parsing is now supported using the data import library of the BioJava project [Holland *et al.*, 2008]. Loaded sequence reads are independently aligned to the corresponding reference sequence using the bisulfite-modified semi-global implementation of Needleman-Wunsch algorithm, which was further tweaked to improve robustness of the modification site alignment (see Supplementary Text and [Lutsik *et al.*, 2011] for more details). The set of filtering criteria was extended to include sequencing quality scores, and default thresholds are set to ensure that only high quality bases are considered. The full list of quality metrics and DNA modification statistics are given in the Supplementary Text. All processing and analysis options remain fully customizable and can be refined after the initial analysis.

5.2.4 Quantification of modification levels

The sequencing results obtained with BiQ HiMod after conventional bisulfite, oxidative--bisulfite, Tet-assisted bisulfite sequencing or fCAB, redBS-seq are at comparably high resolution. A quantitative assessment of the relative abundances of different modifications depends on the comparison of two independent sequencing results. This comparative mode is a core feature implemented in BiQ HiMod. However, these comparisons have two main limitations: i) except for one modification type one can obtain estimates of modification differences at single CpG positions only as an average across all sequences and not per sequence ii) in most comparisons combinations of several modification types are not separable from each other. For illustration, we consider the results of the oxBS-seq method. Here, the reads from the ordinary bisulfite treatment contain unconverted cytosines which were methylated or hydroxymethylated in the original DNA molecule. The read cytosines from the oxidative bisulfite treatment can only be mapped to the 5mC of the original DNA. All in all, oxBS-seq yields a single-molecule resolution map of 5mC, bulk per-CpG average levels of 5hmC and a combination of 5fC, 5caC and unmodified cytosines. Table 1 summarizes the information which can be extracted from each of the data types. For every type of analysis BiQ HiMod takes into account all potential modifications and summarizes the information accordingly.

The estimation of bulk modification levels in BiQ HiMod is performed by a simple subtraction of converted cytosine frequencies. In case of low abundance of the target modification, in particular 5fC and 5caC, the true biological changes of modification levels may overlap with the experimental and the technical error range, sometimes leading to negative DNA modification levels. To overcome such difficulties, on the experimental side we recommend gathering

data from several independent experiments using high and ultra-high sequencing coverage. Moreover, the statistical significance of the observed variation should be tested. Summarized data can be exported in bedGraph format and loaded into downstream statistical software, e.g. MLML or RnBeads (http://rnbeads.mpi-inf.mpg.de), for correction and proper statistical evaluation.

5.2.5 Visualization and data export

The frontend of BiQ HiMod provides capabilities for the top-down level-by-level exploration of a sequencing experiment. On the top level the global overview of the analysis project is given by the samples vs. loci heat maps visualizing average modification abundances for each sample-locus pair (Figure 5.2, a). More detailed profiles of each studied locus are provided by the locus-wide bar plots visualizing the modification levels at each CpG position with realistic relative genomic distances (Figure 5.2, b). On the bottom level, sequence pileup, read-level cytosine pattern map and data tables are available for the most in-depth exploration and quality control similar to the analysis mode of BiQ Analyzer HT. The newly introduced diagnostic plots simplify the selection of the quality control thresholds by visualizing read batch-wide distributions of major analysis metrics, e.g., score and sequence identity of the read-reference alignment, bisulfite conversion rate, number of missing or mutated modification sites etc. (Figure 5.2, c).

Read-level and summarized modification data can be exported as tab-delimited files for downstream statistical analysis and visualization in favorite statistical software packages, e.g., R/Bioconductor, SPSS, and Excel, or custom bioinformatic pipelines. Furthermore, BiQ HiMod feeds into large-scale experimental projects by exporting its results as genome browser tracks. Furthermore, methylation values for each analyzed CpG-site can be saved in bedGraph format importable into genome viewers and browsers, such as IGV [Robinson *et al.*, 2011] or UCSC Genome Browser [Kent *et al.*, 2002] (`http://genome.ucsc.edu/index.html`) (see example in Figure 5.2, d). This allows for a convenient display of amplicon validation data in the context of genome-wide epigenetic data tracks, such as DNA methylation, histone modification, chromatin accessibility maps etc.

5.2.6 Software architecture, GUI improvements and the new graphics engine

BiQ HiMod builds upon the modular and robust design of its predecessor tool, BiQ Analyzer HT, with a multistep data processing pipeline, as its backend, combined with an interactive graphical user interface at the frontend. The backend takes raw sequencing reads and reference sequence information as input, performs a series of preprocessing steps – data loading, alignment, quality filtering, cytosine conversion calling – and generates the resulting tables and graphics for each sample-locus combination. The backend output is written directly to the hard drive and can be found in the project directory. This results in efficient use of the operating memory and guards against information loss. The interactive frontend, implemented in Java standard AWT/Swing framework, enables the setup of a large-scale multiplexed NGS analysis project for one of the supported profiling methods, coordination of the pipeline runs, and summarization of the results at the sample, locus, and project levels as well as export of the results.

BiQ HiMod expands the basic form of the GUI design of the predecessor tool, BiQ Analyzer HT, by introducing several major improvements. First, the project summary view has been

Figure 2

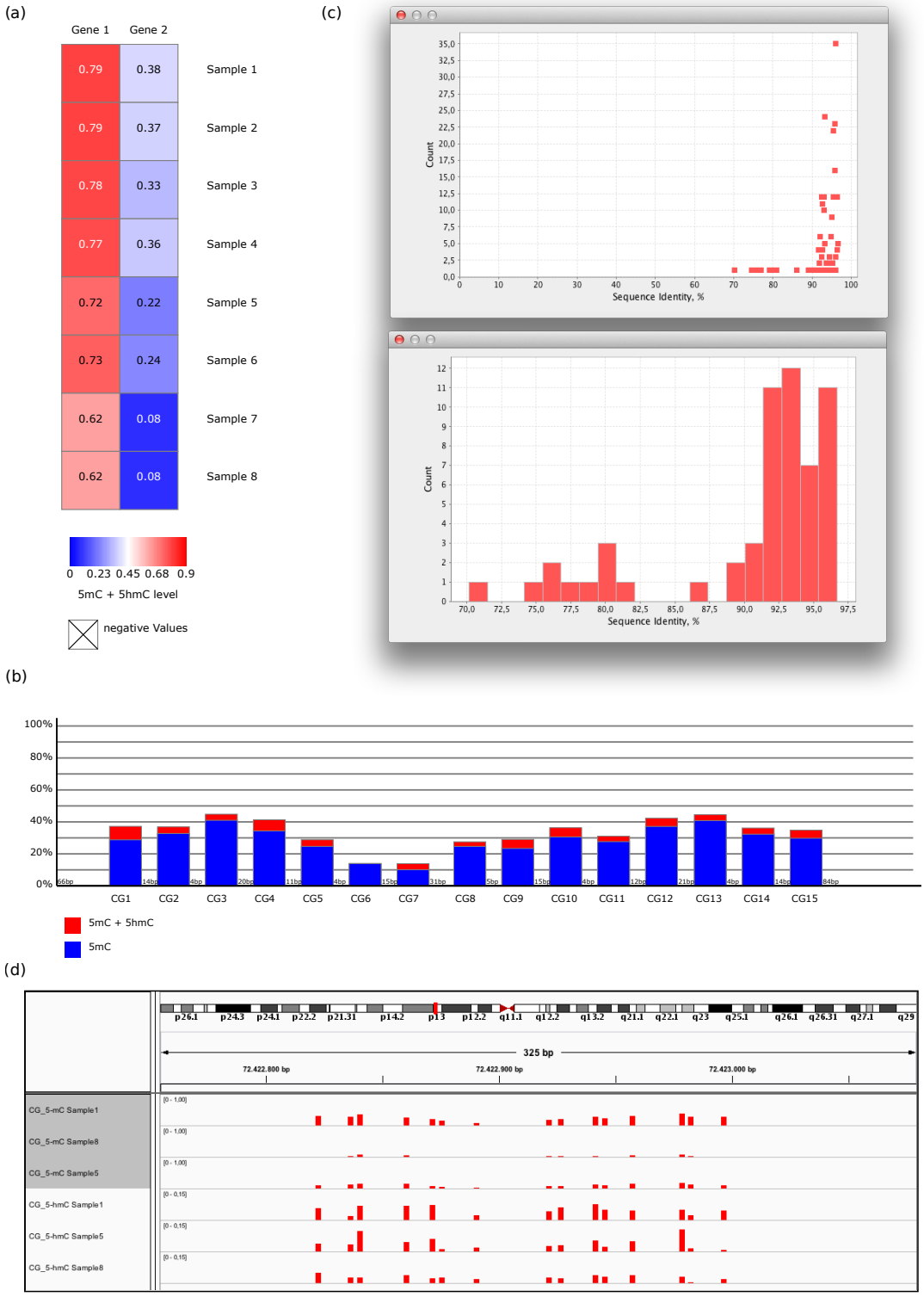


Figure 5.2: BiQ Analyzer HiMod visualization features. **a**. Global heat map of average bulk (5mC + 5hmC) modification levels at sequenced loci across the samples. The default color code is given in the color legend (see Figure 5.S2 for more details). **b**. Bar chart displaying stacked levels of two modifications at each CpG position in a single locus (see also Figure 5.S3). **c**. Diagnostic histogram visualizing the distribution of a quality control metric – sequence identity of each read-reference alignment – simplifying the sequence identity cutoff selection. **d**. Integration with the IGV genome browser showing genomic tracks with 5mC and 5hmC levels of three samples (the value range is much smaller for the 5hmC tracks).

Table 5.2: Benchmarking of BiQ Analyzer HiMod on artificially generated read batches of variable size.

Number. of reads in each batch[a]		1,000	2,000	5,000	10,000	20,000	50,000
Running time, min:s	normal	0:15	0:32	1:52	5:20	16:48	155:15
	w/o pattern maps[b]	0:15	0:26	1:06	2:13	4:25	25:14

[a] Given are numbers of reads in each of two read batches (oxBS and conventional bisulfite) processed in every benchmarking step. Therefore the actual number of processed reads was twice the given number in each step.
[b] The high computation time for large read sets is based on the size of the produced single read-resolution pattern maps. The user can decide which graphics to generate and by choosing to disable pattern maps the computation time can be significantly reduced.

turned into an ultimate control panel, affording an exhaustive overview without the need of frequent panel switching. It includes a project summary table, presenting the most crucial information about each read batch, the project-level heat maps visualizing average modification levels, and locus-wide bar charts dynamically generated for each selected read batch. Second, the detailed views of individual read batches were updated and now include the summary pages presenting the levels of each modification or group of modifications as tables and plots.

BiQ HiMod features a completely new custom graphical engine generating high-quality vector images in SVG format. Most of the plotting parameters, including the flexible color scheme, are fully user-adjustable. The exported SVG files can be further edited via one of the popular image packages without any loss of quality or directly incorporated into manuscripts and reports.

5.3 Validation on artificial and real biological data and performance assessment

In order to test BiQ HiMod and assess its performance we generated artificial read batches for each supported type of experiment by simulation (see Supplementary Text 1 for details). We estimated the accuracy and robustness of BiQ HiMod results under different conditions.

We then applied BiQ HiMod to reprocess the earlier published oxBS-seq data set obtained by sequencing of repetitive elements in two types of mouse ES cells [Ficz *et al.*, 2013]. In brief, serum cultured murine E14 ES cells and 2i-medium cultured ES cells from three different time points – 1, 3 and 7 days of cultivation – were sampled, two replicates of each. After oxidative and conventional bisulfite treatment, hairpin amplicons of regions within two repeat types – intracisternal. A particle, also known as IAP, and long interspersed elements, also known as LINE-1 or L1 – were constructed as described in an earlier study [Arand *et al.*, 2012]. The amplicons were sequenced on the MiSeq platform from Illumina using paired end sequencing (2 x 150). The BiQ HiMod results document an accumulation of 5hmC at LINE1 but not at IAP elements during the first 72 hours of cultivation in 2i-medium (Figure 5.S7).

We also benchmarked the runtime of BiQ Analyzer HiMod by analyzing batches of 1,000 to 50,000 artificially generated reads. The tests were carried out on a machine with a 2.4 GHz Intel Core i5 Haswell dual core processor and 2 GB of RAM. The results of the benchmark, given in Table 2, show that moderate-sized analyses can be performed in a reasonable amount of time on an average commodity laptop. For data from large-scale experiments amounting to several millions or more sequence reads, we recommend running BiQ HiMod on industrial-level workstations with several multicore CPUs and a few dozen GB of RAM.

5.4 Conclusions and outlook

BiQ Analyzer HiMod is the first interactive software package for preprocessing, quality control and initial analysis of various DNA modifications from four different experimental methods. Developed on the basis of BiQ Analyzer HT, it features a thoroughly reworked GUI with a large number of analysis plots as well as a new customizable vector graphics engine, additional data import and export formats. The tool enables a full-blown analysis starting from raw sequence reads up to publication-quality plots and genome browser tracks. BiQ HiMod exploits modular and flexible software design allowing for painless further extensions to support upcoming experimental assays, such as MAB-Seq [Hu and Tzeng, 2014]. The standardized, comprehensive and user-friendly software paradigm implemented by BiQ HiMod will help the tool find numerous applications in basic biology and biomedical research.

5.5 Supplementary Material

Supplementary Text

Artificial test data

In order to have controlled data for sanity checks and testing we developed a simulation framework that generates artificial sequencing reads for each of the supported methods. The simulation took the following parameters as input:

- one or more reference sequences in FASTA format
- mean methylation (m), hydroxymethylation (h), formylation (f) and carboxylation (c)rate $\mu_{mod},\ mod \in m, h, f, c,$
- standard deviation for each modification rate σ_{mod},
- the average number of reads, μ_{reads} and
- standard deviation of the number of reads σ_{reads},
- number of samples to be constructed for each reference.

The number of reads N in each read set was determined by a gaussian distribution with mean μ_{reads} and variance σ^2_{reads} to create some more difference between the data sets.

The modification levels for each of S CpG position in the reference sequence were obtained by generating random samples from a [0,1]-truncated Gaussian distribution with mean μ_{mod} and variance σ^2_{mod} of size C. The sampling was repeated for each of the four modifications, to generate $S \times 4$ matrix M of probabilities of each modification occurring at each CpG position. The input parameter values were chosen such that the sums of the rows did not exceed 1. The modification levels for CpGs for which $\sum_{mod} M_{j,mod} > 1$ were resampled until this criterion was fulfilled.

To create a read set first N simple copies of the reference sequence were generated. Then, the precise state was defined for each CpG position $j \in \{1 \dots S\}$ by sampling N times from a uniform distribution on [0,1] and mapping the outcomes to the discrete modification states for each read $k \in \{1 \dots N\}$ using the following rule:

$$modification_{j,k} = \begin{cases} \text{5mC}, & 0 < r_k < M_{j,m} \\ \text{5hmC}, & M_{j,m} < r_k < M_{j,m} + M_{j,h} \\ \text{5fC}, & M_{j,m} + M_{j,h} < r_k < M_{j,m} + M_{j,h} + M_{j,f} \\ \text{5caC}, & M_{j,m} + M_{j,h} + M_{j,f} < r_k < M_{j,m} + M_{j,h} + M_{j,f} + M_{j,c} \\ \text{C}, & M_{j,m} + M_{j,h} + M_{j,f} + M_{j,c} < r_k < 100 \end{cases}$$

r - a sample of size N from the uniform distribution on $[0, 1]$.

The modified cytosines were finally translated into thymine or cytosine based on their modification state and the selected experimental method, to simulate the effects of the (modified) bisulfite treatment and subsequent PCR amplification. A uniformly distributed random value on [0,1] was used to simulate errors in this process. If the random value was smaller p_{mod} the translation did not take place. To simulate incomplete reads, a gaussian distributed amount of bases were set to gap characters in the beginning or at the end of each read. This gaussian distribution had a mean of $\mu_{edgeGap}$ and a variance of $\sigma^2_{edgeGap}$. An error probability p_{seq} was introduced to simulate sequencing errors, and defined the probability with which a reference sequence base was replaced with a random base. Furthermore a gaussian distributed conversion probability with a mean of μ_{conv} and a variance of σ^2_{conv} was used to convert cytosines outside of the modification context to thymine. If the gaussian distributed value was smaller than a given threshold p_{conv}, the current base was converted.

The specific test sets, distributed with BiQ HiMod, were produced based on real reference sequences of two repetitive elements, analyzed in Ficz et al. study (see main text for more details). The variable positions of the reference sequences, which were abundant in the reference sequences, were mapped to one of their actual possible bases using a uniformly distributed random number on [0,1] and the same probability for each possibility.

The the following parameter values were used:

- $\mu_m = \frac{50}{2^{s-1}}$
- σ_m = 10
- $\mu_h = \frac{20}{2^{s-1}}$
- σ_h = 10
- $\mu_f = \frac{5}{2^{s-1}}$
- σ_f = 10
- $\mu_c = \frac{5}{2^{s-1}}$
- σ_c = 10
- p_{mod} = 0.02
- μ_{reads} = 1000
- σ_{reads} = 10
- $\mu_{edgeGap}$ = 0
- $\sigma_{edgeGap} = l \cdot 0.05$
- p_{seq} = 0.05
- μ_{conv} = 0
- σ_{conv} = 1
- p_{conv} = 0.8

where s is the index of the simulated "biological" sample and l the length of the current read. The parameter values were chosen with the goal that the simulated data were close to the expected real biological data.

Alignment

The pairwise alignment of each read to the amplicon reference sequence in BiQ HiMod is performed by a further modified version of the Needleman-Wunsch algorithm used in BiQ Analyzer HT. In brief, the alignment algorithm implementation features a semi-global gap penalty to account for variable read length, and extended alphabet of the reference sequence to aid the alignment of the potentially modified sites (see [Lutsik *et al.*, 2011] for an in-depth description). In BiQ HT a new wildcard base was introduced to further improve the alignment

of the modification sites. Cytosines within a methylation context are converted to X both in the reference sequence and in the sequence reads (CG cytosines in the reference and [CT]G cytosines in the reads). Matching Xs return a high score in the alignment to promote the alignment of methylation context positions.

The default parameters lead to good alignments in most of the cases but can be adjusted by the user if needed.

M - Cost Matrix =

$$\begin{array}{c} \\ A \\ G \\ C \\ T \\ - \\ N \\ X \end{array} \begin{array}{c} \begin{array}{ccccccccccccccccc} A & G & Y & T & - & N & Z & R & S & W & K & M & B & D & H & V & X \end{array} \\ \begin{pmatrix} 5 & -5 & -5 & -5 & -5 & -5 & -5 & 5 & -5 & 5 & -5 & 5 & -5 & 5 & 5 & 5 & -5 \\ -5 & 5 & -5 & -5 & -5 & -5 & -5 & 5 & 5 & -5 & 5 & -5 & 5 & 5 & -5 & 5 & 5 \\ -5 & -5 & 5 & -5 & -5 & -5 & 8 & -5 & 5 & -5 & -5 & 5 & 5 & -5 & 5 & 5 & -5 \\ -5 & -5 & 5 & 5 & -5 & -5 & 8 & -5 & -5 & 5 & 5 & -5 & 5 & 5 & 5 & -5 & -5 \\ -5 & -5 & -5 & -5 & 0 & 0 & -5 & -5 & -5 & -5 & -5 & -5 & -5 & -5 & -5 & -5 & -5 \\ -5 & -5 & -5 & -5 & 0 & 0 & -5 & -5 & -5 & -5 & -5 & -5 & -5 & -5 & -5 & -5 & -5 \\ -5 & 5 & -5 & -5 & -5 & -5 & -5 & 5 & 5 & -5 & -5 & -5 & 5 & -5 & -5 & 5 & 8 \end{pmatrix} \end{array}$$

g - Gap extension penalty = -7.0

b - CpG Alignment Bonus = 0.0

Metrics

Sequencing quality BiQ Analyzer HiMod controls the sequencing quality by permitting at most 5% of the bases in each sequence read to have sequence quality score below a user-specified threshold. The reads with low quality are discarded. The mean quality of all bases is calculated by dividing the score of each base by the number of bases of the sequence:

$$Q = \frac{\sum_{i=1}^{l_{read}} f_i}{l_{read}}$$

f_i - FastQ score of the base *i*

l_{read} - The length of the read

Sequence identity

The sequence identity is the fraction of bases which are the same in the reference and the read sequence including the converted cytosines. Based on the length of the shorter sequence the corresponding parts of both sequences are compared.

$$SI = \frac{n_{conv} + n_{id}}{min(l_{ref}, l_{read})}$$

n_{conv} - The number of bases converted in one sequence but with the same origin

n_{id} - The number of identical bases

l_{ref} - The length of the reference sequence

l_{read} - The length of the read

Conversion rate

The conversion rate is the ratio is calculated for each read as the number of converted cytosines outside of the modification sites to the number of not converted cytosines outside of the modification sites.

$$CR = \frac{t - t_{in}}{c_{out}}$$

t - the number of all converted cytosines

t_{in} - the number of converted cytosines in the modification sites

c_{out} - the total number of cytosines out of modification sites

Modification rates

The modification rate is calculated by dividing the number of converted modification sites by the total number of the reads in a read batch. This is done for each type of modification (5mC, 5hmC, 5fC, 5caC):

$$MR = \frac{n_c}{n}$$

n_c - the number of converted modification sites

n - the total number of the reads

Output

BiQ Analyzer HiMod has rich export capabilities. The results can be saved as tables in tab-separated value (TSV) format, as vector images in SVG format, as genome-browser tracks etc. The instant and optional output files are summarized in the Tables 5.S1 to 5.S3.

Supplementary Figures

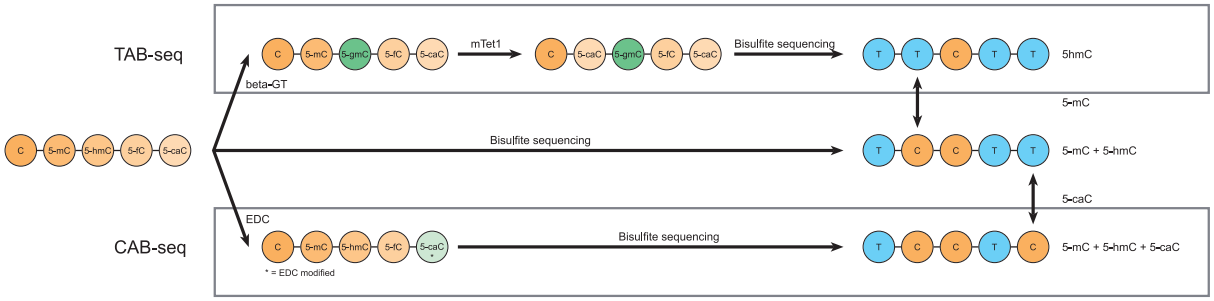


Figure 5.S1: Principal scheme of TAB-seq and CAB-seq methods. **a** TAB-seq applies enzymatic oxidation by Tet1 protein to convert 5mC to 5hmC and further to 5fC and 5caC, while the initially present 5hmC is protected by glycosylation. The 5mC level is calculated by subtraction from conventional bisulfite sequencing. **b** In CAB-seq 5caC is protected from the conversion and together with 5mC and 5hmC appears as cytosine after PCR. The bulk 5caC abundance is calculated by subtracting the cumulative levels of 5mC + 5hmC pair obtained from ordinary bisulfite sequencing.

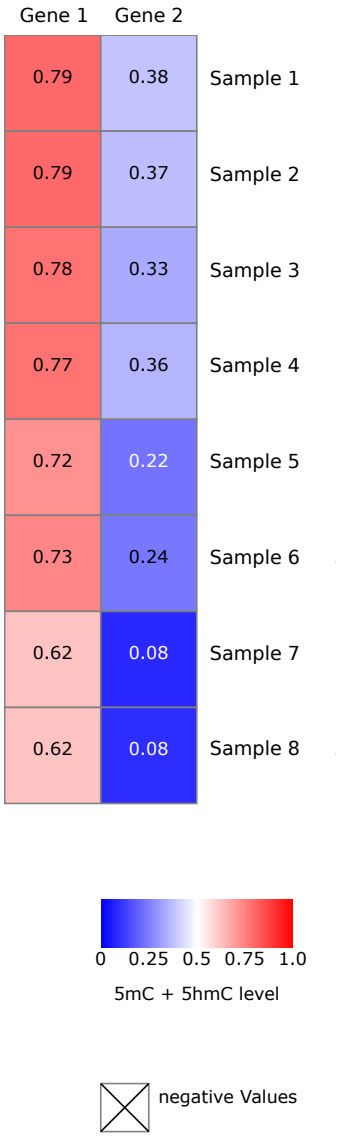


Figure 5.S2: A project heatmap displays a matrix of mean modification level for each analyzed amplicon in every sample. In case of an analysis bases on two sequencing approaches, there are three project heatmaps: one for each sequencing approach and one for the difference between the returned values. Each entry is colored according to the mean modification level of the amplicon. Furthermore the mean modification level is given as a decimal between 0 and 1. In case of difference heatmaps, negative values are possible. Large negative values are, as a rule, indicative of problems with the experimental procedure. Such cases are marked with an X-crossed cells.

Figure 5.S3: A locus heatmap visualizes a matrix of mean modification levels for every potential modification site of very sample for one specified amplicon. In case of an analysis bases on two sequencing approaches, there are three reference heatmaps: one for each sequencing approach and one for the difference between the returned values. The graphical code is the same as for the project heatmap (see Supplementary Fig. 5.S2).

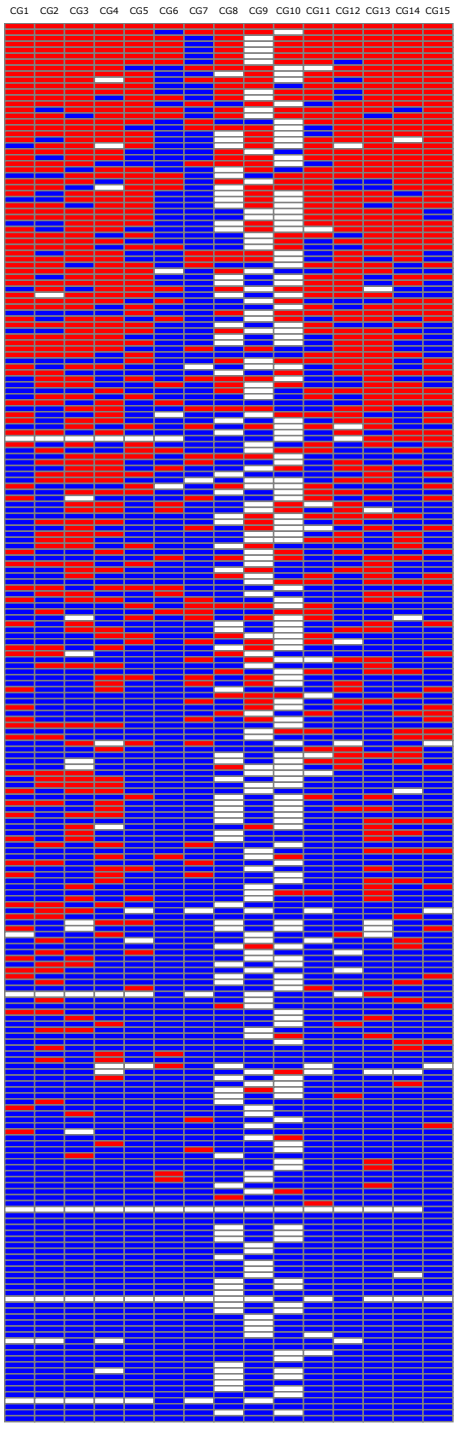


Figure 5.S4: A DNA modification pattern map displays a matrix which shows the conversion state of every potentially modifiable cytosine in every read of a processed read set. Each column corresponds to a modification site (e.g., a CpG) while each row represents a single read. The colors representing converted and unconverted states of the modification site-cytosines as well as the color for mutated modification sites can be customized.

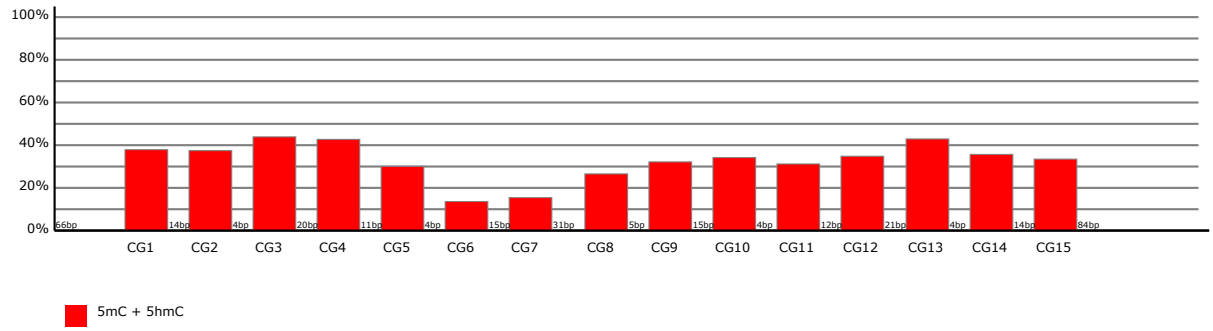


Figure 5.S5: A locus-wide bar chart shows the modification level for each site in a given locus for one selected sample. For an analysis project with two given readsets (i.e., for all modified bisulfite methods) there are two bar charts for each amplicon in every sample. A difference bar chart shows the calculated difference of the modification levels of the two readsets. For instance, in case of an oxidative bisulfite analysis project the difference between the modification levels represents the bulk 5hmC level. The values represented by this bar chart are expected to fall into [0,1] interval, while large negative values point at technical problems of the assay.

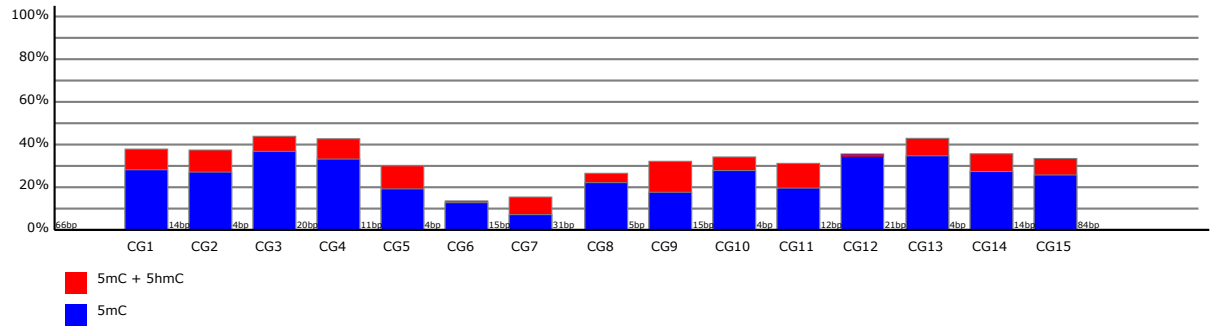


Figure 5.S6: A stacked locus-wide bar chart shows the modification levels for both, modified and conventional bisulfite read sets in one plot. The bars of the readset, in which the modification level is is expected to be higher, are plotted in the background layer while the bars of the other readset are plotted on top of them. Therefore one can read out both values as well as the difference. In case the front row bars are higher than the background ones, the latter are displayed by a thin horizontal line.

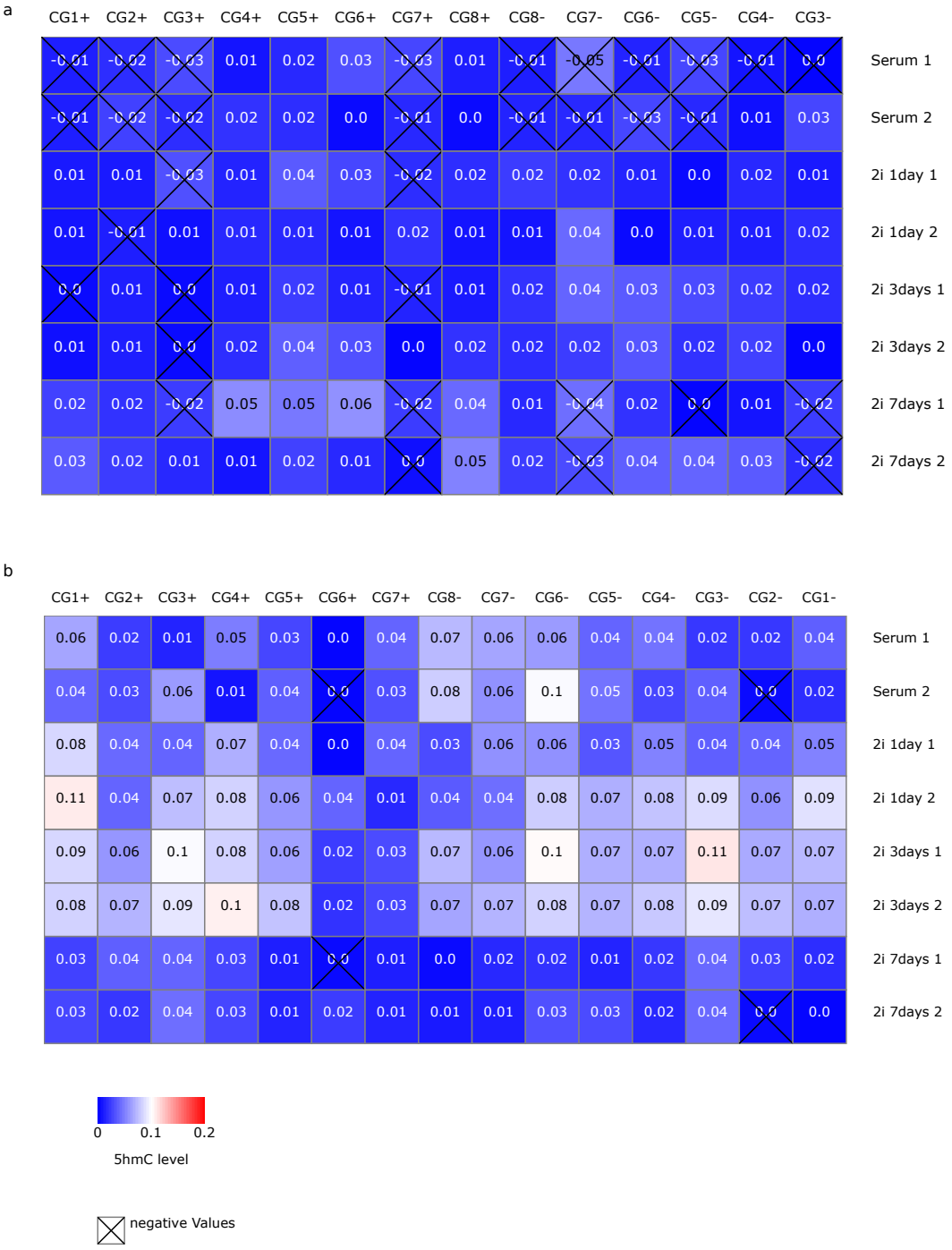


Figure 5.S7: Reprocessing of Ficz et al. oxBS-seq data [Ficz *et al.*, 2013]. **a** 5hmC level locus-wide heatmap for the IAP amplicon. **b** 5hmC level locus-wide heatmap for the LINE1 amplicon. Since the sequencing was done on a hairpin-library, most of the covered CpG sites are called twice: one time for each of the original DNA strands.

Supplementary Tables

Table 5.S1: Main exported graphics

File	Content	Type
Project heatmap	Mean modification data for each amplicon sample combination	SVG
Statistic heatmap / table	Heatmaps and tables containing mean values of important statistics as sequence identity, read length, standard deviation, conversion rate, number of loaded/filtered/exported reads	SVG
Reference heatmap	Mean modification data for each modification context position of every sample	SVG
DNA modification pattern map	Modification pattern for every read of the specified amplicon sample combination	SVG
Single bar chart	Mean modification level for every modification context position	SVG
Difference bar chart	Mean difference modification level for every modification context position	SVG
Comparison bar chart	Mean modification level for every modification context position. Contains modification levels for both readsets	SVG

Table 5.S2: Main exported files

File	Content	Type
Alignment	A pseudo-multiple sequence alignment obtained by merging the pairwise read-reference alignment	FASTA
Result table for a readset	Tab separated information about ID, Alignment score, sequence identity, modification pattern, mean modification level per read, missing CG number, conversion rate reference and sample	TSV
Summary result table	Tab separated information about Reference Sample, CG position and the three mean modification levels	TSV
Summary for a readset	A summary of the analysis	TXT
Comparison Summary	A summary of the read set comparison	TXT

Table 5.S3: Optionally exported files

File	Content	Type
All results table	One tab-delimited table for all samples and amplicons containing ID, alignment score, sequence identity, modification pattern, mean modification, missing sites, conversion rate, reference name and sample name	TSV
Pattern table	One tab-delimited table for the selected sample and amplicon containing ID, tab separated pattern, reference name, sample name and if necessary the readset type	TSV
Pattern table for a reference	Tab-delimited tables for all sample of the selected reference containing ID, tab separated pattern, reference name and sample name. One for each readset type	TSV
GFF file	Genome browser track (one per sample)	GFF
Bedgraph	Genome browser track (one per sample)	bedGraph

References

Arand, J., Spieler, D., Karius, T., Branco, M. R., Meilinger, D., Meissner, A. *et al.* In vivo control of CpG and non-CpG DNA methylation by DNA methyltransferases. *PLoS Genetics*, 8(6), 2012.

Bergman, Y. and Cedar, H. DNA methylation dynamics in health and disease. *Nature structural & molecular biology*, 20(3):274–81, 2013.

Blankenberg, D., Kuster, G. V., Coraor, N., Ananda, G., Lazarus, R., Mangan, M. *et al.* Galaxy: A web-based genome analysis tool for experimentalists. *Current Protocols in Molecular Biology*, Chapter 19(SUPPL. 89):Unit 19 10 1–21, 2010.

Bock, C., Reither, S., Mikeska, T., Paulsen, M., Walter, J., and Lengauer, T. BiQ Analyzer: Visualization and quality control for DNA methylation data from bisulfite sequencing. *Bioinformatics*, 21(21):4067–4068, 2005.

Bock, C., Tomazou, E. M., Brinkman, A. B., Müller, F., Simmer, F., Gu, H. *et al.* Quantitative comparison of genome-wide DNA methylation mapping technologies. *Nature Biotechnology*, 28(10):1106–1114, 2010.

Booth, M. J., Branco, M. R., Ficz, G., Oxley, D., Krueger, F., Reik, W. *et al.* Quantitative Sequencing of 5-Methylcytosine and 5-Hydroxymethylcytosine at Single-Base Resolution. *Science*, 336(6083):934–937, 2012.

Branco, M. R., Ficz, G., and Reik, W. Uncovering the role of 5-hydroxymethylcytosine in the epigenome. *Nature Reviews Genetics*, 13(1):7–13, 2011.

Ficz, G., Hore, T. a., Santos, F., Lee, H. J., Dean, W., Arand, J. *et al.* FGF signaling inhibition in ESCs drives rapid genome-wide demethylation to the epigenetic ground state of pluripotency. *Cell Stem Cell*, 13(3):351–359, 2013.

Giardine, B., Riemer, C., Hardison, R. C., Burhans, R., Elnitski, L., Shah, P. *et al.* Galaxy: A platform for interactive large-scale genome analysis. *Genome Research*, 15(10):1451–1455, 2005.

Goecks, J., Nekrutenko, A., and Taylor, J. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*, 11(8):R86, 2010.

Harris, R. A., Wang, T., Coarfa, C., Nagarajan, R. P., Hong, C., Downey, S. L. *et al.* Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nature Biotechnology*, 28(10):1097–1105, 2010.

Holland, R. C. G., Down, T. a., Pocock, M., Prlic, a., Huen, D., James, K. *et al.* BioJava: an open-source framework for bioinformatics. *Bioinformatics*, 24(18):2096–2097, 2008.

Hu, J. and Tzeng, J. Y. Integrative gene set analysis of multi-platform data with sample heterogeneity. *Bioinformatics*, 30(11):1501–1507, 2014.

Huang, Y., Pastor, W. a., Shen, Y., Tahiliani, M., Liu, D. R., and Rao, A. The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing. *PLoS ONE*, 5(1):e8888, 2010.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, a. M. *et al.* BLAT–the BLAST-like alignment tool. *Genome Res*, 12(4):656–664, 2002.

Kohli, R. M. and Zhang, Y. TET enzymes, TDG and the dynamics of DNA demethylation. *Nature*, 502(7472):472–9, 2013.

Kumaki, Y., Oda, M., and Okano, M. QUMA: quantification tool for methylation analysis. *Nucleic acids research*, 36(Web Server issue):W170–5, 2008.

Lu, X., Song, C. X., Szulwach, K., Wang, Z., Weidenbacher, P., Jin, P. *et al.* Chemical modification-assisted bisulfite sequencing (CAB-seq) for 5-carboxylcytosine detection in DNA. *Journal of the American Chemical Society*, 135(25):9315–9317, 2013.

Lutsik, P., Feuerbach, L., Arand, J., Lengauer, T., Walter, J., and Bock, C. BiQ Analyzer HT: Locus-specific analysis of DNA methylation by high-throughput bisulfite sequencing. *Nucleic Acids Research*, 39(S2):W551–6, 2011.

Quy, J., Zhouy, M., Song, Q., Hong, E. E., and Smith, A. D. MLML: Consistent simultaneous estimates of DNA methylation and hydroxymethylation. *Bioinformatics*, 29(20):2645–2646, 2013.

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G. *et al.* Integrative genomics viewer. *Nature Biotechnology*, 29(1):24–26, 2011.

Rohde, C., Zhang, Y., Reinhardt, R., and Jeltsch, A. BISMA–fast and accurate bisulfite sequencing data analysis of individual clones from unique and repetitive sequences. *BMC bioinformatics*, 11:230, 2010.

Schüler, P. and Miller, A. K. Sequencing the sixth base (5-Hydroxymethylcytosine): Selective DNA oxidation enables base-pair resolution. *Angewandte Chemie - International Edition*, 51(43):10704–10707, 2012.

Smith, Z. D. and Meissner, A. DNA methylation: roles in mammalian development. *Nature Reviews. Genetics*, 14(3):204–20, 2013.

Song, C. X., Szulwach, K. E., Dai, Q., Fu, Y., Mao, S. Q., Lin, L. *et al.* Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming. *Cell*, 153(3):678–691, 2013.

Chapter 6

MeDeCom discovers and quantifies latent components of heterogeneous methylomes

The full text of this chapter is a manuscript by:

Pavlo Lutsik[1,#], Martin Slawski[2,#], Gilles Gasparoni[1], Matthias Hein[3], and Jörn Walter[1,*]

The author of the present thesis suggested the idea of the methylome deconvolution, participated in the design of the MeDeCom model and the algorithms, implemented the algorithms and the supporting infrastructure in R language, selected, prepared and analyzed the simulated and biological data sets. With the other co-authors he wrote the manuscript, created the figures, tables and compiled the supplementary material.

[1]Genetics/Epigenetics, Saarland University, Saarbrücken, Germany
[2]Department of Statistics and Biostatistics and Department of Computer Science, Rutgers University, Piscataway, USA
[3]Machine Learning, Saarland University, Saarbrücken, Germany
[#]These authors contributed equally to this work.
[*]To whom correspondence should be addressed.

Abstract

Background: Large-scale DNA methylation studies on blood or tissue samples are confronted with sample-specific confounding factors such as heterogeneous cell composition and individual genetic variation. An unbiased discovery and exploration of such confounding factors is very important for epigenetic studies.
Results: Here, we present MeDeCom, a reference-free computational framework using non-negative matrix factorization to decompose complex methylome data of cell mixtures. MeDeCom not only generates interpretable latent methylation components (LMCs) but also provides estimates of LMC proportions per sample. LMCs can be used for biological exploration, correlating them to reference epigenomes or to other annotated cell type-specific molecular signatures. Estimated proportions can be used to interpret sample-specific variation, disease or age-related phenotypes. Here we demonstrate the MeDeCom performance on artificial cell mixtures and complex biological methylome data sets generated for whole blood, partially purified cell populations and complex brain tissue.
Conclusions: MeDeCom produces a reference-free deconvolution of complex methylation data across many samples. The latent components defined by MeDeCom can be used for biological exploration. MeDeCom can be applied to any high-resolution WGBS or Illumina 450k/EPIC array data set of mixed cell populations that has a sufficient technical quality.

Keywords: DNA methylation, DNA methylome, cell heterogeneity, deconvolution, matrix factorization, epigenetics

Competing interests

The authors declare that they have no competing interests.

Author's contributions

P.L. conceived the project, did the R implementation and analyzed the data. M.S. and M.H. designed the MeDeCom models (with P.L.) and developed the algorithms implemented and prototyped by M.S. GG provided conceptual advice on neuronal data analysis. MH suggested improvements for primary data processing and co-supervised the project. JW provided key expertise for interpretation of biological results and co-supervised the project. All authors discussed the results at all stages and contributed to writing of the manuscript.

Funding

P.L. obtained funding from the the European Union's Seventh Framework Programme (FP7/2007-2013) grant agreement No. 267038 (NOTOX) M.S. was supported by DFG Cluster of Excellence MMCI. The time on a computational cluster was provided by German Science Ministry grant No. 01KU1216A (DEEP).

Acknowledgements

Our implementation of the matrix factorization Algorithm 1 in this paper is based on extensions of code originally developed by Qinqing Zheng in collaboration with MS and MH. We would like to thank Karl Nordström, Abdulrahman Salhab and the DEEP project for providing and processing the WGBS data of the CD4+ T-cells.

6.1 Background

DNA methylation is one of the most extensively studied epigenetic marks in the human genome. While it is relatively easy to measure, DNA methylation closely mirrors the functional state of a cell [Schübeler, 2015]. Each human cell potentially comes with a characteristic methylation profile (methylome) at roughly 27 million CpG dinucleotides [Pelizzola and Ecker, 2011; Roadmap Epigenomics Consortium *et al.*, 2015]. Our current understanding of DNA methylomes is that: i) they undergo significant global and lineage-related changes during development [Reik *et al.*, 2001]; ii) consequently, they are remarkably cell type-specific [Baron *et al.*, 2006; Ji *et al.*, 2010; Roadmap Epigenomics Consortium *et al.*, 2015]; iii) they reflect the individual (genetic) constitution [Shoemaker *et al.*, 2010] but are also subject to environmental influences [Christiansen *et al.*, 2009; Lee and Pausova, 2013]; iv) they change with aging [Horvath, 2013] and v) they accumulate errors over time and are altered in diseased cells [Baylin, 2005; Esteller, 2007]. DNA methylation can therefore be used to infer the developmental origin, functional and disease-associated changes of cells. However, our methodological capacity to study complete single-cell methylomes on a large scale has been rather limited so far [Schwartzman and Tanay, 2015].

For practical reasons comparative epigenomic studies make use of multi-cellular samples from tissues, organs or body fluids such as blood [Bernstein *et al.*, 2010; Roadmap Epigenomics Consortium *et al.*, 2015]. All these sources are usually composed of several major and minor cell types. Brain tissue and whole blood are two examples of complex primary human resources widely used for comparative DNA methylation studies [Michels *et al.*, 2013]. Whole blood leukocytes, which is the most widely used resource for DNA methylome studies [Michels *et al.*, 2013], include up to ten major and many more minor cell types. Cell population- or cell type-attributed heterogeneity was shown to be a major source of variation in comparative blood-based DNA methylome studies [Lam *et al.*, 2012]. The same holds for studies performed with brain where different neuronal, glial and microglial cell types are present. The compositional change of cells in brain tissues may be due to age, gender and disease state [Kaut *et al.*, 2012; Lunnon *et al.*, 2014; Zhang *et al.*, 2010, 2011]. Therefore, when comparing the brain-specific DNA methylome between individuals it is important to consider such confounding effects [Montaño *et al.*, 2013]. Overall, variable cell composition appears to be one of the strongest confounders in DNA methylome analysis [Adalsteinsson *et al.*, 2012; Houseman *et al.*, 2015; Jaffe and Irizarry, 2014].

In view of these observations DNA methylation studies have been applying cell enrichment or cell separation techniques [Dainiak *et al.*, 2007; Tomlinson *et al.*, 2013] to experimentally homogenize the samples prior to methylation analysis [Bundo *et al.*, 2016; Rakyan *et al.*, 2011]. These methods certainly enhance the interpretability of results, but also come with the risk of introducing undefinable experimental variation, for instance due to non-specific labeling, problems with tissue dissociation, insufficient depletion of non-target cells etc [Kumar and Bhardwaj, 2008; Tomlinson *et al.*, 2013]. In the worst case, cell separation may even exclude unknown but informative cell populations. An ideal analysis would, therefore, avoid cell separation and rather study cell-specific methylomes to understand the complex epigenetic changes occurring in the tissue. Single-cell methylomes would be the gold standard to tackle this problem, but they are still difficult and costly to obtain for studies in which large sample numbers have to be compared [Fang *et al.*, 2012; Kantlehner *et al.*, 2011; Schadt *et al.*, 2013; Schwartzman and Tanay, 2015]. Moreover, non-uniform cell separation or sampling in single-cell approaches may even introduce uncontrollable confounding effects since important changes in rare or difficult-to-recover cell populations may be missed.

Possible approaches to overcome heterogeneity problems is the use of computational estimation or correction (adjustment) methods [Lowe and Rakyan, 2014]. Houseman *et al.* were the first to develop a systematic approach that used reference DNA methylation profiles of purified cell types to infer the cell type proportions in blood via a constrained projection procedure [Houseman *et al.*, 2012]. This method is currently applied for DNA methylation comparisons in whole blood [Accomando *et al.*, 2014; Koestler *et al.*, 2013; Liu *et al.*, 2013]. Similar cell separation-based approaches were also used for complex tissues such as brain [Guintivano *et al.*, 2013; Montaño *et al.*, 2013]. Finally, reference-free methods were developed that adjust for DNA methylation changes caused by cell heterogeneity allowing for the quantification of "direct" methylation effects [Houseman and Ince, 2014; Rahmani *et al.*, 2016; Zou *et al.*, 2014].

So far, no method has been described which would be able to infer full DNA methylomes of cell populations and their mixing proportions in a reference-free manner. Here we present a computational framework called MeDeCom implementing such an unsupervised decomposition approach. MeDeCom builds upon the discreteness of cellular DNA methylation states. It uses a novel matrix factorization approach to decompose methylome data into a set of underlying latent DNA methylation components (LMCs) and infers their relative contribution. LMCs can be used for biological interpretation of complex cell composition and cell states. We demonstrate the performance of MeDeCom in controlled experimental settings and its application in more complex scenarios of cell populations and tissues. MeDeCom allows for an original representation of the DNA methylation data with great relevance to the interpretation of complex biological samples.

6.2 Results and discussion

6.2.1 MeDeCom: a computational framework for decomposition of mixed methylomes

We constructed a novel computational framework MeDeCom to decompose mixed methylomes and recover the hidden signatures of individual cell populations. The concept behind our method is illustrated in Figure 6.1. DNA methylation is a stable base-specific modification. It can be seen as a discrete position-specific value in DNA extracted from cells (Figure 6.1, a). Each cell type has a characteristic pattern of discrete (binary) position-specific methylation states. Most biological samples are mixtures of cells and hence mixtures of homogeneous cell type-specific methylomes with sample-specific frequencies (Figure 6.1, b). Patterns and frequencies are convoluted in DNA methylation data obtained from mixed biological samples (Figure 6.1, c). The variability of methylation between biological samples directly reflects the underlying mixing proportions and numbers of individual and distinguishable cell methylomes. This information can be extracted from varying "intermediate" methylation values in mixed samples (Figure 6.1, d). Our method decomposes the mixed methylomes into latent methylation components (LMCs) approximating the underlying cell type-specific patterns and estimates their frequencies (Figure 6.1, e).

With these premises MeDeCom consists of three key elements (fully described in the Method section). First, we postulate a linear mixture model representing a set of mixture methylomes as a noisy product of matrices of LMCs and mixture proportions (equation (6.1) in Methods). Second, for fitting the model we devise a matrix factorization algorithm which uses a special quadratic regularisation to impose correspondence between LMCs and the hidden (cell-specific) methylomes. Third, we developed a cross-validation heuristic to select the adjustable parameters of the algorithm. The first parameter is the hidden dimension k corre-

sponding to the number of identifiable LMCs. The second parameter is a tunable regularization constant λ which controls the strength of the regularization and helps to obtain accurate reconstructions of latent methylation states (see Figure 6.1, e and f).

In what follows we apply MeDeCom to synthetic and real Infinium 450k data sets, and demonstrate its usefulness to decompose and interpret complex tissue- and cell-based methylation data. The results of MeDeCom can be visualized and inspected by FactorViz, a tool allowing a user to evaluate and interpret results over a range of possible solutions. MeDeCom as well as the interactive FactorViz web-resource are publicly available at `http://public.genetik.uni-sb.de/medecom`.

6.2.2 Validation on synthetic and artificial data

Decomposition of simulated methylation data

We first validated MeDeCom on synthetic DNA methylation mixtures generated by simulation (see Methods for details). The synthetic data sets varied in the numbers of true LMCs, the inter-LMC similarity, the average distribution and the variability of the mixing proportions (see Table 6.S1). Our goal was to assess the robustness of MeDeCom with respect to the recovery of LMCs and mixture proportions.

Figure 6.2, a-f summarizes the results of a single characteristic test case with moderately variable mixing proportions of five blood-derived cell type profiles (see below). The cross-validation error (CVE) started leveling out at $k \geq 5$, indicating that MeDeCom identified the correct number of underlying LMCs (Figure 6.2, a). The optimal range for the regularisation parameter λ was found around $\lambda = 0.01$. Recovered LMCs unambiguously matched the source DNA methylation profiles (Figure 6.2, b). The individual methylation profiles were reconstructed with root mean squared error (RMSE) of 0.064. MeDeCom also accurately reproduced the mixing coefficients (proportions) with mean absolute error (MAE) of 0.0296 (Figure 6.2, c-f). We obtained similar results for other cases with various numbers of underlying components and mixing proportions (see the MeDeCom web-resource).

For extreme mixtures we observed our method approaching its performance limits. The summary plots of the LMC recovery rate (Figure 6.S1) showed that a low number of discriminating data points, the choice of the model and the variability level of the mixing proportions were key factors for the efficiency of LMC reconstruction. We noticed that the test cases with highly non-uniform ("biological") proportions were much more challenging for MeDeCom as compared to the test cases with more similar ("uniform") synthetic cell type composition. Decomposition became impossible when variability of the mixture proportions was very low and, at the same time, was coupled to elevated noise levels (see an example in Figure 6.S2 and the MeDeCom web resource).

Decomposition of reconstructed cell mixtures

Next we analysed the performance of MeDeCom on public data sets in which cell mixtures were experimentally reconstructed, i.e. defined biological samples were mixed in known proportions. A very well analysed and described data set was obtained in a recent study on sorted brain cells [Guintivano *et al.*, 2013] (data set ArtMixN in Table 6.1). In brief, brain cells were separated using fluorescence activated cell sorting (FACS) for a neuron-specific marker NeuN, after which the obtained NeuN^+ (neuronal) and NeuN^- (non-neuronal) fractions were mixed incrementally (Table 6.S2) and the mixed methylomes were measured on a 450k array. We used

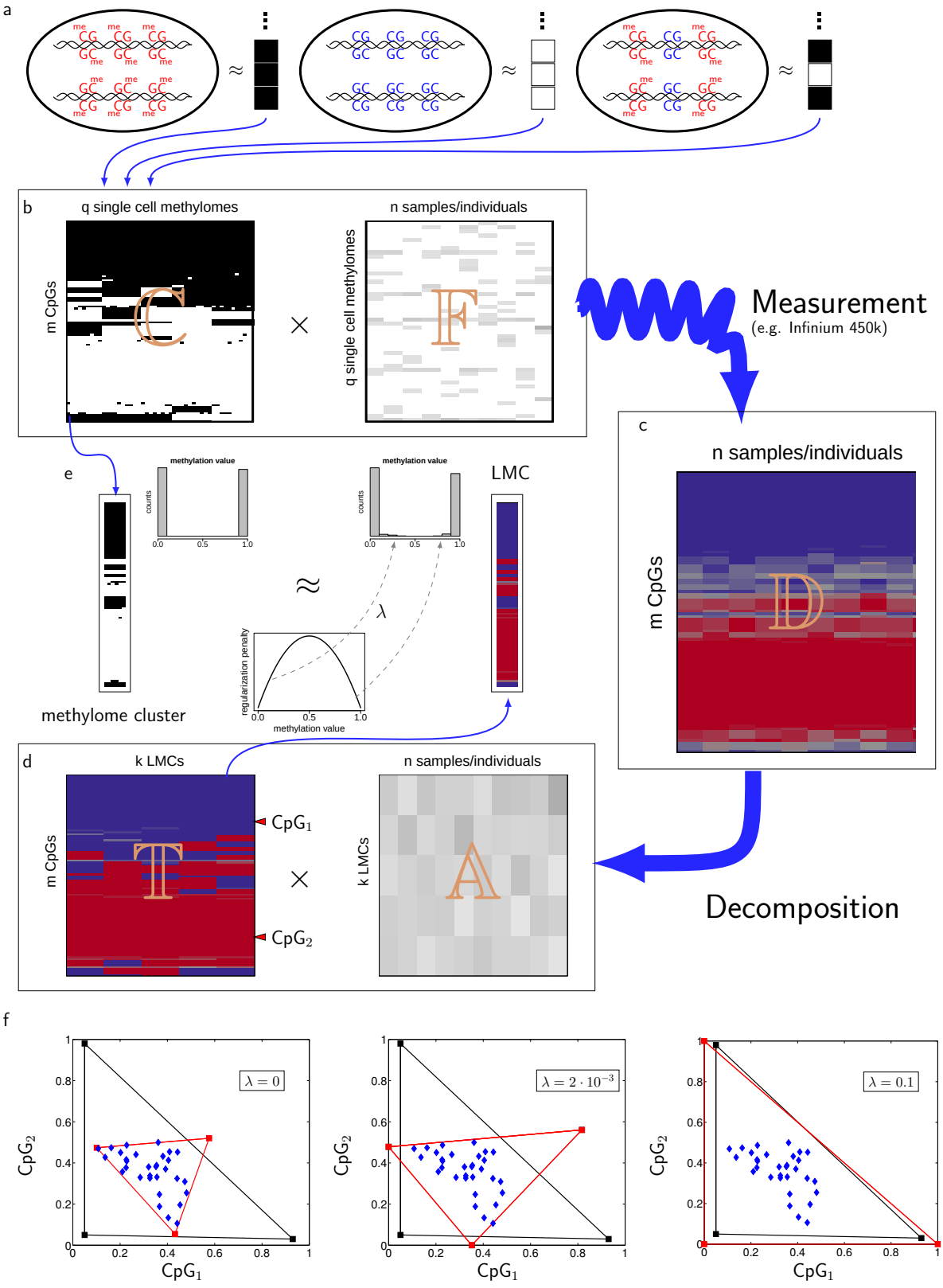

a
b
q single cell methylomes
n samples/individuals
m CpGs
q single cell methylomes
Measurement
(e.g. Infinium 450k)
c
n samples/individuals
m CpGs
e
methylation value
counts
LMC
methylome cluster
regularization penalty
methylation value
λ
d
k LMCs
n samples/individuals
m CpGs
k LMCs
CpG1
CpG2
Decomposition
f
λ = 0
λ = 2 · 10−3
λ = 0.1
CpG2
CpG1

Figure 6.1: (on the previous page) Computational framework of MeDeCom. **a**. At the level of a single cell each genomic CpG site can attain only several discrete DNA methylation states. With a good degree of approximation, single-cell methylomes can be represented by binary vectors. **b**. In one or more related cell samples, e.g. those from a population cohort, all unique single-cell methylomes can be summarized in a binary-valued matrix of unique methylomes C supplemented by a matrix of per-individual pattern frequencies F. Similar single-cell methylomes are expected to form well-separated clusters, corresponding to epigenetically distinct cell populations, such as cell types and subtypes. **c**. Full information in C and F is usually inaccessible in large-scale studies, and a data matrix D for m CpGs in n individuals (e.g. Infinium 450k-based or summarized WGBS profiles) arises as a product of the pattern and frequency matrices, distorted by sampling errors and mmeasurement noise. Inference about the pattern and frequency matrices is unrealistic based only on D. **d**. MeDeCom attempts to represent D as a product with a smaller internal dimension k ($k \ll q$). This gives rise to a matrix T of k methylome prototypes (LMCs) and a matrix A of their relative contributions (mixing proportions) to each sample (columns of A sum up to 1). **e**. The goal is that each LMC closely approximates a cluster of single-cell methylomes corresponding to a distinct cell population. This is by penalizing intermediate methylation values in LMCs by quadratic regularization. Degree of regularization is controlled by means of tunable parameter λ. **f**. A low-dimensional visualization for 2 CpGs ($n = 30, k = 3$) provides a good illustration of the potential ill-posedness of problem (6.2) (see Methods) and influence of the quadratic regularization term. The three plots depict the summarized methylome clusters (black squares), a set of given data (blue dots) and the LMCs returned by Algorithm 1 (red squares) for three different choices of the regularization parameter λ. All three solutions yield a perfect fit to the given data. However, only the solution for $\lambda = 0.1$ comes reasonably close to the ground truth.

ArtMixN data set to test how well MeDeCom could recover the source NeuN$^{+/-}$ methylomes and their mixing ratios.

MeDeCom identified the presence of two major LMCs and detected a sharp CVE minimum close to $\lambda = 1.25 \cdot 10^{-3}$ (Figure 6.2, G; Figure 6.S3). Each of the recovered LMCs showed high CpG-wise correlation to the average profile of either the NeuN^{+} or NeuN^{-} fractions (Figure 6.2, h) with high accuracy (RMSE of 0.0273 for the NeuN^{+}-matching, and 0.0269 for the NeuN^{-}-matching LMC). The mixing proportions were accurately recovered as well (MAE 0.014; Figure 6.2, i). Notably, the accuracy of the proportion recovery became compromised

Figure 6.2: (on the next page) Testing MeDeCom on simulated and artificial cell mixture data. **a-f**. Results for the simulated data example with 5 methylation components, moderately variable mixing proportions and medium noise level. **a**. Selection of parameters k and λ by cross-validation. **b**. Matching of the recovered LMCs to the true underlying profiles. The dendrogram visualizes agglomerative hierarchical clustering analysis with correlation-based distance measure and average linkage. **c-f**. Recovery of the mixing proportions. "Truth" stands for true mixing proportions, "regression" denotes the reference-based proportion estimation as described in the Methods. In each line plot the synthetic samples are sorted by ascending true mixing proportion. **g-i**. Results for the ArtMixN data set. **g**. Selection of parameters k and λ by cross-validation. **h**. Correlation of recovered LMCs to the NeuN$^{+/-}$ profiles. **i**. Recovery of mixing proportions (only NeuN^{+} is shown). Notation is the same as in **c**.-**f**..

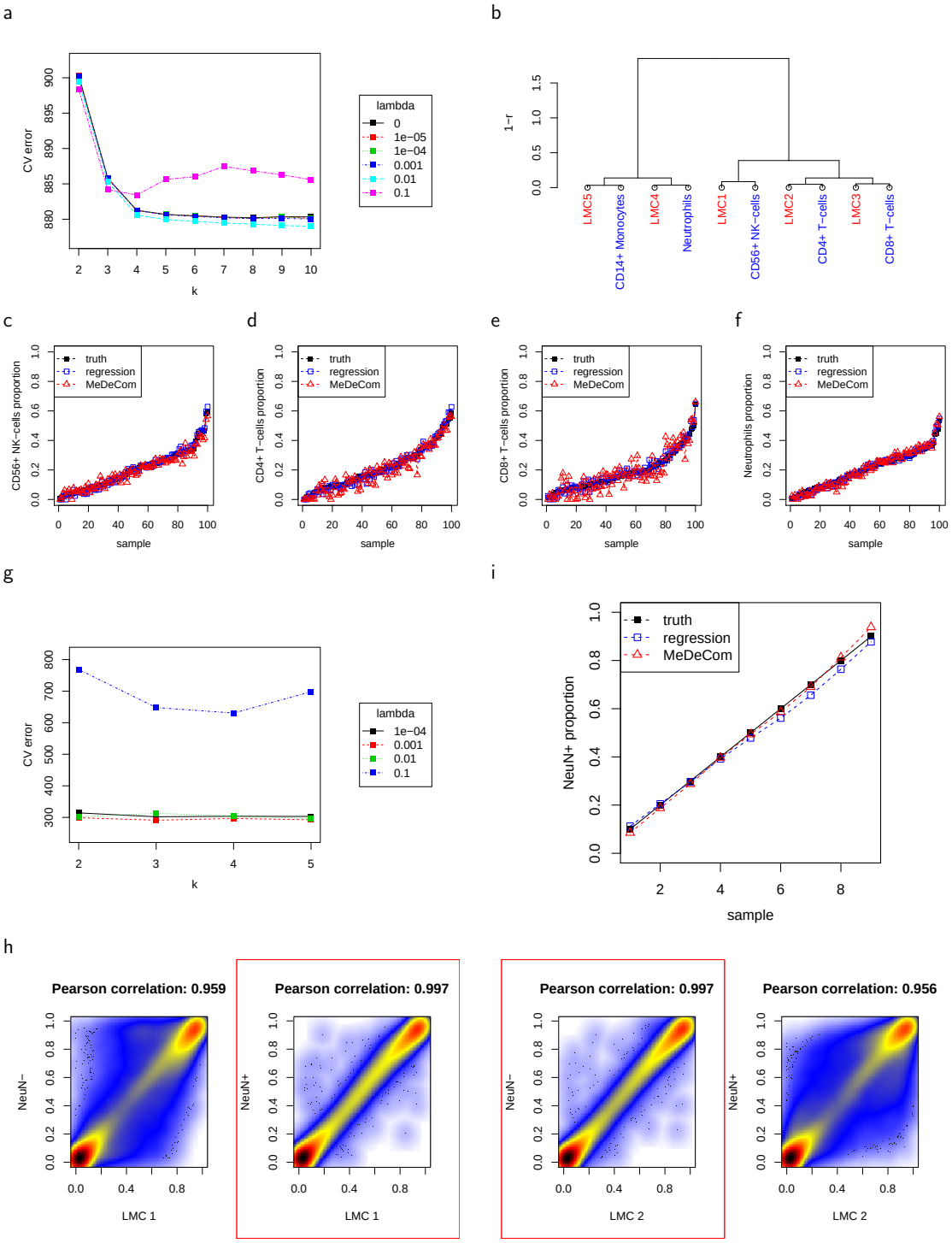

a
CV error
k
lambda
0
1e−05
1e−04
0.001
0.01
0.1
b
1−r
LMC5
CD14+ Monocytes
LMC4
Neutrophils
LMC1
CD56+ NK−cells
LMC2
CD4+ T−cells
LMC3
CD8+ T−cells
c
CD56+ NK−cells proportion
d
CD4+ T−cells proportion
e
CD8+ T−cells proportion
f
Neutrophils proportion
truth
regression
MeDeCom
sample
g
CV error
k
lambda
1e−04
0.001
0.01
0.1
i
NeuN+ proportion
truth
regression
MeDeCom
sample
h
Pearson correlation: 0.959
Pearson correlation: 0.997
Pearson correlation: 0.997
Pearson correlation: 0.956
NeuN−
NeuN+
LMC 1
LMC 2

Table 6.1: Public Infinium 450k data sets used in the study.

ID	Source	GEO Acc.	Brief description	n	Reference
Blood datasets					
PureBC	[Reinius *et al.*, 2012]	GSE35069	7 MACS-purified blood cell types from blood of 6 healthy male donors: Neutrophils, Monocytes, B-cells, CD4+ and CD8+ T-cells, NK-cells, and Eosinophils	42	
WB1	[Liu *et al.*, 2013]	GSE42861	whole blood of healthy controls from an RA study (technical batch II)	87	PureBC
WB2	[Palli *et al.*, 2003]	GSE51032	whole blood of the EPIC Italy study participants which remained cancer-free in 2010	442	PureBC
Neuronal datasets					
PureN	[Guintivano *et al.*, 2013]	GSE15745	cortical NeuN$^{+/-}$ fractions of the 29 healthy controls	58	
ArtMixN	[Guintivano *et al.*, 2013]	GSE15745	9 titration mixtures of the NeuN$^{+/-}$ fractions	9	PureN
FC1	[Guintivano *et al.*, 2013]	GSE15745	frontal cortex of 10 MDD patients and 10 healthy controls	20	PureN
FC2	[Lunnon *et al.*, 2014]	GSE15745	frontal cortex from a large AD study	114	PureN

GEO – Gene Expression Omnibus; MDD – major depression disorder; RA – rheumatoid arthritis; AD – Alzheimer's disese

at lower and higher λ values showing that it is important to determine optimal values for the regularisation parameter (Figure 6.S4). We also complicated the recovery problem by excluding certain mixtures, e.g. those with NeuN^{+} proportion ≥ 0.5. MeDeCom successfully decomposed such data subsets as well (Figure 6.S5).

6.2.3 Methylome decomposition of blood cell samples

Following the successful testing of MeDeCom on synthetic data and reconstructed cell mixtures we applied our method to complex biological samples with unknown composition. Since blood-based methylome profiling is widely used for comparative epigenetic studies we set out to determine whether MeDeCom can help identify and understand cell composition differences in blood cell methylomes.

Whole blood methylome

We used publicly available Infinium 450k profiles of whole blood samples from two independent studies (Table 6.1). This allowed us to test the reproducibility and robustness of our

method in a side-by-side comparison. We first applied MeDeCom to control samples from a large rheumatoid arthritis study [Liu *et al.*, 2013]. As reported in the original publication, the samples grouped into technical batches. This was reflected in the cell type proportions estimated by regression (Figure 6.S6 and 6.S7). To avoid technical variation as a confounding factor we performed our analysis on 87 samples forming a smaller technically homogeneous batch (data set WB1).

In our MeDeCom analysis the CVE continued to decline for $k = 20, \ldots, 30$, implying a potentially large number of distinct LMCs (Figure 6.3, a). We therefore considered factorization results for increasing values of k to understand the relation to underlying major and minor cell populations. To match the identified blood LMCs to known blood cell populations we compared our results to the published reference methylomes of blood cells [Reinius *et al.*, 2012] (data set PureBC).

At $k = 2$ the recovered LMC1 and LMC2 represented populations of the myeloid and the lymphoid lineages, respectively (Figure 6.S8). LMC contributions correlated to the summarized reference-based proportion estimates of lymphoid and myeloid cell types, although the former were noticeably biased (Figure 6.S9).

At increasing values of k, LMCs with high similarities to reference profiles of purified blood cells are still present, and for $k = 20$ and $\lambda = 1.0 \cdot 10^{-3}$ the clustering analysis results in two large clusters corresponding to myeloid and lymphoid lineages (Figure 6.3, b; Figure 6.S10). The 11 LMCs in the "myeloid" cluster together with Monocytes, Eosinophiles and Neutrophils are separated from 9 LMCs clustering with CD4+ T-cells, CD8+ T-cells, NK-cells and B-cells. In the “myeloid” cluster we did not observe direct matches of LMCs to any of the three myeloid cell types. In contrast, within the “lymphoid” cluster we observed one LMC closely matching the CD4+ T-cell profile, and one LMC corresponding to the sub-cluster of CD8+ T-cells and NK-cells, indicating better separability of the T-cell signatures. The analysis also identifies a number of individual-specific LMCs (Figure 6.S11 and 6.S12).

Analysis of the second completely independent whole blood data set from the EPIC Italy study [Palli *et al.*, 2003] (data set WB2) recovered a very similar pattern of LMCs for $k = 20$ (Figure 6.3, b and c). The direct comparison of the LMCs from both whole blood data sets reveals a considerable agreement in terms of recovered components in both data sets (Figure 6.3, d).

An aggregated comparison of LMCs matching to reference cell types in both blood analyses showed a good correspondence to the regression-based estimations of cell proportions (Figure 6.S13). For several LMCs in WB1 and WB2 we observed that their proportions correlated with age, e.g. for WB1 the LMC12 related to CD4+ T-cells (Figure 6.3, e). Although the total number of CD4+ T-cells was reported to change non-significantly with age [Fahey *et al.*, 2000] the T-cell-specific immunological senescence is a well known phenomenon characterized by depletion of the naive T-cell subpopulations [Cossarizza *et al.*, 1996; Romanyukha and Yashin, 2003]. This might imply that LMC12 is rather reflecting the methylation pattern of the naive CD4+ T-cells. Indeed, a comparison to reference methylomes of isolated T-cells supports this suggestion (Figure 6.S14).

Next, we selected the 15,000 marker CpGs with the highest discriminative power between cell types (highest CpG-wise $p = 2.91 \cdot 10^{-44}$, ANOVA F-test). Visual comparison of these CpGs between individual reference populations and whole blood data (Figure 6.S15) shows substantially lower variation of these CpGs in whole blood. Indeed, only a relatively small proportions of marker CpGs also shows high variance across whole blood samples (see the row color code in Figure 6.S15). This means that the signal at the marker CpGs which discriminate cell types of purified myeloid and lymphoid lineages very well is too low as compared to

the level of overall variation in whole blood. This might explain why LMCs recovered in whole blood do not unambiguously match average methylomes of isolated cell populations. To examine this in greater detail we next applied MeDeCom to isolated cell populations.

Purified blood cell populations

Profiles of purified blood cell types used to annotate LMCs in the whole-blood analysis are effectively average methylomes of the cells carrying certain surface marker. However, the global DNA methylation pattern in various blood cells might not directly reflect the expression of the surface markers, even if a certain number of highly discriminative CpGs can be selected for each isolated population. To better understand the complexity of methylation signatures in the reference methylomes we performed a MeDeCom analysis on the seven purified blood cell populations derived from 6 donors (data set PureBC) [Reinius *et al.*, 2012].

The CVE suggested a stable solution at $k = 16$ and $\lambda = 10^{-3}$ (Figure 6.3, f; Figure 6.S16 and 6.S17). A matrix of mixing proportions (Figure 6.3, g) shows that the recovered 16 LMCs could be classified into two distinct groups. In the first group (LMCs 6, 7, 8, 10, 15 and 16) LMCs were associated with individual donors, most likely reflecting donor-specific genetic variation at the informative CpG positions underlying these LMCs. In other words, our analysis directly identified confounding genetic variants. In the second group, LMCs 1, 3, 4, 5, 7, 9 and 11 corresponded to the enriched cell type-associated profiles: e.g. LMC4 was predominantly present in CD4+ T-cells, LMC11 in Neutrophils etc. Nevertheless, we also observed that several LMCs were shared by related cell types. For instance, Eosinophils samples show enrichment of the Neutrophil-specific LMC11, CD8+ T-cells (LMC9) show overlaps with CD4+ T-cells (LMC5). Finally, we observed LMCs which were associated with more than one cell type, but which were not a dominating LMC in any of them. For instance, LMC14 was present at low proportion both in CD8+ T-cells and NK cells. Co-occurrence of two or more LMCs within one isolated cell population, as well as sharing of LMCs between the populations suggests that these cell populations could be either mixtures of still not separated distinct cell types, or that these cell populations share epigenetic features that may indeed co-occur in different cell types.

As a clear-cut example of such within-population heterogeneity, for the CD19+ B-cells we observed a relatively balanced association with two LMCs, LMC2 and LMC13. This is in line with the fact that peripheral blood contains two B-cell subpopulations: naive and memory B-cells with distinct methylomes [Kulis *et al.*, 2015]. We investigated the methylation status of CpGs associated with the two B-cell-specific LMCs in more detail and selected 401 CpG positions which showed a methylation difference of more than 0.33 in LMC2 as compared to LMC13. Many of such CpGs were indeed located in the vicinity of known B-cell associated genes (Additional File 1), such as *PTPRCAP* (Figure 6.S18). We related LMC2- and LMC13-specific CpGs to reference WGBS methylome profiles of memory and naive B-cell samples, obtained by the BLUEPRINT project [Kulis *et al.*, 2015]. The methylation status of 44 matching CpGs (Additional File 4) indeed corresponds to relative methylation states found in memory and naive B-cell subpopulations (Figure 6.3, d). The correspondence remained very stable even when the difference threshold for the selection of LMC-specific CpGs was decreased to 0.25 to cover more matching CpG positions (Figure 6.3, d). In addition, LMC2 and and LMC13 have almost inverse proportions for individual donors indicating that the MeDeCom analysis picks up differences in the sample-specific abundance of memory and naive B-cells highlighting individual- or isolation-attributed variation.

6.2.4 Decomposition of the brain tissue methylomes

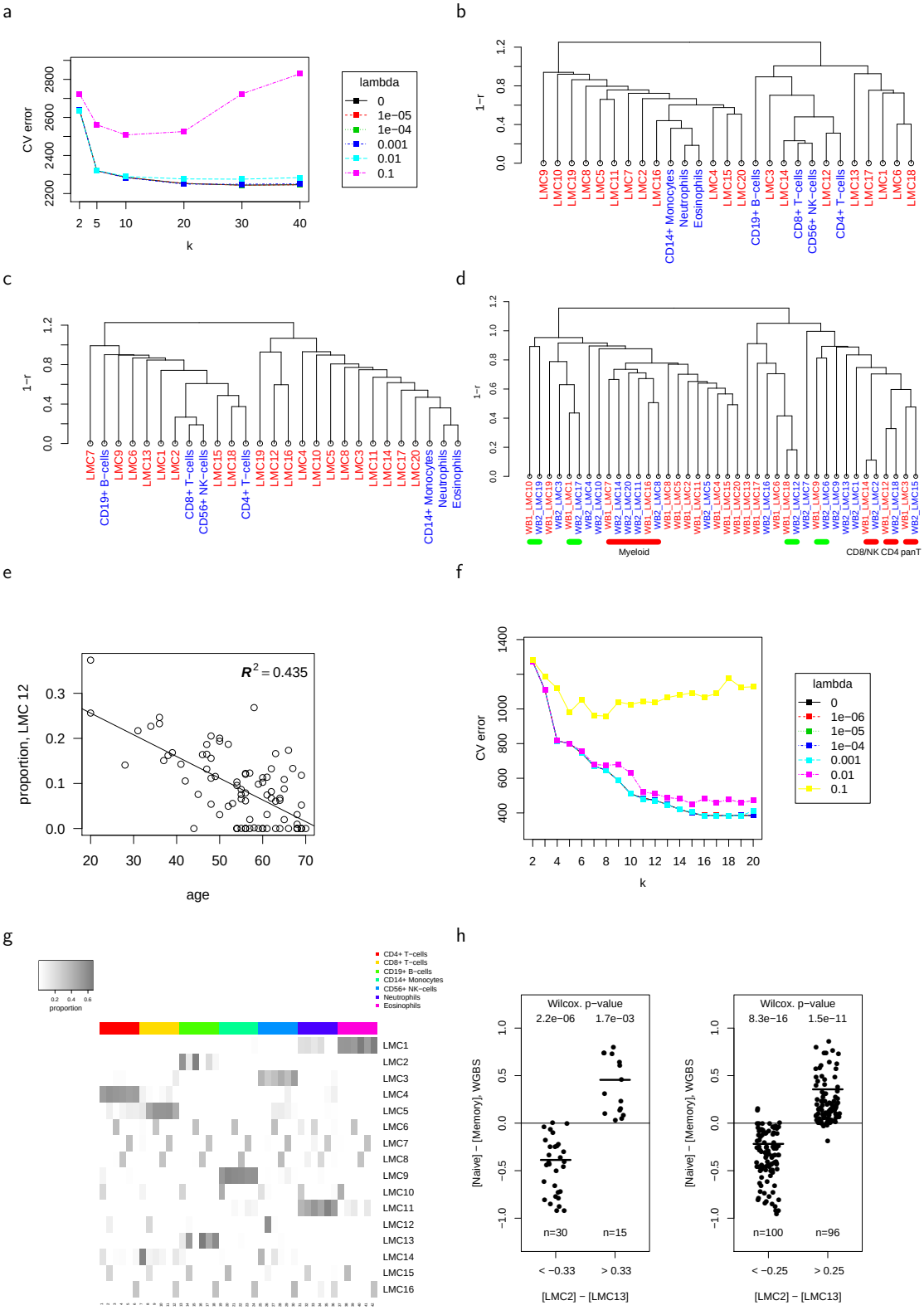
In our final analysis we applied MeDeCom to understand the heterogeneity of brain tissue methylome. The human brain is composed of many neuronal and glial cell types. To separate glial cells and neurons current studies apply FACS-based methods. The RBFOX3 protein localized in the nuclear membrane of most of the neuronal cells (also known as NeuN) is used as a selection marker. While the NeuN-enriched and depleted cell fractions serve as references in methylome analysis, the question remains to which extent these separated methylomes indeed reflect a variable composition of whole brain tissue.

We applied MeDeCom to 20 frontal cortex methylomes from a major depression disorder study [Guintivano *et al.*, 2013] (data set FC1 in Table 6.1). The data set also included NeuN^+ and NeuN^- cell fractions (data set PureN) which we analysed in comparison to total brain tissue. In addition, we examined an independent bulk frontal cortex methylome data set from a recent large-scale Alzheimer's disease study [Lunnon *et al.*, 2014] (data set FC2).

For both FC1 and FC2 data sets the inspection of CVEs showed a substantial change at $k \geq 3$ strongly suggesting three or more LMCs (Figure 6.4, a and Figure 6.S19). We carefully examined the factorization results and compared the three LMCs at $k = 3$ and $\lambda = 5 \cdot 10^{-3}$ to the NeuN^+ and NeuN^- profiles. Clustering analysis (Figure 6.4, b) showed that average NeuN^- reference profile is related to LMC3 while the NeuN^+ profile is more similar to LMC2. The third component LMC1 was truly distinct from both "reference" methylomes retaining a slightly higher similarity to LMC2 and the NeuN^+ methylome. All three LMCs were remarkably well reproduced in the FC2 data set at $k = 3$ (Figure 6.4, c).

Our finding suggests that the FACS separation of brain tissues in NeuN^+ and NeuN^- cells introduces a new confounding variable and does not fully recover the methylomes of total brain tissues. In order to get more clues about biological nature of the LMCs we asked which loci differ in their methylation between the LMCs and examined the biological annotation of genes associated with LMC-specific CpGs. LMC-specific CpGs were selected to have methylation differences more than 0.33 between one LMC against the two other ones (Additional File 2). We then mapped LMC-specific CpG positions to their neighboring genes (Additional

Figure 6.3: (on the next page) Results in blood cell methylomes. **a**-**e**: WB1 data set. **a**. Selection of parameters k and λ by cross-validation. **b**. Matching the WB1 LMCs to PureBC methylomes ($k = 20$, $\lambda = 0.001$). Here and below the dendrogram visualizes agglomerative hierarchical clustering analysis with correlation-based distance measure and average linkage. **c**. Matching the LMCs from the WB2 data set ($k = 20$, $\lambda = 0.001$) to the PureBC methylomes. **d**. Matching the WB1 and WB2 LMCs to each other. Pairs of reproducible LMCs also matching to the reference profiles are highlighted by red segments. Green segments mark reproducible LMCs which are not directly matching to any of the the reference profiles. **e**. Proportion of LMC12 from WB1 correlates with the age of healthy individuals. **f**-**h**. PureBC data. **f**. Selection of parameters k and λ by cross-validation. **g**. Heat map of recovered proportions in PureBC data ($k = 15$, $\lambda = 0.001$). Rows represent LMCs while columns correspond to individual purified samples. The order of blood donors is the same within column sets, corresponding to one cell type. **h**. Methylation differences in naïve versus memory B-cells at CpGs differentially methylated between LMC2 and LMC13 from PureBC data set. WGBS methylation profiles of naïve and memory B-cells are obtained from BLUEPRINT. The value in memory B-cells is an average of three WGBS samples. Wilcoxon ranked sum test was used to test the null that WBGS methylation calls are the same in naive and memory cells at the respective CpG positions.


a
CV error
k
lambda
0
1e−05
1e−04
0.001
0.01
0.1
b
1−r
LMC9
LMC10
LMC19
LMC8
LMC5
LMC11
LMC7
LMC2
LMC16
CD14+ Monocytes
Neutrophils
Eosinophils
LMC4
LMC15
LMC20
CD19+ B−cells
LMC3
LMC14
CD8+ T−cells
CD56+ NK−cells
LMC12
CD4+ T−cells
LMC13
LMC17
LMC1
LMC6
LMC18
c
1−r
LMC7
CD19+ B−cells
LMC9
LMC6
LMC13
LMC1
LMC2
CD8+ T−cells
CD56+ NK−cells
LMC15
LMC18
CD4+ T−cells
LMC19
LMC12
LMC16
LMC4
LMC10
LMC5
LMC8
LMC3
LMC11
LMC14
LMC17
LMC20
CD14+ Monocytes
Neutrophils
Eosinophils
d
1−r
Myeloid
CD8/NK CD4 panT
e
proportion, LMC 12
$R^2 = 0.435$
age
f
CV error
k
lambda
0
1e−06
1e−05
1e−04
0.001
0.01
0.1
g
proportion
CD4+ T−cells
CD8+ T−cells
CD19+ B−cells
CD14+ Monocytes
CD56+ NK−cells
Neutrophils
Eosinophils
LMC1
LMC2
LMC3
LMC4
LMC5
LMC6
LMC7
LMC8
LMC9
LMC10
LMC11
LMC12
LMC13
LMC14
LMC15
LMC16
h
Wilcox. p−value
2.2e−06
1.7e−03
[Naive] − [Memory], WGBS
n=30
n=15
< −0.33
> 0.33
[LMC2] − [LMC13]
Wilcox. p−value
8.3e−16
1.5e−11
[Naive] − [Memory], WGBS
n=100
n=96
< −0.25
> 0.25
[LMC2] − [LMC13]

File 5; see Methods) and performed a functional annotation of the associated genes using GREAT [McLean *et al.*, 2010] (Figure 6.S20). LMC2 ($NeuN^+$)-specific CpGs map to genes with a clear enrichment for neuronal-related terms, while LMC3 ($NeuN^-$)-specific CpGs were close to genes associated with non-neuronal, mostly oligodendrocyte-related categories. LMC1-specific CpGs map close to genes associated with developmental and stem cell-related terms. Strikingly, among the genes associated with LMC1 we found several markers of the neuronal stem cell lineage, such as *PAX6*, *ZIC1*, *ZIC4*, *NEUROG1* (Additional File 2). Notably, the DNA methylation patterns at LMC1-specific CpGs showed significantly higher or lower methylation levels in crude brain tissue than in $NeuN^+$ and $NeuN^+$ reference methylomes (see *PAX6* as an example in Figure 6.S21). Furthermore, a recent study on neuronal heterogeneity in the mouse brain [Mo *et al.*, 2015] provided a reference for the fine cellular subtypes possibly present in the mammalian frontal cortex. We found several of the most significant LMC1-specific genes among the DMRs reported in [Mo *et al.*, 2015] (Table 6.S3).

As we earlier observed that proportions tend to be biased when k is significantly lower than optimal (see the WB1 analysis with $k = 2$ above), we explored MeDeCom results at $k = 4$ and $\lambda = 0.005$ (Figure 6.S22). The analysis revealed that the $NeuN^+$-specific LMC3 rather accurately reproduced a reference-estimated $NeuN^+$ content in most brain samples (Figure 6.4, e; Figure 6.S22, a). However, samples with the highest deviation from the reference-based proportions had the highest proportion of co-purified cells (and methylomes) characteristic of LMC2 (equivalent to LMC1 for $k = 3$; Figure 6.4, f and Figure 6.S22, b). For $k = 4$ two LMCs match to $NeuN^-$. For each of them MeDeCom-recovered proportions deviated significantly from the reference-based estimates for $NeuN^-$ (Figure 6.4, h and i). Nevertheless, the combined proportions largely reflected the reference-estimated $NeuN^-$ content across all samples (Figure 6.4, j). Again, we observe that samples with the lowest correspondence had high contribution of LMC2 (Figure 6.4, g). The proportion analysis shows that using MeDeCom we can infer realistic LMC proportions for $NeuN^+$, $NeuN^-$ in individual samples and a third separate LMC with a distinct cell composition. The latter LMC is variably convoluted into the other main $NeuN^+$ and $NeuN^-$ cell fractions in the reference-based analysis. We conclude that reference-independent decomposition helps to explore, identify and quantify heterogeneity effects across brain tissue samples.

6.3 Conclusions

DNA methylomes of multicellular tissue or cell samples can be modeled as mixtures of several latent variables. Here we present a novel computational framework called MeDeCom which decomposes complex methylation data into latent components and sample-dependent proportions based on a mixture model for methylomes. We show that the method performs reproducibly and with high sensitivity on both synthetic and biological data sets.

MeDeCom provides a few advances compared to all existing methods. First of all, MeDeCom does no require predefined reference cell type measurements. It can be applied to any data set to explore the compositions of mixtures. Note that reference methylome data are not yet available for many cell types, and MeDeCom offers the possibility to explore non-standard data in a reference-free manner. Second, MeDeCom has conceptual differences to other reference-free methods such as the Surrogate Variable Analysis methods (SVA) [Leek and Storey, 2007] [Teschendorff *et al.*, 2011], EWASHER [Zou *et al.*, 2014] or the SVA-inspired RefFreeEWAS [Houseman *et al.*, 2014] method. All these methods focus on the correction of significance analysis for a phenotypic trait of interest by calculating and eliminating confounding heterogeneity effects. In contrast, MeDeCom uses a variant of Non-negative Matrix

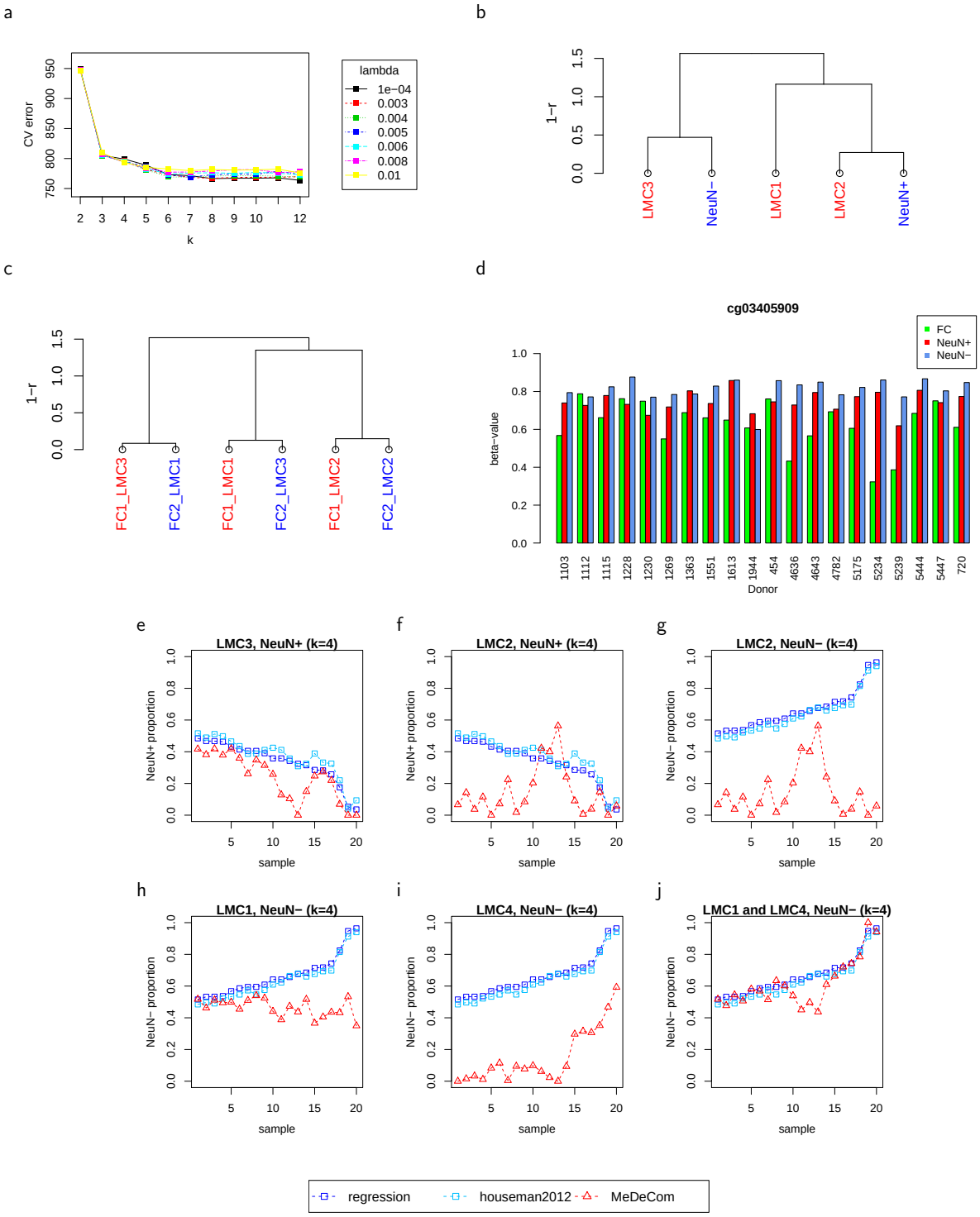


Figure 6.4: **a**,**b** and **e-j**: FC1 data set. **a**. Selection of parameters k and λ by cross-validation. **b**. Matching frontal cortex LMCs to the reference NeuN$^{+/-}$ profiles. The dendrogram visualizes agglomerative hierarchical clustering analysis with correlation-based distance measure and average linkage. **c**. Matching of LMCs between FC1 and FC2. **d**. Example of an LMC1-specific CpG ($k = 3$) in the PAX6 locus. **e-j**. LMC contributions in comparison to the reference-based proportion estimates ($k = 4$). Notation is the same as in 6.2, D. In **i** the proportions are summarized for LMC1 and LMC4.

Factorization (NMF) specifically designed to recover latent DNA methylomes by using suitable constraints and regularization. The imposed constraints on the factorization integrate biological prior knowledge such as non-negativity of the estimated methylation profiles and their proportions. However, we show that these constraints alone are not sufficient to get biologically meaningful methylation profiles and accurate estimates of their proportions. A key element distinguishing MeDeCom from other reference free methods is that we add a regularizer which encodes the prior expectation that most sites in the methylation profiles are close to zero or one. This prior expectation comes from the fact that at the level of a single cell, methylation profiles are binary and for most CpG sites this is true also at the level of a population of cells of a homogeneous group such as a particular cell-type. This allows us to estimate simultaneously methylation profiles and their proportions without any reference profiles. MeDeCom generates series of decomposition pattern models which can be used for supervised biological interpretation and for unsupervised cell compositional corrections.

Our proof-of-concept analysis shows that MeDeCom acts robustly and reliably on complex artificial and natural methylome mixtures measured by Infinium 450k arrays. MeDeCom identifies key signatures of major cell populations present in complex whole blood methylomes without any prior knowledge of references or data adjustment. However, our analysis also reveals the limits of a MeDeCom analysis. The method strongly depends on a fair number of discriminatory methylation positions and a sufficient level of sample-to-sample variation (Figure 6.S15). In complex 450k whole blood methylomes both parameters are affected such that a clean separation and assignment of LMCs specific for blood cell subtypes becomes challenging. Two major aspects are the likely causes of this difficulty. First, the Infinium 450k platform only covers a limited number of CpGs informative for the minor cell subtypes which can easily become indistinguishable from the remaining technical "noise" of the 450k arrays. Second, the proportions of most cell subtypes in blood are too low. MeDeCom factorization requires a certain grade of sample-to-sample variations to identify component (cell type)-specific CpG signals. We already noticed both of these limitations in our artificial mixture analyses. In the future these problems may be partially overcome by using WGBS/RRBS or extended array platforms such as the Methylation EPIC array covering additional cell type-specific "variable" CpG positions. Furthermore, cell enrichment or cell depletion strategies may help to obtain deeper sample-specific compositional insights.

Since MeDeCom does not require predefined references it can be flexibly applied to any level of methylome analysis. We show that MeDeCom can facilitate a deeper insight in cell composition if the sample complexity is experimentally reduced. As one example we investigate the composition of methylomes generated after cell pre-selection e.g. by surface marker-based separation [Reinius *et al.*, 2012]. Our results on pre-sorted CD4+ (T-cell) or CD19+ (B-cell) blood cells clearly show that their methylomes still maintain a substantial level of heterogeneity. We identify a number of additional separable DNA methylation components some of which we can associate with age-dependent changes in T-cell populations or show that they discriminate naive from primed B-cells. In both cases the characteristic CpG signatures vary in their sample-by-sample proportions. Such observations are very important for the biological interpretation of methylation changes across populations of samples. Many of the components identified by MeDeCom are likely to carry such biological information that can be extracted for further exploration.

The decomposition of brain methylomes provided by MeDeCom further supports the usefulness of unsupervised exploratory decomposition for the analysis of complex methylome data. The separation of brain cells into neuronal and non-neuronal fractions has become a new "standard procedure" for brain-specific epigenetic studies. Our first finding shows that

NeuN$^{+/-}$ mixture models do not fully capture the composition of the full brain tissue. In fact, we identify an additional component that differs from the NeuN^{+} (neuron) and NeuN^{-} (non-neuron)-specific components in full brain tissues. This new component is apparently sorted out or even lost in the enrichment procedure. Our analysis shows that the samples denominated as NeuN^{+} and NeuN^{-} contain variable contributions of this unknown cell fraction. Here MeDeCom opens a new possibility to identify the differences in cell composition and hence making data from different NeuN separations more comparable. Moreover, a biological analysis of the CpGs and genes associated with this "new" component reveals a strikingly different association of biological terms as compared to NeuN^{+} and NeuN^{-} fractions.

In summary, our exploratory analysis demonstrates that MeDeCom is a flexible and useful reference-free tool allowing to improve the biological interpretation of large-scale DNA methylation data sets. Although for the pilot demonstration we use Infinium 450k data, MeDeCom is, in principle, applicable to any complex methylome data set. However, since MeDeCom requires a low level of technical noise and a high level of biological variation we suggest that the method is applied to carefully controlled data sets that fulfill such requirements. A high standard technical preprocessing of 450k array data minimizes possible pitfalls of quality, technical batch effects or other non-biological issues. We therefore recommend to use data after filtering through available bioinformatic pipelines (see e.g. [Assenov *et al.*, 2014] or [Aryee *et al.*, 2014]).

6.4 Methods

6.4.1 MeDeCom element I: mixture model for DNA methylation measurements

Let $D \in [0,1]^{m \times n}$ be the matrix of absolute methylation values at m CpGs obtained from n multicellular specimens, with m typically being much larger than n. Here, entry D_{ij} represents methylation level at CpG i for specimen j, $i = 1, \ldots, m$, $j = 1, \ldots, n$. We consider an approximate low-rank model for D which hinges on the assumption that the cell populations underlying the specimens consist of a comparatively small number of homogeneous subpopulations of cells having a similar methylation profile, which is also shared across different specimens. When speaking of homogeneous subpopulations of cells, we typically have in mind cells of different types, e.g. oligodendrocytes of brain tissue or neutrophils of peripheral blood. We assume further that the methylation profiles of the specimen results as a weighted average ('mixture') of the methylation profiles associated with the underlying cell types, where the weights equal the proportions of these cell types within the respective specimen (note that we verified this assumption in our analysis with reconstructed cell mixtures). This lets us propose the matrix factorization model

$$D = TA + E, \tag{6.1}$$

where $T \in [0,1]^{m \times k}$ represents the methylation profiles of k prototypes (in typical cases identifiable with a specific cell type) and $A \in \mathbb{R}_{+}^{k \times n}$ such that $A^{\top}\mathbf{1}_k = \mathbf{1}_n$ (i.e. the entries of A are non-negative and its columns sum up to one). Entry T_{is} equals the methylation profile at CpG i of prototype s, $i = 1, \ldots, m$, $s = 1, \ldots, k$, while A_{si} equals equals the relative abundance (proportion) of prototype s in specimen i. The matrix E represents errors capturing model misspecification and noise arising from the measurement process.

6.4.2 MeDeCom element II: model fitting

Using a least squares approach to fit model (6.1) yields the optimization problem

$$\begin{aligned} \min_{T,A} \|D - TA\|_F^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}(D_{ij} - (TA)_{ij})^2 \\ \text{subject to } \; 0 \leq T_{is} \leq 1 \;\; \forall i, s \\ A_{sj} \geq 0 \;\; \forall s, j \\ \sum_{s=1}^{k} A_{sj} = 1 \;\; \forall j. \end{aligned} \tag{6.2}$$

Here and in the sequel, $\|\cdot\|_F$ denotes the Frobenius norm of a matrix, defined as the square root of the sum of squares of its entries. We may think of the above problem as an instance of 'blind source separation', a task that is well-studied in signal processing [Seungjin Choi, Andrzej Cichocki, Hyung-min Park, 2005]. The attribute 'blind' expresses the fact that the source signals as represented by the columns of the matrix T are unknown, as opposed to the case where they are given in advance and only the mixture coefficients in A need to be recovered.

The minimization problem in (6.2) is not jointly convex in T and A. As a result, one cannot hope to solve it globally; in fact, it has been shown that constrained matrix factorization problem of this form are computationally hard in general [Vavasis, 2007].

Once one of T or A is given, problem (6.2) boils down to a convex quadratic program. This property is the basis of *alternating minimization*, a common (heuristic) approach to fitting matrix factorization models in which minimization w.r.t. T for fixed A and vice versa alternate [Lin, 2007]. While lacking theoretical guarantees, alternating minimization often works rather satisfactorily in practice.

Besides computational hardness, a second major issue of (6.2) is ill-posedness. In general, there are multiple optimal solutions to (6.2) (excluding those generated by column respectively row permutations in T respectively A), as can easily be seen from geometric considerations, cf. Figure 6.1, f.

In geometric terms, problem (6.2) can be re-phrased as follows: find a set of k points $\{t_1, \ldots, t_k\} \subset [0,1]^m$ such that their convex hull $\mathcal{T} = \{y \in \mathbb{R}^m : y = \sum_{s=1}^{k} \lambda_s t_s, \;\; \lambda_s \geq 0 \;\forall s, \;\; \sum_{s=1}^{k} \lambda_s = 1\}$ minimizes the sum of squared Euclidean distances of the data points $\{D_{:,1}, \ldots, D_{:,n}\}$ to that convex hull. As shown in Figure 6.1, I, there exist problem instances for which it is possible to extend or shrink $\mathcal{T}$ while keeping the least squares objective (essentially) unchanged. To deal with such ambiguities, we suggest to complement the least squares objective with a regularizing term pushing the points $\{t_1, \ldots, t_k\}$ towards the vertex set of $[0,1]^m$, i.e. the set of binary vectors $\{0,1\}^m$. The rationale behind this idea is as follows. Recall that the columns of T take the role of methylation profiles of prototypes which in typical cases represent a (near)-homogeneous subpopulation of cells. Depending on the homogeneity of the subpopulation, the methylation profile of the corresponding prototype may be close to binary in view of the fact that at the level of a single cell, methylation profiles are exactly binary (methylated vs. unmethylated) when ignoring the comparatively rare case of half-methylation. It turns out that incorporating this structure contributes significantly to the success in finding biologically meaningful matrices T and A. Specifically, we consider the following regularized

least squares criterion:

$$\begin{aligned} &\min_{T,A} \|D - TA\|_F^2 + \lambda \sum_{i=1}^{m} \sum_{s=1}^{k} \omega(T_{is}), \quad \text{with } \omega(x) = x(1-x), \\ &\text{subject to } \; 0 \leq T_{is} \leq 1 \;\; \forall i, s \\ &\qquad\qquad A_{sj} \geq 0 \;\; \forall s, j \\ &\qquad\qquad \sum_{s=1}^{k} A_{sj} = 1 \;\; \forall j. \end{aligned} \tag{6.3}$$

where $\lambda > 0$ is a hyperparameter. Note that $\omega : [0,1] \to [0,1]$ is a quadratic function symmetric around its mode 0.5 (i.e. $\omega(x) = \omega(1-x)$) and vanishes at the boundary points 0 and 1. The additional regularization term in (6.3) acts as a 'soft binary constraint' depending on the parameter λ. For λ sufficiently large, any minimizer $(\widehat{T}, \widehat{A})$ of (6.3) must satisfy $\widehat{T}_{is} \in \{0,1\}$ for all i, s. We stress that the proposed form of regularization is much better suited to the given problem than the popular lasso (ℓ_1-regularization with $\omega(x) = |x|$), which promotes zeroes, but discourages ones, which is little meaningful for the given problem from a biological perspective.

From the point of view of computation, the extra term in (6.3) poses an additional challenge compared to (6.2) as the function ω is non-convex (in fact, it is concave). As a consequence, when using the alternatization scheme mentioned above, one has to bear in mind that optimizing T for fixed A is no longer a convex quadratic program, but a so-called D.C. (difference of convex) program in virtue of the concavity of ω. The concave-convex procedure [Tao and An, 1997; Yuille and Rangarajan, 2003] can be employed to generate a sequence of iterates ensuring monotonic descent of the objective function before reaching a stationary point. As detailed in Algorithm 1, it is straightforward to integrate this approach into the alternating optimization scheme.

The main computational efforts go into the successive solution of the convex quadratic optimization problems **optT** and **optA** which can be done by a variety of efficient solvers. The update of T follows the concave-convex procedure in which the concave part of the objective (here given by $h(T)$) is repeatedly linearized, yielding a sequence of convex 'surrogate' minimization problems.

6.4.3 MeDeCom element III: parameter selection

The mixture model (6.1) and the fitting algorithm (Algorithm 1) involve two free parameters to be provided by the user. The inner dimension k of the matrix product TA, $k \leq \min\{m, n\}$ in (6.1) equals the number of DNA methylation prototypes used to model the given data. The regularization parameter λ determines how strongly the entries of $\widehat{T}$ are encouraged to take values in $\{0,1\}$. While the choice of k can be guided by prior knowledge about the composition of the underlying mixture to some extent, we developed a cross-validation procedure to select suitable values of both k and λ in a data-driven manner.

Cross-validation

The use of cross-validation in the context of matrix factorization, which is in fact a problem in unsupervised learning, whereas cross-validation is typically used in supervised settings, requires additional explanation. A conceptual difference arises from the fact that as opposed to

Algorithm 1 Alternating minimization algorithm for objective (6.3)

Denote $\mathcal{C}_T = \{T \in \mathbb{R}^{m \times k} : 0 \leq T_{is} \leq 1,\ i = 1, \ldots, m,\ s = 1, \ldots, k\}$,
$\mathcal{C}_A = \{A \in \mathbb{R}^{k \times n} : A_{sj} \geq 0,\ \sum_{s=1}^{k} A_{sj} = 1,\ s = 1, \ldots, k,\ j = 1, \ldots, n\}$,

$g(T, A) = \|D - TA\|_F^2$, $h(T) = \lambda \sum_{i=1}^{m} \sum_{k=1}^{s} \omega(T_{is})$, and $f(T, A) = g(T, A) + h(T)$.

Initialize $T^0 \in \mathcal{C}_T$ and $A^0 \in \mathcal{C}_A$; fix numerical tolerance $\epsilon > 0$.
$t \leftarrow 0$, $f^t \leftarrow f(T^t, A^t)$.
repeat
  Update T:
  $t \leftarrow t + 1, \overline{T} \leftarrow T^{t-1}$
  repeat
    Linearize $h(T)$ around $T = \overline{T}$ to obtain a function
    $\widetilde{h}(T) = h(T^{t-1}) + \sum_{i=1}^{m} \sum_{s=1}^{k} \omega'(T_{is}^{t-1})(T_{is} - T_{is}^{t-1})$.
    $\overline{T} \leftarrow \operatorname{argmin}_{T \in \mathcal{C}_T} g(T, A^{t-1}) + \widetilde{h}(T)$ (**optT**)
  until $(f(\overline{T}, A^{t-1}) - f^{t-1})/f^{t-1} < \epsilon$
  $T^t \leftarrow \overline{T}$.
  Update A:
  $A^t \leftarrow \operatorname{argmin}_{A \in \mathcal{C}_A} g(T^t, A)$. (**optA**)
  $f^t \leftarrow f(T^t, A^t)$.
until $(f^t - f^{t-1})/f^{t-1} < \epsilon$.

the standard supervised setting, where the object to be predicted is a vector (one-dimensional array), one now has to deal with a matrix (two-dimensional array). There are multiple ways of generalizing the principle of leaving out sequentially different portions of the given data when moving from the vector to the matrix case, such as (a) leaving out columns, (b) leaving out rows, (c) leaving out both rows and columns [Owen and Perry, 2009]. We here use (a) mainly because it leads to a straightforward scheme as displayed in Algorithm 2. For each fold, a subset of the samples is left out. The thus column-reduced data matrix D^{in} is factorized as if one were given the full matrix. The resulting left factor $\widehat{T}^{\text{in}}$ is used to fit the left-out columns in D^{out} as $D^{\text{out}} \approx \widehat{T}^{\text{in}} \widehat{A}^{\text{out}}$. The squared error of that approximation or cross-validation error (CVE), is saved and finally combined with the errors from other folds.

Selecting k

The choice of k is canonical as long as the composition of the cell populations is known to a good extent, as is the case e.g. for synthetic mixtures. Cell populations sampled from human tissue tend to be considerably more complex. Prior knowledge about the number of cell types present in the samples may not be available, and even if it is, each cell type may not necessarily correspond to a perfectly homogeneous subpopulation. As a result, multiple similar, yet not identical methylation profiles may exist per cell type, reflecting a hierarchy of cell types and subpopulations. Furthermore, (sub)clusters can emerge from individual-specific DNA methylation effects, like allele-specific methylation and imprinting, or phenotypic effects, e.g. influence of age, gender, disease status etc. It is not feasible to capture such fine-grained structure given a small to moderate number of samples, which are in addition contaminated by noise. As a rule, k should be chosen such that the estimation error and the approximation error in

Algorithm 2 Column-based L-fold cross-validation scheme for validation of model (6.1)

Choose an integer $L \in \{1, \ldots, \lfloor n/2 \rfloor\}$.
Let $\mathcal{I} = \{1, \ldots, n\}$. Randomly partition $\mathcal{I}$ into disjoint subsets $\mathcal{I}_\ell$ so that $\lfloor n/L \rfloor \leq |\mathcal{I}_\ell| \leq \lceil n/L \rceil$ and $\sum_{\ell=1}^{L} |\mathcal{I}_\ell| = n$.
for $\ell \in \{1, \ldots, L\}$ **do**
Form $D^{\text{in}} = D_{:,\mathcal{I}\setminus\mathcal{I}_\ell}$, $D^{\text{out}} = D_{:,\mathcal{I}_\ell}$.

Solve the matrix factorization problem (6.3) with D^{in} in place of D and $\lambda = \lambda_g$. Denote the minimizing T by $\widehat{T}^{\text{in}}$.

Obtain $\widehat{A}^{\text{out}}$ as the minimizer of

$$\min_A \|D^{\text{out}} - \widehat{T}^{\text{in}} A\|_F^2 \text{ subject to } A_{sj} \geq 0 \ \forall s, j, \ \sum_{s=1}^{k} A_{sj} = 1 \ \forall j.$$

$\text{err}_g^{(\ell)} \leftarrow \|D - \widehat{T}^{\text{in}} \widehat{A}^{\text{out}}\|_F^2$
end for
return $\text{err}_g \leftarrow \sum_{\ell=1}^{L} \text{err}_g^{(\ell)}$.

model (6.1) are roughly balanced. The former results from noise and is incurred when fitting the model to the data, while the latter is a consequence of model misspecification, which, as discussed above, is inevitable for limited k given the many possible sources of diversity among methylation profiles.

From a more statistical perspective, the issue of choosing k is related to determining the number of components in principal component analysis (PCA). In fact, the matrix factorization model (6.1) can be seen as a method of linear dimension reduction applied to D. A common computational approach to PCA is the singular value decomposition (SVD) which yields a matrix factorization of rank k of D by discarding all singular vectors not corresponding to the top k singular values. A notable advantage of our model (6.1) over the truncated SVD / PCA is its direct interpretability at a biological level, which is achieved by putting suitable constraints on the two factors T and A.

For a fixed value of the parameter λ, the data-fitting term of the factorization problem (6.3) decreases as k increases. The approximation error of the factorization model decreases since with more columns in T one has a better chance of capturing differences between the cluster methylomes. At the same time, the estimation error increases as the additional degrees of freedom favour over-adaptation to noise. A suitable choice of k balances both effects. The use of cross-validation is intended to achieve this balance by tracing the cross-validation error over a grid of values for k and selecting the one corresponding to the minimum. The final choice of k was made by combining visual inspection of the cross-validation results and available prior information about the most likely number of underlying methylation signatures.

Selecting λ

As illustrated by the example in Figure 6.1B, the regularization parameter λ, which balances the trade-off between the data fidelity term and the data-independent regularization term, has a crucial influence on the solution of the factorization problem (6.1) delivered by Algorithm 1. Since there is in general no objective criterion to assess the suitability of each solution at a

biological level, we embark on cross-validation error as for the parameter k. Determining the value of λ achieving minimum cross-validation error is more difficult as that parameter takes values in a continuous domain, namely the non-negative real line. We perform a two-stage grid search, starting with a coarse grid and then concentrating on a smaller range covered by a finer grid. Details of the procedure are outlined in Algorithm 3. At the beginning of each of the two rounds of grid search, Algorithm 3 is run for each grid point of λ using multiple ($\approx$50) random initializations. As the solutions corresponding to nearby grid points can be expected to be similar, we complement random initializations with a smoothing scheme in which the solutions of the five preceding and the five subsequent grid points are used for initialization.

6.4.4 LMC matching

When interpreting MeDeCom results we matched the recovered methylation prototypes to available reference methylomes. Given a matrix of k prototypes $\widehat{T}$ estimated from a data set D and a matrix of $k^\star$ reference profiles $T^\star$, we first selected a set of rows $\mathcal{R}$ corresponding to the overlap of CpGs present in both $\widehat{T}$ and $T^\star$. We then computed the matrix $S = (S_{i,j})$ of Pearson correlation coefficients between all pairs of vectors $\widehat{T}_{\mathcal{R},i}$ and $T^\star_{\mathcal{R},j}$. We considered prototype $\bar{i}$ as a match to reference profile $\bar{j}$ if $S_{\bar{i},\bar{j}} = \max_i S_{i,\bar{j}}$. We considered the matching unambiguous in case $S_{\bar{i},\bar{j}} = \max_j S_{\bar{i},j} = \max_i S_{i,\bar{j}}$ for all such matching pairs $(\bar{i},\bar{j})$. In most of the cases, we observed better matching when both $\hat{T}$ and $T^\star$ were centered, i.e. $\frac{1}{k}\hat{T}\mathbf{1}_k$ respectively $\frac{1}{k^\star}T^\star\mathbf{1}_{k^\star}$ was subtracted from each column. In order to compare sets of prototypes corresponding to different parameter settings, we normalized the total number of unambiguously matching prototypes by the achievable maximum, which yields a score $\epsilon \in [0,1]$ given by $\epsilon = \frac{1}{\min(k,k^\star)} \left|\left\{(\bar{i},\bar{j}) \in \{1,\ldots,k\} \times \{1,\ldots,k^\star\} : S_{\bar{i},\bar{j}} = \max_j S_{\bar{i},j} \text{ and } S_{\bar{i},\bar{j}} = \max_i S_{i,\bar{j}}\right\}\right|$

We also performed a combined clustering analysis of prototypes and reference profiles. For that we composed a matrix $T^\dagger = [\widehat{T}_{\mathcal{R},:}\ T^\star_{\mathcal{R},:}]$, computed a correlation matrix $S^\dagger$ analogously to S, and used it as a similarity matrix for agglomerative hierarchical clustering with average linkage (procedure `hclust` in the R package *clust*).

6.4.5 Functional annotation of LMC-specific CpG positions

For functional annotation of the recovered prototypes we selected component-specific CpG positions using a fixed methylation difference threshold θ. We considered a CpG position $l \in \{1,\ldots,m\}$ to be specific to component i if $|\widehat{T}_{l,j} - \sum_{j\neq i} \widehat{T}_{l,j}| > \theta$. We investigated each set $\mathcal{L}_j$ of all such CpGs with respect to enrichment of annotation categories using GREAT [McLean *et al.*, 2010]. We used the default definition for a functional domain of a gene, with maximal distance of 10 kb upstream or downstream of the transcriptional start site ("two closest genes" option in GREAT).

6.4.6 Reference-based estimation of the cell type proportions

In case a matrix T of k prototype methylomes is available, e.g. experimentally obtained using cell separation methods, one can estimate a corresponding matrix of mixing proportions by solving sub-problem **optA** in Algorithm 1. From here onwards we refer to this method as "regression", and we apply it for reference-based estimation of mixing proportions whenever the reference methylomes are available. This form of proportion estimation is similar to a method with the name 'Constrained Projection' proposed for the same purpose in [Houseman *et al.*,

Algorithm 3 Choosing the parameter λ by L-fold cross-validation

Let $\Lambda = \{\lambda_1, \ldots, \lambda_G\}$ be a grid of values to be considered for λ, where the spacings between the grid points is typically linear on a log-scale, e.g. $10^{-10}, 10^{-9}, \ldots 10^{-1}$
Choose an integer $L \in \{1, \ldots, \lfloor n/2 \rfloor\}$.
Let $\mathcal{I} = \{1, \ldots, n\}$. Randomly partition $\mathcal{I}$ into disjoint subsets $\mathcal{I}_\ell$ so that $\lfloor n/L \rfloor \leq |\mathcal{I}_\ell| \leq \lceil n/L \rceil$ and $\sum_{\ell=1}^{L} |\mathcal{I}_\ell| = n$.

for $\ell \in \{1, \ldots, L\}$ **do**
  for $g \in \{1, \ldots, G\}$ **do**

(1) Form $D^{\text{in}} = D_{:,\mathcal{I}\setminus\mathcal{I}_\ell}$, $D^{\text{out}} = D_{:,\mathcal{I}_\ell}$.

(2) Solve the matrix factorization problem (6.3) with D^{in} in place of D and $\lambda = \lambda_g$ using N_{it} random initializations. Denote the minimizing T by $\widehat{T}_g^{\text{in}}$ and the corresponding objective value by f_g.

(3) Obtain $\widehat{A}^{\text{out}}$ as the minimizer of

$$\min_A \|D^{\text{out}} - \widehat{T}^{\text{in}} A\|_F^2 \text{ subject to } A_{sj} \geq 0 \ \forall s, j, \ \sum_{s=1}^{k} A_{sj} = 1 \ \forall j.$$

    for $g' \in \{\max(g-5, 1), \ldots, \max(g-1, 1)\}$ **do**

      Solve the matrix factorization problem (6.3) with D^{in} in place of D and $\lambda = \lambda_g$ using $(\widehat{T}_{g'}^{\text{in}}, \widehat{A}_{g'}^{\text{out}})$ for initialization. Denote the solution by $(\bar{T}, \bar{A})$ and its objective value by $\bar{f}$.

      if $\bar{f} < f_g$ **then**
        $(\widehat{T}_g^{\text{in}}, \widehat{A}_g^{\text{out}}) \leftarrow (\bar{T}, \bar{A})$ and $f_g \leftarrow \bar{f}$
      end if

    end for
    $\text{err}_g^{(\ell)} \leftarrow \|D - \widehat{T}^{\text{in}} \widehat{A}^{\text{out}}\|_F^2$
  end for
  for $g \in \{G, \ldots, 1\}$ **do**
    for $g' \in \{\min(g+1, G), \ldots, \min(g+5, G)\}$ **do**

      Solve the matrix factorization problem (6.3) with D^{in} in place of D and $\lambda = \lambda_g$, using $(\widehat{T}_{g'}^{\text{in}}, \widehat{A}_{g'}^{\text{out}})$ for initialization. Denote the solution by $(\bar{T}, \bar{A})$ and its objective value by $\bar{f}$.

      if $\bar{f} < f_g$ **then**
        $(\widehat{T}_g^{\text{in}}, \widehat{A}_g^{\text{out}}) \leftarrow (\bar{T}, \bar{A})$ and $f_g \leftarrow \bar{f}$
      end if

    end for
    $\text{err}_g^{(\ell)} \leftarrow \|D - \widehat{T}^{\text{in}} \widehat{A}^{\text{out}}\|_F^2$
  end for
  $\text{err}_g \leftarrow \sum_{\ell=1}^{L} \text{err}_g^{(\ell)}$.
end for
return $\lambda^* = \lambda_{g^*}$ with g^* defined by $\text{err}_{g^*} = \min_{1 \leq g \leq G} \text{err}_g$.

2012]. The important difference is, however, that the analogue of the matrix T in that method is constructed from a selection of a comparatively small set of cell type-specific marker CpGs. In the sequel we refer to this method as "houseman2012", and we compare to its proportion estimates whenever appropriate.

6.4.7 Simulations

We simulated DNA methylation data by mixing measured profiles of purified cell types with controlled proportions and adding varying levels of Gaussian noise. An $m \times n$ matrix of DNA methylation values D_{sim} was generated according to the model in (6.1). The underlying matrix of LMCs $T \in [0,1]^{m \times k_{\text{sim}}}$ was obtained by averaging methylation profiles for k_{sim} purified blood cell types from 6 donors of the Reinius *et al.* study [Reinius *et al.*, 2012]. We tested four different constellations of blood cell types:

- $k_{\text{sim}} = 2$ with two distant cell types (Neutrophils and CD4+ T-cells),
- $k_{\text{sim}} = 2$ with two similar cell types (Neutrophils and Monocytes),
- $k_{\text{sim}} = 3$ with two similar cell types and one distant from the first two (Neutrophils, Monocytes and CD4+ T-cells),
- $k_{\text{sim}} = 5$ with all major blood cell types, excluding Eosinophils and B-cells.

The columns of the matrix of mixing proportions A were sampled from a Dirichlet distribution commonly used to model distributions over the probability simplex. The distribution had k_{sim} parameters $v\alpha_1, \ldots, v\alpha_{k_{\text{sim}}}$. The simplex base $\alpha_1, \ldots, \alpha_{k_{\text{sim}}}$, $\sum_i \alpha_i = 1$ was chosen to model the prior expectation about the mixing proportions in a typical individual. We tested two scenarios: on average equal ("uniform") proportions across individuals, i.e. $\alpha_i = \frac{1}{k_{\text{sim}}}$, $i = 1 \ldots k_{\text{sim}}$, and a setting where some concentration parameter values were much larger than others, which comes closer to the situation one encounters for whole blood. The scaling factor v was used to control the variability of the mixing proportions, with $v = 1$ yielding highly variable, $v = 10$ moderately variable and $v = 100$ marginally variable proportions across individuals. Finally, the additive noise term E was generated by sampling mn values from a Gaussian distribution with mean 0 and standard deviation 0.05, 0.1, and 0.2 to simulate low, moderate and high levels of noise, respectively.

6.4.8 Infinium 450k data

Public Infinium 450k data sets

The publicly available data sets used for the validation of the factorization approach are summarized in the Table 6.1. For testing MeDeCom in blood-based data we used one reference data set and data from two large whole blood-based studies. The data set by Reinius *et al.* contained profiles of purified blood cell types, as well as mixed samples with known cell counts [Reinius *et al.*, 2012]. Next, we used data from a large rheumatoid arthritis (RA) EWAS with 354 cases and 337 controls [Liu *et al.*, 2013]. Finally, we validated the whole-blood results in the data from the EPIC Italy prospective cohort which provided for 845 Infinium 450k measurements [Palli *et al.*, 2003]. Neuronal data sets were obtained from one reference study, and one large Alzheimer's disese (AD) cohort. As reference we used data from the CETS study [Guintivano *et al.*, 2013] which contained in total 145 Infinium 450k profiles of various neuronal samples from major depression disorder patients and healthy controls, such as

cortical NeuN$^+$ and NeuN$^-$ enriched cell populations, 9 artificial NeuN$^{+/-}$ titration mixtures, as well as 20 intact frontal cortex samples. For validation we used data from a recent AD study [Lunnon *et al.*, 2014].

Processing and preparation of the Infinium 450k data

The raw Infinium 450k data was collected as IDAT files or, if the latter were not available, from probe-wise intensity matrices (Illumina Genome Studio reports). Loading and primary processing, such as intensity summarization and methylation ratio (β-value) calling was performed with the help of the RnBeads package [Assenov *et al.*, 2014]. We used *dasen* as the primary normalization method [Pidsley *et al.*, 2013]. We used several layers of filtering criteria to eliminate low-quality probes. We required each methylation call to be supported by at least 5 Infinium beads. Since too low and too high probe intensity may indicate measurement problems, we discarded CpGs where raw intensity at either of methylated and unmethylated probes was below 0.1 or above 0.9 quantiles of the total intensity distribution in the respective channel. To diminish the effects of genetic variation we also discarded CpGs with probes that overlapped with annotated SNP positions (dbSNP132 entries with MAF> 0.05, as defined in the RnBeads.hg19 annotation) along the whole probe sequence.

6.5 Availability of data and materials

The data sets supporting the conclusions of this article are available in the Gene Expression Omnibus repository under accessions: GSE35069 (PureBC), GSE42861 (WB1), GSE51032 (WB2), GSE15745 (PureN, ArtMixN and FC1) and GSE15745 (FC2). WGBS data of CD4+ T-cells are deposited in EGA as a part of the DEEP project submission (accession EGAS00001001624). WGBS profiles of naive and memory B-cells were downloaded in bedGraph format from the IHEC data portal (sample names S001JP51, C003K951, C003N351 and C0068L51).

6.6 Supplementary Material

Supplementary Notes

“Exact" and “approximate" models for heterogeneous DNA methylation profiles

Model definition Given the negligible measurement error, the methylation data at m CpG positions from multi-cellular samples of n individuals can be represented as:

$$\mathbf{Y} = \mathbf{CF} \tag{6.4}$$

Here, $\mathbf{C}$ is an $m \times q$ matrix of all single-cell DNA methylation profiles existing in all cell populations of all individuals, and $\mathbf{F}$ is an $m \times q$ matrix representing the frequency of given single cell profile in a measured sample.

Theoretical upper bound for q, given the three possible DNA methylation states in the cell $\{0, 0.5, 1\}$, is 3^m. However, the vast majority of potentially possible DNA methylation profiles are not biologically feasible, and $q \ll 3^m$. Nevertheless, in real applications q and m have comparable order of magnitude, and, quite certainly, $q \gg n$. The latter fact makes it impossible to find $\mathbf{C}$ and $\mathbf{Z}$ computationally.

Matrix $\mathbf{C}$ is expected to have a complicated correlation structure. Clustering of its columns should reveal multiple nested groups of single cell DNA methylation patterns, reflecting the

hierarchy of cell types and subtypes. One may than fix k as a number of discrete cell populations with highly similar DNA methylation profiles, depending on a certain similarity threshold. Given a set $\mathcal{C}_s$ of matrix $\mathbf{C}$ columns which represents the s-th such population, one could decrease the largest dimension in eq. (1), by constructing a summary profile $\mathbf{t}_s$ of such population: $\mathbf{t}_s = \frac{1}{n}\mathbf{C}_{:,\mathcal{C}_s}\mathbf{F}_{\mathcal{C}_s,:}\mathbf{W}^{-1}\mathbf{1}_n$, where $\mathbf{W}$ is diagonal matrix with $\mathrm{diag}(\mathbf{W}) = \mathbf{F}^T_{\mathcal{C}_s,:}\mathbf{1}_{|\mathcal{C}_s|}$, and then approximating the measured DNA methylation data as:

$$\mathbf{Y} \approx \mathbf{TA} \tag{6.5}$$

where $\mathbf{a_s} = \mathbf{1}^T_{|C_s|}\mathbf{F}_{\mathcal{C}_s,:}$. In the extreme case of all the columns in $\mathbf{C}_{:,\mathcal{C}_s}$ being identical profile $\mathbf{t}_s$ is obviously also identical to any of these and the approximate model (2) holds exactly. In fact this could be enforced by selecting a subset of rows $\mathcal{R}_s$, such that all columns of $C_{\mathcal{R}_s,\mathcal{C}_s}$ are identical (cell type-specific marker selection). Then the model in (2) augmented to the intersect of all such subsets $\mathcal{R}_s$, $s = 1 \ldots k$, would also hold exactly (see subsection below).

In case $k \leq n$, i.e. the selected number of discrete populations is comparable to the number of profiled individuals, both $\mathbf{T}$ and $\mathbf{A}$ can in theory be recovered computationally.

The approximation error One can estimate the error of this approximation. In case only one the s-th pattern set is substituted, the error is

$$\begin{aligned} err_s &= \left\|\mathbf{t}_s\mathbf{a}_s - \mathbf{C}_{:,C_s}\mathbf{F}_{C_s,:}\right\|_2 = \\ &= \left\|\frac{1}{n}\mathbf{C}_{:,\mathcal{C}_s}\mathbf{F}_{\mathcal{C}_s,:}\mathbf{W}^{-1}\mathbf{1}_n\mathbf{1}^T_{|C_s|}\mathbf{F} - \mathbf{C}_{:,C_s}\mathbf{F}_{C_s,:}\right\|_2 = \\ &= \left\|\mathbf{C}_{:,\mathcal{C}_s}\left(\frac{1}{n}\mathbf{F}_{\mathcal{C}_s,:}\mathbf{W}^{-1}\mathbf{1}_n\mathbf{1}^T_{|\mathcal{C}_s|} - \mathbf{I}_{|\mathcal{C}_s|}\right)\mathbf{F}_{C_s,:}\right\|_2 = \left\|\mathbf{C}_{:,\mathcal{C}_s}\Delta^F_{|\mathcal{C}_s|,:}\right\|_2 \end{aligned}$$

One can show that r, j-th element of matrix $\mathbf{\Delta}^F$ represents the deviation of the r-th pattern frequency from the expected frequency based on the average across all individuals:

$$\mathbf{\Delta}^F_{\mathcal{C}_s} = \left(\tfrac{1}{n}\mathbf{F}_{\mathcal{C}_s,:}\mathbf{W}^{-1}\mathbf{1}_n\mathbf{1}^T_{|\mathcal{C}_s|} - \mathbf{I}_{|\mathcal{C}_s|}\right)\mathbf{F}_{C_s,:} =$$

$$= \left(\begin{vmatrix} \frac{1}{n}\sum_{j'}\frac{f_{1,j'}}{\sum_{r'}f_{r',j'}} & \cdots & \frac{1}{n}\sum_{j'}\frac{f_{1,j'}}{\sum_{r'}f_{r',j'}} \\ \vdots & \ddots & \vdots \\ \frac{1}{n}\sum_{j'}\frac{f_{|\mathcal{C}_s|,j'}}{\sum_{r'}f_{r',j'}} & \cdots & \frac{1}{n}\sum_{j'}\frac{f_{|\mathcal{C}_s|,j'}}{\sum_{r'}f_{r',j'}} \end{vmatrix} - \mathbf{I}_{|\mathcal{C}_s|}\right)\mathbf{F}_{C_s,:} =$$

$$= \begin{vmatrix} \frac{1}{n}\sum_{j'}\frac{f_{1,j'}}{\sum_{r'}f_{r',j'}} - 1 & \cdots & \frac{1}{n}\sum_{j'}\frac{f_{1,j'}}{\sum_{r'}f_{r',j'}} & \cdots & \frac{1}{n}\sum_{j'}\frac{f_{1,j'}}{\sum_{r'}f_{r',j'}} \\ \vdots & & \vdots & & \vdots \\ \frac{1}{n}\sum_{j'}\frac{f_{r,j'}}{\sum_{r'}f_{r',j'}} & \cdots & \frac{1}{n}\sum_{j'}\frac{f_{r,j'}}{\sum_{r'}f_{r',j'}} - 1 & \cdots & \frac{1}{n}\sum_{j'}\frac{f_{1,j'}}{\sum_{r'}f_{r',j'}} \\ \vdots & & \vdots & & \vdots \\ \frac{1}{n}\sum_{j'}\frac{f_{|\mathcal{C}_s|,j'}}{\sum_{r'}f_{r',j'}} & \cdots & \frac{1}{n}\sum_{j'}\frac{f_{|\mathcal{C}_s|,j'}}{\sum_{r'}f_{r',j'}} & \cdots & \frac{1}{n}\sum_{j'}\frac{f_{|\mathcal{C}_s|,j'}}{\sum_{r'}f_{r',j'}} - 1 \end{vmatrix}\mathbf{F}_{C_s,:} =$$

$$= \begin{vmatrix} \frac{1}{n}\sum_{j'}\frac{f_{1,j'}}{\sum_{r'}f_{r',j'}}\cdot\sum_{r''}f_{r'',1} - f_{1,1} & \cdots & \frac{1}{n}\sum_{j'}\frac{f_{1,j'}}{\sum_{r'}f_{r',j'}}\cdot\sum_{r''}f_{r'',j} - f_{1,j} & \cdots & \frac{1}{n}\sum_{j'}\frac{f_{1,j'}}{\sum_{r'}f_{r',j'}}\cdot\sum_{r''}f_{r'',n} - f_{1,n} \\ \vdots & & \vdots & & \vdots \\ \frac{1}{n}\sum_{j'}\frac{f_{r,j'}}{\sum_{r'}f_{r',j'}}\cdot\sum_{r''}f_{r'',1} - f_{r,1} & \cdots & \frac{1}{n}\sum_{j'}\frac{f_{r,j'}}{\sum_{r'}f_{r',j'}}\cdot\sum_{r''}f_{r'',j} - f_{r,j} & \cdots & \frac{1}{n}\sum_{j'}\frac{f_{r,j'}}{\sum_{r'}f_{r',j'}}\cdot\sum_{r''}f_{r'',j} - f_{r,n} \\ \vdots & & \vdots & & \vdots \\ \frac{1}{n}\sum_{j'}\frac{f_{|\mathcal{C}_s|,j'}}{\sum_{r'}f_{r',j'}}\cdot\sum_{r''}f_{r'',1} - f_{|\mathcal{C}_s|,1} & \cdots & \frac{1}{n}\sum_{j'}\frac{f_{|\mathcal{C}_s|,j'}}{\sum_{r'}f_{r',j'}}\cdot\sum_{r''}f_{r'',n} - f_{|\mathcal{C}_s|,j} & \cdots & \frac{1}{n}\sum_{j'}\frac{f_{|\mathcal{C}_s|,j'}}{\sum_{r'}f_{r',j'}}\cdot\sum_{r''}f_{r'',n} - f_{|\mathcal{C}_s|,n} \end{vmatrix}$$

Here the term $\frac{1}{n}\sum_{j'}\frac{f_{r,j'}}{\sum_{r'}f_{r',j'}}\cdot\sum_{r}f_{r,j}$ represents the expected frequency of pattern r in the

individual j, based on the average in all individuals, and each element of $\mathbf{\Delta}_{\mathcal{C}_s}^F$ reflects the deviation of the actually observed frequency from this average value.

Including prior information

Model fitting as outlined in the Methods works satisfactorily for the data sets under consideration herein. In general, performance can be further improved by incorporating prior knowledge about T and/or A if available. Specifically, we have the following two forms of such prior knowledge in mind.

- One or more columns of T may be known in advance given reference profiles obtained from methylation measurements on isolated cell types.
- It is common to have additional knowledge about the cellular composition of the samples the methylation measurements in D are based on. For example, the composition of blood cells is well-studied, and the relative abundance of the underlying major cell types can be narrowed down to intervals.

Both scenarios can be taken advantage of by straightforward modifications of Algorithm 1.

- In case some of the columns of T are known, without loss generality, we may partition $T = [T_0\ \widetilde{T}]$, where T_0 denotes the sub-matrix assumed to be known while $\widetilde{T}$ still needs to be determined. Accordingly, in Algorithm 1 the update of T is confined to $\widetilde{T}$, whereas T_0 is kept fixed over all iterations.
- Lower and upper bounds on the proportions of specific cell types directly translate into identical bounds on the entries of entire rows of A. To give an example, suppose it is known that the proportion of cell type X ranges between 30% and 45% and that of a second cell type Y between 10% and 20%. Unless columns of T are known in advance (see above), problem (6.3) is invariant to the ordering of the rows of A and we may add the following constraints:
$$0.3 \leq A_{1j} \leq 0.45, \qquad 0.1 \leq A_{2j} \leq 0.2, \quad j = 1, \ldots, n.$$
If a subset of the columns of T is known, one proceeds accordingly: the row indices in A associated with the bound constraints either have to match those corresponding to T_0 or can be chosen freely among the rest, depending on whether prior knowledge about cell type proportions concerns cell types with known respectively unknown methylation profiles.

None of the above modifications make the optimization problem (6.3) more difficult. Additional bound constraints on the rows of A come into play for optimization problem **optA** in Algorithm 1 and can be handled in a straightforward manner by most solvers of convex quadratic programs. Partial knowledge of T makes optimization even easier.

Supplementary Figures

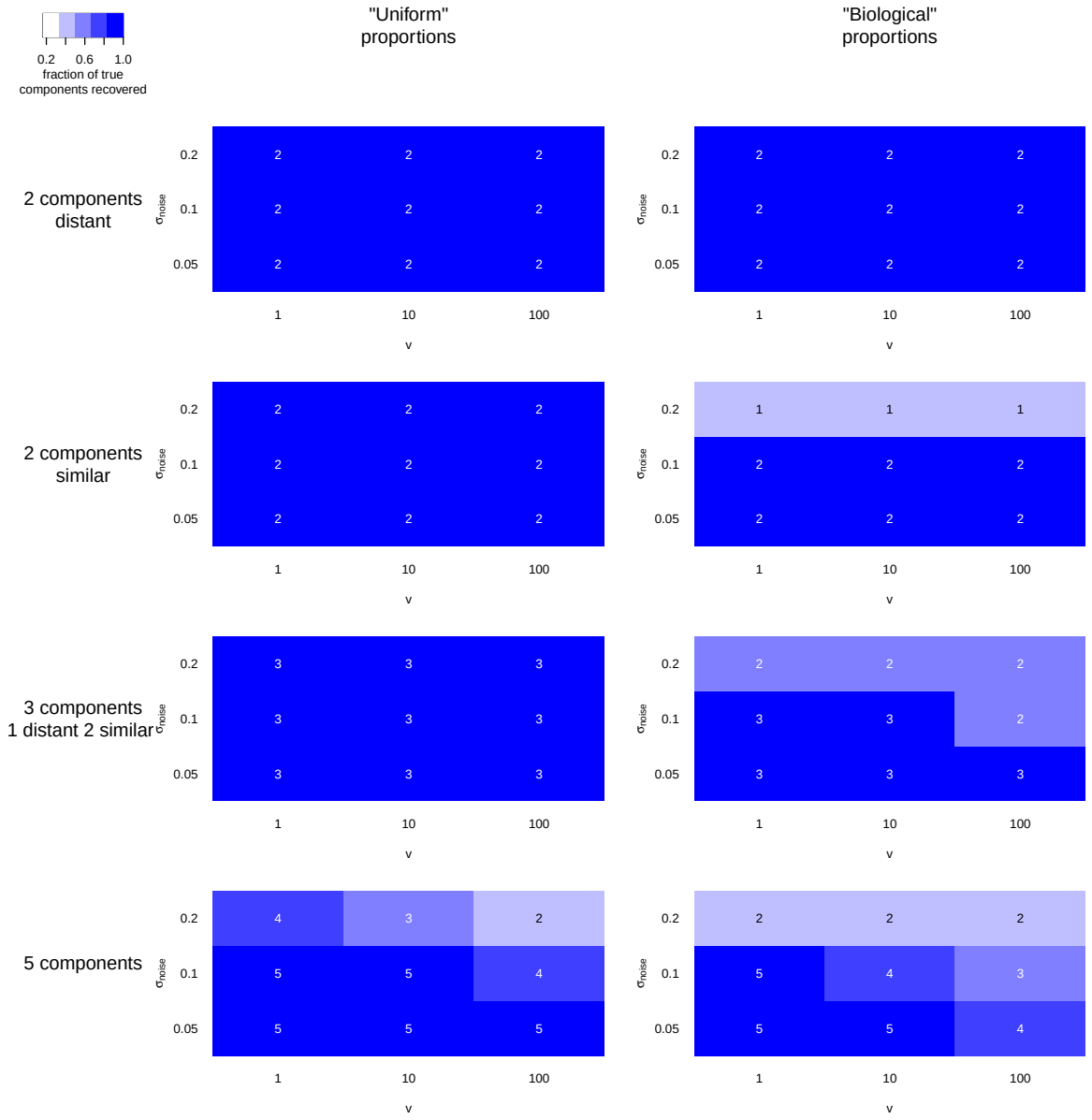


Figure 6.S1: Efficiency of component recovery in all simulated data sets. v is the scaling parameter used in proportion sampling, while σ_{noise} is the standard deviation of the additive Gaussian noise. The values show the maximum number of mutually matching $\hat{T}$ and $T^{\star}$ columns in the simulation analyses achieved for any of the tested k and λ values. The color code shows the efficiency metric on $[0, 1]$, obtained by dividing the above number by k_{sim}.

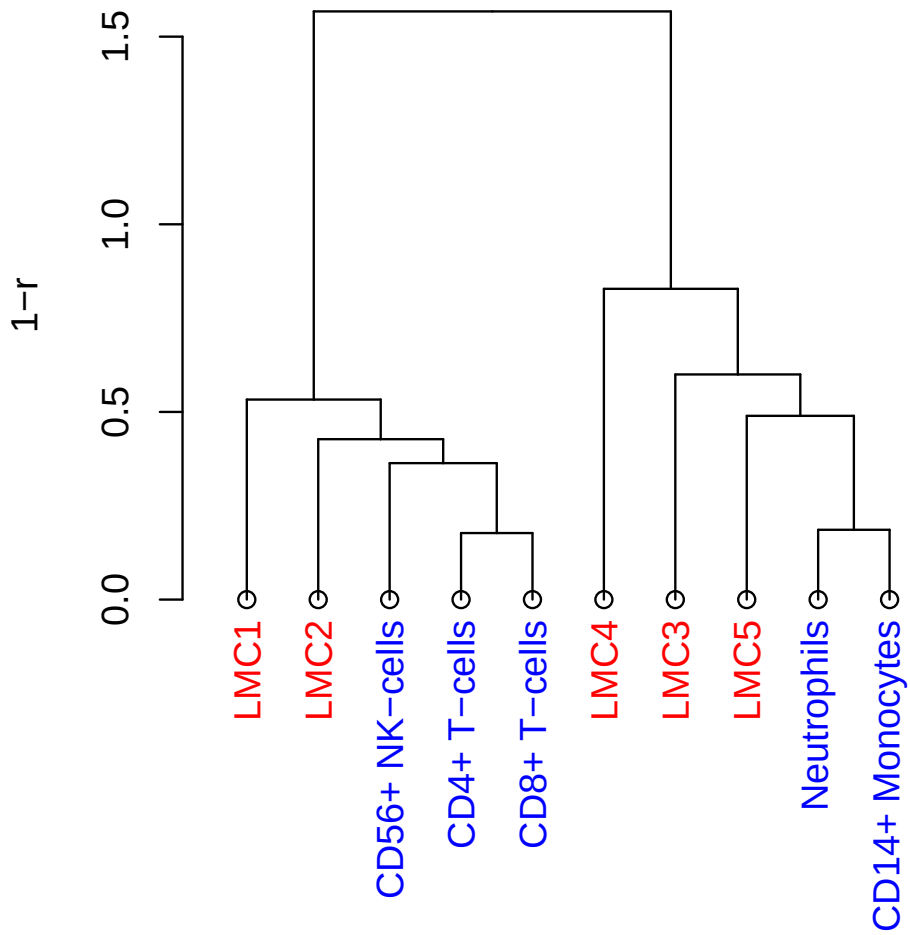


Figure 6.S2: LMC recovery in a hard simulated test case with $k_{\text{sim}} = 5$, "biological" proportions with low variability ($v = 100$) and medium noise ($\sigma_{\text{noise}} = 0.1$), $k = 5$, $\lambda = 0.01$.

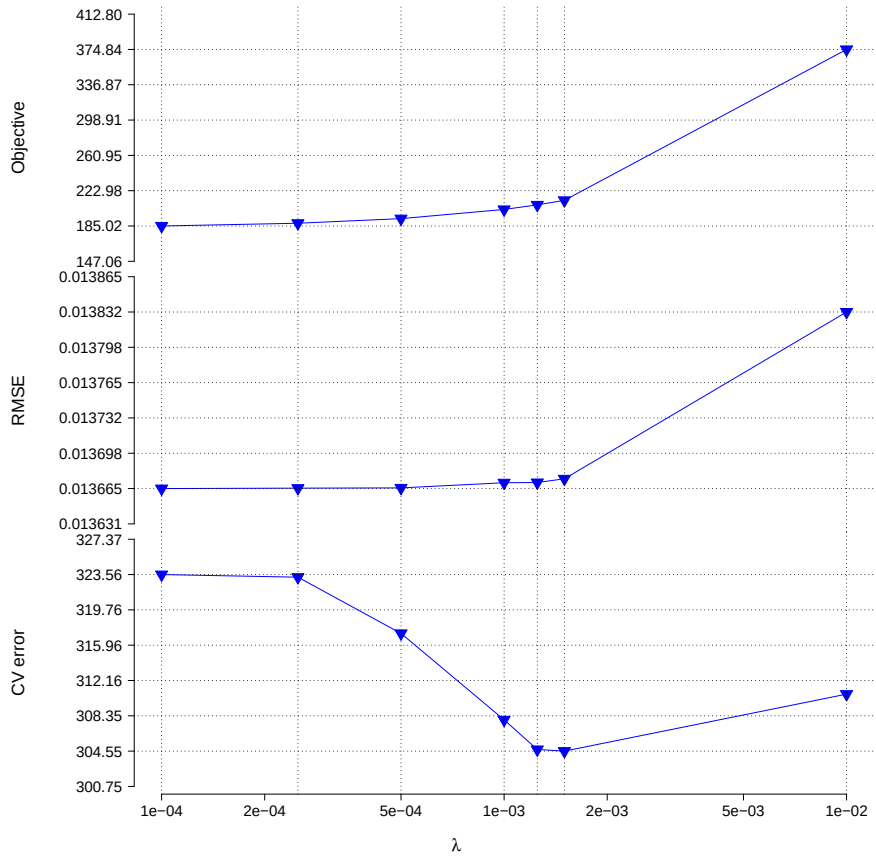


Figure 6.S3: λ selection for the ArtMixN data set ($k = 2$)

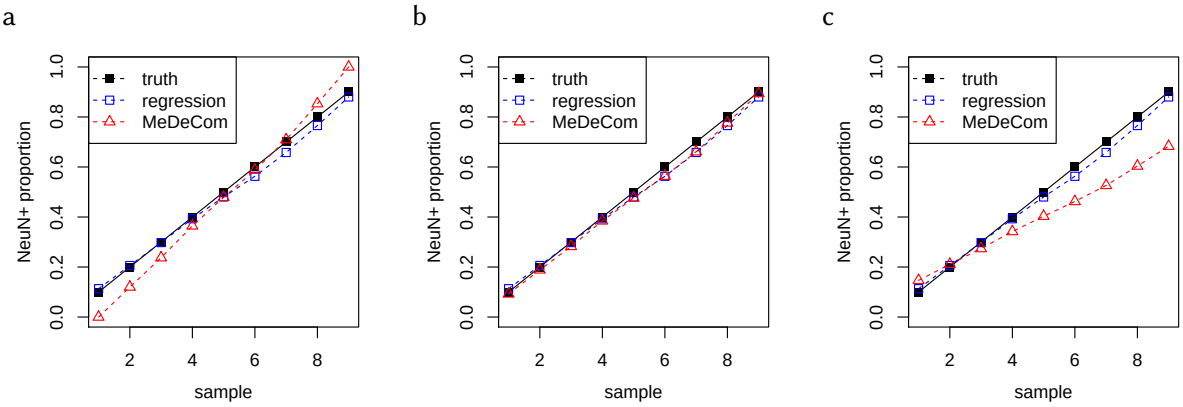


Figure 6.S4: Regularization effects upon proportion recovery in ArtMixN data. **a**. $\lambda = 10^{-4}$. **b**. $\lambda = 10^{-3}$. **c**. $\lambda = 10^{-2}$.

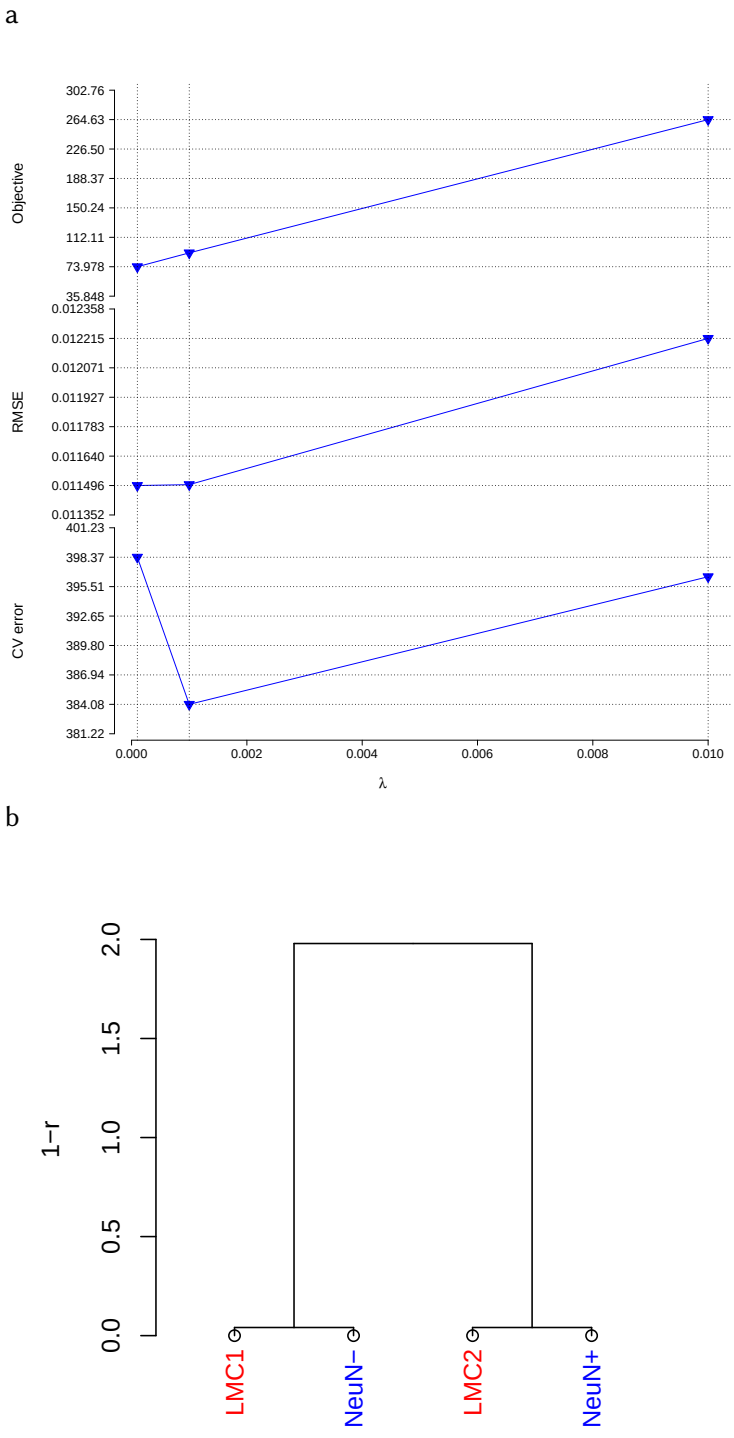


Figure 6.S5: Subset of ArtMixN data with NeuN$^+$ proportion ≥ 0.5. **a**. λ selection ($k = 2$). **b**. LMC matching ($\lambda = 0.001$).

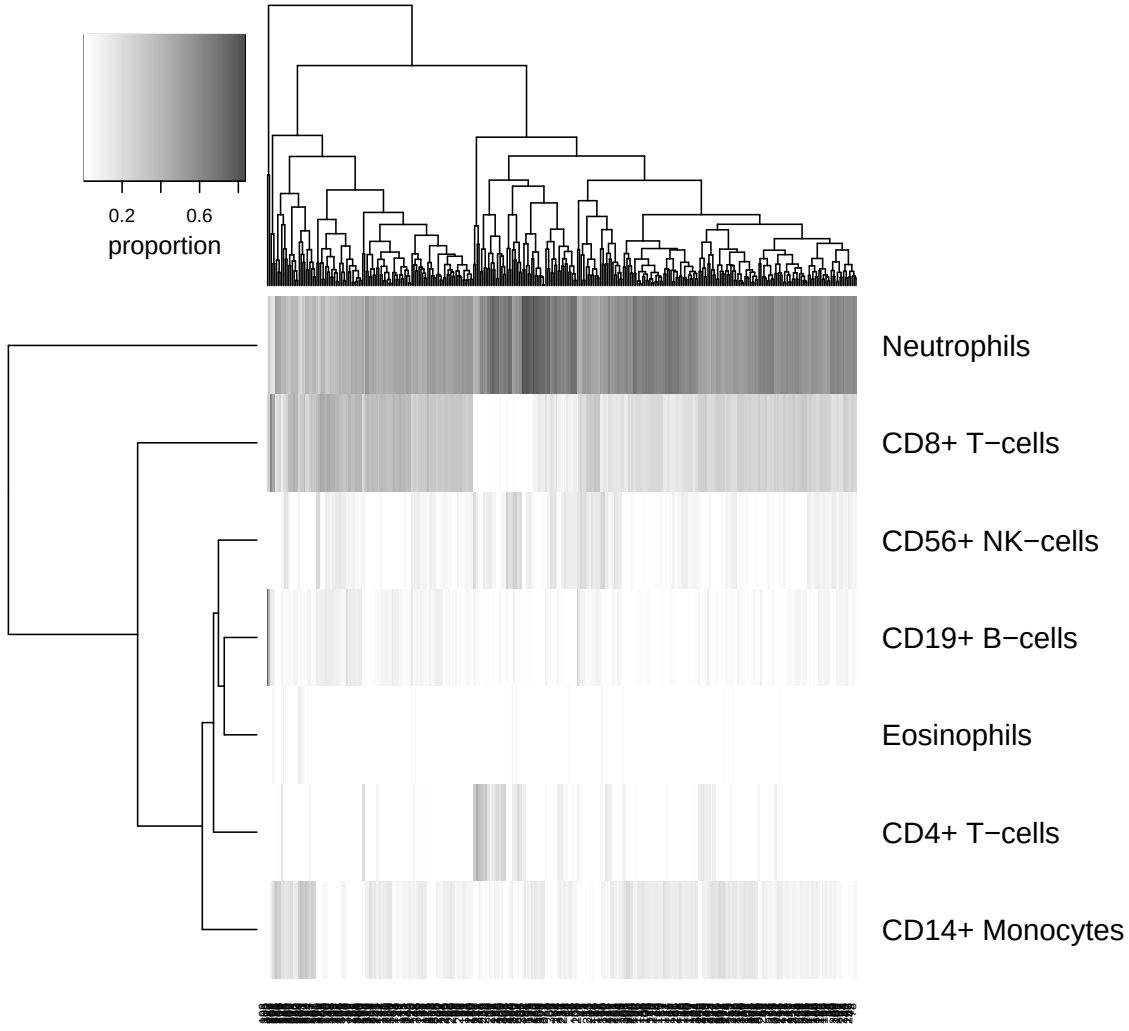


Figure 6.S6: Regression estimated proportions of reference cell types in the control samples of the complete Liu *et al.* data set.

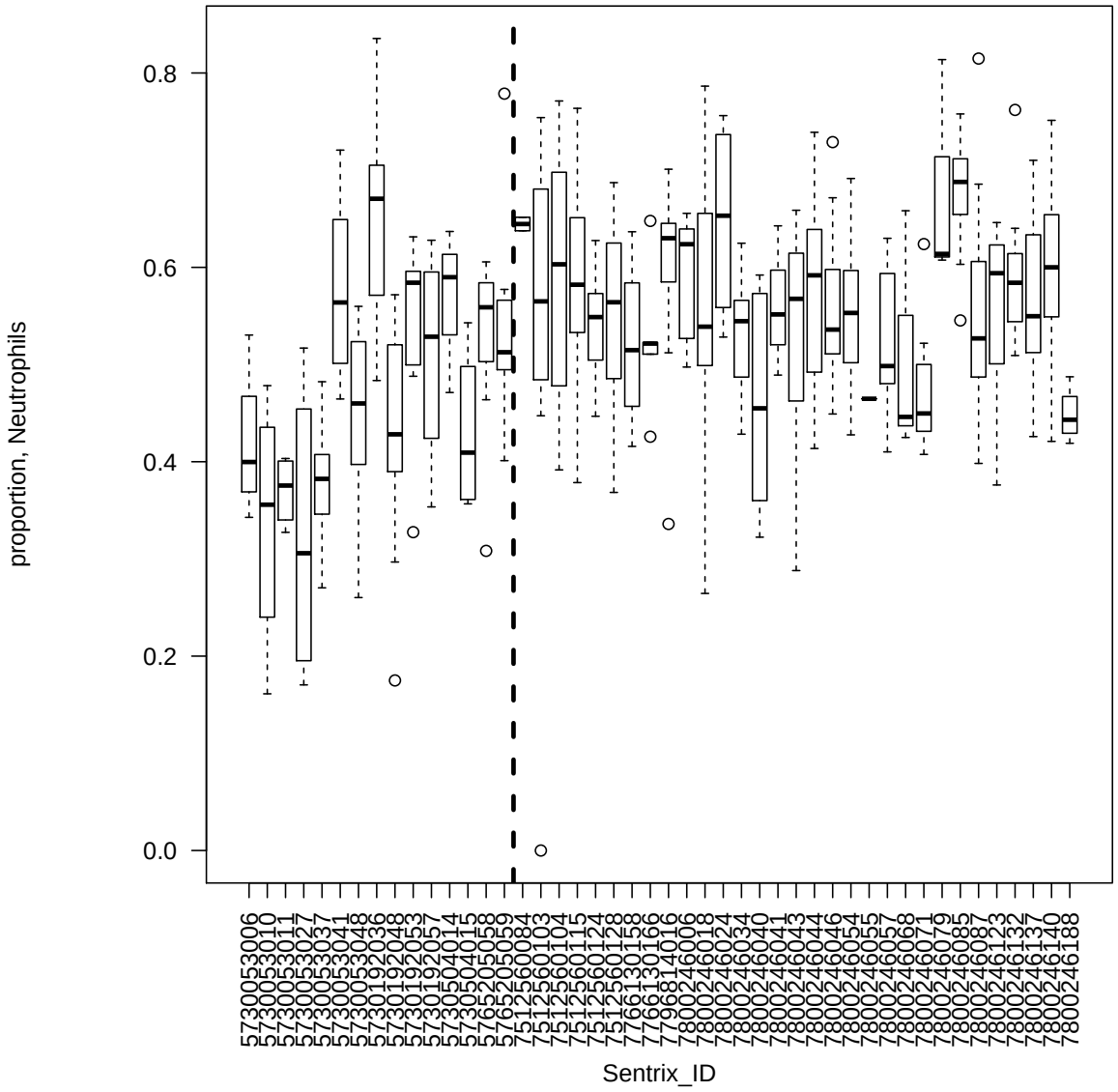


Figure 6.S7: Estimated Neutrophil proportions in the complete Liu *et al.* data set, stratified by the 450k microarray plate (Sentrix_ID). One can easily notice a strong batch effect. The subsequent analysis was performed on a smaller technical batch of 87 samples with Sentrix_ID< 7512560084 (to the left from the dashed vertical line) comprising the WB1 data set.

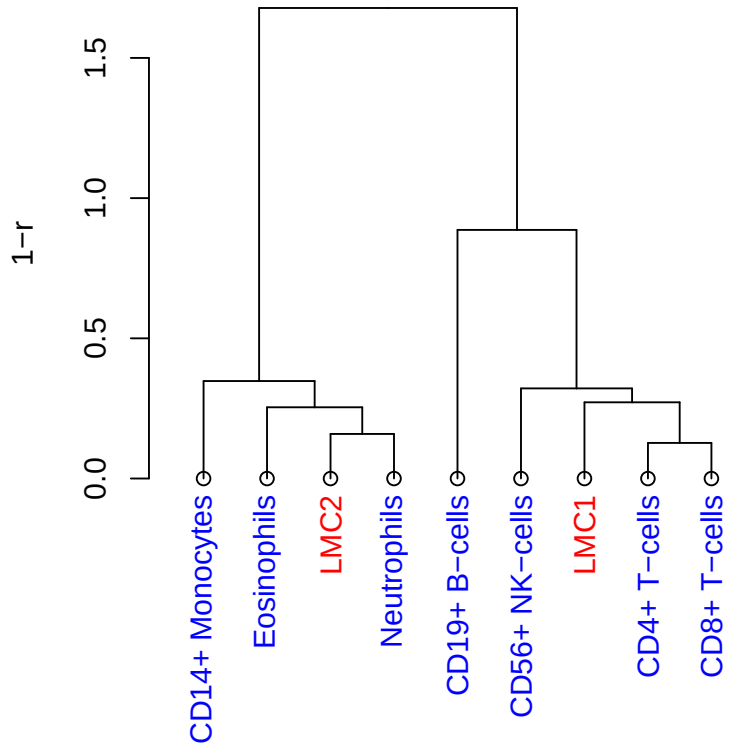


Figure 6.S8: WB1 data set, matching of LMCs recovered with $k = 2$ and $\lambda = 0.01$

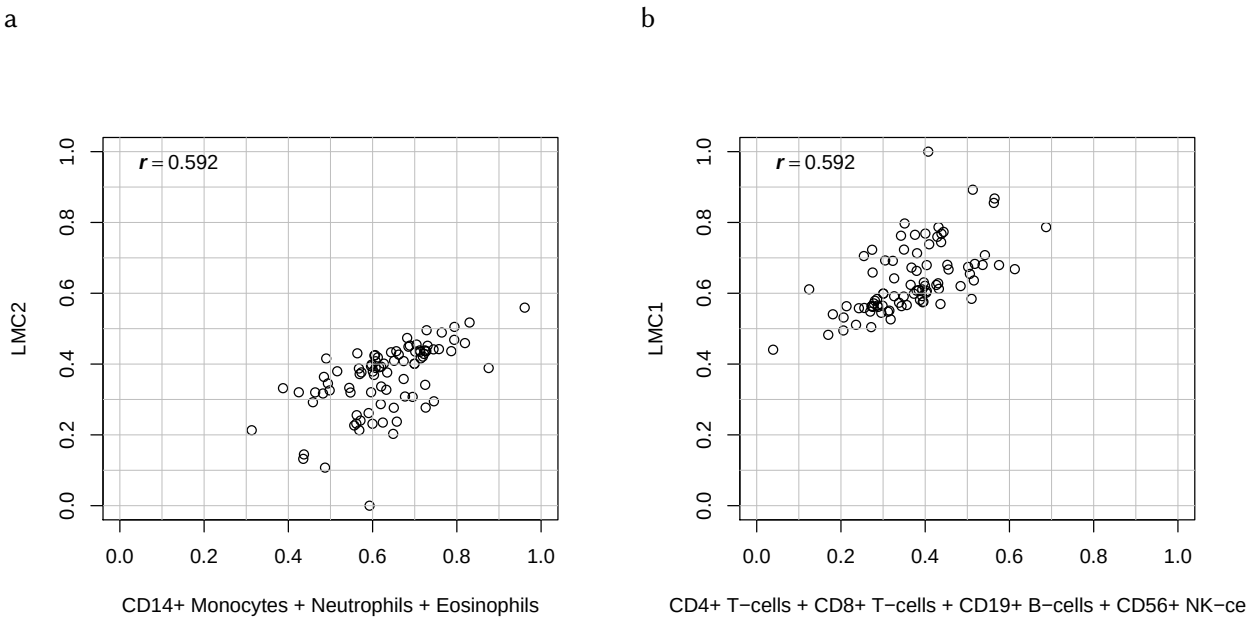


Figure 6.S9: WB1 data set, comparison of the reference-based proportions of myeloid and lymphoid cell types and LMC proportions for the case with $k = 2$ and $\lambda = 0.01$. **a**. Myeloid cell types, including Neutrophils, Eosinophils and Monocytes. **b**. Lymphoid cell types, including T-cells, B-cells and NK-cells.

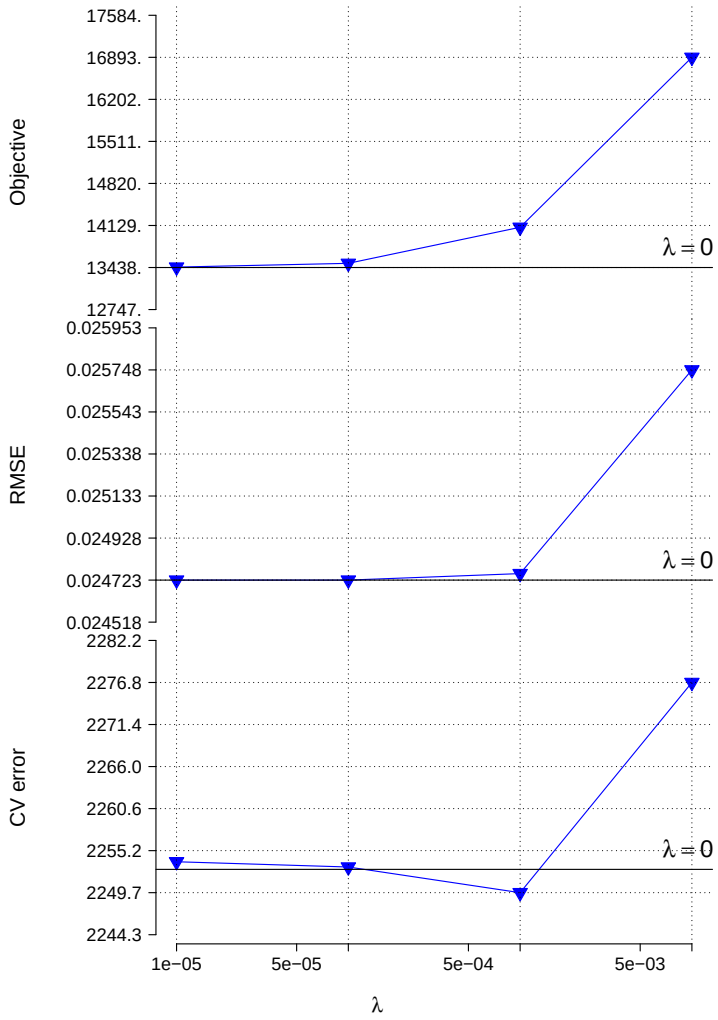


Figure 6.S10: WB1 data set, λ selection ($k = 20$).

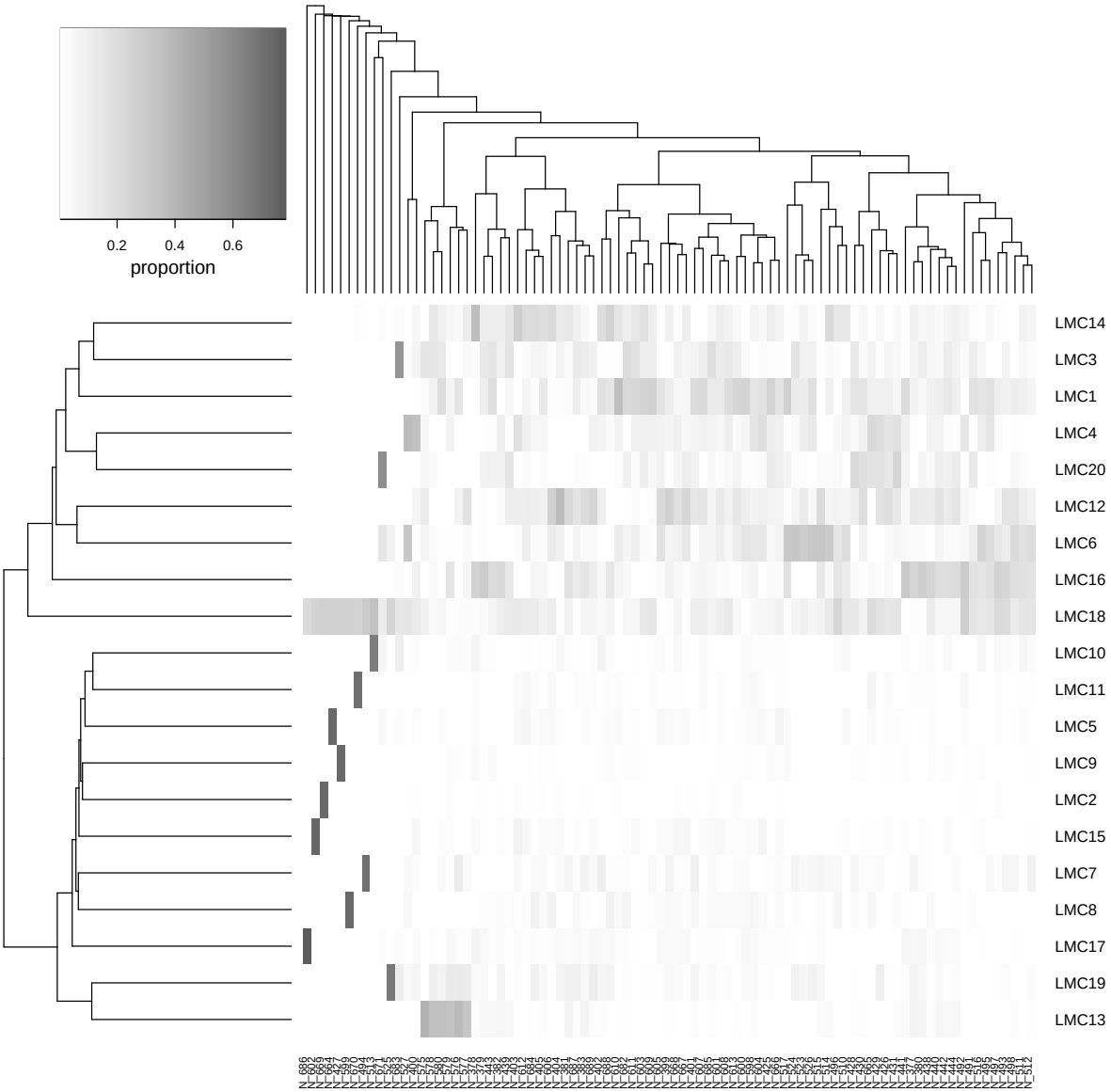


Figure 6.S11: WB1 data set ($k = 20$, $\lambda = 0.001$), heat map of the recovered mixing proportions. Rows and columns were clustered to improve readability (hierarchical clustering with euclidean distance and average linkage).

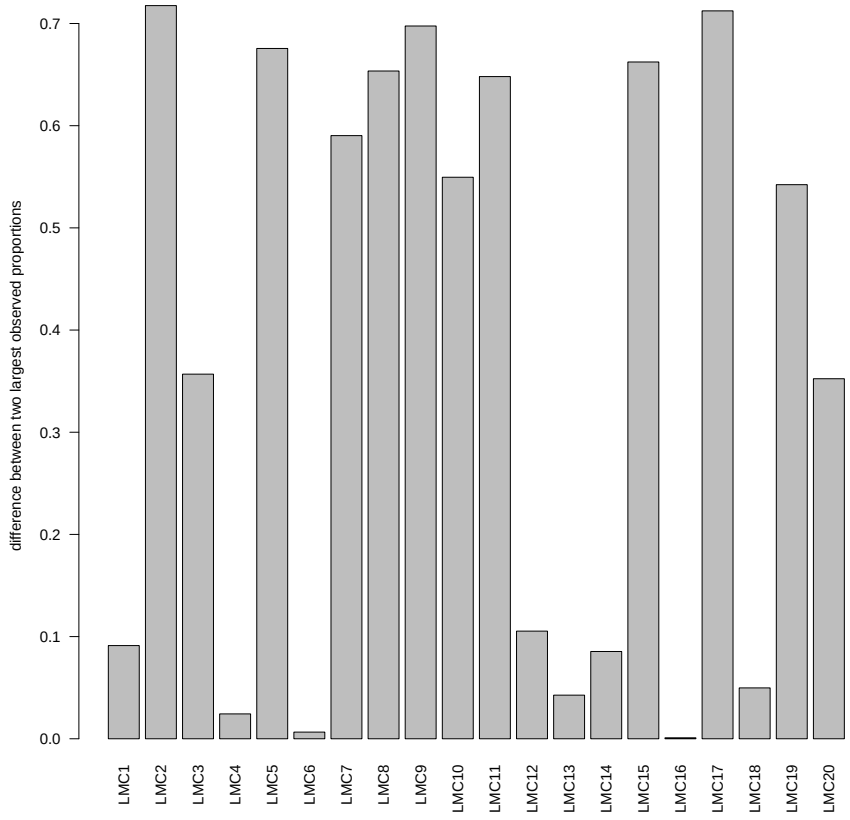


Figure 6.S12: Individual-specific LMCs in the WB1 data set ($k = 20$, $\lambda = 0.001$). For each LMC a difference between the largest and the second largest observed proportion was calculated.

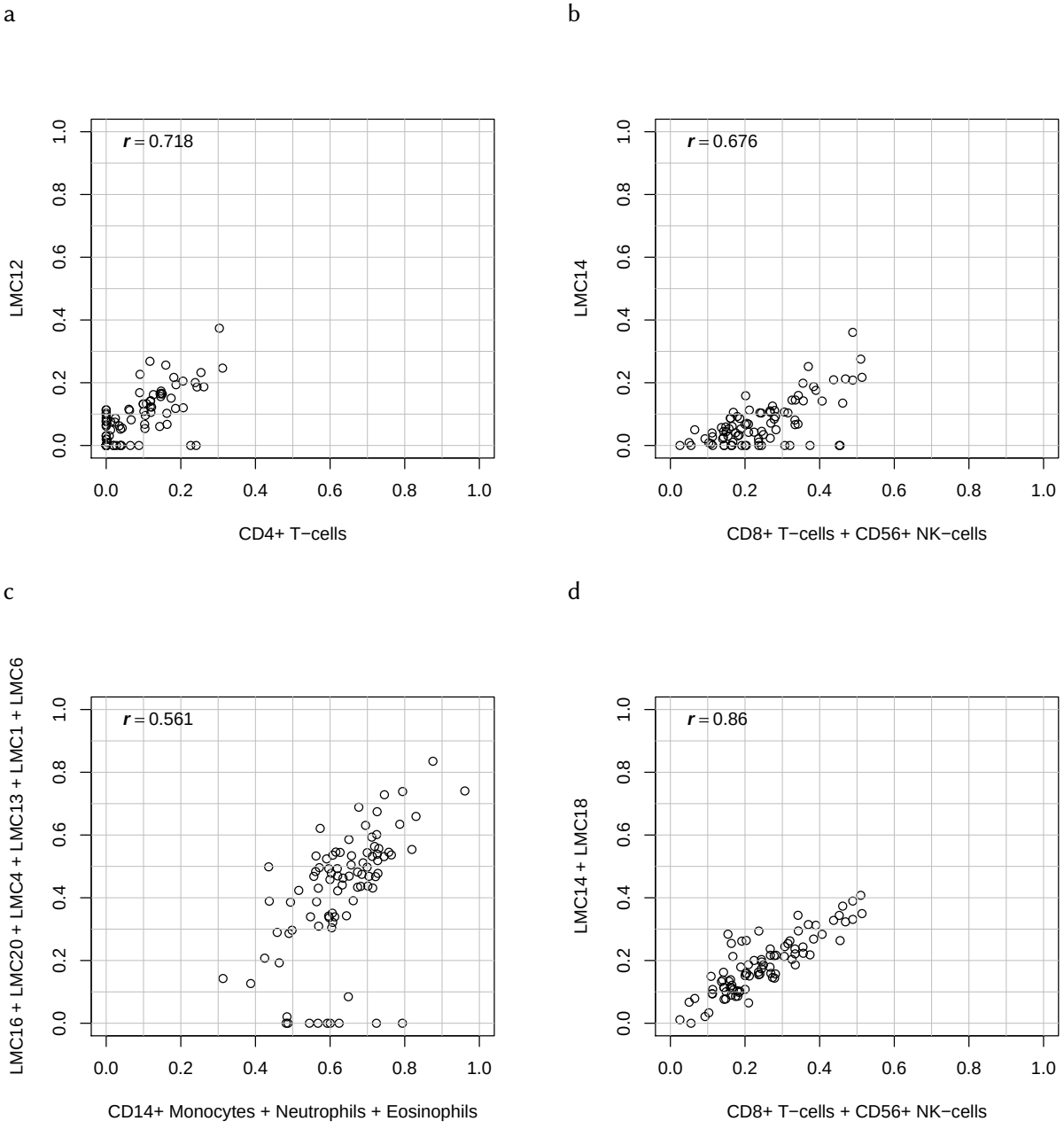


Figure 6.S13: WB1 data set, proportion recovery. **a**. CD4+ T-cells. **b**. CD8+ T-cells and NK-cells. **c**. Myeloid cell types. Here individual-specific LMCs (see Supplementary Figure 6.S11 and 6.S12) matching the “myeloid cluster" were excluded. **d**. CD8+ T-cells and NK-cells, with LMC18 included.

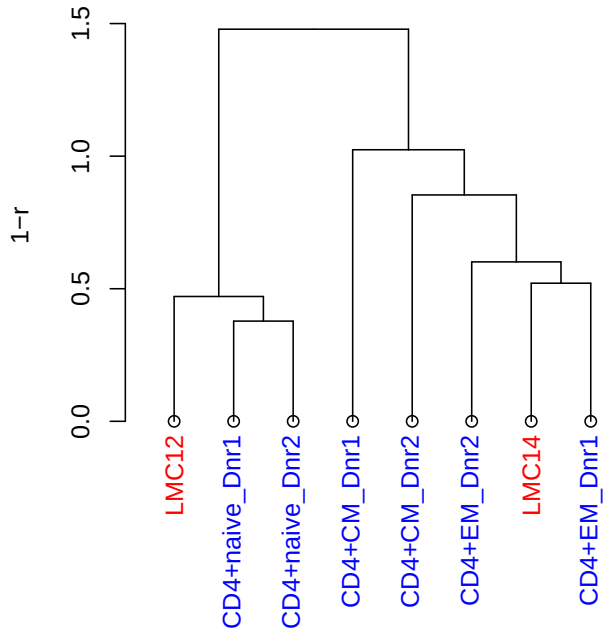


Figure 6.S14: Matching the T-cell-specific LMCs from WB1 data set to reference WGBS-based CD4+ T-cell profiles. CD4+naive – CD4+ naive T-cells; CD4+CM – CD4+ central memory T-cells; CD4+EM – effector memory T-cells.

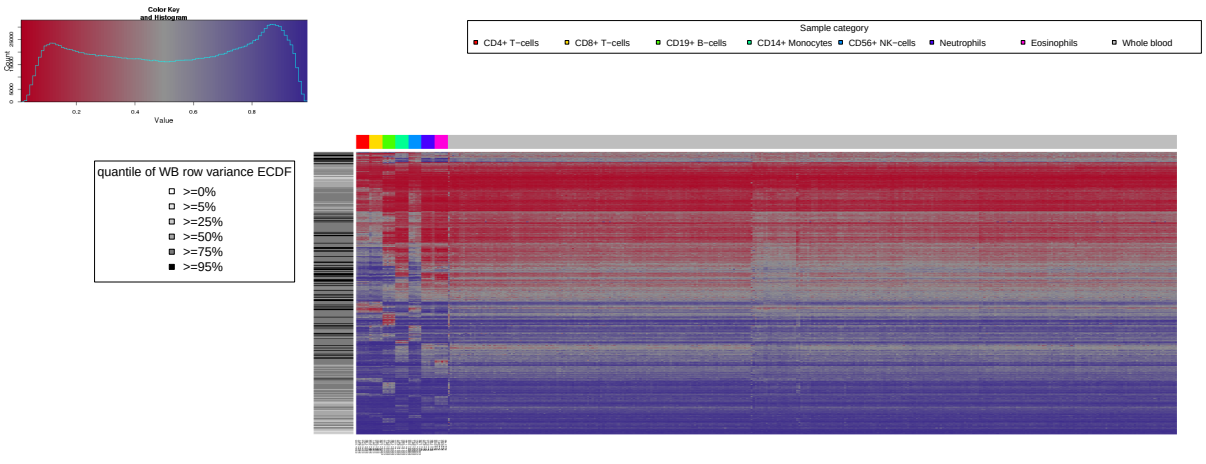


Figure 6.S15: Preprocessed Infinium 450k methylation calls in PureBC and WB1 data at 15,000 CpGs with highest cell type specificity. The rows are ordered based on hierarchical clustering in the purified data only. The whole blood columns are ordered based on hierarchical clustering in the whole blood data only. The row color code reflects the quantile of empirical CpG-wise variance distribution in the whole blood data into which the variance of the corresponding CpG falls.

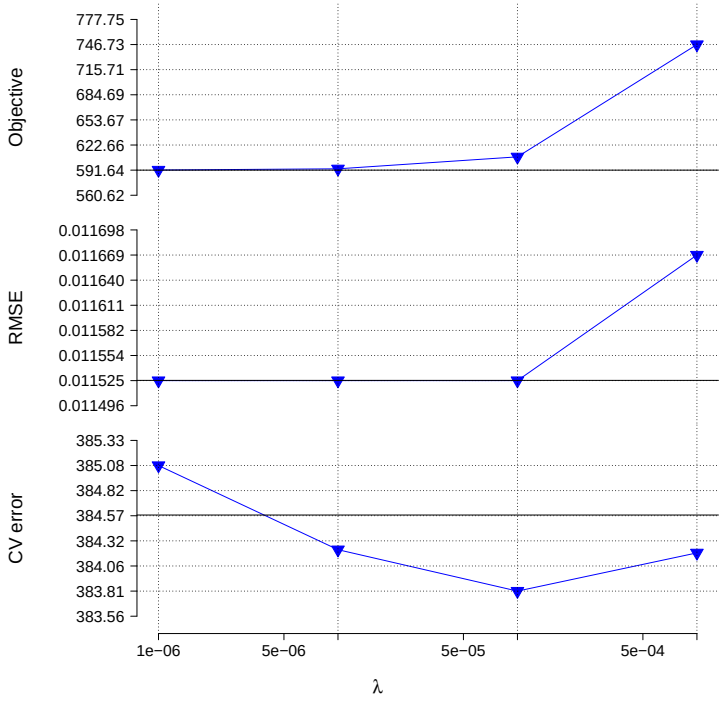


Figure 6.S16: λ selection for the PureBC data set ($k = 16$)

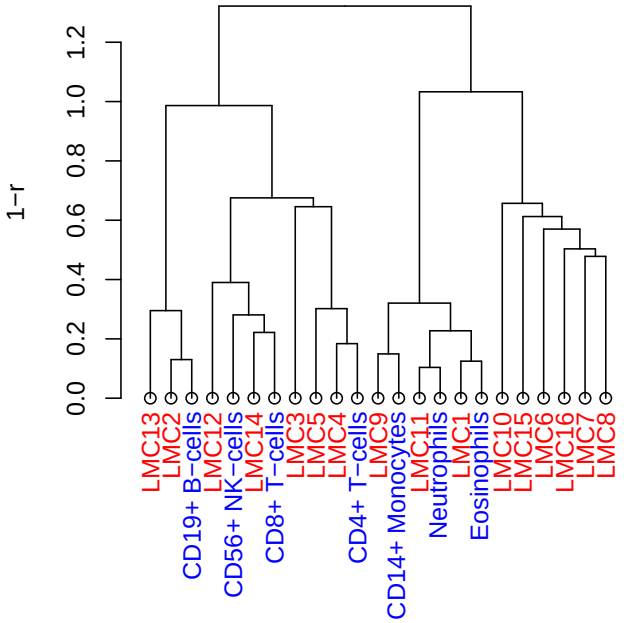


Figure 6.S17: Matching of the LMCs from the PureBC data to average cell type profiles ($k = 16$, $\lambda = 10^{-4}$).

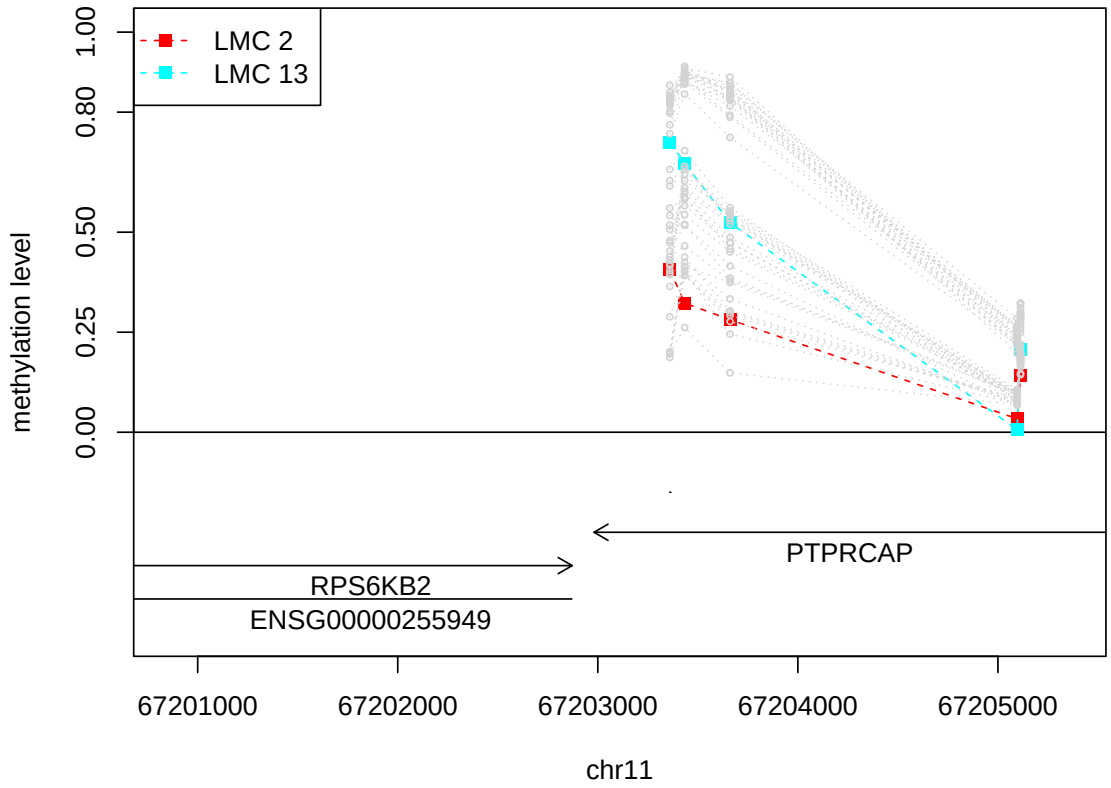


Figure 6.S18: Purified blood cells: methylation level of the *PTPRCAP* locus in different purified blood cells (grey dotted lines), LMC2 and LMC13.

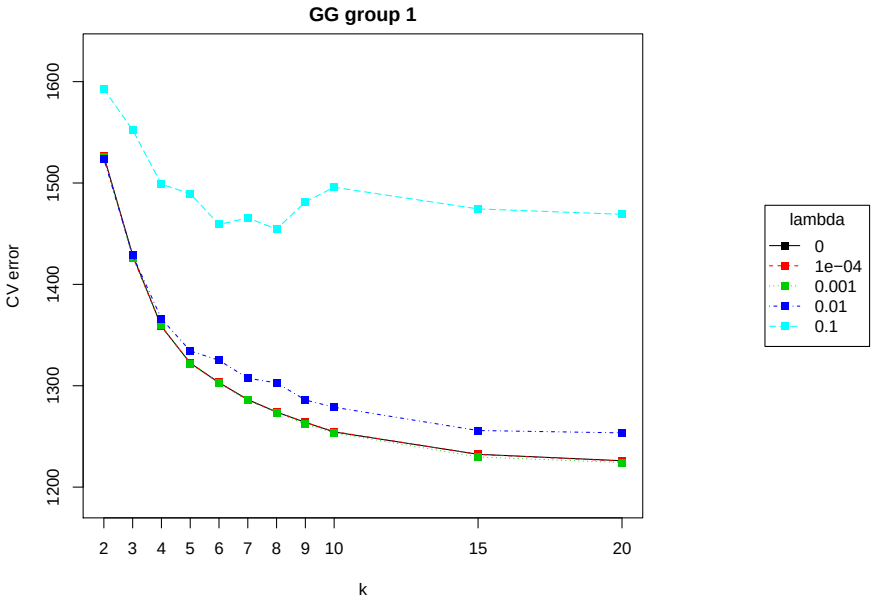
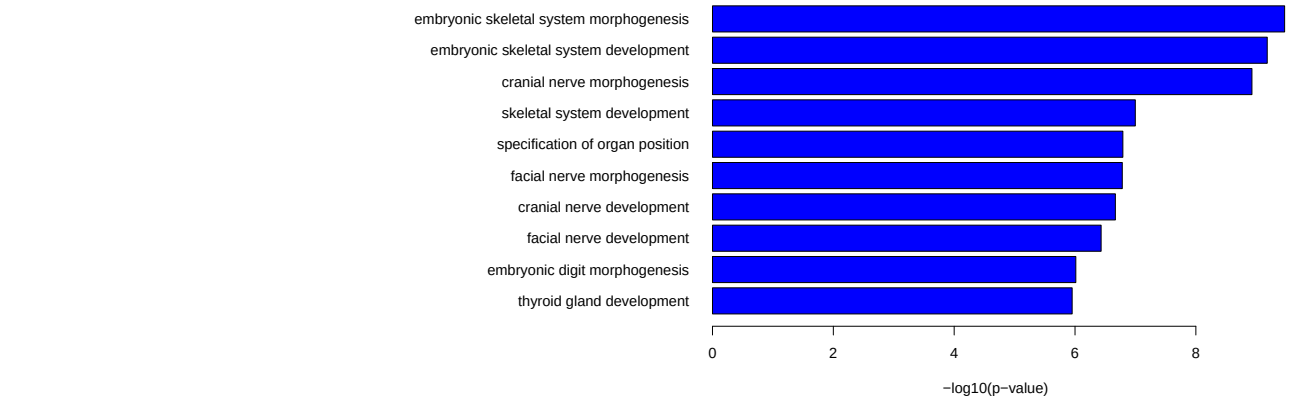
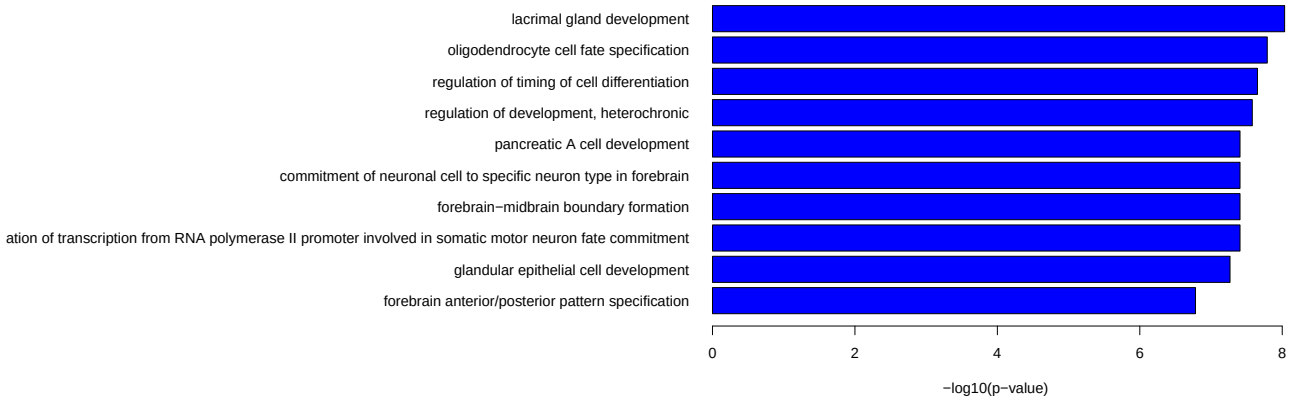


Figure 6.S19: FC2 data set, parameter selection

a



b



c

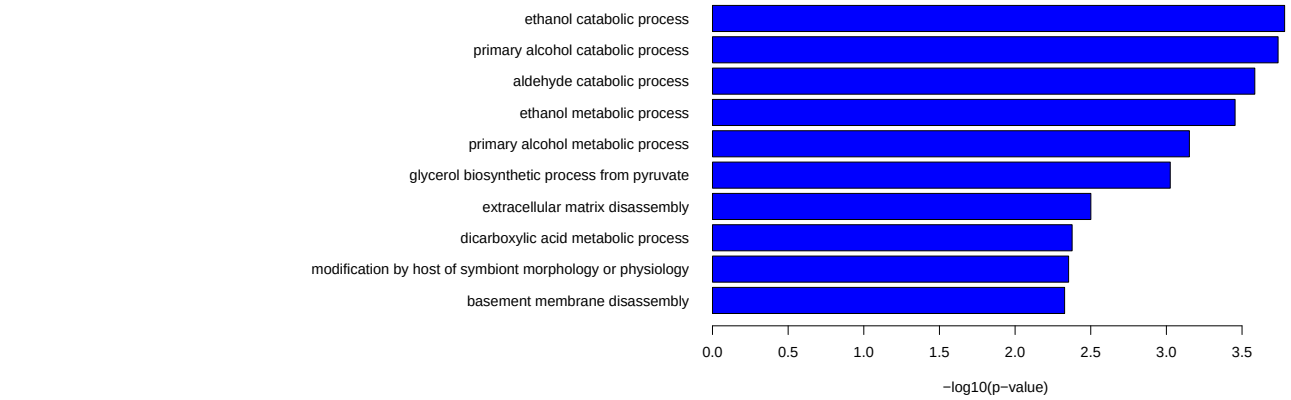
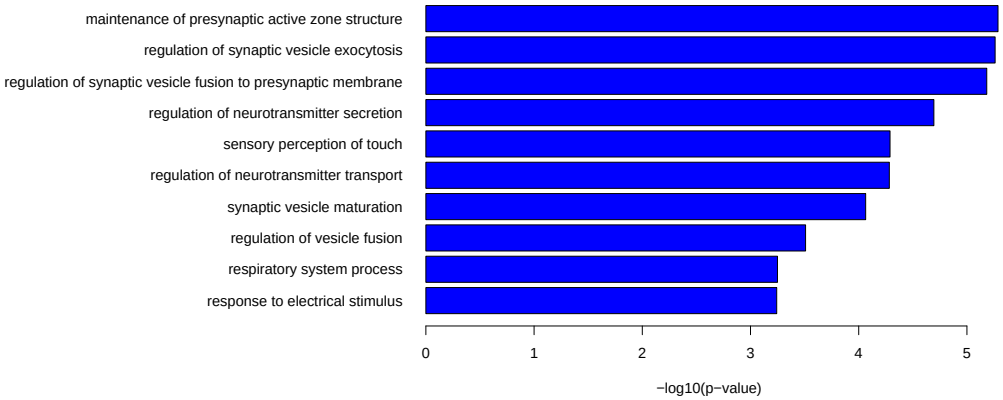
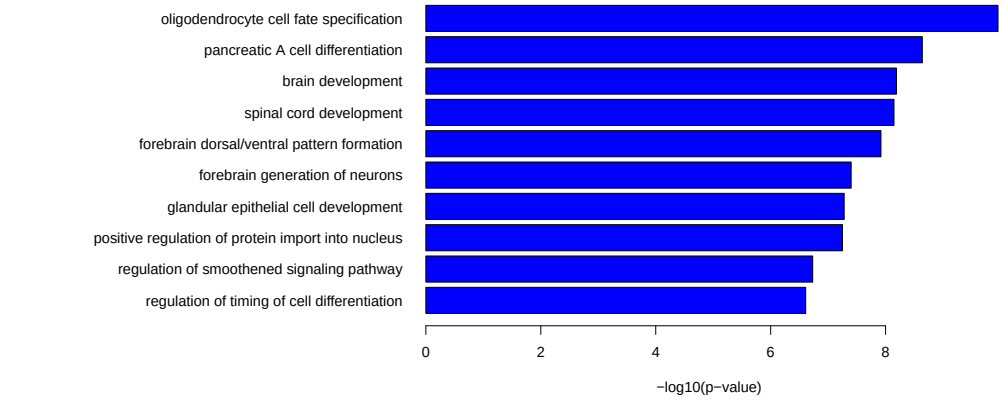


Figure 6.S20: Functional annotation of frontal cortex LMCs. Raw (unadjusted) p-values of the GREAT binomial test are reported. **a**. LMC1, hypermethylated CpGs **b**. LMC1, hypomethylated CpGs. **c**. LMC2, hypermethylated CpGs.

d



e



f

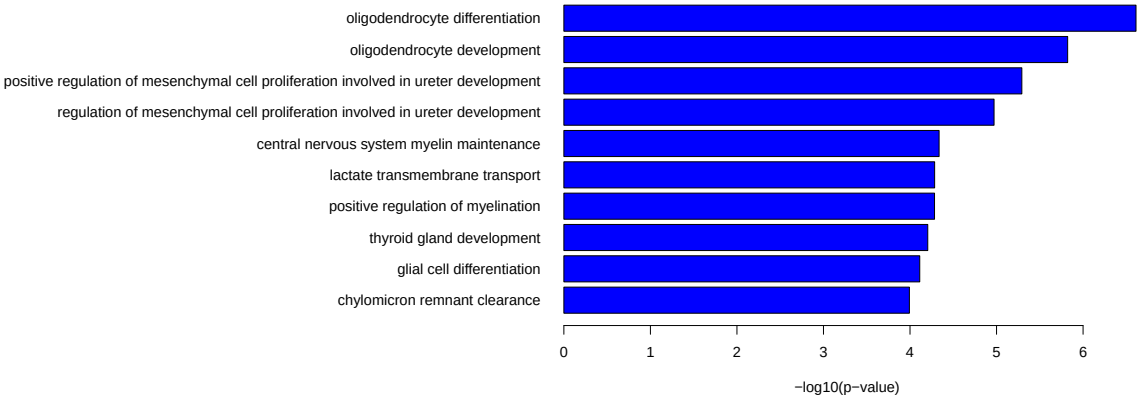


Figure 6.S20: (continued) Functional annotation of frontal cortex LMCs. **d**. LMC2, hypomethylated CpGs. **e**. LMC3, hypermethylated CpGs. **f**. LMC3, hypomethylated CpGs.

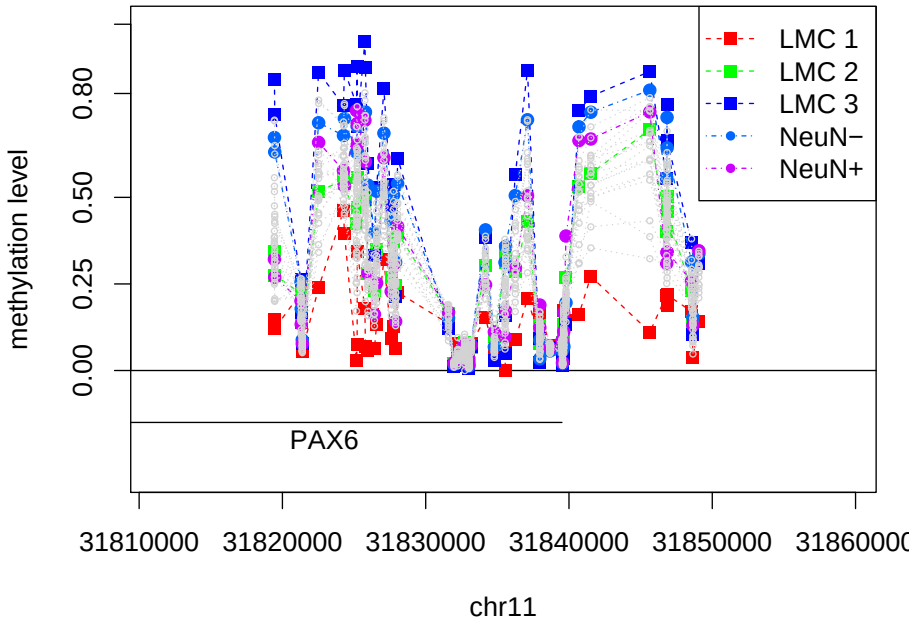


Figure 6.S21: FC1 data set ($k = 3$, $\lambda = 0.003$), example of an LMC1-specific locus *PAX6*. Grey dotted lines correspond to the original frontal cortex profiles.

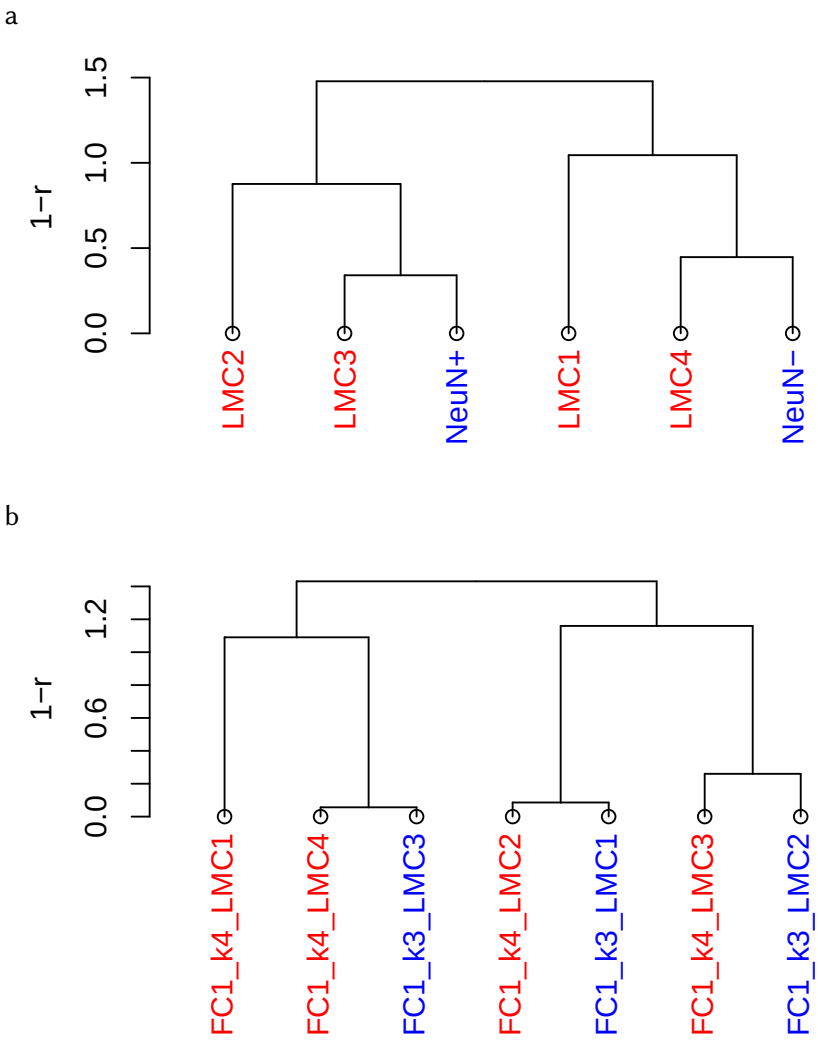


Figure 6.S22: FC1 data set, MeDeCom solution with $k = 4$ and $\lambda = 0.005$ used for the estimation of mixing proportions. **a**. Matching LMCs to the PureN reference. **b**. Matching LMCs to the LMCs of the $k = 3$ case.

Supplementary Tables

Table 6.S1: Parameters for the simulation runs

Methylation component schemes				
Name	2CompDistant	2CompSimilar	3comp1Distant2Similar	5Comp
k_{sim}	2	2	3	5
Cell types	Neutrophils CD4+ T-cells	Neutrophils Monocytes	Neutrophils CD4+ T-cells Monocytes	Neutrophils CD4+ T-cells Monocytes CD8+ T-cells NK-cells

Proportion model ($k_{sim} = 5$)		
Name	"Uniform"	"Biological"
$\alpha_{Neutrophils}$	0.20	0.62
$\alpha_{Monocytes}$	0.20	0.23
$\alpha_{CD4+T-cells}$	0.20	0.05
$\alpha_{CD8+T-cells}$	0.20	0.07
$\alpha_{NK+T-cells}$	0.20	0.03

Proportion variability			
Name	Low	Moderate	High
v	1	10	100

Noise levels			
Name	Low	Medium	High
$\sigma_{\mathbf{noise}}$	0.05	0.1	0.2

Table 6.S2: NeuN+ and NeuN- fraction proportions in the ArtMixN data set

Sample	Propotion	
	NeuN-	NeuN+
121_Mix1(P2E1)_7786915074_R01C01	0.90	0.10
122_Mix2(P2E2)_7786915074_R02C01	0.80	0.20
123_Mix3(P2E3)_7786915074_R03C01	0.70	0.30
124_Mix4(P2E4)_7786915074_R04C01	0.60	0.40
125_Mix5(P2E5)_7786915074_R05C01	0.50	0.50
126_Mix6(P2E6)_7786915074_R06C01	0.40	0.60
127_Mix7(P2E7)_7786915074_R01C02	0.30	0.70
128_Mix8(P2E8)_7786915074_R02C02	0.20	0.80
129_Mix9(P2E9)_7786915074_R03C02	0.10	0.90

Table 6.S3: FC1 data set ($k = 3$, $\lambda = 3 \cdot 10^{-3}$): overlap of the LMC1-specific genes with Moe *et al.* hypo-DMRs in telencephalon development markers from Vicel *et al.*

TF	# CTs	LMC1 status
Ascl1	3	-
Dbx1	3	-
Ebf1	3	Hypermethylated
Ebf3	3	Hypermethylated
Egr3	3	-
Emx1	3	-
Emx2	3	-
Esrrg	3	-
Fezf2	3	Hypomethylated
Foxp2	3	-
Foxp4	3	-
Gbx2	3	Hypermethylated
Gli3	3	Hypermethylated
Gsx1	3	-
Gsx2	3	-
Hes1	3	-
Hes5	3	-
Id2	3	-
Id4	3	-
Isl1	3	-
Lef1	3	-
Lhx2	3	-
Lhx5	3	-
Lhx8	3	-
Lhx9	3	-
Mafb	3	-
Meis1	3	Hypomethylated
Meis2	3	-
Neurod1	3	-
Neurog1	3	Hypomethylated
Neurog2	3	-
Nkx6-2	3	-
Nr2e1	3	Hypomethylated
Nr2f1	3	Hypomethylated
Nr2f2	3	-
Olig1	3	-
Olig2	3	Hypomethylated
Otx1	3	Hypomethylated
Pax6	3	Hypomethylated
Pou3f2	3	Hypomethylated
Pou3f3	3	Hypermethylated
Rara	3	-
Sall3	3	-
Six3	3	-
Sox1	3	Hypomethylated
Sox4	3	-
Sp8	3	Hypomethylated
Sp9	3	-
Tle1	3	-
Tle3	3	-
Tle4	3	-
Tshz1	3	-
Vax1	3	-
Zfhx4	3	-
Zfp521	3	-
Zic1	3	Hypomethylated
Bcl11b	2	-
Dlx1	2	Hypomethylated
Dlx2	2	-
Dlx5	2	-
Dlx6	2	-
Eomes	2	-
Fezf1	2	-
Foxg1	2	-
Hmx3	2	-
Lhx6	2	-
Otx2	2	-
Prox1	2	-
Sox11	2	-
Tbr1	2	-
Zic5	2	Hypomethylated
Bcl11a	1	Hypermethylated
Nkx2-1	1	-
Pbx1	1	-
Zfp503	1	-

Additional Files

Additional file 1— CpGs used for the analysis of memory and naive B-cells

A comma-separated value table file.

Additional file 2 — LMC-specific CpG positions of the FC1 data set

A comma-separated value table file.

References

Accomando, W. P., Wiencke, J. K., Houseman, E. A., Nelson, H. H., and Kelsey, K. T. Quantitative reconstruction of leukocyte subsets using DNA methylation. *Genome Biology*, 15(3):R50, 2014.

Adalsteinsson, B. T., Gudnason, H., Aspelund, T., Harris, T. B., Launer, L. J., Eiriksdottir, G. *et al.* Heterogeneity in white blood cells has potential to confound DNA methylation measurements. *PloS one*, 7(10):e46705, 2012.

Aryee, M. J., Jaffe, A. E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A. P., Hansen, K. D. *et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*, 2014.

Assenov, Y., Müller, F., Lutsik, P., Walter, J., Lengauer, T., and Bock, C. Comprehensive Analysis of DNA Methylation Data With RnBeads. *Nature Methods*, 11(11):1138–1140, 2014.

Baron, U., Türbachova, I., Hellwag, A., Eckhardt, F., Berlin, K., Hoffmuller, U. *et al.* DNA methylation analysis as a tool for cell typing. *Epigenetics*, 1(1):55–60, 2006.

Baylin, S. B. DNA methylation and gene silencing in cancer. *Nature Clinical Practice. Oncology*, 2 Suppl 1:S4–S11, 2005.

Bernstein, B. E., Stamatoyannopoulos, J. a., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nature Biotechnology*, 28(10):1045–1048, 2010.

Bundo, M., Kato, T., and Iwamoto, K. Epigenetic Methods in Neuroscience Research. In N. Karpova, editor, *Neuromethods*, volume 105 of *Neuromethods*, pages 115–123. Springer New York, New York, NY, 2016. ISBN 978-1-4939-2753-1.

Christiansen, J., Kolte, A. M., Hansen, T. O., and Nielsen, F. C. IGF2 mRNA-binding protein 2: biological function and putative role in type 2 diabetes. *J Mol Endocrinol*, 43(5):187–195, 2009.

Cossarizza, A., Ortolani, C., Paganelli, R., Barbieri, D., Monti, D., Sansoni, P. *et al.* CD45 isoforms expression on CD4+ and CD8+ T cells throughout life, from newborns to centenarians: implications for T cell memory. *Mechanisms of ageing and development*, 86(3):173–95, 1996.

Dainiak, M. B., Kumar, A., Galaev, I. Y., and Mattiasson, B. Methods in cell separations. *Advances in biochemical engineering/biotechnology*, 106:1–18, 2007.

Esteller, M. Cancer epigenomics: DNA methylomes and histone-modification maps. *Nature reviews. Genetics*, 8(4):286–298, 2007.

Fahey, J. L., Schnelle, J. F., Boscardin, J., Thomas, J. K., Gorre, M. E., Aziz, N. *et al.* Distinct categories of immunologic changes in frail elderly. *Mechanisms of Ageing and Development*, 115(1-2):1–20, 2000.

Fang, G., Munera, D., Friedman, D. I., Mandlik, A., Chao, M. C., Banerjee, O. *et al.* Genome-wide mapping of methylated adenine residues in pathogenic Escherichia coli using single-molecule real-time sequencing. *Nature Biotechnology*, 30(12):1232–9, 2012.

Guintivano, J., Aryee, M. J., and Kaminsky, Z. a. A cell epigenotype specific model for the correction of brain cellular heterogeneity bias and its application to age, brain region and major depression. *Epigenetics*, 8(3):290–302, 2013.

Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biology*, 14(10):R115, 2013.

Houseman, E. A., Accomando, W. P., Koestler, D. C., Christensen, B. C., Marsit, C. J., Nelson, H. H. *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*, 13(1):86, 2012.

Houseman, E. A. and Ince, T. A. Normal cell-type epigenetics and breast cancer classification: a case study of cell mixture-adjusted analysis of DNA methylation data from tumors. *Cancer informatics*, 13(Suppl 4):53–64, 2014.

Houseman, E. A., Kelsey, K. T., Wiencke, J. K., and Marsit, C. J. Cell-composition effects in the analysis of DNA methylation array data: a mathematical perspective. *BMC Bioinformatics*, 16(1):95, 2015.

Houseman, E. A., Molitor, J., and Marsit, C. J. Reference-free cell mixture adjustments in analysis of DNA methylation data. *Bioinformatics*, 30(10):1431–1439, 2014.

Jaffe, A. E. and Irizarry, R. a. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biology*, 15(2):R31, 2014.

Ji, H., Ehrlich, L. I. R., Seita, J., Murakami, P., Doi, A., Lindau, P. *et al.* Comprehensive methylome map of lineage commitment from haematopoietic progenitors. *Nature*, 467(7313):338–42, 2010.

Kantlehner, M., Kirchner, R., Hartmann, P., Ellwart, J. W., Alunni-Fabbroni, M., and Schumacher, A. A high-throughput DNA methylation analysis of a single cell. *Nucleic Acids Research*, 39(7):E44–U68, 2011.

Kaut, O., Schmitt, I., and Wüllner, U. Genome-scale methylation analysis of Parkinson's disease patients' brains reveals DNA hypomethylation and increased mRNA expression of cytochrome P450 2E1. *Neurogenetics*, 13(1):87–91, 2012.

Koestler, D. C., Christensen, B. C., Karagas, M. R., Marsit, C. J., Langevin, S. M., Kelsey, K. T. *et al.* Blood-based profiles of DNA methylation predict the underlying distribution of cell types: A validation analysis. *Epigenetics*, 8(8):816–826, 2013.

Kulis, M., Merkel, A., Heath, S., Queirós, A. C., Schuyler, R. P., Castellano, G. *et al.* Whole-genome fingerprint of the DNA methylome during human B cell differentiation. *Nature genetics*, 47(7):746–756, 2015.

Kumar, A. and Bhardwaj, A. Methods in cell separation for biomedical application: cryogels as a new tool. *Biomedical materials (Bristol, England)*, 3(3):034008, 2008.

Lam, L. L., Emberly, E., Fraser, H. B., Neumann, S. M., Chen, E., Miller, G. E. *et al.* Factors underlying variable DNA methylation in a human community cohort. *Proceedings of the National Academy of Sciences*, 109(Supplement_2):17253–17260, 2012.

Lee, K. W. K. and Pausova, Z. Cigarette smoking and DNA methylation. *Frontiers in genetics*, 4:132, 2013.

Leek, J. T. and Storey, J. D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS genetics*, 3(9):1724–35, 2007.

Lin, C.-J. Projected gradient methods for nonnegative matrix factorization. *Neural computation*, 19(10):2756–2779, 2007.

Liu, Y., Aryee, M. J., Padyukov, L., Fallin, M. D., Hesselberg, E., Runarsson, A. *et al.* Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nature Biotechnology*, 31(2):142–7, 2013.

Lowe, R. and Rakyan, V. K. Correcting for cell-type composition bias in epigenome-wide association studies. *Genome medicine*, 6(3):23, 2014.

Lunnon, K., Smith, R., Hannon, E., De Jager, P. L., Srivastava, G., Volta, M. *et al.* Methylomic profiling implicates cortical deregulation of ANK1 in Alzheimer's disease. *Nature neuroscience*, 17(9):1164–70, 2014.

McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nature Biotechnology*, 28(5):495–501, 2010.

Michels, K. B., Binder, A. M., Dedeurwaerder, S., Epstein, C. B., Greally, J. M., Gut, I. *et al.* Recommendations for the design and analysis of epigenome-wide association studies. *Nature Methods*, 10(10):949–55, 2013.

Mo, A., Mukamel, E. A., Davis, F. P., Luo, C., Henry, G. L., Picard, S. *et al.* Epigenomic Signatures of Neuronal Diversity in the Mammalian Brain. *Neuron*, 86(6):1369–84, 2015.

Montaño, C. M., Irizarry, R. a., Kaufmann, W. E., Talbot, K., Gur, R. E., Feinberg, A. P. *et al.* Measuring cell-type specific differential methylation in human brain tissue. *Genome Biology*, 14(8):R94, 2013.

Owen, A. B. and Perry, P. O. Bi-cross-validation of the SVD and the nonnegative matrix factorization. *Annals of Applied Statistics*, 3(2):564–594, 2009.

Palli, D., Berrino, F., Vineis, P., Tumino, R., Panico, S., Masala, G. *et al.* A molecular epidemiology project on diet and cancer: the EPIC-Italy Prospective Study. Design and baseline characteristics of participants. *Tumori*, 89(6):586–93, 2003.

Pelizzola, M. and Ecker, J. R. The DNA methylome. *FEBS Letters*, 585(13):1994–2000, 2011.

Pidsley, R., Y Wong, C. C., Volta, M., Lunnon, K., Mill, J., Schalkwyk, L. C. *et al.* A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics*, 14(1):293, 2013.

Rahmani, E., Zaitlen, N., Baran, Y., Eng, C., Hu, D., Galanter, J. *et al.* Sparse PCA corrects for cell type heterogeneity in epigenome-wide association studies. *Nature Methods*, 2016.

Rakyan, V. K., Down, T. a., Balding, D. J., and Beck, S. Epigenome-wide association studies for common human diseases. *Nature Reviews. Genetics*, 12(8):529–541, 2011.

Reik, W., Dean, W., and Walter, J. Epigenetic reprogramming in mammalian development. *Science (New York, N.Y.)*, 293(5532):1089–93, 2001.

Reinius, L. E., Acevedo, N., Joerink, M., Pershagen, G., Dahlén, S.-E., Greco, D. *et al.* Differential DNA Methylation in Purified Human Blood Cells: Implications for Cell Lineage and Studies on Disease Susceptibility. *PLoS ONE*, 7(7):e41361, 2012.

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–30, 2015.

Romanyukha, A. A. and Yashin, A. I. Age related changes in population of peripheral T cells: towards a model of immunosenescence. *Mechanisms of ageing and development*, 124(4):433–43, 2003.

Schadt, E. E., Banerjee, O., Fang, G., Feng, Z., Wong, W. H., Zhang, X. *et al.* Modeling kinetic rate variation in third generation DNA sequencing data to detect putative modifications to DNA bases. *Genome Research*, 23(1):129–141, 2013.

Schübeler, D. Function and information content of DNA methylation. *Nature*, 517(7534):321–326, 2015.

Schwartzman, O. and Tanay, A. Single-cell epigenomics: techniques and emerging applications. *Nature Reviews Genetics*, 16(12):716–726, 2015.

Seungjin Choi, Andrzej Cichocki, Hyung-min Park, S.-y. L. Blind Source Separation and Independent Component Analysis: A Review. *Neural Information Processing - Letters and Reviews*, 6(1):1–57, 2005.

Shoemaker, R., Deng, J., Wang, W., and Zhang, K. Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. *Genome Research*, 20(7):883–889, 2010.

Tao, P. and An, L. Convex analysis approach to dc programming: Theory, algorithms and applications. *Acta Mathematica Vietnamica*, 1997.

Teschendorff, A. E., Zhuang, J., and Widschwendter, M. Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies. *Bioinformatics (Oxford, England)*, 27(11):1496–505, 2011.

Tomlinson, M. J., Tomlinson, S., Yang, X. B., and Kirkham, J. Cell separation: Terminology and practical considerations. *Journal of tissue engineering*, 4:2041731412472690, 2013.

Vavasis, S. a. On the complexity of nonnegative matrix factorization. *SIAM Journal on Optimization*, 20(3):1–12, 2007.

Yuille, A. L. and Rangarajan, A. The concave-convex procedure. *Neural computation*, 15(4):915–936, 2003.

Zhang, D., Cheng, L., Badner, J. a., Chen, C., Chen, Q., Luo, W. *et al.* Genetic Control of Individual Differences in Gene-Specific Methylation in Human Brain. *American Journal of Human Genetics*, 86(3):411–419, 2010.

Zhang, Z., Tang, H., Wang, Z., Zhang, B., Liu, W., Lu, H. *et al.* MiR-185 Targets the DNA Methyltransferases 1 and Regulates Global DNA Methylation in human glioma. *Molecular Cancer*, 10(1):124, 2011.

Zou, J., Lippert, C., Heckerman, D., Aryee, M., and Listgarten, J. Epigenome-wide association studies without the need for cell-type composition. *Nature Methods*, 11(3):309–11, 2014.

Chapter 7

General Discussion, Conclusions and Outlook

The present thesis is a collection of several data analysis advances for DNA methylation mapping. In this concluding chapter the results presented above will be put into a common context and discussed in respect of their benefits for DNA methylation analysis and epigenetics in general. Section 7.2 conveys a unifying view of how all the contributions of the thesis are together composing a data analysis framework of a typical DNA methylation profiling study. Next, Section 7.2 deals with a more specific problem of methylome deconvolution. Finally, Section 7.3 gives an outlook at the upcoming development of the field and provides for general considerations about bioinformatic tool development and perspectives of this research field.

7.1 An analytical framework of a large DNA methylation study

Collectively, the contributions presented in Chapters 2 to 6 comprise different data analysis aspects of a typical DNA methylation study. The relationships between particular results are given in a conceptual diagram (Figure 7.1). In the center is a genome-scale DNA methylation profiling study exemplified by a birth-weight EWAS in birth-weight discordant monozygotic twins (Chapter 2). A study of such kind usually performs a screen of DNA methylation in multiple individuals resulting in a high-dimensional DNA methylation data set. The data are then preprocessed, quality controlled and normalized using the RnBeads package (Chapter 3). The preprocessed data can be adjusted for heterogeneity effects using the third-party methods implemented in RnBeads, as well as a marker-based adjustment procedure described in Chapter 2. The heterogeneity-corrected statistical analysis results in a set of candidate differentially methylated positions or loci. These candidates are, as a rule, verified using a different DNA methylation profiling technology. Locus-specific deep bisulfite sequencing is a method of choice due to its high sensitivity and resolution. In addition to normal bisulfite sequencing, the protocols for detection of oxidative modifications can be applied to investigate methylation dynamics at the candidate loci. The obtained verification data set is processed using BiQ Analyzer HT (Chapter 4) and HiMod (Chapter 5). Finally, DNA methylation heterogeneity in the genome-scale profiles can be exhaustively explored in an unsupervised manner using the reference-free deconvolution by MeDeCom (Chapter 6). In the remainder of this section each of these aspects is thoroughly discussed.

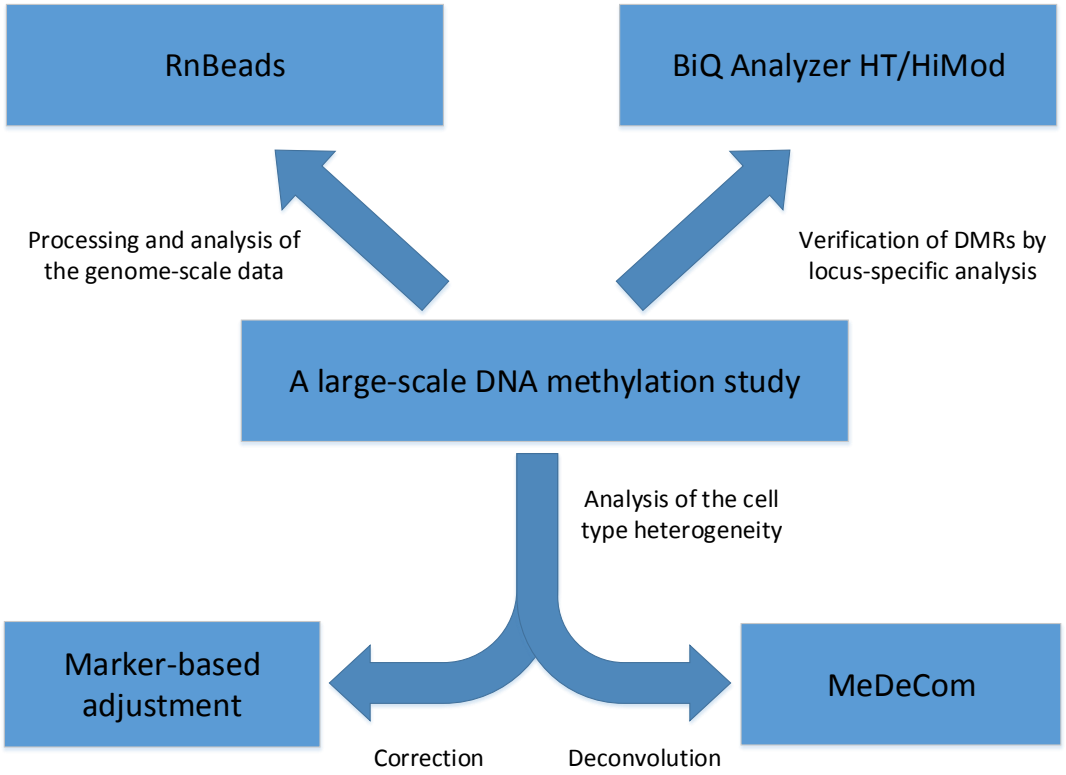


Figure 7.1: Conceptual diagram of the main results.

Study preparation

Proper design and planning are deterministic for the success of a DNA methylation study [Michels *et al.*, 2013; Rakyan *et al.*, 2011]. The EWAS in birth-weight discordant monozygotic twins presented in Chapter 2 confirmed that the more thinking is invested into the stages prior to data retrieval the easier it is to analyze the obtained data and the more reliable are the findings. Adequate consideration of variability sources, confounding factors, power limits and pitfalls of the selected technology allows to avoid potentially deteriorating data preprocessing and excessively complicated data analysis methods. For the particular case of an EWAS there are several design aspects requiring thoughtful consideration, namely the selection of a cohort in the context of potential genetic confounding, heterogeneity-aware choice of the target tissue or cell type, and the design of the profiling procedure minimizing technical batch effects.

Genetic information has a strong influence upon the methylome, and, therefore, the studies in case-control cohorts are severely confounded by the genetic variability. Low sample numbers, still common for many EWAS, are aggravating the problem since the genetic effects, such as ASM or methQTLs, have very large effects sizes and require a lot of observations to be eliminated by averaging in the compared groups. Bead arrays, used predominantly in EWAS, are specifically influenced by SNPs. A popular method of dealing with this problem is to remove all potentially affected probes. In case a rather liberal threshold of 3% for the minor allele frequency is used, such an approach would result in a removal of circa potentially informative 40,000 Infinium 450k CpGs, without solving a problem of methQTLs. In this respect, DNA methylation studies with a genetically matched design, such as those based on monozygotic twins, successfully avoid this type of confounding and are inherently more powerful.

The issue of confounding by cell type heterogeneity discussed below, can be avoided or, at least, minimized by a proper consideration at the early design stages. In due of the costs and the sample collection efforts EWAS have to rely upon easily accessible tissues and biofluids. The choice of a sampled biological material should, if possible, avoid severely heterogeneous tissues, in case this is not necessary for testing the main hypothesis. As an example relevant for the blood-based studies, if one can exclude that the sought effects do not take place in granulocytes, sampling the whole blood is not a reasonable choice, since granulocytes contribute for more than a half of thereby obtained DNA. A comparatively simple and non-invasive PBMC isolation procedure might be preferable in this case. A pan-somatic effect is better studied using saliva or buccal epithelium swabs carrying less sources of cell type heterogeneity compared to whole blood. One should also carefully consider the possibility that the target phenotype or condition being associated with strong changes of the cell type composition in the sampled material. In this case the confounding is the most severe and has to be corrected if the "direct" effects across the sampled tissue are the primary target.

Technical confounding can be as deteriorating as genetic and cell type heterogeneity effects are, and design precautions are necessary to minimize such effects. Although the noise level is usually inherent to the particular DNA methylation profiling technology and cannot be influenced, the various types of technical biases can and should be accounted for. For instance, the non-linear intensity increase across the Infinium/EPIC microarray plate can translate into a global bias of methylation values. Hence, any periodicity in sample positioning will inevitably confound an association analysis of a case-control study, and, therefore, it is highly recommended to randomize the sample layout. On the opposite, under a matched design the sample pairs should be positioned at neighboring slots to minimize the intra-pair differences.

Standardized and comprehensive data handling

The results of the EWAS example in Chapter 2 as well as other similar studies mentioned above emphasize the importance of appropriate data handling. Chapter 3 presents a universal package RnBeads for the preprocessing and analysis of high-throughput bisulfite sequencing and DNA methylation bead array data. The contribution of the present thesis is encompassing the core infrastructure of RnBeads, the data import, quality control and normalization modules, as well as the cell type confounding estimation and adjustment functionality.

One particular problem is the large data volumes typically occupied by the DNA methylation data. The biggest bisulfite sequencing and bead array data sets are approximately equal in size, due to orders of magnitude differences in numbers of profiled samples. In this respect, the operating memory is comprising a resource bottleneck, and a large analysis is only possible when significant portions of the data are stored on the hard drive space in an easily accessible form. Such a solution was implemented in RnBeads facilitating simultaneous processing of the complete set of Infinium arrays generated by TCGA.

Quality control is an essential part of any analysis. Low quality samples have to be detected and removed from the analysis. In case the quality of the whole data acquisition was low, one should consider repeating it. In many cases, in particular for DNA methylation microarrays, the data should be properly normalized. RnBeads collects the best available methods for quality control and normalization into a standardized comprehensive pipeline allowing for easy testing and comparison of the methods. The modular architecture allows seamless extension of the pipeline and addition of new methods.

Correction of the cell-type heterogeneity effects

The issue of methylome heterogeneity is very topical for the current DNA methylation research. It will maintain its actuality until the single-cell methods reach final development, making it feasible to obtain methylomes of a representative cell sample from each individual in a large study cohort. The single-molecule sequencing methods which will be able to discriminate the DNA modifications are also promising.

The birth-weight EWAS presented in Chapter 2 offers a clear-cut example of confounding by heterogeneity issues. Several samples with a very different cellular composition of saliva inflated the effect size at a large bulk of CpG positions. If not accounted for they could have led to spurious reporting of methylation effects of the restricted *in utero* growth.

To address this problem in this particular study we developed a simple regression-based method presented in the same chapter (see Supplementary Methods). The method has two essential steps. The first step includes marker selection based on values in the reference methylomes as well as on linearity properties in the target data set. In the second step the marker profiles observed in the actual data are used to model the complete target data set by the standard least-squares regression. The intercept-scaled residuals of the model fit are treated as an adjusted version of the DNA methylation data set, and are directly submitted to an association analysis using standard statistical techniques. Our method helped to avoid false positive findings and spurious associations.

Our correction approach belongs to the reference-based heterogeneity estimation and adjustment methods reviewed in Section 1.3. Most of them use reference methylomes directly for computing the estimates of cell type proportions which is associated with potential risks. First, reference methylomes are usually obtained in completely independent profiling experiments and can be drastically different from the target data from the technical standpoint. These effects can be partially solved by combined preprocessing of the target and reference data, which is implemented in available pipelines, e.g. RnBeads or minfi [Aryee *et al.*, 2014]. However, even if the technical differences are minimal, the reference data is typically obtained via the cell purification experiments in an independent cohort of individuals. Second, the estimation method assumes that that the cell type methylomes of the reference and target individuals are indistinguishable. Failure to fulfill this assumption, for instance, due to genetic or environmental differences between the target and the reference cohorts, the estimates obtained by the reference-based methods are inherently biased. In this respect, using the reference methylomes only for marker selection, as suggested in Chapter 2, appears to be a more reliable procedure, free of such hardly verifiable assumptions. This, of course, requires careful examination of the selected quantitative marker CpGs for potential genetic or technical confounding.

An additional advantage of selecting a few adjustment markers is that the latter can be used in the subsequent analysis with a targeted locus-specific technology in a large validation cohort. The selected markers can be profiled together with the loci selected in the primary analysis and used for the correction of the generated locus-level data.

Based on the considerations above, the reference-free methods for heterogeneity correction, in particular RefFreeEWAS [Houseman *et al.*, 2014], EWASHER [Zou *et al.*, 2014] and ReFACTor [Rahmani *et al.*, 2016], are expected to be more robust and bias-tolerant. The comparison of all three methods shows almost equivalent performance on the same EWAS data sets [Rahmani *et al.*, 2016]. All theses methods, except for the very recent ReFACTor, are implemented in RnBeads.

Verification by locus-specific analysis

A typical genome-scale DNA methylation study aims to find the candidate loci, potentially associated with the target biological condition. Once the regions of interest are stratified, a verification by a high-resolution technology, such as deep bisulfite sequencing, should follow. Furthermore, with respect to the heterogeneity problem, locus-specific deep bisulfite sequencing allows for direct visualization of the underlying cell population methylomes as demonstrated in Chapter 2. Chapters 4 and 5 present two software packages for processing and analysis of locus-specific DNA methylation data of various types. Both tools are intended for the life science users and implement an interactive application model allowing a biological researcher to visually explore the results from the level of aligned sequence reads down to summarized results for a complete analysis project. One of the major advances of these tools compared to their predecessor BiQ Analyzer and similar third-party tools is the project-oriented architecture which supports a typical investigation scheme of a candidate gene or EWAS-verification study. BiQ Analyzer HT (Chapter 4) and its successor BiQ Analyzer HiMod (Chapter 5) are successfully used by DNA methylation researchers worldwide.

7.2 Deconvolution of the mixture methylomes

The specific issue of methylome heterogeneity has received a particular attention here. Apart from the first more practical aspect of deconfounding, thoroughly discussed in the previous section, the methylome heterogeneity problem has another inferential dimension. This latter aspect can be defined as a deconvolution problem introduced at the end of Section 1.3.

Mathematically this problem was formulated in Chapter 6 (Supplementary Note 1). It is important to understand that the true entities which are mixed in the average methylomes are the unique methylomes of single cells (the "exact" model of mixture methylomes). This is why a model representing measured data as a combination of cell type methylomes is conceptually ill-posed and can only be seen as an approximation.

Even if the mixture methylomes could be measured by an ideal method, free of any technical biases and noise, the "exact" problem is computationally intractable until the number of profiled mixtures gets the same order of magnitude as the number of underlying unique single-cell methylomes. Although, completely unrealistic at the moment, it can be reached in the near future as the numbers of profiled samples grow (currently approaching ten thousand and more), and the profiling methods get increasingly low-input. Once computationally feasible, the exact problem can be provably solved by the method we earlier developed [Slawski *et al.*, 2013].

At the moment the majority of large DNA methylation studies exploit methods taking hundreds of thousands to millions of cells as input. In this situation the methylome deconvolution problem can only be pursued in its approximated variant. A meaningful approximation, preserving most of the information about the underlying distribution of the single-cell methylation patterns, can be seen as a series of nested models where groups (clusters) of highly similar single-cell methylomes are substituted by their average profiles. Since the cell type is one of the major determinants of the global DNA methylation landscapes, one can expect that in the low-rank approximation the similarity clusters will be close to the average methylomes of cell types and cell populations. Nevertheless, other large sources of DNA methylation variability affecting many CpG positions in multiple cells will inevitably affect the similarity clustering of the single-cell profiles.

Importantly, such a structure-preserving approximation model does not have to be (and

will most likely not be) the best one in terms of its reconstruction performance, i.e. the distance between the predicted and the observed data. For the sake of contrast one can consider PCA. PCA is based on singular value decomposition and is guaranteed to deliver the best possible low-rank approximation of a data matrix. However, the orthogonality of the returned principal components is incompatible with the properties of the sought approximation, since the average methylomes of cell populations are far from being linearly independent due to the global structure of the DNA methylation landscape. On the opposite, the data of the reference profiling studies shows that even the methylomes of relatively distant cell types maintain a considerable level of similarity [Horvath *et al.*, 2012].

The computational framework MeDeCom presented in Chapter 6 was designed to achieve the goal of finding a biologically meaningful, structure-preserving approximation. The model constraints and the regularization of the fitting algorithm limit the reconstruction performance, but are supposed to aid in recovering the approximation solution that are more biologically meaningful compared to those obtained with standard methods. To describe the elements of the MeDeCom approximation model we introduced the concept of latent methylation components (LMCs). The aim of MeDeCom is that LMCs correspond to clusters of similar single-cell patterns.

In practice the convoluted methylome matrix is only available through an unideal measurement. The particular profiling method has a decisive influence upon the computational properties of the deconvolution problem. Currently, reasonably large data sets of multi-cellular samples are obtained with Infinium/EPIC bead arrays. While this technology is comparatively robust and each called methylation value has a sufficient support, the specific problems of the Infinium/EPIC bead arrays such as the augmented probe-specific value range, proneness to batch effects and sensible noise levels are comprising a substantial limits upon the grade of deconvolution. Bisulfite sequencing methods may prove more useful in this respect. Although often used as an averaged quantitative profile, conventional bisulfite sequencing essentially delivers sub-sampled snapshots of the single-cell methylome distribution. This information can be used to recover the cell population methylomes with a much higher efficiency and resolution. That said, even the summarized bisulfite sequencing profiles are void of many pitfalls seen in the microarray data. More specifically, the methylation calls, as a rule, have the same effective value range at all positions, the background is practically absent, allowing for efficient removal of invariable CpGs, simultaneously retrievable genetic variants can be used to eliminate the genetically affected CpGs.

7.3 Outlook

Its involvement into the key regulatory processes and favourable properties of 5-methylcytosine as an easily accessible mark turn DNA methylation into an attractive target to study regardless of how causal its role is in controlling the gene activity. Hence, DNA methylation profiling will doubtlessly maintain its high importance for epigenetics and the biological research in general. Although the mapping methodology is constantly improving, the data analysis problems will unlikely get deprecated in the near future. Consequently, a need for new computational methods, as well as comprehensive and convenient bioinformatic tools will persist. Ideally, the major bioinformatic effort will be aimed at transforming the cutting edge methods developed by mathematicians, computer scientists and computational biologists into comprehensive and user-friendly tools helping a life scientist to get a full understanding of their data and lead the data analysis to a conclusive result. RnBeads, BiQ Analyzer HT, BiQ Analyzer HiMod and MeDeCom presented above were all created with this goal in mind.

It is quite apparent that the current issue of methylome heterogeneity, also addressed in this thesis, will be substantially simplified and transformed in the long to medium term. The large-scale profiling efforts, such as the International Human Epigenome Consortium [Abbott, 2010], will generate numerous reference methylomes. This will allow massive annotation of cell type-specific marker CpGs that will further aid the reference-based heterogeneity correction methods. Furthermore, the anticipated advances in low input and single-cell bisulfite sequencing together with the long-awaited breakthroughs of the third generation sequencing approaches. The current problem of analyzing the average cell sample methylomes will be succeeded by a more large-scale setting with hundreds to thousands single-cell DNA methylation patterns per each biological specimen. Despite that this setting seems much more informative and easy to analyze, the data analysis focus will most likely shift towards the more intricate issues, such as estimation of the sampling biases.

One can foresee that the other types of data will become increasingly important as the respective profiling methods are constantly improving. The methodology derived here for the problem of methylome deconvolution is mathematically generic and can, in theory, be applied to a range of other epigenomic signals sharing the discreteness properties of DNA methylation. These include all epigenetic marks and other data types backed by binary phenomena at the single cell level for instance open chromatin states (open vs. closed), transcription factor binding profiles (bound vs. free), genome physical contact maps (absence vs. presence of a contact) etc. It is clear, however, that a certain adaptation of the algorithm as well as specific data preparation will be required.

Last but not least, there are a lot of sings that the very way large epigenomic studies are carried out will transform in respect of how life scientists collaborate with bioinformaticians and computational biologist. Current practice in the majority of the DNA methylation studies is that the division of labor between between them follows the study stages. Typically, life science researchers conceive, plan and design the studies and generate high-throughput data, while bioinformaticians enter the study after the data has been generated and take over the processing and analysis. This model is flawed in several respects. First, bioinformaticians can provide necessary expertise about specific and often subtle issues which have to be considered during the study design and planning, and should be involved from the start. Second, the life-science researchers are the most informed about all aspects and subtleties of their study, and as such should play the most active role in the data analysis. Third, bioinformatic members of life science groups, which are usually in minority, are overwhelmed with data analysis tasks which they have little affinity with, leading to an overall delay of the study cycle and wastes of the human and material resources. The issue of perceiving bioinformatic work as a highly skilled service has earlier received attention from the social standpoint. It is speculated for creating tensions in the field [Lewis and Bartlett, 2013].

The situation is gradually improving, as the large epigenomic consortia, such as DEEP project and its IHEC peers, introduce a qualitatively new collaborative spirit and involve bioinformaticians and computational biologists at all stages. One is also tempted to speculate that in the upcoming future these collaboration flaws will be overcome by a change in the profiles of the involved researchers towards the individual interdisciplinarity, as was already expected at the verge of the Human Genome Project, but comes with a noticeable delay [Lewis and Bartlett, 2013]. Knowledge of a required statistical minimum, as well as of basic programming should become necessary skills of every life science experimentalist, just as it is currently the case for many other natural sciences.

References

Abbott, A. Project set to map marks on genome. *Nature*, 463(7281):596–597, 2010.

Aryee, M. J., Jaffe, A. E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A. P., Hansen, K. D. *et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*, 2014.

Horvath, S., Zhang, Y., Langfelder, P., Kahn, R. S., Boks, M. P. M., van Eijk, K. *et al.* Aging effects on DNA methylation modules in human brain and blood tissue. *Genome Biology*, 13(10):R97, 2012.

Houseman, E. A., Molitor, J., and Marsit, C. J. Reference-free cell mixture adjustments in analysis of DNA methylation data. *Bioinformatics*, 30(10):1431–1439, 2014.

Lewis, J. and Bartlett, A. Inscribing a discipline: tensions in the field of bioinformatics. *New Genetics and Society*, 32(3):243–263, 2013.

Michels, K. B., Binder, A. M., Dedeurwaerder, S., Epstein, C. B., Greally, J. M., Gut, I. *et al.* Recommendations for the design and analysis of epigenome-wide association studies. *Nature Methods*, 10(10):949–55, 2013.

Rahmani, E., Zaitlen, N., Baran, Y., Eng, C., Hu, D., Galanter, J. *et al.* Sparse PCA corrects for cell type heterogeneity in epigenome-wide association studies. *Nature Methods*, 2016.

Rakyan, V. K., Down, T. a., Balding, D. J., and Beck, S. Epigenome-wide association studies for common human diseases. *Nature Reviews. Genetics*, 12(8):529–541, 2011.

Slawski, M., Hein, M., and Lutsik, P. Matrix factorization with Binary Components. In *Advances in Neural Information Processing Systems*, pages 3210–3218. 2013.

Zou, J., Lippert, C., Heckerman, D., Aryee, M., and Listgarten, J. Epigenome-wide association studies without the need for cell-type composition. *Nature Methods*, 11(3):309–11, 2014.

List of publications

First and second author

- **Lutsik*, P.**, Slawski*, M., Gasparoni, G., Vedeneev, N., Hein, M., and Walter, J. MeDeCom: discovery and quantification of latent components in heterogeneous methylomes. *Genome Biology*, 2017, in press
- Assenov*, Y., Müller*, F., **Lutsik*, P.**, Walter, J., Lengauer, T., and Bock, C. Comprehensive Analysis of DNA Methylation Data With RnBeads. *Nature Methods*, 11(11):1138–1140, 2014
- Chen, J., **Lutsik, P.**, Akulenko, R., Walter, J., and Helms, V. AKSmooth: Enhancing low-coverage bisulfite sequencing data via kernel-based smoothing. *Journal of Bioinformatics and Computational Biology*, 12(06):1442005, 2014
- Becker, D., **Lutsik#, P.**, Ebert, P., Bock, C., Lengauer, T., and Walter#, J. BiQ Analyzer HiMod: An interactive software tool for high-throughput locus-specific analysis of 5-methylcytosine and its oxidized derivatives. *Nucleic Acids Research*, 42(W1):W501–7, 2014
- Souren, N. Y., **Lutsik, P.**, Gasparoni, G., Tierling, S., Gries, J., Riemenschneider, M., Fryns, J.-P., Derom, C., Zeegers, M. P., and Walter, J. Adult monozygotic twins discordant for intra-uterine growth have indistinguishable genome-wide DNA methylation profiles. *Genome Biology*, 14(5):R44, 2013
- **Lutsik, P.**, Feuerbach, L., Arand, J., Lengauer, T., Walter, J., and Bock, C. BiQ Analyzer HT: Locus-specific analysis of DNA methylation by high-throughput bisulfite sequencing. *Nucleic Acids Research*, 39(S2):W551–6, 2011

*shared first author
#shared corresponding author

Contributing author

- Souren, N. Y. P., Gerdes, L. A., Kümpfel, T., **Lutsik, P.**, Klopstock, T., Hohlfeld, R., and Walter, J. Mitochondrial DNA Variation and Heteroplasmy in Monozygotic Twins Clinically Discordant for Multiple Sclerosis. *Human Mutation*, 2016
- Dyke, S. O. M., Cheung, W. A., Joly, Y., Ammerpohl, O., **Lutsik, P.**, Rothstein, M. A., Caron, M., Busche, S., Bourque, G., Rönnblom, L., Flicek, P., Beck, S., Hirst, M., Stunnenberg, H., Siebert, R., Walter, J., and Pastinen, T. Epigenome data release: a participant-centered approach to privacy protection. *Genome Biology*, 16(1):142, 2015

- Slawski, M., Hein, M., and **Lutsik, P.** Matrix factorization with Binary Components. In *Advances in Neural Information Processing Systems*, pages 3210–3218. 2013
- Gries, J., Schumacher, D., Arand, J., **Lutsik, P.**, Markelova, M. R., Fichtner, I., Walter, J., Sers, C., and Tierling, S. Bi-PROF: Bisulfite profiling of target regions using 454 GS FLX Titanium technology. *Epigenetics*, 8(7):765–771, 2013
- Tierling, S., Souren, N. Y., Gries, J., Loporto, C., Groth, M., **Lutsik, P.**, Neitzel, H., Utz-Billing, I., Gillessen-Kaesbach, G., Kentenich, H., Griesinger, G., Sperling, K., Schwinger, E., and Walter, J. Assisted reproductive technologies do not enhance the variability of DNA methylation imprints in human. *Journal of medical genetics*, 47(6):371–376, 2010