Translational Bioinformatics 7

Series Editor: Xiangdong Wang, MD, PhD, Prof

Andrew E. Teschendorff *Editor*

Computational and Statistical Epigenomics

Springer

Translational Bioinformatics

Volume 7

Series editor
Xiangdong Wang, MD, Ph.D.
Professor of Medicine, Zhongshan Hospital, Fudan University Medical School, China
Director of Shanghai Institute of Clinical Bioinformatics, (www.fuccb.org)
Professor of Clinical Bioinformatics, Lund University, Sweden

Aims and Scope

The Book Series in Translational Bioinformatics is a powerful and integrative resource for understanding and translating discoveries and advances of genomic, transcriptomic, proteomic and bioinformatic technologies into the study of human diseases. The Series represents leading global opinions on the translation of bioinformatics sciences into both the clinical setting and descriptions to medical informatics. It presents the critical evidence to further understand the molecular mechanisms underlying organ or cell dysfunctions in human diseases, the results of genomic, transcriptomic, proteomic and bioinformatic studies from human tissues dedicated to the discovery and validation of diagnostic and prognostic disease biomarkers, essential information on the identification and validation of novel drug targets and the application of tissue genomics, transcriptomics, proteomics and bioinformatics in drug efficacy and toxicity in clinical research.

The Book Series in Translational Bioinformatics focuses on outstanding articles/chapters presenting significant recent works in genomic, transcriptomic, proteomic and bioinformatic profiles related to human organ or cell dysfunctions and clinical findings. The Series includes bioinformatics-driven molecular and cellular disease mechanisms, the understanding of human diseases and the improvement of patient prognoses. Additionally, it provides practical and useful study insights into and protocols of design and methodology.

Series Description

Translational bioinformatics is defined as the development of storage-related, analytic, and interpretive methods to optimize the transformation of increasingly voluminous biomedical data, and genomic data in particular, into proactive, predictive, preventive, and participatory health. Translational bioinformatics includes research on the development of novel techniques for the integration of biological and clinical data and the evolution of clinical informatics methodology to encompass biological observations. The end product of translational bioinformatics is the newly found knowledge from these integrative efforts that can be disseminated to a variety of stakeholders including biomedical scientists, clinicians, and patients. Issues related to database management, administration, or policy will be coordinated through the clinical research informatics domain. Analytic, storage-related, and interpretive methods should be used to improve predictions, early diagnostics, severity monitoring, therapeutic effects, and the prognosis of human diseases.

Recently Published and Forthcoming Volumes

Bioinformatics of Human Proteomics
Editor: Xiangdong Wang
Volume 3

Single Cell Sequencing and Systems Immunology
Editors: Xiangdong Wang, Xiaoming Chen, Zhihong Sun, Jinglin Xia
Volume 5

Bioinformatics for Diagnosis, Prognosis and Treatment of Complex Diseases
Editor: Bairong Shen
Volume 4

Genomics and Proteomics for Clinical Discovery and Development
Editor: György Marko-Varga
Volume 6

More information about this series at http://www.springer.com/series/11057

Andrew E. Teschendorff
Editor

Computational and Statistical Epigenomics

Springer

Editor
Andrew E. Teschendorff
CAS Key Laboratory of
Computational Biology
Chinese Academy of Sciences
and Max-Planck Gesellschaft Partner
Institute for Computational Biology
Shanghai, China

UCL Cancer Institute
University College London
London, UK

ISSN 2213-2775 ISSN 2213-2783 (electronic)
Translational Bioinformatics
ISBN 978-94-017-9926-3 ISBN 978-94-017-9927-0 (eBook)
DOI 10.1007/978-94-017-9927-0

Library of Congress Control Number: 2015936464

Springer Dordrecht Heidelberg New York London

Printed on acid-free paper

Springer Science+Business Media B.V. Dordrecht is part of Springer Science+Business Media (www.springer.com)

Contents

Part I
Normalization and Analysis Methods for DNA Methylation and ChIP-Seq Data

Chapter 1
Introduction to Data Types in Epigenomics

Francesco Marabita, Jesper Tegnér, and David Gomez-Cabrero

Abstract The epigenome is the collection of all epigenetic modifications occurring on a genome. To properly generate, analyze, and understand epigenomic data has become increasingly important in basic and applied research, because epigenomic modifications have been broadly associated with differentiation, development, and disease processes, thereby also constituting attractive drug targets. In this chapter, we introduce the reader to the different aspects of epigenomics (e.g., DNA methylation and histone marks, among others), by briefly reviewing the most relevant underlying biological concepts and by describing the different experimental protocols and the analysis of the associated data types. Furthermore, for each type of epigenetic modification we describe the most relevant analysis pipelines, data repositories, and other resources. We conclude that any epigenomic investigation needs to carefully align the selection of the experimental protocols with the subsequent bioinformatics analysis and vice versa, as the effect sizes can be small and thereby escape detection if an integrative design is not well considered.

Keywords Epigenomics • DNA methylation • Histone modifications • ChIP-seq • Bioinformatics

1.1 Epigenomics

In eukaryotes, the DNA is stored in the nucleus through mechanisms allowing DNA packaging in condensed structures. This packaging allows a level of compression such that the DNA of a human diploid cell – which would linearly span for about 2 m – can be condensed efficiently in the space of a cell nucleus, typically 2–10 μm. The uncovering of the minimal unit of such condensation (Kornberg 1974), the nucleosome, showed DNA is tightly packed around a protein octamer (histones), with a left-handed superhelical turn of approximately 147 base pairs. The histone octamer consists of two copies of four histones: H2A, H2B, H3 and H4 and

F. Marabita (✉) • J. Tegnér • D. Gomez-Cabrero
Department of Medicine, Unit of Computational Medicine, Karolinska Institutet, Center for Molecular Medicine, Karolinska University Hospital, Solna, Stockholm, Sweden
e-mail: francesco.marabita@ki.se; jesper.tegner@ki.se; david.gomezcabrero@ki.se

A.E. Teschendorff (eds.), *Computational and Statistical Epigenomics*,
Translational Bioinformatics 7, DOI 10.1007/978-94-017-9927-0_1

A fifth histone (H1) binds the nucleosome and the linker DNA region and increases the stability. Higher-order packaging structures contribute to the final level of compression.

The nucleosome structure is inherently linked to gene expression, as it is intuitive that nucleosomes have to be displaced to allow gene expression to occur. The structure of the chromatin fulfills the role of condensing and protecting the DNA but it also preserves genetic information and controls gene expression. Therefore, this mechanism represents a process control and the accessibility of the DNA is regulated by chemical modifications that occur at the chromatin level, both for DNA and proteins. In this sense, nucleosomes contribute to regulatory mechanisms because they forbid or allow access for essential processes such as gene transcription or DNA replication (Fyodorov and Kadonaga 2001). For instance, DNA located near entry or exit points of the nucleosome are more accessible than those located centrally (Anderson and Widom 2000). Additionally, nucleosomes regulate DNA breathing (or fraying), that is, the spontaneous local conformational fluctuations within DNA of exit and entry points of nucleosome (Fei and Ha 2013), depending on the sequence wrapped around the histones and the covalent histone modifications. On the other hand, the DNA itself may be chemically modified, without associated changes in its sequence, generating important marks for regulation of gene expression, including DNA methylation. The collection of covalent changes to the DNA and histone proteins in the chromatin is called "epigenome." Changes in the epigenome are observed during development and differentiation and can be mitotically stable, modulate gene expression patterns in a cell and preserve cellular states. However, we have started to understand that also environmental factors can contribute to reshape the epigenome, potentially providing a mechanism to alter the gene expression program of a cell both in normal and disease conditions.

High-throughput technologies, including next-generation sequencing, offer the unprecedented opportunity of assaying epigenetic alteration usually in a hypothesis-free approach, by looking at multiple sites in the genome and verifying their association with the biological phenomenon observed. Therefore, bioinformatics and biostatistics represent key disciplines for obtaining solid results and are required in each phase of an epigenomics project, from study design to data analysis, visualization, and storage. In this chapter, we will give an overview of the two most studied epigenetic modifications, namely, DNA methylation and histone modifications; we will present the major steps in their respective experimental and data analysis pipelines, briefly discussing the associated challenges and opportunities.

1.2 DNA Methylation

1.2.1 Introduction to DNA Methylation

DNA methylation results from the addition of a methyl group to cytosine residues in the DNA to form 5-methylcytosine (5-mC) and in mammals it is predominantly restricted to the context of CpG dinucleotides, although other sequences might be methylated in some tissues (Lister et al. 2009, 2013; Ziller et al. 2013). CpG methylation has not only been observed during development or differentiation and in association with diseases, but has also been proposed as a prerequisite to understand disease pathogenesis in complex phenotypes (Petronis 2010). DNA methylation was initially identified as an epigenetic mark for gene repression (Riggs 1975; Holliday and Pugh 1975). Currently, although the silencing mechanism remains valid, we know that methylation in CpG-rich promoter regions is associated with gene repression, while CpG-poor regions show a less simple connection with transcription (Jones 2012). Therefore, the relationship between DNA methylation and transcriptional activation/repression is more complex than initially portrayed and dependent on the genomic and cellular context.

In the human genome, 70–80 % of CpG sequences are methylated (Ziller et al. 2013); however, both the distribution of CpG dinucleotides and the DNA methylation mark are not evenly distributed. CpG islands (CGI) are sequences with high C + G content that are generally unmethylated and colocalize with more than half of the promoters of human genes (Illingworth and Bird 2009). Housekeeping genes generally contain a CGI in the neighborhood of their TSS (Transcription Start Site), concordantly with the notion that chromatin at promoter with CGI shows a transcriptionally permissive state (Deaton and Bird 2011).

In addition to CGIs and TSSs, methylation at other classes of genomic elements has gained further attention over time. For example, CpG shores are genomic regions up to 2 kb distant from CGI, which show lower CpG density but increased variability in DNA methylation, and are found “to be among the most variable genomic regions” (Ziller et al. 2013). Most of tissue-specific DNA methylation in fact, as well as methylation differences between cancer and normal tissue, occur at CpG shores (Irizarry et al. 2009). DNA methylation at enhancers is also highly dynamic (Ziller et al. 2013; Stadler et al. 2011), has been shown to vary in physiological and pathological contexts (Aran and Hellman 2013; Lindholm et al. 2014; Rönnerblad et al. 2014), and methylation levels at enhancers are more closely associated with gene expression alterations than promoter methylation in cancer (Aran et al. 2013). Enhancers represent crucial determinants of tissue-specific gene expression and their identification methods include the analysis of epigenomic data (ChIP-seq, DNase-seq), since enhancer chromatin shows characteristic marks (Calo and Wysocka 2013). Moreover, DNA methylation at enhancer elements can influence the binding of Transcription Factors (TFs) (Stadler et al. 2011; Wiench et al. 2011), providing a direct link between CpG hypomethylation and target gene expression. However, it remains unsolved how this complex interplay is regulated and whether DNA methylation changes are a consequence of TF binding or whether they drive enhancer activity through exclusion of TF.

1.2.2 The Axes of DNA Methylation Variability

The role of DNA methylation variation has been investigated in many different contexts. Below we will give a brief overview of the phenotypes, settings, and major domains that together constitute the "axes" along which variability in DNA methylation has been studied.

Development Early studies proposed DNA methylation as a mechanism involved in X-chromosome inactivation and developmental programs (Riggs 1975; Holliday and Pugh 1975). Since then, the dynamics of DNA methylation during developmental changes has been studied extensively, and technological advances now render possible the study of methylomes of single cells (Smallwood et al. 2014; Guo et al. 2014), with manifest implications for the study of early embryos.

Imprinting and X Chromosome Inactivation Through the phenomenon of imprinting, genes that are expressed in allele-specific manner have regions showing parent-of-origin specific DNA methylation. When measured at an imprinted region, methylation is expected to approach a theoretical 50 % level. X-chromosome inactivation in females is also achieved through methylation, in order to transcriptionally silence the inactivated X chromosome, which is random in humans, and obtain gene dosage similar to males. Therefore, measured levels of DNA methylation differ by gender at X chromosome.

Disease The study of DNA methylation variability in common complex diseases is the focus of Epigenome-Wide Association Studies (EWAS), which aim at associating phenotypic traits to interindividual epigenomic variation, and in particular DNA methylation. A notable example is represented by cancer EWAS, which not only aim at understanding the molecular changes of tumorigenic pathways and disease risk, but also exploit DNA methylation profiling for disease diagnosis and prognosis. It is also thought that a combination of environmental, genetic, and epigenetic interactions contribute to the problem of the "missing heritability" (Eichler et al. 2010; Feinberg 2007).

Space and Time When designing and analyzing EWAS, it should be carefully considered that CpG methylation is subjected to spatial and temporal variability. One could consider the *genome space* as the main axis of variability, because different genomic elements have different methylation levels and show different degree of inter-sample variability. Alternatively, the *tissue/cell type space* represents another important axis of variation, as it is extensively established that different cell types possess their characteristic methylome. Other cases illustrate perfectly the extent of *temporal variability*. Monozygotic twins, for example, accumulate variability over time in their epigenome, such that older monozygous twins have higher differences in CpG methylation than younger twins (Fraga et al. 2005). Moreover, an "epigenetic drift" has been generally observed during aging (Bjornsson et al. 2008; Teschendorff et al. 2013b), confirming that both hypomethylation and hypermethylation are occurring over time, with acceleration dependent on disease or tissue factors (Horvath 2013; Horvath et al. 2014; Hannum et al. 2013).

Genotype Genotype is a strong source of interindividual variability in DNA methylation (Bell et al. 2011). Such genetic variants are defined as methylation quantitative trait loci (meQTLs) and they have been described in blood and other tissues (Bell et al. 2011; Drong et al. 2013; Shi et al. 2014). It is possible that some genotype-dependent CpGs mediate the genetic risk of common complex diseases (Liu et al. 2013).

Environment Accumulating evidence shows that several environmental factors can influence DNA methylation. For example, dietary factors have the potency to alter the degree of DNA methylation in different tissues (Feil and Fraga 2011; Lim and Song 2012). Cigarette smoking and pollution represent other known epigenetic modifiers (Lee and Pausova 2013; Feil and Fraga 2011). Moreover, short- or long-term physical exercise have also been proposed as physiological stimuli which can cause changes in DNA methylation (Barrès et al. 2012; Rönn et al. 2013; Lindholm et al. 2014).

1.2.3 Methods for DNA Methylation Profiling

Classically, methods for measuring DNA methylation have been divided into three major classes, including enrichment-based methods, digestion with methylation-sensitive restriction enzymes, and methods using bisulfite (BS) treatment. When coupled with DNA sequencing, affinity-based enrichment of methylated DNA fragments allows the interrogation of methylation of genomic regions with a methyl-binding protein (MBD-seq) (Serre et al. 2010) or an antibody (MeDIP-seq) (Down et al. 2008). These measurements do not give an absolute estimation of the methylation levels, but rather a relative enrichment that is dependent on the CpG density and the quality of the affinity assay (i.e., immunoprecipitation). Furthermore, the length of the DNA fragments determines their resolution. Similarly, methods based on restriction enzymes measure the relative enrichment after digesting the DNA with endonucleases that are sensitive to cytosine methylation (MRE-seq) (Maunakea et al. 2010), and they are therefore influenced by the genomic frequency of the recognition site for the selected enzyme. In this chapter, we will focus on methods using bisulfite conversion to assay the cytosine methylation status. After treatment with sodium bisulfite, unmethylated cytosines (C) in the genomic DNA are selectively converted to uracil (U), which are replaced by thymine (T) following PCR amplification (Fig. 1.1b). Methylated Cs are however protected from being converted. Afterward, the methylation levels can be quantified using microarrays or sequencing. Bisulfite treatment may be combined with digestion using methylation-insensitive restriction enzymes, in a technique called Reduced-Representation Bisulfite Sequencing (RRBS) (Meissner et al. 2008), to reduce the amount of reads to a fraction of the genome and thus reduce the cost. As opposed to Whole Genome Bisulfite Sequencing (WGBS) (Lister et al. 2009), this approach has reduced genome-wide coverage, but the coverage is higher for CpG islands (Harris et al. 2010). It is alternatively possible to capture targeted DNA fragments, in order

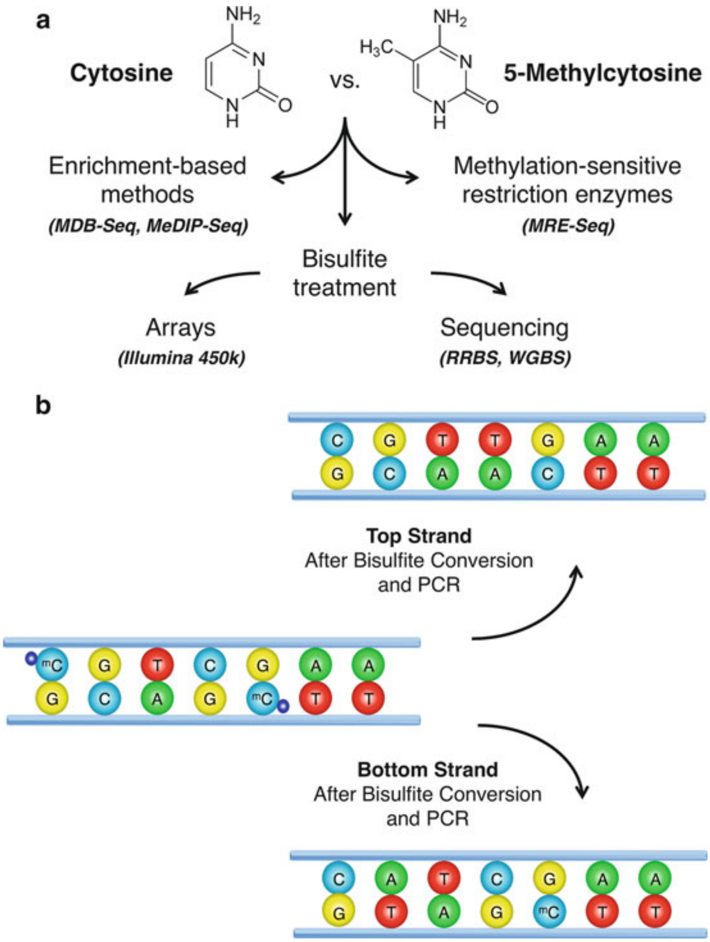


Fig. 1.1 Overview of experimental methods and bisulfite treatment for the analysis of DNA methylation. (**a**) The experimental methods to assay are shown, as further explained in the text. (**b**) The bisulfite treatment and PCR reactions will result in the conversion of unmethylated cystosines (*C*) into thymines (*T*), while 5-methylcitosine (*mC*) will be protected from the bisulfite-induced conversion

to restrict the sequencing to specific regions (Lee et al. 2011). Both sequencing and microarray technology offer single-base resolution. While microarray platforms have a lower cost per sample and limited genome-wide coverage, WGBS has the most comprehensive genome-wide coverage but at a higher cost. In the next sections we will better elucidate microarray- and sequence-based approaches, together with an overview of the analysis pipelines and softwares.

1.3 Bisulfite Microarrays

Commercially available microarray platforms conditioned the growing availability of EWAS, allowing a large sample size at an affordable cost. The sample size issue is relevant since, in many cases, changes in DNA methylation are mild and the biological variability may be high. Illumina Infinium HumanMethylation27 (27k) and HumanMethylation450 BeadChip (450k) are the most common types of oligonucleotide microarrays used for DNA methylation studies; at the date of writing (December 2014) >16,000 27k and >21,000 450k samples are deposited on GEO (Gene Expression Omnibus) database. The 450k arrays are based on the Infinium chemistry and contain 485,512 probes, targeting 99 % of genes and 96 % of CpG island regions (Bibikova et al. 2011). Oligonucleotide probes are attached to beads and deposited on an array, where the detection of the methylation status occurs through fluorescence reading. They represent an extension of the previous 27k platform, which was biased toward promoter regions. This extension resulted in wider coverage (but still limited compared to sequencing methods), specially toward other genomic regions like gene bodies and CpG shores (Bibikova et al. 2011; Sandoval et al. 2011). However, this also resulted in the introduction of two different bead types associated to two different chemical assays, Infinium I and Infinium II. Infinium I consists of two bead types (Methylated and Unmethylated) for the same CpG locus, both sharing the same color channel, whereas Infinium II utilizes a single bead type and two color channels (green and red) (Bibikova et al. 2011). Infinium II assays have larger variance and are less sensitive for the detection of extreme methylation values, which is probably associated to the dual-channel readout, thus rendering the Infinium I assay a better estimator of the true methylation state (Dedeurwaerder et al. 2011; Teschendorff et al. 2013a; Marabita et al. 2013). Moreover, different genomic elements (promoters, CpG islands, gene bodies, etc.) have different relative fraction of type I or type II probes (Dedeurwaerder et al. 2011). Methods have been introduced to correct for probe-type bias (see below for discussion).

The C methylation status for single CpG sites at each allele is always binary (0 or 1); however, the measured methylation levels can, in principle, take any value between 0 and 1 when averaging over many cells, or when the methylation status differs between the two alleles (imprinting, X-chromosome inactivation). For bisulfite microarrays, the methylation level is usually measured in two different scales, the β-value and the M-value. The β-value is calculated from signal intensities and can be interpreted as the percentage of methylation (it ranges from 0 to 1). It is related to the M-value through a logistic transformation. See reference (Du et al. 2010) for a detailed description of the two quantities. Even if M-values cannot be directly interpreted as methylation percentages, they offer several advantages, including the possibility of employing downstream association models that rely on the assumption of Gaussianity, as β-values appear compressed in the high and low range and often display heteroscedasticity. Moreover, when the sample size is relatively large the use of β- or M-values has been shown to give similar results, but with a limited sample size, M-values allow more reliable identification of true positives (Zhuang et al. 2012). However, from a pragmatic point of view and to allow biological interpretation, it is always advisable to report the final effect size in terms of median or mean β-value change, even if the feature selection step has been performed in the M-value space.

Independently of the scale used, the methylation profile for each sample shows a bimodal distribution, with two peaks corresponding to the unmethylated and methylated CpG positions. Because of the technical differences in probe design, a correction method is advisable. It could be argued that for CpG-level methylation difference analysis, the comparison will involve only probes of the same type. However, several indications suggest that it is advantageous to perform probe-type correction: (a) when the fold change (or effect size) is used in combination with the p-value for feature selection, as otherwise a bias may result from the dissimilar range between probes of different type (Marabita et al. 2013); (b) when dimensionality reduction or clustering algorithms are used, the pattern of variability between probe types may bias the grouping of CpG sites; (c) when DMR identification is anticipated, the methylation estimates along subsequent genomic positions will be dependent on probe type.

Methods for reducing the probe-type bias include a peak-based correction (Dedeurwaerder et al. 2011), SWAN method (Maksimovic et al. 2012), subset quantile normalization (Touleimat and Tost 2012), and BMIQ (Teschendorff et al. 2013a). In a benchmarking work (Marabita et al. 2013), BMIQ resulted as the best algorithm for reducing probe design bias. BMIQ, which employs a beta-mixture and quantile dilation intra-array normalization strategy, is available through several R packages (ChAMP (Morris et al. 2014), RnBeads (Assenov et al. 2014), WateRmelon (Schalkwyk et al. 2013)). Briefly, it first applies a beta-mixture model to assign probes of a given design type to methylation states and subsequently and uses state-membership probabilities to reassign the quantiles of the type2 probes according to the type1 distribution. Finally, for the probes with intermediate methylation values (which are not well described by a beta-distribution), a methylation-dependent dilation transformation is used, which also preserves the monotonicity and continuity of the data.

While probe-type normalization is a form of within-array normalization, between-array normalization is intended to remove part of the technical variability that is not associated with any biological factor, but which can be considered as caused by experimental procedures. For 450k data, there is no consensus on the best approach (Wilhelm-Benartzi et al. 2013; Dedeurwaerder et al. 2014), although a comparison of different normalization pipelines has been performed in recent works (Marabita et al. 2013; Pidsley et al. 2013). Many of the proposed approaches employ a form of quantile normalization (QN), which has been shown to perform well for gene expression studies (Irizarry et al. 2003). The goal of QN is to produce identical distribution of probe intensities for all the arrays and it has been applied to 450k data in several forms (Dedeurwaerder et al. 2014). While forcing the distribution of the methylation estimates to be the same for all the samples is a reasonably too strong an assumption for many biological comparisons, normalizing signal intensities appears a valid alternative in reducing technical variability in several contexts (Marabita et al. 2013; Dedeurwaerder et al. 2014). However, examination of the signal intensities and the study design should guide the application of this level of between-samples normalization, in order not to harm the integrity of the biological signal. A recent extension of QN, termed functional normalization (Fortin et al. 2014), uses control probes from the array to remove unwanted variation, assuming that summarized control probes function as surrogates of the nonbiological variation, which may include batch effects (see below).

Several comprehensive R packages have been developed for the processing and the analysis of 450k data (such as lumi (Du et al. 2010), methylumi (Davis et al. 2014), minfi (Aryee et al. 2014), wateRmelon (Schalkwyk et al. 2013), ChAMP (Morris et al. 2014), and RnBeads (Assenov et al. 2014)), and the reader is referred elsewhere for detailed discussion on popular pipelines and packages (Morris and Beck 2015; Wilhelm-Benartzi et al. 2013; Marabita et al. 2013; Dedeurwaerder et al. 2014).

Another type of unwanted variation in 450k data is represented by batch effects, which contaminate many high-throughput experiments including 450k arrays (Leek et al. 2010; Sun et al. 2011). We define a batch as a subgroup of samples or experiments exhibiting a systematic nonbiological difference that is not correlated with the biological variables under study. For example, different batches are represented by groups of samples that are processed separately, on different days or by a different operator. However, the definition of a batch results from careful examination of the data set, in order to identify what is an appropriate batch variable other than the processing group, as the slide or the position on the slide (i.e., the array), which represent known sources of batch effect for 450k arrays (Sun et al. 2011; Marabita et al. 2013; Harper et al. 2013).

Batch effects can only affect a subset of probes instead of generating artifacts globally; therefore, many normalization methods fail in eliminating or reducing batch effects. Specific methods have been developed to deal with this source of variability, including ComBat (Johnson et al. 2006), SVA (Surrogate Variable Analysis) (Leek and Storey 2007), ISVA (Independent Surrogate Variable Analysis) (Teschendorff et al. 2011), RUV (Remove Unwanted Variation) (Gagnon-Bartsch and Speed 2012; Fortin et al. 2014). The above methods aim at removing the unwanted variation that remains in high-throughput assays despite the application of between-sample normalization procedures. They rely on the explicit specification of the experimental design, in order to maintain the variability associated to a biological factor, while removing variability associated to either known or unknown batch covariates. For example, the ComBat method directly removes known batch effects and returns adjusted methylation data, by using an empirical Bayes procedure. However, when the sources of unwanted variation are unknown, surrogate variables can be identified by SVA directly from the array data. This method does not directly adjust the methylation data; however, in a second step, the latent variables can be included as covariates into a statistical model, in order to identify differential methylation while correcting for batch effect. Similarly, ISVA, an extension of SVA, does not adjust data but identifies features associated with the phenotype of interest in the presence of potential confounding factors. However, the methods indicated above may still fail or be inapplicable. Therefore, it is important to remember that the best safeguard against problematic batch effects is a careful experimental design (Leek et al. 2010), coupled with a random assignment of the samples to the arrays, the inclusion of a method to account for batch effect and possibly the presence of technical replicates, one for each processing subgroup, if the samples cannot be processed together in the case of large cohorts.

Whole blood is one of the most extensively used tissues for EWAS studies because it is easily accessible and minimally invasive, allowing large cohorts to be characterized prospectively and retrospectively, in contrast to most disease-relevant tissues that are hard to collect. However, cellular heterogeneity is an important factor to consider in the analysis of 450k data, particularly when blood is the source of DNA. In fact, cellular composition can explain a large fraction of the variability in DNA methylation (Reinius et al. 2012; Jaffe and Irizarry 2014). It can thus represent an important confounder in the association analysis when the phenotype under study alters cellular composition in blood, therefore resulting in spurious associations. Statistical methods are available to adjust for cellular composition. The popular Houseman method (Houseman et al. 2012) requires the availability of reference data measuring DNA methylation profiles for individual cell types in order to estimate cell proportions, which can be used to adjust a regression model (Liu et al. 2013). Alternatively, reference-free approaches (Zou et al. 2014; Houseman et al. 2014) can be employed to deconvolute DNA methylation when a reference data set is not available or extremely difficult to obtain.

A critical goal of most experimental designs is to identify DNA methylation changes that correlate with the phenotype of interest, for example, by comparing cases and controls. A detailed discussion of the available methods is beyond the scope of this chapter; however, we will briefly describe some of the most popular methods. We will first consider the identification of Differentially Methylated Positions (DMPs). The first and very simple approach consists in the calculation of a $\Delta\beta$ as the difference between the median β-values of two experimental groups, and selecting probes whose absolute $\Delta\beta$ exceeds a threshold. A $|\Delta\beta| > 0.2$ corresponds to the recommended difference that can be detected with 99 % confidence according to Bibikova et al. (Bibikova et al. 2011). Many works identify DMPs using a threshold on a p-value from a statistical test (t-test, Mann–Whitney test), including a correction method for multiple hypothesis testing (Bonferroni or False Discovery Rate correction). Moreover, to decrease the false positive rate, a second threshold on the effect size is recommended (Marabita et al. 2013; Dedeurwaerder et al. 2014). For example, a minimal fold change could be considered (if working with log-ratios), or a minimal difference in the β-values (5–10 %). Another popular method is represented by the moderated statistical tests as implemented in limma (Smyth 2004), which uses a moderated t-statistic and an empirical Bayes approach to shrink the estimated sample variances toward a pooled estimate across sites, resulting in better inference when the number of samples is small. In this latter case, M-values are appropriate, as limma expects log-ratios and the Gaussiainty assumption is violated by the bounded nature of β-values.

An alternative approach for feature selection consists in assessing differential variability between sites instead of using statistics based on differential methylation. In epigenomics of common diseases, this notion has been proposed to be relevant for understanding and predicting diseases (Feinberg et al. 2010; Feinberg and Irizarry 2010), by assuming that common disease involves a combination of genetic and epigenetic factors and that DNA methylation variability could either mediate genetic effects or be mediators of environmental effects. Methods are available to analyze differential variability and associate it with a phenotype of interest (Teschendorff and Widschwendter 2012; Jaffe et al. 2012a).

While the best approach for the identification of Differentially Methylated Regions (DMRs) is today represented by bisulfite sequencing, 450k arrays are a possible alternative and methods have been developed to deal with their characteristics, including Probe Lasso (Butcher and Beck 2015), Bump hunting (Jaffe et al. 2012b), DMRcate (Peters and Buckley 2014), and A-clustering (Sofer et al. 2013). The genomic coverage of 450k arrays is uneven, with a bias toward CpG islands, promoters, and genic regions; moreover, neighboring CpG sites have highly correlated methylation levels. These characteristics complicate the application of fixed window-based approaches for the identification of DMRs, and methods like Probe Lasso apply a flexible window based on probe density to call DMR and calculate a p-value by combining individual p-values, weighting by the underlying correlation structure of methylation level. The Bump hunting method (which is not restricted to 450k arrays) is another approach that was developed to deal with the spatial correlation of CpG positions, and which finds genomic regions where there is statistical evidence of an association.

1.4 Bisulfite Sequencing

Bisulfite sequencing (BS) has been thoroughly compared with other sequence- and array-based approaches (Bock et al. 2010; Harris et al. 2010; Li et al. 2010) and it currently represents the gold-standard technology for a quantitative and accurate genome-wide measurement of DNA methylation at single base-pair resolution. Although a less cost-attractive option, sequencing technologies and experimental protocols have advanced recently and it is becoming advantageous to use BS in many settings. For example, the profiling of the methylome in single cells has been recently achieved (Smallwood et al. 2014; Guo et al. 2014).

The use of next-generation sequencing has not only represented a technological improvement, but it has also contributed conceptual developments in our understanding of the biological role of DNA methylation (Rivera and Ren 2013). For example, the traditional view of DNA methylation favored a mitotically stable modification, characteristic of repressed chromatin. However, sequencing technologies have expanded our view on DNA methylation, and we have started to understand the complexity of this epigenetic modification and its dynamical patterns, the relationship with other marks (including 5-hydroxymethylcytosine (5hmC) or 5-formylcytosine (5fC)), and the distribution of non-CpG methylation in embryonic or adult tissues, for example.

Several protocol variants exist for performing genome-wide BS (Lister and Ecker 2009; Laird 2010), and here we focus on two of the most widely used, namely, WGBS and RRBS. The two strategies use bisulfite treatment to infer the methylation status of the Cs in the genome; however, they noticeably differ for their genome-wide coverage and costs. RRBS libraries are prepared by digesting genomic DNA with the methylation-insensitive restriction enzyme MspI, which cut at the CCGG sites. After end-repair and adapter ligation, DNA is size-selected and treated with sodium bisulfite. Then, purified DNA is PCR-amplified and sequenced. RRBS provides single-base resolution measurements of DNA methylation, with good coverage for CpG-rich regions (as CpG islands), but low genome-wide coverage. Therefore, this method increases the depth and reduces the cost per CpG for cytosines in CpG islands (Harris et al. 2010). Instead, WGBS has larger genome-wide coverage, but increased cost. WGBS libraries are generated from fragmented genomic DNA, which is adapter-ligated, size-selected, bisulfite-converted, and finally amplified by PCR amplification. However, modifications of this experimental workflow have been introduced in order to expand the applicability of this approach to many settings. For example, Post-Bisulfite Adaptor Tagging (PBAT) has been developed to reduce the loss of amplifiable DNA caused by degradation during bisulfite conversion, and therefore to reduce the amount of input DNA (Miura et al. 2012). Alternatively, a "tagmentation" protocol (Tn5mC-seq) allows the production of libraries from reduced amount of starting DNA (Adey and Shendure 2012).

The recent work by Ziller et al. (2013) observed that roughly only 20 % of CpG methylation in the genome can be considered "dynamic," and that therefore a substantial part of WGBS reads are potentially uninteresting, resulting in a combined loss of around 80 % of sequencing depth due to noninformative reads and static regions. Therefore, capture protocols that sequence target regions would appear to be advantageous if a flexible design could allow one to focus on representative, dynamic, or regulatory regions only. For example, the Agilent SureSelect platform allows BS on a selected panel of regions using hybridization probes (Ivanov et al. 2013; Miura and Ito 2015). The predefined regions include 3.7 M CpGs on CpG islands and promoters, cancer and tissue DMRs, DNAseI hypersensitive sites, and other regulatory elements.

The percentage of methylation after sequencing is calculated by counting the reads supporting a methylated or unmethylated C, and this is achieved by aligning reads to a reference genome. However, the bisulfite treatment converts unmethylated Cs into Ts, resulting in libraries of reduced complexity and reads that do not exactly match the reference genome sequence. Therefore, a method is needed to incorporate the possible conversion into the alignment procedure. Several alternative strategies and aligners have been proposed and their different features have been reviewed elsewhere (Bock 2012; Krueger et al. 2012; Tran et al. 2014). Bismark (Krueger and Andrews 2011), for example, represents one of the most popular mapping tools. It converts in silico all the Cs both in the reads and in the reference genome; then a standard aligner (Bowtie or Bowtie2) is used to map the reads to each strand of the genome. This method therefore uses only three letters for alignment, and the reduced complexity is compensated by the lack of bias toward methylated regions. To avoid decreased mapping efficiency, special care should be taken by an initial quality control and it is recommended to perform both sequence adapter trimming and adaptive quality trimming at the read 3′ end (Krueger et al. 2012). Indeed, some libraries may show both reduced quality scores and the presence of adapter sequences at the end of the reads (if the read length is longer than the DNA fragment), causing a dramatic decrease in the percentage of mapped reads.

After mapping, the DNA methylation levels are calculated from the aligned reads, counting the number of reads containing a C or a T in the genome, for each C independently of the context. Usually, only CpG methylation is further retained for downstream analysis; however, non-CpG methylation can be analyzed as well using BS, if required in the biological context (Lister et al. 2009, 2013). At this stage, M-bias plots (Hansen et al. 2012) can help in identifying any bias in methylation levels toward the beginning or the end of the reads. For example, many library preparation protocols include an end-repair step after DNA fragmentation. This enzymatic reaction will introduce unmethylated Cs, which will align to the genome, but without preserving their original methylation. Therefore, if detected with the M-bias plot, this effect should be removed by excluding the biased positions from the methylation call. If desired, the Bis-SNP package (Liu et al. 2012) can perform base quality recalibration, indel calling, genotyping, and methylation extraction from BS data.

Fragments aligning exactly to the same genomic position could be the result of PCR amplification. However, the execution of de-duplication step is dependent on the exact experimental protocol. For example, in RRBS libraries it is expected that a higher fraction of fragments will all start at the same genomic location, given the initial MspI enzymatic digestion, and therefore the de-duplication step could remove large fraction of valid reads. For other protocols, including WGBS and target enrichment, de-duplication is suitable to prevent multiple counting of the same fragment, which will cause methylation bias.

Similar to microarrays, the analysis of BS data allows site- and region-level differential methylation analysis. While some aspects are common to all DNA methylation studies, specific considerations and statistical tools apply only to BS data. The simplest test for assessing differential methylation is Fisher's exact test. This method uses read counts to assess statistical significance; however, it is not able to completely model the biological variability. If biological replicates are present, the counts are pooled together to apply this method, thus removing the within-group variation that is a requisite to evaluate significant difference given the observed biological differences between samples of the same group. Therefore, logistic (MethylKit (Akalin et al. 2012)) or beta-binomial (methylSig (Park et al. 2014)) models have been used to account for sampling (read coverage) and biological variability.

For the identification of DMR, the abovementioned Bump hunting method could be extended to deal with sequencing data (Jaffe et al. 2012b). Similarly, BSmooth (Hansen et al. 2012) (available through bbseq) identifies regions as groups of consecutive CpGs where an absolute score (similar to t-statistics) is above a selected threshold. The approach is based on the application of a local regression to smooth the methylation profiles using weights that are also influenced by the coverage. In this way, the algorithm improves the precision and allows the use of a lower coverage threshold, by assuming that the methylation estimates vary smoothly along the genome. This method is therefore mainly applicable for WGBS in the presence of biological replicates, from which variability is modeled. Local smoothing is also used by another algorithm, BiSeq (Hebestreit et al. 2013), which instead was developed for targeted BS approaches such as RRBS. BiSeq first finds clusters of CpGs and applies local smoothing before testing for differential methylation, using a beta regression model and a Wald test. The algorithm also provides a hierarchical method for calculating an FDR on clusters and sites, and therefore allows defining DMR boundaries.

In order to functionally annotate the discovered DMPs/DMRs, pathway or gene ontology analysis is commonly used. In a typical enrichment analysis, DMRs are first mapped to their nearest genes and then the fraction of annotated genes with a DMR for a given pathway/ontology is compared to the total fraction of genes annotated with that category in the genome. To this purpose, numerous tools are available, which use different algorithms to define enrichment (Huang et al. 2009). Otherwise, a region-based enrichment analysis for cis-regulatory regions is possible through the GREAT tool (http://great.stanford.edu/). This software defines gene regulatory domains with an adjustable "association rule" to connect a TSS (transcriptional start site) of a gene with its cis-regulatory region, such that all DMRs (or other noncoding sequences) that lie within the regulatory domain are assumed to regulate that gene. Then, a genomic region-based enrichment significance test is performed, accounting for the length of gene regulatory domains. Thus, the functional enrichment is carried using regions as input, instead of genes. This approach has been shown to improve the functional interpretation of regulatory regions (McLean et al. 2010). However, even assuming the mapping problem has been solved, it is important to remember that for regions changing in DNA methylation, there is no absolute and unequivocal link between the direction of change and the corresponding change in gene expression. For example, for promoter regions with CpG islands, a methylation event corresponds to gene repression; however, opposite examples have been reported for other regions (Jones 2012). Moreover, when the probes on the array are not evenly distributed across the genome, the use of the proper background is important not to bias the pathway/ontology enrichment analysis. For all the abovementioned reasons, care should be included in performing and interpreting functional enrichment analysis with DNA methylation data, in order to avoid potential biases.

1.5 Histone Modifications

While DNA methylation was the first uncovered epigenetic regulatory mechanism, several other mechanisms have been uncovered. Arguably, "histone modification" is among the most relevant epigenetic marks. In this third section, we provide an introduction to histone modifications, an overview of profiling experimental protocols and data analysis.

1.5.1 Introduction to Histone Modifications

Histones are key players because, through covalent modifications of their residuals (e.g., lysine), they have a crucial role in the regulation of transcription, DNA repair, and replication. These modifications are dynamically regulated by chromatin-modifying enzymes (Kouzarides 2007); an enzyme first recognizes available docking sites in histones and then recruits additional chromatin modifiers and remodeling enzymes. Enzymes are associated with specific histone modifications. During the last decades, major efforts have been devoted to the experimental, functional, and regulatory characterization of the different covalent modifications (Tollefsbol 2010). Most relevant experimental protocols and data analysis procedures are described in the next subsection. Table 1.1 summarizes the most relevant types of histone modifications such as methylation, sumoylation, ubiquitination, and acetylation.

The histone modifications selected in the ENCODE project are among the most well characterized (Consortium et al. 2012) and include H2A.Z, H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K9me1, H3K9me3, H3K27ac, H3K27me3, H3K36me3, H3K29me2, H4K20me1. For the interested reader we also recommend to consider chromatin modifications associated with nucleosome regulation (Tessarz and Kouzarides 2014; Becker and Workman 2013), the role of histone variants (Henikoff and Smith 2015) and its association to disease (Maze et al. 2014).

1.5.2 Profiling Histone Modifications: Experimental Protocol and Data Analysis

1.5.2.1 Protocol

The idea behind genome-wide histone modification profiling is the generation of DNA fragments enriched with the selected histone mark of interest. Once the DNA fragments are obtained, microarray-based or sequencing-based technologies can be applied to quantify the histone marks. The widely accepted protocol for histone mark DNA fragment enrichment detection is Chromatin Immunoprecipitation (ChIP) (Solomon et al. 1988).

Table 1.1 Histone modifications

Histone modification	Mechanism	Affected residuals	Functional role	Proteins and protein families associated
Sumoylation	Addition of a small ubiquitin-related modified protein	Lysine	Transcription repression	E1,E2,E3
Ubiquinitation	Covalent attachment of one or more ubiquitin monomers	Lysine	Transcriptional activation	E1,E2,E3, PRC1, UBP
ADP-ribosylation	Addition of a ADP-ribose moiety	Lysine	Chromatin condensation, DNA repair	ART, PARPs
Phosphorylation	Addition of a phosphate group	Serine, tyrosine	Transcription regulation (activation, repression), DNA repair	PI3K, WSTF,
Methylation	Addition of a methyl group	Arginines (mono,di), lysines (mono,di,tri)	Transcription regulation (activation, elongation, repression), DNA repair	LSD1, JMD2, JARID1
Acetylation	Addition of an acetyl functional group	Lysine	Transcriptional activation, DNA repair	HAT (GNAT, MOTYF), p300, CBP

ChIP is a powerful tool for studying protein–DNA interactions. Briefly, ChIP consists of two experimental parts:

1. *DNA–protein fragment generation*. First protein–DNA complexes are cross-linked in living cells. This is usually achieved by the addition of formaldehyde. Next, cells are lysed and chromatin is mechanically sheared in order to obtain fragments of 0.2–2 kb depending on later requirements (array or sequencing). In the context of histone modification, DNA digestion without cross-linking or sonication is preferred for fragmenting the DNA.
2. *Enrichment for selected marker*. Antibodies are used to immunoprecipitate cross-linked protein–DNA complexes enriched with a selected epitope. Then cross-links are reversed and DNA is recovered.

The DNA recovered can be then processed in two different ways:

(a) *Array-based profiling*: this technique is named ChIP-on-chip and consists in the labeling and hybridization of enriched DNA fragments to tiling DNA microarrays. ChIP-on-chip allowed the first genome-wide study of DNA–protein binding interactions (Ren et al. 2000; Blat and Kleckner 1999). Before Next-Generation Sequencing (NGS) became widely affordable, ChIP-on-chip was the standard methodology for genome-wide histone profiling. However, with the advent of sequencing technologies ChIP-on-chip has been replaced by ChIP-seq because the latter produces better signal-to-noise ratios, allows a better detection of marks (Ho et al. 2011), has higher resolution, fewer artifacts, greater coverage, and larger dynamic range (Park 2009). For this reason, for the rest of the chapter we will discuss mainly sequencing-based analysis.

(b) *Sequencing-based profiling*: similar to DNA methylation, NGS provided novel and better tools for histone genome-wide profiling. Interestingly, ChIP-seq was one of the earliest applications of NGS (Johnson et al. 2007; Barski et al. 2007). Nowadays, sample preparation kits for ChIP-seq are commercially available, thus facilitating the preparation of libraries for sequencing.

In Fig. 1.2, the ChIP-seq protocol is detailed. The outcome of ChIP-seq is a set of (millions of) DNA sequences that require processing in order to identify the regions associated to the mark of interest. ChIP-seq has been widely used for profiling histone marks, transcription factor binding, and DNA methylation; in each case, the experimental and data analysis procedures are adapted accordingly. When doing ChIP-seq, it is critical to generate a control ChIP-seq experiment (Landt et al. 2012), which is necessary to account for possible biases, because DNA digestion may not be completely uniform. Two methods for the generation of control libraries are considered: (1) "Input": DNA from the same sample is processed as any ChIP-seq library but without the immune-precipitation step; and (2) "mock ChIP-seq": DNA from the same sample is processed similarly but using instead a "control antibody" expected to react only with an irrelevant nonnuclear antigen. Several works have shown the benefits of using control libraries (Landt et al. 2012; Liang and Keleş 2012); interestingly, the possibility of using immunoprecipitation of histone H3 as a background has been proposed (Flensburg et al. 2014).

1.5.2.2 Data Analysis

The aim of data analysis is to identify the genomic regions associated with the mark of interest. In the case of ChIP-on-chip, Negre et al. (2006) and Huebert et al. (2006) provide an integrated overview of experimental procedures and data analysis methods while Benoukraf et al. (2009) provide an analysis suite for ChIP-on-chip data analysis.

In the case of ChIP-seq, the starting material is a set of (millions of) DNA sequences and for each sequence, a string of quality score for each base. The analysis of ChIP-seq data involves several steps, some of which are shared among several NGS-based data analysis pipelines:

Step (1) Quality Control The first step of the analysis is to assess the quality of the data from the set of sequences. Several tools do exist, but arguably FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) from Babraham Institute and Picard (http://broadinstitute.github.io/picard/) from the Broad Institute are two of the most common. The most relevant quality measures are shown in Fig. 1.3 and

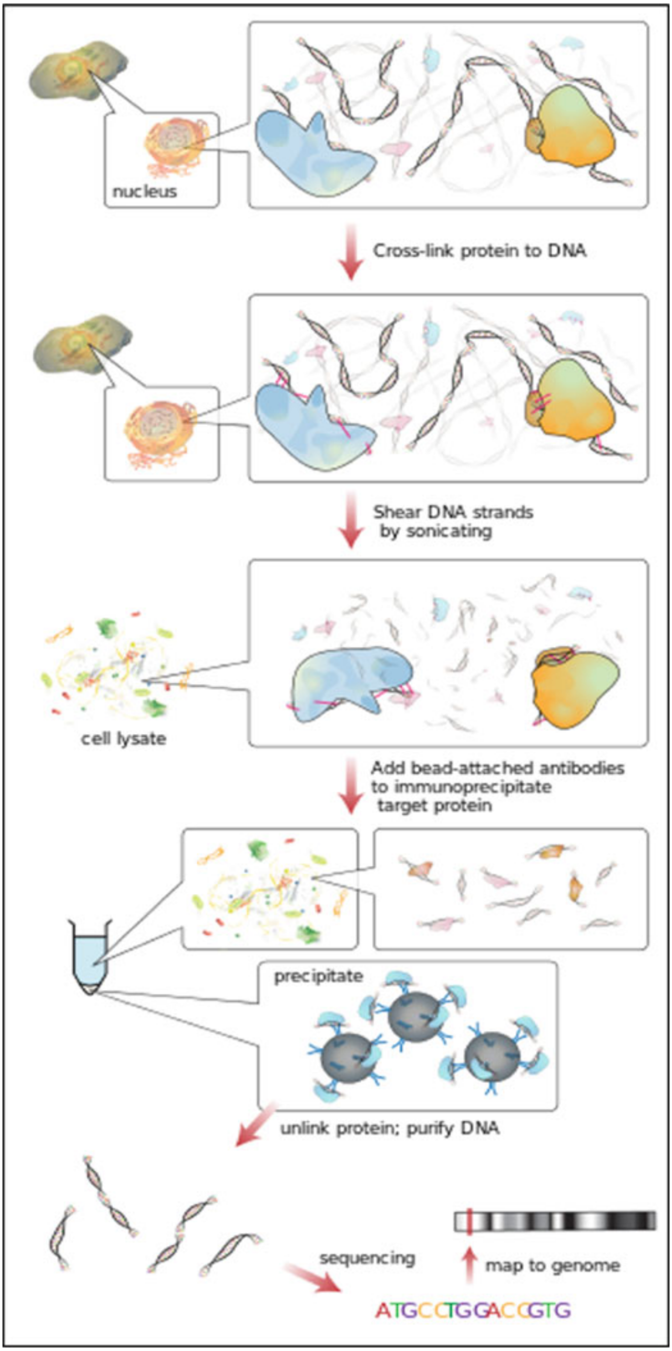


Fig. 1.2 ChIP-seq protocol. The figure depicts the different experimental steps of ChIP-seq as described in the text (Figure generated by Jkwchiu under CreativeCommons3.0)

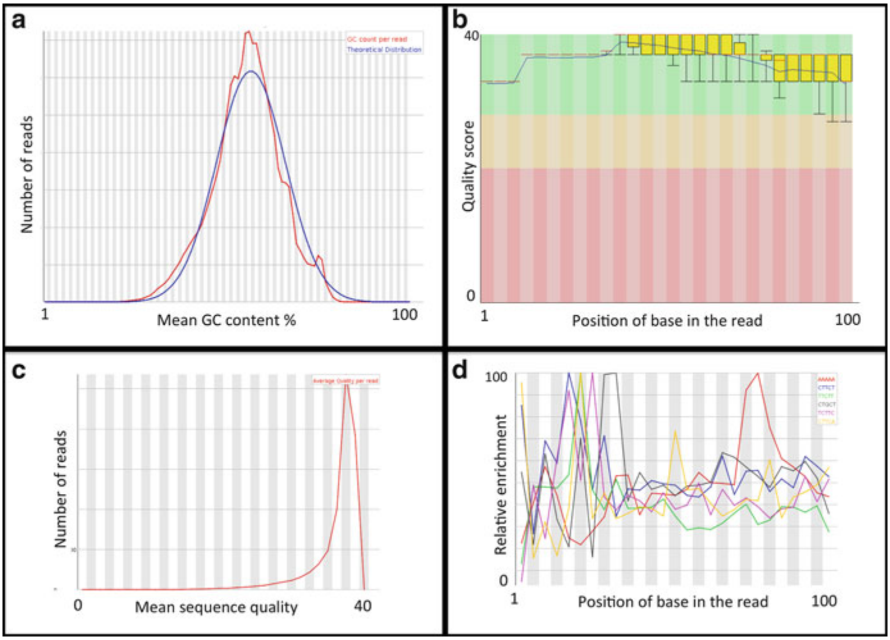


Fig. 1.3 Example of FastQC output. (**a**) Per sequence GC content: expected (*blue*) versus observed (*red*). *Y*-axis presents the number of reads and *X*-axis shows the mean GC content. (**b**) Quality scores across all bases. *Y*-axis presents the quality score and *X*-axis shows position in the read. (**c**) Quality score distribution over sequences. *Y*-axis presents the number of reads and *X*-axis shows the mean sequence quality. (**d**) K-mer enrichment. *Y*-axis presents the relative enrichment of a k-mer and *X*-axis shows the position in the read

briefly described here. We highly recommend the reader to visit the online tutorial material of mentioned tools; some of the measures explained below are estimated differently by the different tools.

1. *Percentage of duplications*: percentage of sequences that are not unique in the set of DNA sequences. An elevated duplication level may argue for PCR artifacts or DNA contamination.
2. *Per sequence GC content*: the level of GC content is expected to be similar to that of the entire genome. This is not true when considering DNA methylation but is commonly considered valid when doing histone mark analysis. An example is provided in Fig. 1.3a.
3. *Quality scores across all bases*: we observe that the base quality score is degraded in the last bases and this is expected because sequencing chemistry degrades with increased read length. However, when the quality is below a certain threshold (e.g., median for any base below 25) the quality of the sequences is under question. Figure 1.3b provides the distribution of quality score along reads, while Fig. 1.3c provides the density of the median quality score.

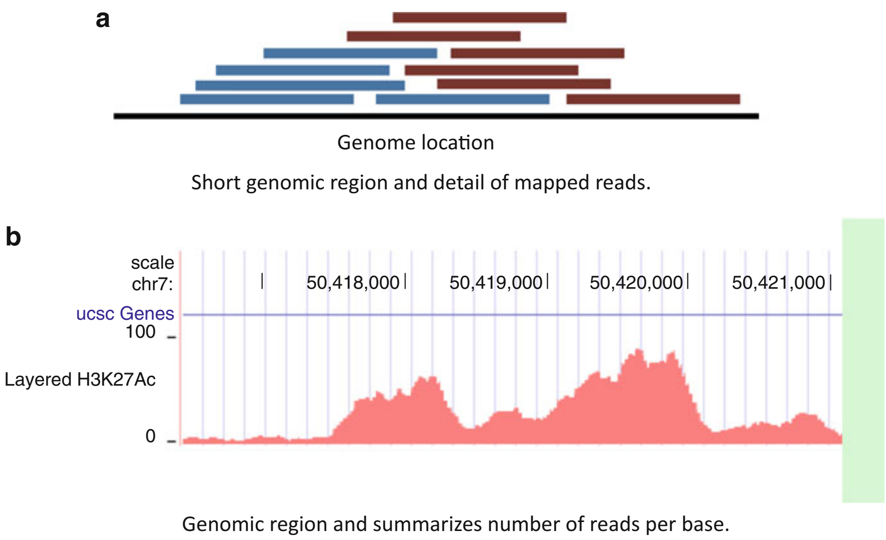


Fig. 1.4 ChIP-seq mapped data. The figure presents an example of mapped reads into the genome. (**a**) Short genomic region where it is superposed the mapping of several reads to be positive (*blue*) or negative (*red*) strand. (**b**) USCS display of a genomic region where the enrichment of H3K27Ac enrichment score is shown at a base resolution

4. *k-mer content*: investigate if a k-mer (a sequence of length k) is overrepresented at different locations of the sequences. It is usual to investigate if the initial part of the sequences contains an overrepresentation of adapter sequences, because those will require trimming. Figure 1.3d provides an overview.

FastQC provides certain thresholds to raise warnings on the different quality controls; however, as important as those thresholds is that quality measures are homogeneous among all samples under consideration. In addition, software that provides comprehensive quality controls on ChIP-seq data includes CHANCE (Diaz et al. 2012) or even user-friendly tools such as CLCbio software (CLCbio 2014). Additionally, ChIPQC Bioconductor package provides an R-based tool for quality metrics generation of ChIP-seq data (Carroll et al. 2014b).

Step (2) Mapping to a reference genome The next task is to map reads (sequences) to the genome of reference. To this end, several softwares exist, with one of the most widely used tools being, arguably, bowtie2 (Langmead and Salzberg 2012). The general output of a mapper assigns each read to a genomic location and a quality of the mapping; a summarized example is depicted in Fig. 1.4a, where reads are mapped to genomic regions, either to the positive or negative strand. It is always recommended to, at least, visually investigate selected regions; among the visualization tools UCSC Genome Browser (Karolchik et al. 2014) and Broad Institute's IGV (Robinson et al. 2011) are commonly used; we also recommend the use of Bioconductor package tracktables (Carroll et al. 2014a) to generated customized visualizations and dynamic IGV reports. Figure 1.4b provides an example of a UCSC Genome Browser histone mark summary visualization where for each base the enrichment is provided. Since ChIP-seq reads are sequenced from both ends of a signal, the positive and negative strands will enrich each one at different ends; for this reason, the signal is to be considered bimodal when considering positive and negative strands simultaneously (Zhang et al. 2008). We denote the distance between those ends as d.

Step (3) Peak Identification Histone marks are identified across the genome as "peaks." Figure 1.4b provides an intuitive idea of the signal as a peak: we are interested in finding genomic regions where several consecutive bases show significant signal enrichment. In Fig. 1.4b, the left part of the signal (pink) shows low enrichment while to the right there are several regions (peaks) with higher enrichment. Many algorithms, usually called peak-finders, have been developed in order to identify significant enriched regions (peaks) from sequencing data. It is out of the scope of this chapter to present a comprehensive review of them, but we shortly characterize them:

- *Generic peak-finders*: among the first peak-finder algorithms, the most successful one was the Model-based analysis of ChIP-seq data (MACS) algorithm from Zhang et al. (2008), which uses some concepts inherited from algorithms developed for ChIP-on-chip data analysis. Briefly, MACS first performs a linear scaling of the control library to be the same as the signal ChIP-seq library. Subsequently, MACS models the distribution of the number of reads per base as a Poisson distribution and then considers all reads a $d/2$ number of bases across the genome. Finally, a search for significantly enriched regions is conducted through a sliding window of $2*d$ size. The use of a control library allows FDR estimation. In Bailey et al. (2013), a discussion among current methodologies is provided.
- *Histone-specific peak-finder*: many of the methodologies developed considered the peaks of interest to be narrow, such as those observed from most Transcription Factor (TF) ChIP-seq data. However, in the case of histone marks, histone modification enzymes, chromatin remodeling complexes, or RNAPII, we expect a spreading of the signal over larger regions; those are defined as broad-source factors by Landt et al. (2012). For this reason, methodologies such as SICER (Zang et al. 2009), which aims at the identification of statistically significant spatial cluster of signals, were developed. Methods aiming at uncovering both broad and narrow peaks also exist (Peng and Zhao 2011), which would be optimal for mixed-source factors, that is, marks that can be broad or narrow.

The output of most peak-finders provides similar type of information. Most common outputs include the following:

- *Genomic location*: chromosome, start and ending site.
- *Summit*: in many cases, the base of the peak with the highest enrichment is also identified.

- *Signal strength*: number of reads or number of reads per million are also usually provided.
- *Statistical significance*: *p*-values and FDRs are provided. This allows the use of different thresholds during follow-up analyses.

In Landt et al. (2012), and as part of the ENCODE consortia, the authors recommend the use of ChIP-seq replicates; it is recommended to generate a control library for every chromatin preparation and sonication batch. When more than one library is prepared and analyzed against the control, we will obtain a set of peaks for every replication; in those cases, irreproducible discovery rate (IDR) (Li et al. 2011) allows assessing agreement between replicates and also provides FDR estimates for peaks.

Step (4) Peak analysis Once signals have been uncovered, many follow-up analyses are possible. We enumerate the most common ones:

- *Motif discovery*: denotes the identification of transcription factor binding sites in peaks. When applied to TF ChIP-seq, it allows the uncovering of associated TF motifs; however, when applied to histone marks, the identification of motifs and its characterization through motif databases such as TRANSFAC (Matys et al. 2006) or JASPAR (Portales-Casamar et al. 2010) may provide insights into histone-associated TFs for the system under study. MD tools are Homer (Heinz et al. 2010), MEME suite (Bailey et al. 2009), CisFinder (Sharov and Ko 2009), and rGADEM (Droit et al. 2014) among many others. Tran and Huang (2014) is a recent survey on MD web tools.
- *Pathway enrichment analysis*: similarly to gene expression analysis, it is important to reveal if the signal from the peaks can be associated, for instance, with specific pathways, diseases, or gene ontology terms. Mapping peaks to genes and then applying classical gene set analyses is an option. However, this option may not be optimal because biases are introduced by gene length (higher probability of having peaks) or by peaks from intergenic regions (such peaks may be associated with genes 10–20 kb away, and which are therefore possibly not closest to the peak itself). ChIP-Enrich (Welch et al. 2014) was developed to correct for gene lengths, while GREAT (McLean et al. 2010) introduces different definitions of gene domains to correct for the uncertainty of the gene-peak mappings.
- *Mapping to genes*: because different histone marks may act at different genomic locations, the characterization of peaks as being intergenic, or associated with promoter, gene body, intron, exon, or start/end of the gene (among others) may also provide insights into a histone's genomic location preferences and association mechanisms. ENCODE provided relevant examples of such characterization in (ENCODE Project Consortium et al. 2007).

1.6 Repositories and Other Resources

A common task in current data analysis is represented by the integration of different public available data with own experimental data. A first use is, for instance, the overlap of a given histone mark, that is, H3K4me1, in a specific system, that is, CD4 T cells, with H3K4me1 profiles of other cell or tissue types. A second possible usage is to conduct integrative analysis with different epigenetic marks in order to gain functional insight into the regulatory network that is active in the studied biological process.

Typically, histone marks are analyzed in combination with gene expression or DNA methylation data. Furthermore, the researcher has now the availability of a growing selection of epigenomic data (Table 1.2), produced by several international consortia and projects. The size of the epigenomic data sets and publications has grown a lot in recent years, resulting in the availability of different data types that are essential to define the function of the regions under study, and which can be visualized using online or local tools (Table 1.3).

Large data sets allow the research of regulatory mechanisms, impossible to perform in smaller samples. For instance, the idea that histone marks act in a combinatorial manner was considered by different researchers when ChIP-on-chip experiments were first generated; the ENCODE's pilot project (Thurman et al. 2007) identified higher-order patterns of active and repressed functional domains in human chromatin, through the integration of histone modifications, RNA output, and DNA replication timing. Only when Zho's laboratory generated ChIP-seq data for several histones and for the same system (CD4+ T cells) was it possible to obtain more robust insights into the cooperation among histone marks (Wang et al. 2008). Interestingly, Karlic et al. (2010) showed that specific combination of histone marks was predictive of gene expression; later the prediction of gene expression was also conducted in new ENCODE data by Dong et al. (2012). Histone acetylation dynamics were also investigated by Zho's laboratory by profiling HDACs and HATs again in CD4+ T cells.

Over the years, the ENCODE project has generated larger sets of histone mark profiles for several histone marks and several cell types. Interestingly, the generation of such large data sets motivated the use of unsupervised learning methods (Hoffman et al. 2013; Ernst and Kellis 2010; Ernst et al. 2011) in order to identify functional regions and classify them into a small number of labels. In the analysis, data from histone modifications, DNase-seq, FAIRE, RNA polymerase 2, and CTCF were considered. Labels were annotated in a post hoc analysis step; those were further summarized into summary states (Transcription Start Site, Promoter Flanking, Enhancer, Weak Enhancer, CTCF binding, Transcribed Region, and Repressed or Inactive Region).

Table 1.2 Epigenome projects and other data repositories

Name	Description	URL
ENCODE	A project aimed at identifying functional elements in the human genome. Assays include: ChIP-seq, RNA-seq, DNase-seq, gene expression arrays, 450k arrays, RRBS, Repli-seq, CAGE, Genotype, RNA Bind-n-Seq, WGBS, FAIRE-seq, RAMPAGE, RIP-chip, RNA-PET, Repli-chip, MRE-seq, ChIA-PET, protein sequencing by tandem mass spectrometry, 5C, and more. Samples include mainly immortalized cell lines but also tissues and primary cells	https://www.encodeproject.org/
NIH Roadmap Epigenomics	A collection of normal epigenomes to provide a reference for the normal counterparts of tissues and organ systems frequently involved in human disease. Assays include DNA methylation (MeDIP-Seq, MRE-Seq, RRBS, WGBS), histone modifications (ChIP-seq), chromatin accessibility (DNase-seq), and RNA expression (mRNA-Seq, smRNA-Seq). Samples include Embryonic Stem Cells and primary ex vivo fetal and adult tissues	http://www.roadmapepigenomics.org/
Blueprint Epigenome	A project focused on obtaining reference epigenomes from cells of the hematopoietic system. Assays include: RNA expression (RNA-seq), DNA methylation (WGBS), chromatin accessibility (DNase-seq), and histone modifications (ChIP-seq). Samples include primary cells from healthy individuals and patients (hematopoietic neoplasias, chronic autoimmune diseases, type 1 diabetes)	http://www.blueprint-epigenome.eu/
International Human Epigenome Consortium	A consortium with goal of providing access to human epigenomes and coordinate their production for key cellular states relevant to health and diseases. It gathers data from different projects (Blueprint, CEEHRC, CREST/IHEC, DEEP, ENCODE, NIH Roadmap)	http://www.ihec-epigenomes.org/
GEO	Public data repository of high-throughput genomic data, including array- and sequence-based assays	http://www.ncbi.nlm.nih.gov/geo/

(continued)

Table 1.2 (continued)

Name	Description	URL
The Cancer Genome Atlas (TCGA)	A project focusing on cancer genomics, with the primary goal of understanding the molecular basis of cancer. Assays include: genome and exome sequencing, DNA methylation (450k array and WGBS), gene expression (mRNA-seq, miRNA-seq, Total RNA-seq, arrays), CNV (arrays and DNA-seq), protein expression, and more	http://cancergenome.nih.gov/
FANTOM	Although not an epigenomic project in the strict sense, it focuses on transcriptome analysis toward an understanding of the transcriptional regulatory network and the identification of functional elements in mammalian genomes. The FANTOM5 phase used CAGE to map the sets of transcripts, transcription factors, promoters, and enhancers active in diverse mammalian primary cell types	http://fantom.gsc.riken.jp/

Table 1.3 Genome browsers and other software tools for the visualization and the analysis of epigenomes

Name	Description	URL
UCSC Genome Browser	Online genomic browser that contains the reference sequences of a large collection of genomes. It also provides access to ENCODE data. Both the browser and the data can be downloaded for local runs	http://genome.ucsc.edu/
WashU Epigenome Browser	Online genomic browser that provides access and visualization of ENCODE, NIH Roadmap, and other data. Several visualizations are available	http://epigenomegateway.wustl.edu/browser/
Roadmap Epigenome Browser	Online genomic browser providing visualization of NIH Roadmap assays	http://epigenomegateway.wustl.edu/browser/roadmap/
IGV	Integrative Genomics Viewer that can be downloaded and run locally for interactive exploration of large genomic data sets. Java Web Start or binary download are available	http://www.broadinstitute.org/igv/
IGB	Integrated Genome Browser that can be downloaded and run locally for interactive exploration of large genomic data sets. Java Web Start or binary download are available	http://bioviz.org/igb/
Ensembl Genome Browser	Online genomic browser that contains the reference sequences of a large collection of genomes. It also provides access to ENCODE data. Both the browser and the data can be downloaded for local run	http://www.ensembl.org/index.html
Galaxy	Web-based application to analyze genomic data. Custom data can be uploaded and a web interface is used to execute command line applications. It provides direct access to ENCODE data through UCSC table browser. It can be downloaded and installed locally	http://galaxyproject.org/

1.7 Conclusions

We have presented a brief overview of epigenomics and provided the newcomer with information of available tools for the analysis of epigenomic data sets. However, the methodologies are in continuous development especially in the context of data integration.

References

Adey A, Shendure J. Ultra-low-input, tagmentation-based whole-genome bisulfite sequencing. Genome Res. 2012;22(6):1139–43.

Akalin A, et al. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. Genome Biol. 2012;13(10):R87.

Anderson JD, Widom J. Sequence and position-dependence of the equilibrium accessibility of nucleosomal DNA target sites. J Mol Biol. 2000;296(4):979–87.

Aran D, Hellman A. DNA methylation of transcriptional enhancers and cancer predisposition. Cell. 2013;154(1):11–3.

Aran D, Sabato S, Hellman A. DNA methylation of distal regulatory sites characterizes dysregulation of cancer genes. Genome Biol. 2013;14(3):R21.

Aryee MJ, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics (Oxford, Engl). 2014;30(10):1363–9.

Assenov Y, et al. Comprehensive analysis of DNA methylation data with RnBeads. Nat Methods. 2014;11:1138–40.

Bailey T, et al. Practical guidelines for the comprehensive analysis of ChIP-seq data. PLoS Comput Biol. 2013;9(11):e1003326.

Bailey TL, et al. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 2009;37(Web Server issue):W202–8.

Barrès R, et al. Acute exercise remodels promoter methylation in human skeletal muscle. Cell Metab. 2012;15(3):405–11.

Barski A, Cuddapah S, Cui K, Roh T-Y, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K. High-resolution profiling of histone methylations in the human genome. Cell. 2007;129(4):823–37. doi:10.1016/j.cell.2007.05.009.

Becker PB, Workman JL. Nucleosome remodeling and epigenetics. Cold Spring Harb Perspect Biol. 2013;5(9). pii: a017905.

Bell JT, et al. DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. Genome Biol. 2011;12(1):R10.

Benoukraf T, et al. CoCAS: a ChIP-on-chip analysis suite. Bioinformatics (Oxford, Engl). 2009;25(7):954–5.

Bibikova M, et al. High density DNA methylation array with single CpG site resolution. Genomics. 2011;98(4):288–95.

Bjornsson HT, et al. Intra-individual change over time in DNA methylation with familial clustering. JAMA. 2008;299(24):2877–83.

Blat Y, Kleckner N. Cohesins bind to preferential sites along yeast chromosome III, with differential regulation along arms versus the centric region. Cell. 1999;98(2):249–59.

Bock C. Analysing and interpreting DNA methylation data. Nat Rev Genet. 2012;13(10):705–19.

Bock C, et al. Quantitative comparison of genome-wide DNA methylation mapping technologies. Nat Biotechnol. 2010;28(10):1106–14.

Butcher LM, Beck S. Probe Lasso: A novel method to rope in differentially methylated regions with 450K DNA methylation data. Methods. 2015;72:21–8.

Calo E, Wysocka J. Modification of enhancer chromatin: what, how, and why? Mol Cell. 2013;49(5):825–37.

Carroll T, et al. tracktables: build IGV tracks and HTML reports. R package version 1.0.0; 2014a.

Carroll TS, et al. Impact of artifact removal on ChIP quality metrics in ChIP-seq and ChIP-exo data. Front Genet. 2014b;5:75.

CLCbio, CLC shape-based peak caller. White paper. 2014. http://www.clcbio.com/files/whitepapers/whitepaper-chip-seq-analysis.pdf.

Consortium, T.E.P. et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;488(7414):57–74.

Davis S, et al. methylumi: Handle Illumina methylation data. R package version 2.12.0; 2014.

Deaton AM, Bird A. CpG islands and the regulation of transcription. Genes Dev. 2011;25(10):1010–22.

Dedeurwaerder S, et al. Evaluation of the Infinium Methylation 450K technology. Epigenomics. 2011;3(6):771–84.

Dedeurwaerder S, et al. A comprehensive overview of Infinium HumanMethylation450 data processing. Brief Bioinform. 2014;15:929–41.

Diaz A, Nellore A, Song JS. CHANCE: comprehensive software for quality control and validation of ChIP-seq data. Genome Biol. 2012;13(10):R98.

Dong X, et al. Modeling gene expression using chromatin features in various cellular contexts. Genome Biol. 2012;13(9):R53.

Down TA, et al. A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. Nat Biotechnol. 2008;26(7):779–85.

Droit A, et al. rGADEM: de novo motif discovery. R package version 2.14.0; 2014.

Drong AW, et al. The presence of methylation quantitative trait loci indicates a direct genetic influence on the level of DNA methylation in adipose tissue. PLoS One. 2013;8(2):e55923.

Du P, et al. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics. 2010;11:587.

Eichler EE, et al. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet. 2010;11(6):446–50.

ENCODE Project Consortium, et al. Identification and analysis of functional elements in 1 % of the human genome by the ENCODE pilot project. Nature. 2007;447(7146):799–816.

Ernst J, Kellis M. Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat Biotechnol. 2010;28(8):817–25.

Ernst J, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature. 2011;473(7345):43–9.

Fei J, Ha T. Watching DNA breath one molecule at a time. Proc Natl Acad Sci U S A. 2013;110(43):17173–4.

Feil R, Fraga MF. Epigenetics and the environment: emerging patterns and implications. Nat Rev Genet. 2011;13(2):97–109.

Feinberg AP. Phenotypic plasticity and the epigenetics of human disease. Nature. 2007;447(7143):433–40.

Feinberg AP, Irizarry RA. Evolution in health and medicine Sackler colloquium: stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. Proc Natl Acad Sci U S A. 2010;107 Suppl 1:1757–64.

Feinberg AP, et al. Personalized epigenomic signatures that are stable over time and covary with body mass index. Sci Transl Med. 2010;2(49):49ra67.

Flensburg C, et al. A comparison of control samples for ChIP-seq of histone modifications. Front Genet. 2014;5:329.

Fortin J-P, et al. Functional normalization of 450 k methylation array data improves replication in large cancer studies. Genome Biol. 2014;15(11):503.

Fraga MF, et al. Epigenetic differences arise during the lifetime of monozygotic twins. Proc Natl Acad Sci U S A. 2005;102(30):10604–9.

Fyodorov DV, Kadonaga JT. The many faces of chromatin remodeling: SWItching beyond transcription. Cell. 2001;106(5):523–5.

Gagnon-Bartsch JA, Speed TP. Using control genes to correct for unwanted variation in microarray data. Biostatistics. 2012;13(3):539–52.
Guo H, et al. The DNA methylation landscape of human early embryos. Nature. 2014;511(7511):606–10.
Hannum G, et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. Mol Cell. 2013;49(2):359–67.
Hansen KD, Langmead B, Irizarry RA. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. Genome Biol. 2012;13(10):R83.
Harper KN, Peters BA, Gamble MV. Batch effects and pathway analysis: two potential perils in cancer studies involving DNA methylation array analysis. Cancer Epidemiol Biomarkers Prev. 2013;22(6):1052–60.
Harris RA, et al. Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. Nat Biotechnol. 2010;28(10):1097–105.
Hebestreit K, Dugas M, Klein H-U. Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. Bioinformatics (Oxford, Engl). 2013;29(13):1647–53.
Heinz S, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010;38(4):576–89.
Henikoff S, Smith MM. Histone variants and epigenetics. Cold Spring Harb Perspect Biol. 2015;7(1):a019364.
Ho JWK, et al. ChIP-chip versus ChIP-seq: lessons for experimental design and data analysis. BMC Genomics. 2011;12:134.
Hoffman MM, et al. Integrative annotation of chromatin elements from ENCODE data. Nucleic Acids Res. 2013;41(2):827–41.
Holliday R, Pugh JE. DNA modification mechanisms and gene activity during development. Science (New York, NY). 1975;187(4173):226–32.
Horvath S. DNA methylation age of human tissues and cell types. Genome Biol. 2013;14(10):R115.
Horvath S, et al. Obesity accelerates epigenetic aging of human liver. Proc Natl Acad Sci U S A. 2014;111(43):15538–43.
Houseman EA, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics. 2012;13:86.
Houseman EA, Molitor J, Marsit CJ. Reference-free cell mixture adjustments in analysis of DNA methylation data. Bioinformatics (Oxford, Engl). 2014;30(10):1431–9.
Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009;37(1):1–13.
Huebert DJ, et al. Genome-wide analysis of histone modifications by ChIP-on-chip. Methods. 2006;40(4):365–9.
Illingworth RS, Bird AP. CpG islands–'a rough guide'. FEBS Lett. 2009;583(11):1713–20.
Irizarry RA, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 2003;4(2):249–64.
Irizarry RA, et al. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nat Genet. 2009;41(2):178–86.
Ivanov M, et al. In-solution hybrid capture of bisulfite-converted DNA for targeted bisulfite sequencing of 174 ADME genes. Nucleic Acids Res. 2013;41(6):e72.
Jaffe AE, Irizarry RA. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. Genome Biol. 2014;15(2):R31.
Jaffe AE, Feinberg AP, et al. Significance analysis and statistical dissection of variably methylated regions. Biostatistics. 2012a;13(1):166–78.
Jaffe AE, Murakami P, et al. Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. Int J Epidemiol. 2012b;41(1):200–9.
Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2006;8(1):118–27.

Johnson DS, et al. Genome-wide mapping of in vivo protein-DNA interactions. Science (New York, NY). 2007;316(5830):1497–502.
Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet. 2012;13(7):484–92.
Karlic R, et al. Histone modification levels are predictive for gene expression. Proc Natl Acad Sci U S A. 2010;107(7):2926–31.
Karolchik D, et al. The UCSC Genome Browser database: 2014 update. Nucleic Acids Res. 2014;42(Database issue):D764–70.
Kornberg RD. Chromatin structure: a repeating unit of histones and DNA. Science (New York, NY). 1974;184(4139):868–71.
Kouzarides T. Chromatin modifications and their function. Cell. 2007;128(4):693–705.
Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics (Oxford, Engl). 2011;27(11):1571–2.
Krueger F, et al. DNA methylome analysis using short bisulfite sequencing data. Nat Methods. 2012;9(2):145–51.
Laird PW. Principles and challenges of genome-wide DNA methylation analysis. Nat Rev Genet. 2010;11(3):191–203.
Landt SG, et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res. 2012;22(9):1813–31.
Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357–9.
Lee KWK, Pausova Z. Cigarette smoking and DNA methylation. Front Genet. 2013;4:132.
Lee E-J, et al. Targeted bisulfite sequencing by solution hybrid selection and massively parallel sequencing. Nucleic Acids Res. 2011;39(19):e127.
Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. 2007;3(9):1724–35.
Leek JT, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet. 2010;11(10):733–9.
Li N, et al. Whole genome DNA methylation analysis based on high throughput sequencing technology. Methods. 2010;52(3):203–12.
Li Q, et al. Measuring reproducibility of high-throughput experiments. Ann Appl Stat. 2011;5(3):1752–79.
Liang K, Keleş S. Normalization of ChIP-seq data with control. BMC Bioinformatics. 2012;13:199.
Lim U, Song M-A. Dietary and lifestyle factors of DNA methylation. In: Methods in molecular biology. Methods in molecular biology. Totowa: Humana Press; 2012. p. 359–76. Available at: http://link.springer.com/10.1007/978-1-61779-612-8_23.
Lindholm ME, et al. An integrative analysis reveals coordinated reprogramming of the epigenome and the transcriptome in human skeletal muscle after training. Epigenetics. 2014;9(12):1557–69.
Lister R, Ecker JR. Finding the fifth base: genome-wide sequencing of cytosine methylation. Genome Res. 2009;19(6):959–66.
Lister R, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature. 2009;462(7271):315–22.
Lister R, et al. Global epigenomic reconfiguration during mammalian brain development. Science (New York, NY). 2013;341(6146):1237905.
Liu Y, et al. Bis-SNP: combined DNA methylation and SNP calling for Bisulfite-seq data. Genome Biol. 2012;13(7):R61.
Liu Y, et al. Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. Nat Biotechnol. 2013;31(2):142–7.
Maksimovic J, Gordon L, Oshlack A. SWAN: subset quantile within-array normalization for illumina infinium HumanMethylation450 BeadChips. Genome Biol. 2012;13(6):R44.

Marabita F, et al. An evaluation of analysis pipelines for DNA methylation profiling using the illumina HumanMethylation450 BeadChip platform. Epigenetics. 2013;8(3):333–46.

Matys V, et al. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res. 2006;34(Database issue):D108–10.

Maunakea AK, et al. Conserved role of intragenic DNA methylation in regulating alternative promoters. Nature. 2010;466(7303):253–7.

Maze I, et al. Every amino acid matters: essential contributions of histone variants to mammalian development and disease. Nat Rev Genet. 2014;15(4):259–71.

McLean CY, et al. GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol. 2010;28(5):495–501. pp.nbt.1630–9.

Meissner A, et al. Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature. 2008;454(7205):766–70.

Miura F, Ito T. Highly sensitive targeted methylome sequencing by post-bisulfite adaptor tagging. DNA Res. 2015;22:13–8.

Miura F, et al. Amplification-free whole-genome bisulfite sequencing by post-bisulfite adaptor tagging. Nucleic Acids Res. 2012;40(17):e136.

Morris TJ, Beck S. Analysis pipelines and packages for Infinium Human Methylation 450 BeadChip (450k) data. Methods. 2015;72:3–8.

Morris TJ, et al. ChAMP: 450k chip analysis methylation pipeline. Bioinformatics. 2014;30:428–30.

Negre N, et al. Mapping the distribution of chromatin proteins by ChIP on chip. Methods Enzymol. 2006;410:316–41.

Park PJ. ChIP–seq: advantages and challenges of a maturing technology. Nat Rev Genet. 2009;10(10):669–80.

Park Y, et al. methylSig: a whole genome DNA methylation analysis pipeline. Bioinformatics (Oxford, Engl). 2014;30:2414–22.

Peng W, Zhao K. An integrated strategy for identification of both sharp and broad peaks from next-generation sequencing data. Genome Biol. 2011;12(7):120.

Peters T, Buckley M. DMRcate: illumina 450K methylation array spatial analysis methods. R package version 1.2.0; 2014.

Petronis A. Epigenetics as a unifying principle in the aetiology of complex traits and diseases. Nature. 2010;465(7299):721–7.

Pidsley R, et al. A data-driven approach to preprocessing illumina 450K methylation array data. BMC Genomics. 2013;14(1):293.

Portales-Casamar E, et al. JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. Nucleic Acids Res. 2010;38(Database issue):D105–10.

Reinius LE, et al. Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility A. H. Ting, ed. PLoS One. 2012;7(7):e41361.

Ren B, et al. Genome-wide location and function of DNA binding proteins. Science (New York, NY). 2000;290(5500):2306–9.

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

Rivera CM, Ren B. Mapping human epigenomes. Cell. 2013;155(1):39–55.

Robinson JT, et al. Integrative genomics viewer. Nat Biotechnol. 2011;29(1):24–6.

Rönn T, et al. A six months exercise intervention influences the genome-wide DNA methylation pattern in human adipose tissue J. M. Greally, ed. PLoS Genet. 2013;9(6):e1003572.

Rönnerblad M, et al. Analysis of the DNA methylome and transcriptome in granulopoiesis reveals timed changes and dynamic enhancer methylation. Blood. 2014;123(17):e79–89.

Sandoval J, et al. Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. Epigenetics. 2011;6(6):692–702.

Schalkwyk LC, et al. wateRmelon: Illumina 450 methylation array normalization and metrics. R package version 1.5.1; 2013.

Serre D, Lee BH, Ting AH. MBD-isolated Genome Sequencing provides a high-throughput and comprehensive survey of DNA methylation in the human genome. Nucleic Acids Res. 2010;38(2):391–9.
Sharov AA, Ko MSH. Exhaustive search for over-represented DNA sequence motifs with CisFinder. DNA Res. 2009;16(5):261–73.
Shi J, et al. Characterizing the genetic basis of methylome diversity in histologically normal human lung tissue. Nat Commun. 2014;5:3365.
Smallwood SA, et al. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. Nat Methods. 2014;11(8):817–20.
Smyth GK. Linear models and empirical Bayes methods for assessing differential expression in 1053 microarray experiments. Stat Appl Genet Mol Biol. 2004;3(1):1–25.
Sofer T, et al. A-clustering: a novel method for the detection of co-regulated methylation regions, and regions associated with exposure. Bioinformatics. 2013;29:2884–91.
Solomon MJ, Larsen PL, Varshavsky A. Mapping protein-DNA interactions in vivo with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene. Cell. 1988;53(6):937–47.
Stadler MB, et al. DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature. 2011;480(7378):490–5.
Sun Z, et al. Batch effect correction for genome-wide methylation data with Illumina Infinium platform. BMC Med Genomics. 2011;4(1):84.
Teschendorff AE, Widschwendter M. Differential variability improves the identification of cancer risk markers in DNA methylation studies profiling precursor cancer lesions. Bioinformatics (Oxford, Engl). 2012;28(11):1487–94.
Teschendorff AE, Zhuang J, Widschwendter M. Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies. Bioinformatics (Oxford, Engl). 2011;27(11):1496–505.
Teschendorff AE, Marabita F, et al. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450k DNA methylation data. Bioinformatics (Oxford, Engl). 2013a;29(2):189–96.
Teschendorff AE, West J, Beck S. Age-associated epigenetic drift: implications, and a case of epigenetic thrift? Hum Mol Genet. 2013b;22(R1):R7–15.
Tessarz P, Kouzarides T. Histone core modifications regulating nucleosome structure and dynamics. Nat Rev Mol Cell Biol. 2014;15(11):703–8.
Thurman RE, et al. Identification of higher-order functional domains in the human ENCODE regions. Genome Res. 2007;17(6):917–27.
Tollefsbol T, editor. Handbook of epigenetics. San Diego: Academic; 2011.
Touleimat N, Tost J. Complete pipeline for Infinium ®Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. Epigenomics. 2012;4(3):325–41.
Tran NTL, Huang C-H. A survey of motif finding Web tools for detecting binding site motifs in ChIP-Seq data. Biol Direct. 2014;9:4.
Tran H, et al. Objective and comprehensive evaluation of bisulfite short read mapping tools. Adv Bioinformatics. 2014;2014:472045.
Wang ZB, et al. Combinatorial patterns of histone acetylations and methylations in the human genome. Nat Genet. 2008;40:897–903.
Welch RP, et al. ChIP-Enrich: gene set enrichment testing for ChIP-seq data. Nucleic Acids Res. 2014;42(13):e105.
Wiench M, et al. DNA methylation status predicts cell type-specific enhancer activity. EMBO J. 2011;30(15):3028–39.
Wilhelm-Benartzi CS, et al. Review of processing and analysis methods for DNA methylation array data. Br J Cancer. 2013;109(6):1394–402.
Zang C, et al. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. Bioinformatics (Oxford, Engl). 2009;25(15):1952–8.

Zhang Y, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9(9):R137.

Zhuang J, Widschwendter M, Teschendorff AE. A comparison of feature selection and classification methods in DNA methylation studies using the Illumina Infinium platform. BMC Bioinformatics. 2012;13(1):59.

Ziller MJ, et al. Charting a dynamic DNA methylation landscape of the human genome. Nature. 2013;500(7463):477–81.

Zou J, et al. Epigenome-wide association studies without the need for cell-type composition. Nat Methods. 2014;11(3):309–11.

Chapter 2
DNA Methylation and Cell-Type Distribution

E. Andrés Houseman

Abstract Epigenetic processes form the principal mechanisms by which cell differentiation occurs. Consequently, DNA methylation measurements are strongly influenced by the DNA methylation profiles of constituent cell types as well as by their mixing proportions, raising the potential for confounding of direct molecular associations at single CpG dinucleotides by associations between overall cell-type distribution with phenotype or exposure. In this chapter we review the literature on epigenetics and cell mixture; we then present techniques for deconvolution of DNA methylation measurements, either in the presence or in the absence of reference data. Finally, we present several data analysis examples.

Keywords Cell composition • Confounding • DMP • DMR • Immune • Mediation

2.1 Introduction

In the last decade, numerous published articles have demonstrated associations between DNA methylation profiles and disease or exposure phenotypes. For example, DNA methylation profiles measured in blood have been shown to correlate with ovarian cancer (Teschendorff et al. 2009), bladder cancer (Marsit et al. 2011), cardiovascular disease (Kim et al. 2010), obesity (Dick et al. 2014), and environmental exposures (Kile et al. 2014; Koestler et al. 2013a; Joubert et al. 2012). These associations have led to an interest in *epigenome-wide association studies* (EWAS), which aim to investigate associations between DNA methylation and health or exposure phenotypes across the genome (Rakyan et al. 2011a). Many of these epidemiologic studies have employed the Infinium platforms by Illumina, Inc. (San Diego, CA): the older HumanMethylation27 (27K) platform, which interrogates

E.A. Houseman (✉)
School of Biological and Population Health Sciences, College of Public Health and Human Sciences, Corvallis, OR, USA
e-mail: andres.houseman@oregonstate.edu

A.E. Teschendorff (eds.), *Computational and Statistical Epigenomics*, Translational Bioinformatics 7, DOI 10.1007/978-94-017-9927-0_2

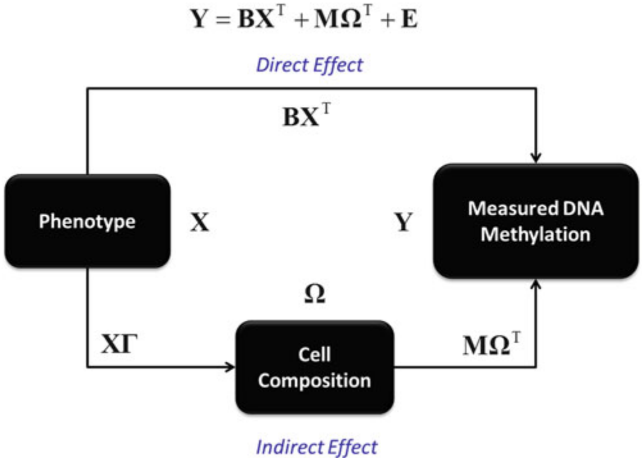


Fig. 2.1 Mediation by cell composition

27,578 CpG loci, and the newer HumanMethylation450 (450K) platform, which interrogates 485,412 CpG loci. Both of these platforms measure locus-specific DNA methylation on an *average beta* scale, which is confined to the unit interval $[0, 1]$ and roughly represents the fraction of methylated molecules in the given sample at the genomic position represented by the locus.

However, DNA methylation, associated with chromatin alterations, is partially responsible for coordination of gene expression in individual cells (Ji et al. 2010; Khavari et al. 2010; Natoli 2010). Consequently, normal tissue differentiation and cellular lineage is regulated by epigenetic mechanisms (Khavari et al. 2010), and DNA methylation shows substantial variation across tissue types (Christensen et al. 2009) as well as individual cell types, particularly distinct types of leukocytes (Ji et al. 2010). This understanding has led to a search for *differentially methylated regions* (DMRs) that distinguish specific cell lineages with high sensitivity and specificity (Baron et al. 2006). Figure 2.1 illustrates the consequence of heterogeneity in DNA methylation profile across cell types as it pertains to epidemiologic analysis of DNA methylation. In particular, DNA methylation measured in a tissue sample will be influenced both by cellular heterogeneity and by direct locus-specific phenotype effects. If the phenotype alters the composition of cells in the sample, then the *total effect* of phenotype on measured DNA methylation will be partially mediated by effects of phenotype on cell composition. For example, if a phenotype alters the immune system, then DNA methylation measured in blood will register both the indirect effects of the phenotype on the immune system as well as any direct effect not mediated by cell composition. When the direct effects are of principal interest in a study, then the cell-composition effects will represent a confound of the direct effects if they are not taken into account. This issue has been highlighted in numerous recent publications (Jaffe and Irizarry 2014; Koestler et al. 2012; Langevin et al. 2012, 2014; Li et al. 2014).

2.2 Fundamental Concepts

Much has been written about mediation and confounding, which are interrelated but distinct concepts (Robins and Greenland 1992; Pearl 2009; VanderWeele 2009). However, linear analysis is sufficient to untangle direct and mediated effects when (1) there is no modification of the effect of the independent variable (phenotype) on dependent variable (DNA methylation) by the mediator (cell composition) and (2) errors in the measurement of mediator (cell composition) and dependent variable (DNA methylation) are uncorrelated. Under these assumptions, several techniques are currently available for analyzing DNA methylation data while accounting for cellular heterogeneity. All of them assume essentially the following linear model for m CpG loci measured on n subjects:

$$\mathbf{Y} = \mathbf{B}\mathbf{X}^{\mathrm{T}} + \mathbf{M}\boldsymbol{\Omega}^{\mathrm{T}} + \mathbf{E}, \tag{2.1}$$

where $\mathbf{Y}$ is an $m \times n$ matrix of average beta values, $\mathbf{X}$ is an $n \times d$ design matrix of phenotype variables and potential confounders (for a total of d covariates including an intercept), $\mathbf{B}$ is the $m \times d$ matrix of regression coefficients representing direct effects, $\mathbf{M}\boldsymbol{\Omega}^{\mathrm{T}}$ represents a linear mixture effect, with $\mathbf{M}$ an $m \times k$ matrix representing m CpG-specific methylation states for k cell types, $\boldsymbol{\Omega}$ is an $n \times k$ matrix representing subject-specific cell-type distributions (each row representing the cell-type proportions for a given subject), and $\mathbf{E}$ is an $m \times n$ matrix of errors with $E(\mathbf{E}) = \mathbf{0}_{m \times n}$. Note that the value k is assumed to be known in advance. Note also that the entries of $\mathbf{Y}$, of $\mathbf{M}$, and of $\boldsymbol{\Omega}$ are assumed to lie in the unit interval, and that the rows of $\boldsymbol{\Omega}$ sum to one. In addition, we assume $\boldsymbol{\Omega}$ is a random variable that is potentially associated with $\mathbf{X}$. Although a Dirichlet model would most appropriately model the rows of $\boldsymbol{\Omega}$, we assume the following linear model as a computationally efficient approximation:

$$\boldsymbol{\Omega} = \mathbf{X}\boldsymbol{\Gamma} + \boldsymbol{\Xi}, \tag{2.2}$$

where $\boldsymbol{\Gamma}$ is a $d \times k$ matrix of covariate effects upon cell proportion and $\boldsymbol{\Xi}$ is an $n \times k$ error matrix. Figure 2.1 depicts these quantities in the context of mediation. Note that Eq. (2.1) explicitly omits interaction between $\mathbf{X}$ and $\boldsymbol{\Omega}$. With the additional assumption that $\mathbf{E}$ and $\boldsymbol{\Xi}$ are independent (and independent of $\mathbf{X}$), linear regression is sufficient for studying the mediation of phenotype effects on DNA methylation by cell composition. In particular, substituting (2.2) in (2.1),

$$\mathbf{Y} = \mathbf{B}\mathbf{X}^{\mathrm{T}} + \mathbf{M}\boldsymbol{\Omega}^{\mathrm{T}} + \mathbf{E} = \left(\mathbf{B} + \mathbf{M}\boldsymbol{\Gamma}^{\mathrm{T}}\right)\mathbf{X}^{\mathrm{T}} + \left(\mathbf{M}\boldsymbol{\Xi} + \mathbf{E}\right), \tag{2.3}$$

the total effect of $\mathbf{X}$ upon $\mathbf{Y}$ is $E\left(\mathbf{Y}\middle|\mathbf{X}\right) = \left(\mathbf{B} + \mathbf{M}\boldsymbol{\Gamma}^{\mathrm{T}}\right)\mathbf{X}^{\mathrm{T}}$, the direct effect is $\mathbf{B}\mathbf{X}^{\mathrm{T}}$, and the mediated, or *cell-composition effect*, is $\boldsymbol{\Delta}\mathbf{X}^{\mathrm{T}}$, where $\Delta = \mathbf{M}\boldsymbol{\Gamma}^{\mathrm{T}}$. Note that the error term for the total effects model is $\mathbf{M}\boldsymbol{\Xi} + \mathbf{E}$, which includes a term that depends on the cell-type-specific coefficient matrix $\mathbf{M}$.

In the remainder of this chapter, we present methods for estimating the total, direct, and cell-composition effects. We present both *reference-based* methods, that is, those relying on the availability of an external reference data set for estimating the matrix **M**, and *reference-free* methods, those that do not require such reference data and treat **M** as essentially unknown.

2.3 Reference-Based Methods

When $\boldsymbol{\Omega}$ is known through explicitly measured cell counts, then it can be absorbed into the covariate matrix after deleting one of the cell types (in order to circumvent over-parameterization of the design matrix); subsequently, simple linear model methods such as limma (Smyth 2004) can be employed for analysis. For example, when a single cell type is being analyzed, $\boldsymbol{\Omega} = \mathbf{1}_n$, and cell type can effectively be ignored. Examples of single-cell-type studies include an analysis of DNA methylation associations with diabetes in CD14+ monocytes (Rakyan et al. 2011b) as well as associations between DNA methylation and autism in ectodermal cells (Berko et al. 2014). Alternatively, leukocyte counts may be available through standard complete blood count (CBC) methods and converted to proportions to obtain $\boldsymbol{\Omega}$, although standard methods will typically provide only coarse categories, for example, grouping all lymphocyte types together. Generally, finely differentiated cell counts can be obtained using cell sorting methods such as fluorescence-activated cell sorting (FACS) or magnetic-activated cell sorting (MACS). DNA methylation in a community cohort was characterized for peripheral blood mononuclear cells (PBMCs), accompanied by CBC counts (Lam et al. 2012). Another recent example demonstrated associations of DNA methylation with depression in postmortem brains using proportions of neuron and glial cells (Guintivano et al. 2013). Note that some mRNA expression analyses of blood have incorporated FACS measurements of individual leukocyte counts (Shen-Orr et al. 2010), but to date there are no major analyses of DNA methylation data in whole blood or PBMCs that have incorporated comprehensive FACS or MACS counts.

In many studies, it may be infeasible to obtain direct measures of cell counts. Fortunately, DNA methylation measurements themselves may be used to obtain approximate cell proportion estimates, as long as a reference data set is available for measuring the cell-type-specific mean methylation for a set of CpG loci that differentiate the types with a high degree of sensitivity and specificity. We have referred to such loci as *pseudo*-DMRs, since they are single locus markers rather than regions, although they are also commonly known as differentially methylated *positions* (DMPs). Interest in the detection of DMRs and DMPs for specific types of leukocytes has arisen from the study of tumor infiltration by lymphocytes (Accomando et al. 2012; Wiencke et al. 2012); this, in turn has led to more comprehensive characterization of genome-wide DNA methylation profiles for major types of leukocytes. Existing reference sets include an Infinium 27K data set (Houseman et al. 2012) as well as an Infinium 450K data set (Reinius et al.

2012). These data sets can be deployed to obtain estimates $\widehat{\boldsymbol{\Omega}}$ of cell proportions, as Houseman et al. (2012) have shown. The method is briefly described as follows.

Suppose S is an ordered set of DMP loci for distinguishing k cell types, $\mathbf{y}_l^{(S)}$ is a DNA methylation measurement on the set S for a purified sample of type $l \in \{1, \ldots, k\}$, and $E\left(\mathbf{y}_l^{(S)}\right) = \boldsymbol{\mu}_l^{(S)}$ for a vector $\boldsymbol{\mu}_l^{(S)}$ whose elements fall in the unit interval. If $\mathbf{M}^{(S)} = \left[\boldsymbol{\mu}_1^{(S)}, \ldots, \boldsymbol{\mu}_k^{(S)}\right]$ and $\mathbf{y}_*^{(S)}$ is a vector of DNA measurements on S for a heterogeneous tissue sample of mixed cell types, type l representing proportion $\omega_l \geq 0$ of the tissue sample $\left(\sum_{l=1}^{k} \omega_l \leq 1\right)$, then $E\left(\mathbf{y}_*^{(S)}\right) = \mathbf{M}^{(S)}\boldsymbol{\omega}$, where $\boldsymbol{\omega}^{\mathrm{T}} = [\omega_1, \ldots, \omega_k]$. It follows that $\boldsymbol{\omega}$ can be estimated by minimizing the quantity $\left\|\mathbf{y}_*^{(S)} - \mathbf{M}^{(S)}\boldsymbol{\omega}\right\|^2$; although this problem is easily solved by computing the least squares estimator for $\boldsymbol{\omega}$, slightly better results can be obtained by imposing the natural constraints $\omega_l \geq 0$ and $\sum_{l=1}^{k} \omega_l \leq 1$ onto the solution space. Quadratic programming (Goldfarb and Idnani 1983) can easily be employed to obtain an estimate $\widehat{\boldsymbol{\omega}}$ that obeys these constraints. This *cellular deconvolution* method was initially shown to work well in recovering proportions of artificial blood mixtures (Houseman et al. 2012). Subsequent validation studies have demonstrated acceptable performance of cellular deconvolution of DNA methylation data. Comparing estimated proportions of monocytes within PBMC samples (which lack granulocytes) obtained from a community cohort (Lam et al. 2012) to their corresponding CBC-derived quantities, Koestler et al. (2013b) measured a root-mean-square-error (RMSE) of approximately 5 percentage points (Koestler et al. 2013b). In a comprehensive analysis of six donor blood samples with counts measured using three distinct FACS techniques, Accomando et al. (2014) estimated a RMSE of about 3.0–4.3 percentage points for six distinct leukocyte subtypes; when compared with each other, the FACS methods produced RMSE values of approximately 2 percentage points (i.e., only slightly smaller magnitude) (Accomando et al. 2014). First popularized in a study of rheumatoid arthritis (Liu et al. 2013), the method has become a widely adopted method for estimating cell proportions when individual count data are unavailable.

The method is available in the R/Bioconductor package *minfi* (function *EstimateCellCounts*). The *minfi* library also supports mutual normalization of reference and target data sets, which leads to some improvement in the estimation of cell proportions. The R/bioconductor package *FlowSorted.Blood.450k* encapsulates the 450K leukocyte reference data set published by Reinius et al. (2012); a 27K leukocyte reference data set is available on Gene Expression Omnibus (GEO), accession number GSE39981.

Note that $\mathbf{M}^{(S)}$ should represent a reasonably exhaustive characterization of the cell types comprising the tissue to be analyzed, in that the k profiled types represent the major portion of each sample (Houseman et al. 2012). Under these circumstances, the sum $\sum_{l=1}^{k} \omega_l$ will typically lie close to 1 for each sample.

Consequently, when incorporated into the design matrix $\mathbf{X}$ of Eq. (2.1) for data analysis, the matrix $\widehat{\boldsymbol{\Omega}}$ derived from these measures should omit one of the types, otherwise the resulting design matrix will exhibit poor conditioning and lead to unstable estimates. For example, in analyzing whole blood, the granulocyte proportion might be omitted, and in analyzing PBMC samples, the monocyte proportion might be omitted.

Note also that the cell-composition term $\mathbf{M}\boldsymbol{\Omega}^{\mathrm{T}}$ in Eq. (2.1) entails a linear mixing assumption that is most plausible for measurement scales which correspond to fractions of cells or molecules. Consequently, cellular deconvolution should always be performed on the average beta scale instead of a popular alternative, the *M-value* scale obtained by logit-transforming the average beta. In addition, genome-wide application of Eq. (2.1) is likely to produce slightly better fit to data when beta values are used instead of M-values. However, use of average beta values in regression analysis is complicated slightly by the non-normal nature of the error term. For mid-range values, beta values and M-value will covary in an approximately linear fashion, so that both scales will return similar results for loci that exhibit great sensitivity to cell composition (i.e., DMPs). An alternative to applying Eq. (2.1) directly is to remove the cell-composition effects on the beta scale before implementing genome-wide regression analysis on the M-value scale. This strategy is consistent with *removal of unwanted variability* (RUV) (Gagnon-Bartsch and Speed 2012; Jaffe and Irizarry 2014). In this approach, $\widehat{\mathbf{M}}$ is obtained by fitting the genome-wide DNA methylation data to the equation $\mathbf{Y} = \mathbf{M}\widehat{\boldsymbol{\Omega}}^{\mathrm{T}} + \mathbf{E}$, each column $\mathbf{y}$ of $\mathbf{Y}$ is adjusted for cell composition via $\mathbf{y}^{(adj)} \leftarrow \mathbf{y} - \widehat{\mathbf{M}}\left(\widehat{\boldsymbol{\omega}} - \overline{\boldsymbol{\omega}}\right)$ (where $\widehat{\boldsymbol{\omega}}$ is the corresponding column of $\widehat{\boldsymbol{\Omega}}$, $\overline{\boldsymbol{\omega}} = n^{-1}\widehat{\boldsymbol{\Omega}}^{\mathrm{T}}\mathbf{1}_n$ is the average cell proportion profile), and each resulting adjusted value is logit-transformed to an M-value, $m_j^{(adj)} \leftarrow \log_2\left(\max\left\{y_j^{(adj)}, \varepsilon\right\}\right) - \log_2\left(\max\left\{1 - y_j^{(adj)}, \varepsilon\right\}\right)$, with ε a small value chosen to avoid infinite M-values. Note that centering $\widehat{\boldsymbol{\omega}}$ by $\overline{\boldsymbol{\omega}}$ is necessary to avoid a non-negligible proportion of adjusted values $y_j^{(adj)}$ lying outside the unit interval, as the resulting values of $y_j^{(adj)}$ will be centered around the average DNA methylation value.

Finally, we note that associations between $\mathbf{X}$ and $\boldsymbol{\Omega}$ may be of scientific interest. Analysis is straightforward when $\boldsymbol{\Omega}$ is measured directly. However, when $\boldsymbol{\Omega}$ is estimated via cellular deconvolution, it is desirable to account for all sources of variability, including the contribution of measurement error from the reference data set. Houseman et al. (2012) describe a comprehensive method for conducting such analysis.

2.4 Reference-Free Methods

Although the method of Houseman et al. (2012) provides an algorithm for estimating cell proportions $\boldsymbol{\Omega}$ from DNA methylation data, it requires the existence of a reference data set. To date, such data sets exist only for blood (Accomando et al.

2012; Houseman et al. 2012; Reinius et al. 2012) and, to a limited extent, brain tissue (Guintivano et al. 2013). However, other tissues are of interest in EWAS. For example, population-based studies of DNA methylation have been published with DNA methylation measured in placenta (Banister et al. 2011; Suter et al. 2011; Wilhelm-Benartzi et al. 2012), umbilical cord tissue (Teh et al. 2014), and (with sparser arrays) buccal swabs (Breton et al. 2009; Kaminsky et al. 2009); no reference sets currently exist for these and other tissues of interest (e.g., adipose tissue).

To circumvent this problem, Houseman et al. (2014) propose a method for approximating the 2012 method. This method also assumes Eq. (2.1), but treats the matrix $\mathbf{M}$ as unknown. The method works by first fitting the model for total effects,

$$\mathbf{Y} = \mathbf{B}^{*}\mathbf{X}^{\mathrm{T}} + \mathbf{E}^{*}$$

where $\mathbf{B}^{*} = \mathbf{B} + \mathbf{M}\boldsymbol{\Gamma}^{\mathrm{T}}$ and $\mathbf{E}^{*} = \mathbf{M}\boldsymbol{\Xi}^{\mathrm{T}} + \mathbf{E}$, as evident from Eq. (2.3). Note that

$$\mathbf{R} = \left[\mathbf{B}^{*}, \mathbf{E}^{*}\right] = \mathbf{M}\left[\boldsymbol{\Gamma}^{\mathrm{T}}, \boldsymbol{\Xi}^{\mathrm{T}}\right] + [\mathbf{B}, \mathbf{E}]\,. \tag{2.4}$$

With k, the number of assumed cell types, chosen in advance by prior biological knowledge or using a method for estimating the number of factors in a factor-analytic model [e.g., using random matrix theory (Teschendorff et al. 2011)], the method associates the largest k singular values of $\mathbf{R}$ with cell-composition effects. Specifically, applying a singular value decomposition (SVD) to $\widehat{\mathbf{R}} = \left[\widehat{\mathbf{B}}^{*}, \widehat{\mathbf{E}}^{*}\right]$, $\widehat{\mathbf{R}} = \mathbf{U}_1\boldsymbol{\Lambda}_1\mathbf{V}_1^{\mathrm{T}} + \mathbf{U}_2\boldsymbol{\Lambda}_2\mathbf{V}_2^{\mathrm{T}}$, where $\mathbf{U}_1$ is an orthogonal $m \times k$ matrix, $\mathbf{U}_2$ is an orthogonal $m \times (n-k)$ matrix, $\mathbf{U}_1^{T}\mathbf{U}_2 = \mathbf{0}_{k\times(n-k)}$, $\mathbf{V}_1$ is an orthogonal $n \times k$ matrix, $\mathbf{V}_2$ is an orthogonal $n \times (n-k)$ matrix, $\boldsymbol{\Lambda}_1$ is a diagonal $k \times k$ matrix, $\boldsymbol{\Lambda}_2$ is a diagonal $(n-k) \times (n-k)$ matrix, and the two terms separate the k largest singular values from the $n-k$ smallest ones (i.e., every diagonal element of $\boldsymbol{\Lambda}_1$ is larger than every diagonal element of $\boldsymbol{\Lambda}_2$), it is assumed that $\mathbf{M}\left[\boldsymbol{\Gamma}^{\mathrm{T}}, \boldsymbol{\Xi}^{\mathrm{T}}\right] = \mathbf{U}_1\boldsymbol{\Lambda}_1\mathbf{V}_1^{\mathrm{T}}$ and $[\mathbf{B}, \mathbf{E}] = \mathbf{U}_2\boldsymbol{\Lambda}_2\mathbf{V}_2^{T}$. Note that the two terms on the right hand side of Eq. (2.4) must be orthogonal in order for this identity to hold; to ensure orthogonality it is sufficient to assume $\mathbf{M}^{T}\mathbf{E} = \mathbf{0}_{k\times n}$ and $\mathbf{M}^{T}\mathbf{B} = \mathbf{0}_{k\times d}$. The former condition is an essential assumption entailed by the linear regression represented by Eq. (2.1); the latter assumption, that “indirect” effects lie in a space orthogonal to the cell-type-specific profiles, represents an unverifiable biological condition also necessary for the deconvolution method of Houseman et al. (2012), although the Supplement of the 2012 paper argues that orthogonality will approximately hold if the effects in $\mathbf{B}$ are relatively sparse. Note also that association of the k largest singular values with cell-composition effects represents another biological assumption, that the cell-composition effects will dominate the linear associations evident in the array. Under the assumptions just described, $\widehat{\mathbf{B}}$ is obtained by selecting the first d columns of $\mathbf{U}_2\boldsymbol{\Lambda}_2\mathbf{V}_2^{\mathrm{T}}$. Note that $\widehat{\Delta} = \widehat{\mathbf{B}}^{*} - \widehat{\mathbf{B}}$ represents the matrix of coefficients that explain the cell-mediated associations between $\mathbf{X}$ and $\mathbf{Y}$, which may be of interest in some studies.

Houseman et al. (2014) also propose a method for generating bootstrap samples from the sampling distribution of $\widehat{\mathbf{B}}^*$ and $\widehat{\mathbf{B}}$, from which standard errors for $\widehat{\mathbf{B}}^*$, $\widehat{\mathbf{B}}$, and $\widehat{\boldsymbol{\Delta}}$ can be estimated. Briefly, the method generates a bootstrap sample $\mathbf{Y}^{(b)}$ of DNA methylation average beta values as $\mathbf{Y}^{(b)} = \widehat{\mathbf{B}}^*\mathbf{X}^T + \mathbf{E}^{(b)}$, where $\widehat{\mathbf{B}}^*$ is the estimated coefficient of total effects and $\mathbf{E}^{(b)}$ is a bootstrap error matrix constructed element-wise as $e_{ij}^{(b)} = q_{ij}^{(b)}\sqrt{\widehat{\mu}_{ij}\left(1-\widehat{\mu}_{ij}\right)}$, where $\widehat{\mu}_{ij}$ is the element of $\widehat{\mathbf{B}}^*\mathbf{X}^T$ corresponding to the i^{th} column and j^{th} row, $q_{ij}^{(b)}$ is the element of the matrix obtained by sampling with replacement from the columns of $\mathbf{Q}$, each of whose elements q_{ij} were obtained from $\widehat{\mathbf{E}}^* = \left(\widehat{e}_{ij}\right)$ and $\widehat{\mathbf{B}}^*\mathbf{X}^T$ as $q_{ij} = \widehat{e}_{ij}^*/\sqrt{\widehat{\mu}_{ij}\left(1-\widehat{\mu}_{ij}\right)}$. The method factors the error $\mathbf{E}^*$ element-wise as the product of a mean-dependent scaling factor $\sqrt{\widehat{\mu}_{ij}\left(1-\widehat{\mu}_{ij}\right)}$ and a "dispersion" value q_{ij}; this strategy respects the approximate beta distribution of $\mathbf{Y}$, while simultaneously preserving correlation across the rows (CpGs). The estimation method, as well as its corresponding bootstrap generation procedure, is publicly available in an R package entitled *RefFreeEWAS*.

The 2014 method is similar to *surrogate variable analysis* (SVA) (Leek and Storey 2007; Teschendorff et al. 2011), which uses a factor-analytic decomposition similar to Eq. (2.1) but applies SVD or *independent components analysis* (ICA) to the error term $\hat{\mathbf{E}}^*$ rather than $\widehat{\mathbf{R}} = \left[\widehat{\mathbf{B}}^*, \widehat{\mathbf{E}}^*\right]$, thus potentially missing linear effects that are explicitly the result of cell composition. It is also similar in spirit to the recently published *Ewasher* method (Zou et al. 2014); this method models the phenotype as a function of methylation and potentially other confounding covariates, instead of modeling methylation as a function of phenotype and potential confounders. Specifically, the following model is assumed:

$$\mathbf{x} = \beta_j^{(Y)}\mathbf{y}_j + \mathbf{Z}^T\boldsymbol{\beta}_j^{(Z)} + m^{-1/2}\tilde{\mathbf{Y}}^T\mathbf{u} + \mathbf{e}_j, \tag{2.5}$$

where $\mathbf{x}$ is the $n \times 1$ matrix of subject phenotypes (dichotomous or continuous), $\mathbf{y}_j$ is the $n \times 1$ matrix of DNA methylation value measured for each at CpG j, $\mathbf{Z}$ is a $n \times d'$ matrix of potential confounders for each subject (including an intercept term), $\tilde{\mathbf{Y}}$ is the $m \times n$ matrix of standardized DNA methylation values obtained from $\mathbf{Y}$ by standardizing each row (CpG), $\mathbf{u}$ is an $m \times 1$ matrix of Gaussian random effects, each having variance σ_u^2 and uncorrelated across entries, $\mathbf{e}_j$ is an $n \times 1$ matrix of independent errors having variance $\sigma_{e,j}^2$, and $\beta_j^{(Y)}$ and $\boldsymbol{\beta}_j^{(Z)}$ are coefficients to be estimated. Estimation proceeds by considering the multivariate distribution of $\mathbf{x}$, whose variance–covariance matrix is $\boldsymbol{\Sigma}_x = m^{-1}\sigma_u^2\tilde{\mathbf{Y}}^T\tilde{\mathbf{Y}} + \sigma_{e,j}^2\mathbf{I}_{n\times n}$. Note that if $\tilde{\mathbf{Y}} = \tilde{\mathbf{M}}\boldsymbol{\Omega}^T$ captures the rescaled cell-composition effects, then $\tilde{\mathbf{Y}}^T\tilde{\mathbf{Y}} = \boldsymbol{\Omega}\tilde{\mathbf{M}}^T\tilde{\mathbf{M}}\boldsymbol{\Omega}^T$, which is essentially the contribution to $\boldsymbol{\Sigma}_x$ that would result from substituting the explicit cell-composition effect $\tilde{\mathbf{M}}\boldsymbol{\Omega}^T$ for $m^{-1/2}\tilde{\mathbf{Y}}^T\mathbf{u}$ in Eq. (2.5). Thus, the term $m^{-1/2}\tilde{\mathbf{Y}}^T\mathbf{u}$ captures cell-composition effects in a manner similar to the approach based on Eq. (2.1).

Note that these reference-free methods entail strong linearity assumptions and, ultimately, assumptions about the relationship between measured DNA methylation and the actual proportion of methylated cytosine molecules among the specific targeted loci. Consequently, the technical properties of the assay to be used should be considered carefully, and analysis should be preceded by the execution of a pre-processing pipeline that results in DNA measurements that are as comparable as possible across loci. For example, use of the popular 450K assay should entail proper normalization (Marabita et al. 2013), alignment of the biochemically distinct Type I and Type II probes (Dedeurwaerder et al. 2011; Teschendorff et al. 2013), and removal of loci whose probes contain common variants or cross-hybridize across the genome (Chen et al. 2013).

2.5 Data Examples

Several published analyses of DNA methylation data have employed the methods described above to adjust for heterogeneity in cell composition. Guintivano et al. (2014) incorporated blood count data to adjust for cellular heterogeneity in association between DNA methylation measured in blood and postpartum depression (Guintivano et al. 2014). Liu et al. (2013) published the first analysis that employed the Houseman et al. (2012) method of estimating cell proportions from DNA methylation data, demonstrating marked attenuation of significance in association of DNA methylation measured in blood with rheumatoid arthritis after adjusting for estimated cell proportions (Liu et al. 2013). Similarly, in a perinatal study of arsenic exposure in Bangladesh, Kile et al.(2014) demonstrated marked attenuation of significance in association of DNA methylation measured in cord blood with ingestion of inorganic arsenic via drinking water after adjusting for cell proportions, additionally suggesting that arsenic exposure could alter the proportion of CD4+ and CD8+ T lymphocytes (Kile et al. 2014). Koestler et al. (2013a, b) demonstrated association of cord blood methylation and urinary inorganic arsenic concentration after adjusting for cell proportion (Koestler et al. 2013a). Finally, Jaffe and Irizarry (2014) employed several methods including the Houseman et al. (2012) method to demonstrate that the commonly acknowledged association between age and DNA methylation can be explained in large part by age-related changes in cell composition (Jaffe and Irizarry 2014).

Using two data sets, we briefly compare and contrast some of the methods described in this chapter: the community cohort data published by Lam et al. (2012) and re-analyzed by Koestler et al. (2013a, b), and the rheumatoid arthritis data set published by Liu et al. (2013) and re-analyzed by Houseman et al. (2014) and Zou et al. (2014). See Houseman et al. (2014) for additional details.

For 26,486 autosomal CpG sites assayed by the 27K array, Fig. 2.2 shows quantile-quantile (QQ) plots on a logarithmic scale comparing a uniform distribution against nominal p-values obtained using several different methods: unadjusted ("Unadj", representing total effect $\mathbf{B}^*$), reference-based ["Ref", representing direct

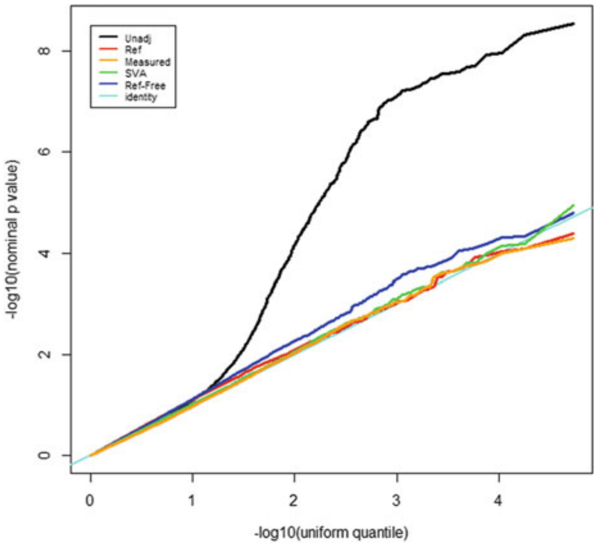


Fig. 2.2 Analysis of DNA methylation and IL-6 response bioassay in a community cohort

effect **B** obtained by applying the method of (Houseman et al. 2012), to obtain cell proportion estimates $\widehat{\boldsymbol{\Omega}}$], a direct effect based on monocyte/lymphocyte proportions measured by CBC ("Measured"), a direct effect estimate based on SVA ("SVA") with $k = 11$ assumed surrogate variables, and a direct effect estimate based on the reference-free approach of Houseman et al. (2014) with $k = 10$ (see the original article for details on the choice of k). Each p-value represents significance of association between DNA methylation in PBMCs measured on an average beta scale and IL-6 response to phorbol-12-myristate-13-acetatein. All methods except the unadjusted method result in p-values that are effectively uniform (i.e., characteristic of a null effect). This suggests that there may be a strong total effect of the IL-6 phenotype on DNA methylation, but that this effect is explained by alterations in monocyte/lymphocyte proportions and accounted for using the reference-based and SVA methods. Note that Fig. 2.2 suggests a small number of CpGs with slightly elevated significance for the reference-free method; however, the distribution of p-values across the 26,486 CpGs is consistent with a uniform distribution, as Fig. 2.3 implies. Figure 2.3 shows the QQ plots for unadjusted and reference-free methods, superimposed upon 95 % probability bands representing their corresponding null distributions obtained from 1,000 bootstrap estimates using a method suggested in the supplementary material of Houseman et al. (2014). This plot suggests significant modification of total DNA methylation by the IL-6 phenotype, but no significant alteration after accounting for covariation in monocytes. Figure 2.4 compares significance of the $\widehat{\Delta}$ coefficients from the reference-free method with significance of the monocyte coefficients from the linear model using only the measured monocyte proportions. There is high concordance in significance between the two methods; by Fisher's exact test, concordance of p-values less than 0.001 is quite high (odds ratio = 47.5, 95 % confidence interval: 21.1–106, Fisher $p < 10^{-16}$). Thus, this

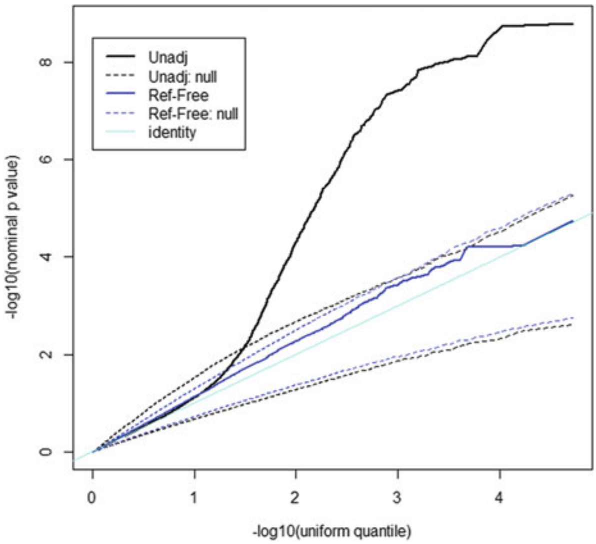


Fig. 2.3 Analysis of DNA methylation and IL-6 response bioassay in a community cohort: comparison with bootstrap-based null sampling distribution

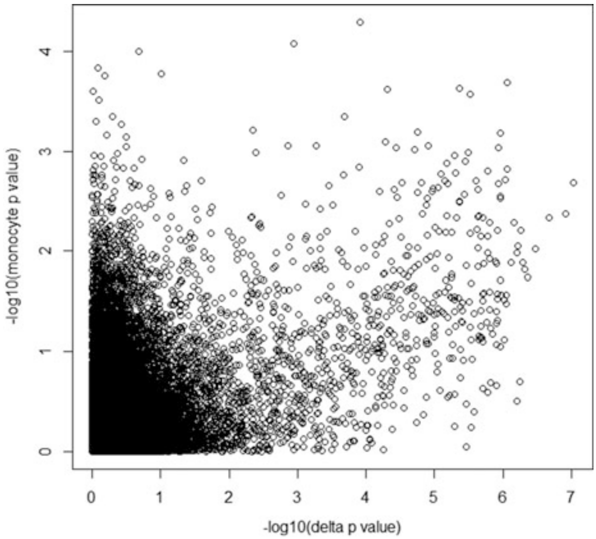
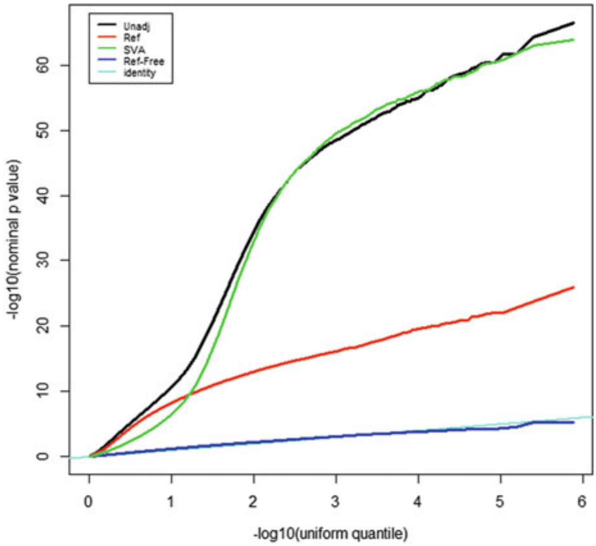


Fig. 2.4 Analysis of DNA methylation and IL-6 response bioassay in a community cohort: comparison of significance of cell-composition effects from reference-free methods with significance of effects of known monocyte proportions

analysis demonstrates how Δ coefficients can be used to identify DMPs for distinct cell types within a sample. This strategy was used in a recent article evaluating the effect of cellular heterogeneity on breast tissue (Houseman and Ince 2014).

For 384,410 autosomal CpG sites assayed by the 450K array and having probes free of common variants, Fig. 2.5 shows QQ plots on a logarithmic scale comparing a uniform distribution against nominal p-values obtained using the same methods as for Fig. 2.2, except for the "Measured" method since measured cell counts were unavailable for this data set. Additionally, for SVA, $k = 53$ surrogate variables were

Fig. 2.5 Analysis of DNA methylation and rheumatoid arthritis



assumed, and for the reference-based method, $k = 37$ cell types were assumed; these values were based on application of appropriate dimension-estimating algorithms (Houseman et al. 2014). Each p-value represents significance of association between rheumatoid arthritis case status and DNA methylation in whole blood measured on an average beta scale. The unadjusted and SVA-adjusted methods result in QQ plots reflecting strong significance; the QQ plot from the reference-based approach reflects attenuated but still moderately strong significance; and the reference-free approach reflects null association. As previously suggested (Houseman et al. 2014), the reference-free approach may be capturing subtle shifts in proportions of cell types not profiled in the reference data set used for the reference-based adjustment. Note that while SVA was adequate for cell-composition adjustment in the previous analysis, it was insufficient for the present one.

2.6 Conclusions

Heterogeneity in cell type is an important consideration in the analysis of DNA methylation measured from complex tissues. In many applications, the phenotype of interest may alter the composition of cell types within the target tissue, thus altering DNA methylation profile independently of specific molecular alterations that are not mediated by cell type. Therefore, proportions of each cell type should be included in models for phenotypic effects of DNA methylation. In the best-case scenario, proportions of each cell type will be available for each sample. However, since the cell sorting techniques necessary for measuring these proportions can be costly, many studies lack these measurements. In such a situation, the cell proportions can be estimated directly from DNA methylation data if a reference data set exists for the cell types that constitute the target tissue. If no such reference data set exists, recently published reference-free methods can be used to account for cellular heterogeneity when estimating phenotype associations with DNA methylation, although more work is needed to validate these new methods.

References

Accomando WP, Wiencke JK, Houseman EA, Butler RA, Zheng S, Nelson HH, Kelsey KT. Decreased NK cells in patients with head and neck cancer determined in archival DNA. Clin Cancer Res. 2012;18(22):6147–54. doi:10.1158/1078-0432.CCR-12-1008.

Accomando WP, Wiencke JK, Houseman EA, Nelson HH, Kelsey KT. Quantitative reconstruction of leukocyte subsets using DNA methylation. Genome Biol. 2014;15(3):R50. doi:10.1186/gb-2014-15-3-r50.

Banister CE, Koestler DC, Maccani MA, Padbury JF, Houseman EA, Marsit CJ. Infant growth restriction is associated with distinct patterns of DNA methylation in human placentas. Epigenetics. 2011;6(7):920–7. doi:10.4161/epi.6.7.16079.

Baron U, Türbachova I, Hellwag A, Eckhardt F, Berlin K, Hoffmuller U, Gardina P, Olek S. Research paper DNA methylation analysis as a tool for cell typing. Epigenetics. 2006;1(1):55–60.

Berko ER, Suzuki M, Beren F, Lemetre C, Alaimo CM, Calder RB, Ballaban-Gil K, Gounder B, Kampf K, Kirschen J, Maqbool SB, Momin Z, Reynolds DM, Russo N, Shulman L, Stasiek E, Tozour J, Valicenti-McDermott M, Wang S, Abrahams BS, Hargitai J, Inbar D, Zhang Z, Buxbaum JD, Molholm S, Foxe JJ, Marion RW, Auton A, Greally JM. Mosaic epigenetic dysregulation of ectodermal cells in autism spectrum disorder. PLoS Genet. 2014;10(5):e1004402. doi:10.1371/journal.pgen.1004402.

Breton CV, Byun HM, Wenten M, Pan F, Yang A, Gilliland FD. Prenatal tobacco smoke exposure affects global and gene-specific DNA methylation. Am J Respir Crit Care Med. 2009;180(5):462–7. doi:10.1164/rccm.200901-0135OC.

Chen YA, Lemire M, Choufani S, Butcher DT, Grafodatskaya D, Zanke BW, Gallinger S, Hudson TJ, Weksberg R. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. Epigenetics. 2013;8(2):203–9. doi:10.4161/epi.23470.

Christensen BC, Houseman EA, Marsit CJ, Zheng S, Wrensch MR, Wiemels JL, Nelson HH, Karagas MR, Padbury JF, Bueno R, Sugarbaker DJ, Yeh RF, Wiencke JK, Kelsey KT. Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CpG island context. PLoS Genet. 2009;5(8):e1000602. doi:10.1371/journal.pgen.1000602.

Dedeurwaerder S, Defrance M, Calonne E, Denis H, Sotiriou C, Fuks F. Evaluation of the Infinium Methylation 450 K technology. Epigenomics. 2011;3(6):771–84. doi:10.2217/epi.11.105.

Dick KJ, Nelson CP, Tsaprouni L, Sandling JK, Aissi D, Wahl S, Meduri E, Morange PE, Gagnon F, Grallert H, Waldenberger M, Peters A, Erdmann J, Hengstenberg C, Cambien F, Goodall AH, Ouwehand WH, Schunkert H, Thompson JR, Spector TD, Gieger C, Tregouet DA, Deloukas P, Samani NJ. DNA methylation and body-mass index: a genome-wide analysis. Lancet. 2014;383(9933):1990–8. doi:10.1016/S0140-6736(13)62674-4.

Gagnon-Bartsch JA, Speed TP. Using control genes to correct for unwanted variation in microarray data. Biostatistics. 2012;13(3):539–52. doi:10.1093/biostatistics/kxr034.

Goldfarb D, Idnani A. A numerically stable dual method for solving strictly convex quadratic programs. Math Program. 1983;27(1):1–33.

Guintivano J, Aryee MJ, Kaminsky ZA. A cell epigenotype specific model for the correction of brain cellular heterogeneity bias and its application to age, brain region and major depression. Epigenetics. 2013;8(3):290–302. doi:10.4161/epi.23924.

Guintivano J, Arad M, Gould TD, Payne JL, Kaminsky ZA. Antenatal prediction of postpartum depression with blood DNA methylation biomarkers. Mol Psychiatry. 2014;19(5):560–7. doi:10.1038/mp.2013.62.

Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, Wiencke JK, Kelsey KT. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics. 2012;13:86. doi:10.1186/1471-2105-13-86.

Houseman EA, Molitor J, Marsit CJ. Reference-free cell mixture adjustments in analysis of DNA methylation data. Bioinformatics. 2014. doi:10.1093/bioinformatics/btu029.

Houseman EA, Ince TA. Normal cell-type epigenetics and breast cancer classification: a case study of cell mixture–adjusted analysis of DNA methylation data from tumors. Cancer Inform. 2014;13 Suppl 4:53.

Jaffe AE, Irizarry RA. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. Genome Biol. 2014;15(2):R31.

Ji H, Ehrlich LI, Seita J, Murakami P, Doi A, Lindau P, Lee H, Aryee MJ, Irizarry RA, Kim K, Rossi DJ, Inlay MA, Serwold T, Karsunky H, Ho L, Daley GQ, Weissman IL, Feinberg AP. Comprehensive methylome map of lineage commitment from haematopoietic progenitors. Nature. 2010;467(7313):338–42. doi:10.1038/nature09367.

Joubert BR, Haberg SE, Nilsen RM, Wang X, Vollset SE, Murphy SK, Huang Z, Hoyo C, Midttun O, Cupul-Uicab LA, Ueland PM, Wu MC, Nystad W, Bell DA, Peddada SD, London SJ. 450 K epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy. Environ Health Perspect. 2012;120(10):1425–31. doi:10.1289/ehp.1205412.

Kaminsky ZA, Tang T, Wang SC, Ptak C, Oh GH, Wong AH, Feldcamp LA, Virtanen C, Halfvarson J, Tysk C, McRae AF, Visscher PM, Montgomery GW, Gottesman II, Martin NG, Petronis A. DNA methylation profiles in monozygotic and dizygotic twins. Nat Genet. 2009;41(2):240–5. doi:10.1038/ng.286.

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

Kile ML, Houseman EA, Baccarelli AA, Quamruzzaman Q, Rahman M, Mostofa G, Cardenas A, Wright RO, Christiani DC. Effect of prenatal arsenic exposure on DNA methylation and leukocyte subpopulations in cord blood. Epigenetics. 2014;9(5):774–82. doi:10.4161/epi.28153.

Kim M, Long TI, Arakawa K, Wang R, Yu MC, Laird PW. DNA methylation as a biomarker for cardiovascular disease risk. PLoS One. 2010;5(3):e9692. doi:10.1371/journal.pone.0009692.

Koestler DC, Marsit CJ, Christensen BC, Accomando W, Langevin SM, Houseman EA, Nelson HH, Karagas MR, Wiencke JK, Kelsey KT. Peripheral blood immune cell methylation profiles are associated with nonhematopoietic cancers. Cancer Epidemiol Biomarkers Prev. 2012;21(8):1293–302. doi:10.1158/1055-9965.EPI-12-0361.

Koestler DC, Avissar-Whiting M, Houseman EA, Karagas MR, Marsit CJ. Differential DNA methylation in umbilical cord blood of infants exposed to low levels of arsenic in utero. Environ Health Perspect. 2013a;121(8):971–7. doi:10.1289/ehp.1205925.

Koestler DC, Christensen B, Karagas MR, Marsit CJ, Langevin SM, Kelsey KT, Wiencke JK, Houseman EA. Blood-based profiles of DNA methylation predict the underlying distribution of cell types: a validation analysis. Epigenetics. 2013b;8(8):816–26. doi:10.4161/epi.25430.

Lam LL, Emberly E, Fraser HB, Neumann SM, Chen E, Miller GE, Kobor MS. Factors underlying variable DNA methylation in a human community cohort. Proc Natl Acad Sci U S A. 2012;109 Suppl 2:17253–60. doi:10.1073/pnas.1121249109.

Langevin SM, Koestler DC, Christensen BC, Butler RA, Wiencke JK, Nelson HH, Houseman EA, Marsit CJ, Kelsey KT. Peripheral blood DNA methylation profiles are indicative of head and neck squamous cell carcinoma: an epigenome-wide association study. Epigenetics. 2012;7(3):291–9. doi:10.4161/epi.7.3.19134.

Langevin SM, Houseman EA, Accomando WP, Koestler DC, Christensen BC, Nelson HH, Karagas MR, Marsit CJ, Wiencke JK, Kelsey KT. Leukocyte-adjusted epigenome-wide association studies of blood from solid tumor patients. Epigenetics. 2014;9(6):884–95.

Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. 2007;3(9):1724–35. doi:10.1371/journal.pgen.0030161.

Li H, Zheng T, Chen B, Hong G, Zhang W, Shi T, Li S, Ao L, Wang C, Guo Z. Similar blood-borne DNA methylation alterations in cancer and inflammatory diseases determined by subpopulation shifts in peripheral leukocytes. Br J Cancer. 2014. doi:10.1038/bjc.2014.347.

Liu Y, Aryee MJ, Padyukov L, Fallin MD, Hesselberg E, Runarsson A, Reinius L, Acevedo N, Taub M, Ronninger M, Shchetynsky K, Scheynius A, Kere J, Alfredsson L, Klareskog L, Ekstrom TJ, Feinberg AP. Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. Nat Biotechnol. 2013;31(2):142–7. doi:10.1038/nbt.2487.

Marabita F, Almgren M, Lindholm ME, Ruhrmann S, Fagerstrom-Billai F, Jagodic M, Sundberg CJ, Ekstrom TJ, Teschendorff AE, Tegner J, Gomez-Cabrero D. An evaluation of analysis pipelines for DNA methylation profiling using the Illumina HumanMethylation450 BeadChip platform. Epigenetics. 2013;8(3):333–46. doi:10.4161/epi.24008.

Marsit CJ, Koestler DC, Christensen BC, Karagas MR, Houseman EA, Kelsey KT. DNA methylation array analysis identifies profiles of blood-derived DNA methylation associated with bladder cancer. J Clin Oncol. 2011;29(9):1133–9. doi:10.1200/JCO.2010.31.3577.

Natoli G. Maintaining cell identity through global control of genomic organization. Immunity. 2010;33(1):12–24. doi:10.1016/j.immuni.2010.07.006.

Pearl J. Causal inference in statistics: an overview. Stat Surv. 2009;3:96–146.

Rakyan VK, Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. Nat Rev Genet. 2011a;12(8):529–41. doi:10.1038/nrg3000.

Rakyan VK, Beyan H, Down TA, Hawa MI, Maslau S, Aden D, Daunay A, Busato F, Mein CA, Manfras B, Dias KR, Bell CG, Tost J, Boehm BO, Beck S, Leslie RD. Identification of type 1 diabetes-associated DNA methylation variable positions that precede disease diagnosis. PLoS Genet. 2011b;7(9):e1002300. doi:10.1371/journal.pgen.1002300.

Reinius LE, Acevedo N, Joerink M, Pershagen G, Dahlén S-E, Greco D, Söderhäll C, Scheynius A, Kere J. Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. PLoS One. 2012;7(7):e41361.

Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. Epidemiology. 1992;3(2):143–55.

Shen-Orr SS, Tibshirani R, Khatri P, Bodian DL, Staedtler F, Perry NM, Hastie T, Sarwal MM, Davis MM, Butte AJ. Cell type-specific gene expression differences in complex tissues. Nat Methods. 2010;7(4):287–9. doi:10.1038/nmeth.1439.

Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol. 2004;3(1):3.

Suter M, Ma J, Harris A, Patterson L, Brown KA, Shope C, Showalter L, Abramovici A, Aagaard-Tillery KM. Maternal tobacco use modestly alters correlated epigenome-wide placental DNA methylation and gene expression. Epigenetics. 2011;6(11):1284–94. doi:10.4161/epi.6.11.17819.

Teh AL, Pan H, Chen L, Ong ML, Dogra S, Wong J, MacIsaac JL, Mah SM, McEwen LM, Saw SM, Godfrey KM, Chong YS, Kwek K, Kwoh CK, Soh SE, Chong MF, Barton S, Karnani N, Cheong CY, Buschdorf JP, Stunkel W, Kobor MS, Meaney MJ, Gluckman PD, Holbrook JD. The effect of genotype and in utero environment on interindividual variation in neonate DNA methylomes. Genome Res. 2014. doi:10.1101/gr.171439.113.

Teschendorff AE, Menon U, Gentry-Maharaj A, Ramus SJ, Gayther SA, Apostolidou S, Jones A, Lechner M, Beck S, Jacobs IJ, Widschwendter M. An epigenetic signature in peripheral blood predicts active ovarian cancer. PLoS One. 2009;4(12):e8274. doi:10.1371/journal.pone.0008274.

Teschendorff AE, Zhuang J, Widschwendter M. Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies. Bioinformatics. 2011;27(11):1496–505. doi:10.1093/bioinformatics/btr171.

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

VanderWeele TJ. Marginal structural models for the estimation of direct and indirect effects. Epidemiology. 2009;20(1):18–26. doi:10.1097/EDE.0b013e31818f69ce.

Wiencke JK, Accomando WP, Zheng S, Patoka J, Dou X, Phillips JJ, Hsuang G, Christensen BC, Houseman EA, Koestler DC, Bracci P, Wiemels JL, Wrensch M, Nelson HH, Kelsey KT. Epigenetic biomarkers of T-cells in human glioma. Epigenetics. 2012;7(12):1391–402. doi:10.4161/epi.22675.

Wilhelm-Benartzi CS, Houseman EA, Maccani MA, Poage GM, Koestler DC, Langevin SM, Gagne LA, Banister CE, Padbury JF, Marsit CJ. In utero exposures, infant growth, and DNA methylation of repetitive elements and developmentally related genes in human placenta. Environ Health Perspect. 2012;120(2):296–302. doi:10.1289/ehp.1103927.

Zou J, Lippert C, Heckerman D, Aryee M, Listgarten J. Epigenome-wide association studies without the need for cell-type composition. Nat Methods. 2014;11(3):309–11. doi:10.1038/nmeth.2815.

Chapter 3
A General Strategy for Inter-sample Variability Assessment and Normalisation

Zhen Yang and Andrew E. Teschendorff

Abstract The sources of inter-sample variability in omic studies are not only biological but often also technical. Assessment of the relative magnitude of biological and technical sources of variation is therefore of paramount importance, especially in the context of epigenome-wide association studies (EWAS) where biological signals are quantitative and may be of a relatively small magnitude. This chapter introduces the reader to a general strategy for determining the number and nature of the sources of variation in an omic data set. It further presents guidelines for inter-sample normalisation. Techniques and tools are illustrated throughout with examples from cancer epigenome and EWAS studies.

3.1 Introduction

Assessment of the sources of inter-sample variation is a key step in the analysis of any omic data set (Leek 2010). It is often the case that not all inter-sample variation is biological, with technical sources of variation also present, which may confound statistical analyses (Leek 2010). Indeed, not adjusting for confounding variation could dramatically skew estimates of statistical significance (Leek 2010; Leek and Storey 2007, 2008).

Confounding variation is most often technical; for instance, samples may be processed by different laboratories, on different dates, by different personnel or on different plates/chips. However, it is important to point out that in any study confounding variation can also be biological; for instance, when comparing normal

Z. Yang
Key Laboratory of Computational Biology, Chinese Academy of Sciences-Max Planck Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 320 Yue Yang Road, Shanghai 200031, China

A.E. Teschendorff (✉)
CAS Key Laboratory of Computational Biology, Chinese Academy of Sciences and Max-Planck Gesellschaft Partner Institute for Computational Biology, Shanghai, China

UCL Cancer Institute, University College London, London, UK
e-mail: a.teschendorff@ucl.ac.uk

A.E. Teschendorff (eds.), *Computational and Statistical Epigenomics*, Translational Bioinformatics 7, DOI 10.1007/978-94-017-9927-0_3

to cancer tissue, it could happen that the normal samples are not age matched. There are also circumstances when the confounding sources of variation may be unknown to the experimentalist or only known with error. For instance, we might detect a source of variation which correlates with the season in which the sample was collected (de Jong et al. 2014). Very likely, however, season is only a very imperfect surrogate of the real factor (perhaps temperature, ozone level, etc.) causing the variation, yet detailed information about the exposure of the sample to this factor may not have been recorded, or indeed the real causal factor may remain unknown. Thus, dealing with confounding variation constitutes a statistical as well as a biological challenge.

Confounding variation is particularly pertinent in the context of epigenome-wide association studies (EWAS), where the effect sizes of interest could be small in relation to the confounding sources of variation (Rakyan et al. 2011; Teschendorff et al. 2009). There are many aspects to consider when assessing if confounding variation could pose a problem in a specific EWAS study. Prominent among these is the tissue type in which the study is being performed. In contrast to genome-wide association studies, the mixture of cell types, which make up common tissues like blood, can obscure potential associations, since no two cell types have identical epigenomes and cell-type composition of tissues could change in response to the phenotype (Houseman et al. 2012, 2014; Langevin et al. 2014; Teschendorff et al. 2009). Thus, dealing with intra-sample heterogeneity is especially important in the context of EWAS studies (Houseman et al. 2014), and a whole chapter of this book is devoted to this problem (see Chap. 2 by Houseman).

However, tissue type is also important in relation to the specific phenotype of interest (POI) being considered. For instance, if one is interested in studying DNA methylation changes between normal and cancer tissues, it is very likely that technical factors (e.g. batch effects) can be ignored, because DNA methylation differences between normal and cancer tissues are normally quite large, typically around 50 %, if not higher. On the other hand, if one compares blood samples from healthy individuals to those of age-matched epithelial cancer patients, one may find that the magnitude of the DNA methylation changes in the blood are much smaller (on the order of 5–10 %) and thus comparable to the effects caused by technical factors like beadchip or plate (Teschendorff et al. 2009). That tissue type plays such a key role should be obvious, since in one case we are looking at the cell of origin for the disease in question, while in the other case we are studying a tissue (blood) which is unrelated to the cells that give rise to the epithelial cancer. EWAS studies conducted in easily accessible tissues such as blood or buccal epithelial cells are of particular interest, however, since it is plausible that these tissues may record epigenetic fingerprints in response to exposure to environmental cancer risk factors. For instance, recent EWAS studies have identified specific genomic sites which exhibit significant and reproducible, yet small, differences in DNA methylation between heavy smokers and nonsmokers (Philibert et al. 2012; Shenker et al. 2013; Zeilinger et al. 2013). These changes may typically represent shifts in average DNA methylation of only 1–5 %, if not lower. Thus, given the growing number of EWAS studies conducted in easily accessible tissues like blood, it has become of paramount importance to be able to critically assess the nature and relative magnitude of the sources of inter-sample variation.

To date, most epigenomic data generated are of two types: ChIP-Seq (or the older ChIP-chip) binding profiles of key chromatin marks and transcription factors and DNA methylation data (Gerstein et al. 2012). Owing to its biochemical stability and the fact that it can be measured genome wide in a high-throughput manner from limited amounts of DNA (thus easily amenable to the analysis of clinical samples), DNA methylation has emerged as the epigenetic mark of choice for cancer epigenome and EWAS studies (Rakyan et al. 2011). For this reason, we will focus in this chapter on DNA methylation data and specifically on the data generated using the Illumina 450k DNA methylation beadarrays, which offer so far the best compromise between cost, genome coverage and scalability (Sandoval et al. 2011). However, it is worth pointing out that the techniques and tools described in this chapter are applicable to any omic data type, including RNA-Seq, ChIP-Seq or whole-genome bisulfite sequencing (WGBS) data.

Thus, given the problem that confounding factors pose in EWAS studies, our aim with this chapter is to introduce the reader to a general strategy for dealing with this challenge. This strategy has been successfully used by us in numerous studies on DNA methylation and gene expression (see, e.g. Bell et al. 2010; Lechner et al. 2013; Teschendorff et al. 2009, 2010) and has been incorporated into an existing Bioconductor package (ChAMP, Morris et al. 2014). The chapter is organised as follows. In the next section, we address the dimensionality estimation problem, which is an important preliminary step when assessing the sources of inter-sample variation. Subsequently, we give a practical introduction to the singular value decomposition and illustrate how it provides a powerful framework in which to critically assess the sources of inter-sample variation. We then provide some brief guidelines as to how to perform inter-sample normalisation, referring to two of the most popular algorithms available for this purpose: the ComBat algorithm (Johnson et al. 2007) and SVA/ISVA (Leek and Storey 2007; Teschendorff et al. 2011). The overall strategy to inter-sample normalisation is then illustrated with a specific example from an EWAS study on smoking in blood tissue. We summarise the strategy in the final section.

3.2 Estimating the Dimensionality of Your Data Matrix

Because this chapter is devoted to inter-sample variation, we shall assume that our DNA methylation data matrix has already undergone basic preprocessing and quality control at the intra-sample level. Specifically, we assume that we have a correctly normalised data matrix at the intra-sample level. In order to assess the nature of the sources of variation in the data matrix, we must first determine the number of significant components of variation. Determining this number is a common problem in omic data analysis known informally as “dimensionality estimation”.

One of the most powerful statistical frameworks in omic data analysis, which we can use to address this dimensionality estimation problem, is the singular value decomposition (SVD) or principal component analysis (PCA). These algorithms infer components of iteratively maximal variation in the data, where the individual components are required to be linearly uncorrelated to each other. Thus, the components/singular vectors align along directions associated with maximal variance subject to the constraint that the variation they pick out is in some sense nonredundant, or orthogonal, to that of the other components.

Clearly, one of the assumptions in applying SVD or PCA to a data matrix is that the components of maximal variation are biologically interesting. Although this is the usual scenario, it is important to be aware that this is not always the case. We will return later to this important point. For the time being, let us consider what SVD does algebraically: it takes as input a $p \times n$ data matrix, X, where we assume that p labels the genomic features and n the number of samples, and so typically $p \gg n$. SVD then decomposes the data matrix X into a sum of components of variation (in the SVD context called singular vectors), with the components linearly uncorrelated to each other. Specifically, SVD on X decomposes X as the product of three matrices UDV^T, which in the matrix entry form is written as:

$$X_{ij} = \sum_{k=1}^{n} U_{ik} D_{kk} V_{jk} \tag{3.1}$$

with U and V as orthogonal matrices (i.e. $U^T U = I$ and $V^T V = I$) and D a diagonal matrix. Note that we are assuming that X is of full rank, i.e. rank n and not lower, which means that the columns of X are linearly independent and span an n-dimensional subspace of $\mathbb{R}^p$. Since typically $n \ll p$, there can be at most n linearly independent components of variation. Thus, U is of dimension $p \times n$, whereas D and V are both of dimension $n \times n$. The orthogonality constraints on U and V embody the "linear" decorrelation of the data covariance matrix. What is also important to note about this decomposition is that it is exact. In other words, if we wrote the SVD as

$$X = UDV^T + \epsilon \tag{3.2}$$

then the error term ϵ is exactly zero (i.e. $\epsilon = 0$).

It is key to understand what SVD does in practice and, in particular, how to interpret the matrices U, D and V. In this subsection, the most important of these matrices is D which represents a diagonal matrix $D = \mathrm{diag}\{d_1, d_2, \ldots, d_n\}$, with the diagonal entries always positive and ranked, i.e. $d_1 \geq d_2 \geq \ldots \geq d_n$. The normalised square of these diagonal entries approximate the fractional variation carried by the corresponding singular vector or component of variation. In fact, the fractional variation fV carried by component h would be

$$fV_h = \frac{d_h^2}{\sum_{k=1}^{n} d_k^2} \tag{3.3}$$

That the matrix D provides a ranking of the components in terms of the fraction of the total data variance they account for is one of the great features of SVD since it allows us to automatically assess the signal-to-noise ratio of our study.

The other matrices in the decomposition (U, V) tell us how the components/singular vectors vary across the genomic features (U) and samples (V). The columns of V are therefore particularly important, since they tell us how the components vary in relation to the biological and technical factors, i.e. our phenotypes of interest and batch effects.

The fact that the decomposition is exact ($\epsilon = 0$) means that in practice SVD yields a model which is overfitted to the data. Thus, SVD has been extended to probabilistic, Bayesian versions (Bishop 2006) which perform a reduced SVD by projecting onto a latent subspace, i.e. they try to estimate U, D, V matrices

$$X_{ij} = \sum_{k=1}^{K} U_{ik} D_{kk} V_{jk} + \epsilon_{ij} \tag{3.4}$$

by minimising the error term ϵ in a least-squares sense, where now $K < n$ and where U, D, V are of dimension $p \times K$, $K \times K$ and $n \times K$, respectively. In a Bayesian SVD formulation, it would be possible to then estimate the best possible value for K, the number of components or singular vectors used in the decomposition. We can think of K as providing us with an estimate of the dimensionality of the data matrix.

We stress that for most real data matrices observed in nature and in particular for those arising from biological (omic) data, the true dimensionality is not only not known but also very hard to define. Indeed, its value will also depend largely on the model used to estimate it. Nevertheless, obtaining an *approximate* value for the dimensionality of a data matrix is important for downstream statistical inference. For instance, it helps to know approximately how many columns of V we need to correlate to the biological and technical factors. A common fallacy is to focus on only the top 2 or 3 singular vectors, assuming (wrongly so) that lower-ranked components carry insignificant variance. To illustrate this point, an omic study performed with Illumina DNA methylation beadarrays on whole blood samples reported that a component of DNA methylation variation, which ranked only 5th in the SVD, correlated with the age of the patient the blood sample was taken from (Teschendorff et al. (2010)). The biological significance of this age component, which only carries a small fraction of the data variance, has been validated by many independent studies (see, e.g. Horvath 2013; Teschendorff et al. 2013). Thus, components of variation which may seem to carry little variance may nevertheless be of upmost biological importance. Hence, estimating the statistical significance of the variability carried by each component is an important endeavour.

Although, as mentioned earlier, we could use a probabilistic Bayesian SVD/PCA (Bishop 2006) to try to estimate K, it is also of interest to consider estimating K from within the deterministic SVD formulation. Indeed, ordinary SVD provides a fast means of estimating ranked and nonredundant components of decreasing variance, and so it is common to allow some degree of overfitting. It is therefore also natural to try to estimate data dimensionality within this deterministic SVD framework. While a technique, known as the Buja-Eyuboglu algorithm, based on explicit randomisation is possible (Buja and Eyuboglu 1992; Leek and Storey 2007), we here briefly describe a much faster analytical procedure which works remarkably well. It borrows a technique from a branch of nuclear physics and is known as random matrix theory (RMT). In a nutshell, RMT provides an analytical framework in which one can estimate the eigenvalue distribution of a sufficiently large "random" cross-correlation data matrix. Specifically, by a "random" cross-correlation matrix R of dimension $n \times n$, we mean a symmetric matrix of the form $R = \frac{1}{p} G^T G$ where G is of dimension $p \times n$ and where each column of G can be thought of as a random normal deviate of mean zero and unit standard deviation. According to RMT, the probability density of the eigenvalue distribution of a "random" cross-correlation matrix R takes the form

$$P_{rm}(\lambda) = \frac{Q}{2\pi} \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{\lambda} \tag{3.5}$$

where $Q = p/n$ and where we are also assuming that p and n are in some sense large (i.e. $p, n \gg 1$). In typical omic applications, $p \sim 10^5$ and $n \sim$ 50–100, which certainly satisfy the RMT requirements. In the above expression, λ_- and λ_+ represent the smallest and largest eigenvalues, which themselves are also determined analytically by

$$\lambda_\pm = 1 + \frac{1}{Q} \pm 2\sqrt{\frac{1}{Q}} \tag{3.6}$$

Given a real data matrix X of dimension $p \times n$, we can always normalise each feature to have mean zero and unit standard deviation and then construct the covariance matrix $C = \frac{1}{p} X^T X$. The eigenvalues μ_i of this data covariance matrix C (which will be related to the singular values of X through $\mu_i = d_i^2$) which are larger than λ_+ will finally give us an estimate of the dimensionality of the data matrix X. In other words, the observed eigenvalues μ_i, which are larger than the maximum expected under RMT, can be used as a means of selecting the significant components of variation.

As an example, let us consider an Illumina 450k DNA methylation data matrix consisting of 32 head and neck cancer samples analysed previously in Lechner et al. (2013). After imputation of missing data and intra-sample normalisation, the application of the RMT procedure using the EstDimRMT function of the *isva* R-package resulted in nine significant components (Fig. 3.1). As shown in Fig. 3.1, there are nine peaks (indicated as red vertical lines) in the density distribution of the observed eigenvalues, which are larger than the maximum of the null density (indicated in green). We shall see later, when discussing the sources of inter-sample variation, how this estimate of "9" significant components is sensible.

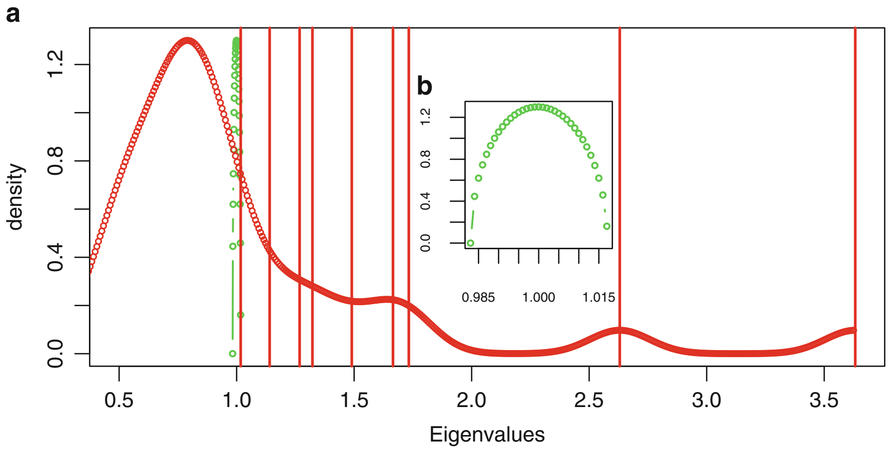


Fig. 3.1 (**a**) The density distribution of the observed eigenvalues (*red*) is compared to that of a random covariance matrix of the same dimension (*green points* and *curve*), as estimated by RMT. The *vertical lines* indicate the positions of peaks in the observed eigenvalue distribution which are larger than the maximum of the null distribution. These peaks correspond to the locations of the observed eigenvalues. Plot generated using the *EstDimRMT* function of the *isva* R-package. (**b**) Zoomed in version of the eigenvalue density of the random covariance matrix, as estimated by RMT

At this stage, several notes regarding RMT are in order: (1) First, it does assume that the data matrix is sufficiently large. Thus, RMT is not applicable to small data matrices, for instance, a data matrix of only 10 or fewer samples. In such a scenario, the permutation procedure of Buja-Eyuboglu is preferable (Buja and Eyuboglu 1992). (2) It is important, before applying the EstDimRMT from the *isva* R-package, to centre the data matrix, so that each row (genomic feature) has mean zero. Otherwise, the top component could just capture the trivial variation in the mean levels of the different genomic features, which is normally not of interest. This is particularly pertinent for DNA methylation studies, where CpGs often exhibit large differences in mean methylation depending on their genomic location and only much smaller variations in DNA methylation across samples. (3) Third, RMT takes as the reference/null eigenvalue distribution the one derived from a Gaussian random matrix. It could be argued that null components of variation might exist which are not Gaussianly distributed and that therefore this Gaussian assumption could lead to biased dimensionality estimates. While RMT may introduce some bias, in practice, this bias is small and does not present a problem. Moreover, a direct evaluation of this bias on real biological data is hard because the true dimensionality of biological data sets is rarely known.

3.3 Assessing the Sources of Inter-sample Variation: The SVD Heatmap

Having estimated the number of significant components of variation in the SVD, the next step is to assess the nature of these components. This is best achieved by generating a SVD P-value heatmap. The matrix that gives rise to this heatmap represents the P-values of association between the significant components of the SVD, i.e. the columns of the V matrix, with the various biological and technical factors. For instance, among the biological factors, we might have clinical outcome as our phenotype of interest (e.g. if the study is a cancer study aiming to find prognostic markers) and potential biological confounders such as age. Among the technical factors, we might have chip effects, processing date or, in the case of DNA methylation data, bisulfite conversion efficiency (BSCE).

The power of the SVD heatmap is that it tells us immediately what the top components of variation correlate with, i.e. are they biological or technical? Importantly, it also informs us how much of the data variation is accounted for by each of the significant components. Moreover, a given component could correlate simultaneously with a phenotype of interest and a technical factor, immediately raising the concern that the association with the POI could be driven by the technical factor. In combination with the fractional variance plot, we thus obtain an overall picture of the major sources of variation in the data, their relative contributions to data variation and whether adjustment for technical factors is necessary.

Returning to our previous head and neck cancer example, this Illumina 450k study aimed to assess whether DNA methylation profiles are different between HPV+ and HPV− head and neck cancer subtypes (Lechner et al. 2013). Thus, HPV status is our phenotype of interest. The SVD heatmap is shown in Fig. 3.2. As we can see, the top component of variation, accounting for just over 12 % of the total data variation, correlates with HPV status, although we can see that it also correlates with another biological factor (lymph node stage) as well as a technical factor, bisulfite conversion efficiency (as assessed by a specific technical control probe). Thus, although it is encouraging that the top PC correlates with our phenotype of interest, we must express caution since the SVD heatmap suggests that this association could be driven by one of these other confounding factors. We can also see, for instance, that the 4th component of variation is correlating with age. An association with age was only expected (Horvath 2013; Teschendorff et al. 2010), even despite the relatively small sample size of this particular study. Incidentally, this shows that RMT is correctly predicting that this component's variation is of significance. Another important observation to make is that technical probes measuring the efficiency of BSCE account for the variation seen in the lower-ranked components, notably for all components deemed significant by RMT. Although we can see that some of the components not deemed significant by RMT are correlating with other factors, these associations are generally speaking less significant and thus more likely to be false positives. We point out that while it is impossible to assess if the RMT dimensionality estimate of 9 is truly correct or not, based on the heatmap

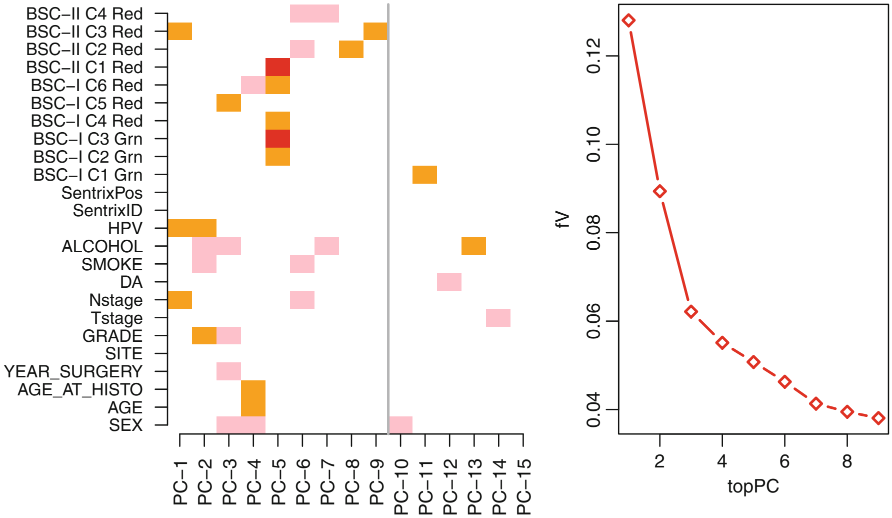


Fig. 3.2 *Left panel* depicts the SVD heatmap for an Illumina 450k study of 32 head & neck cancers. RMT predicted a total of 9 significant components, indicated by the *grey vertical line*. The x-axis of the heatmap label the top singular vectors (principal components) of a SVD, whilst the y-axis label technical and biological factors. The phenotype of interest is HPV status (HPV). Beadchip is indicated by SentrixID, Nstage indicates nodal stage. BSC indicates bisulfite conversion controls. The colours in the heatmap represent significance levels between the singular vectors and the factors. Colour codes: *dark red* ($P < 1e-10$), *red* ($P < 1e-5$), *orange* ($P < 0.001$), *pink* ($P < 0.05$), *white* (n.s) = not significant. *Right panel* depicts the fraction of data variation explained by the top singular vectors/principal components

figure we can see that the estimate appears within a range of biologically plausible values (6–13).

3.4 Inter-sample Normalisation Methods

Having generated the SVD heatmap of our study, this heatmap can now be used as a guide to assess whether inter-sample normalisation is necessary and, if so, what the best procedure might be. Before proceeding with our head and neck cancer example, it is important to make a distinction between three types of confounding factors, because the type of confounding variation we observe will influence our choice of inter-sample normalisation method.

One type of confounding variation (type I) is driven by factors which are known to the experimentalist. For instance, the beadchip or plate on which a sample was profiled is usually always well known to the experimentalist, so if we observe variation correlating with the beadchip, then explicit adjustment by inclusion of a covariate in the supervised regression model will work reasonably well

(Teschendorff et al. 2009). Empirical Bayes inter-sample normalisation methods like ComBat (Johnson et al. 2007) deal particularly well with confounding factors of this type. A second type of confounding variation (type II) could be driven by underlying factors which are unknown to the experimentalist or which have not been recorded but for which there is an approximate surrogate. For instance, in a study we might observe variation associated with bisulfite conversion efficiency, as assessed by an in-built technical control probe. It might be tempting to treat this as a confounding variation of the first type; however, the estimation of BSCE is itself subject to measurement error. Another example could be the season in which a sample has been collected. Variation associated with season likely indicates some other underlying causal factor, such as temperature or ozone levels, but the level of exposure of the sample to these factors may be unknown or may not have been recorded. How to deal best with this type of confounding variation is unfortunately still unclear. While we may use the surrogate (e.g. "season") to adjust the data using an algorithm such as ComBat, another possibility is to construct a surrogate variable from the data itself using a method like surrogate variable analysis (SVA) (Leek and Storey 2007, 2008) or independent surrogate variable analysis (ISVA) (Teschendorff et al. 2011). Indeed, there is some evidence that SVA or ISVA would work better in this scenario (Teschendorff et al. 2011). A third type of confounding variation (type III) is driven by factors which are unknown and for which there is also no known surrogate. For this type of confounding variation, ComBat or adjustment with explicit covariates is clearly not possible. In this kind of scenario, we need to use SVA or ISVA, which does not require prior knowledge of the confounding factors (Teschendorff et al. 2014).

It is important to elaborate further on the differences between ComBat, SVA and ISVA. SVA and ISVA represent *supervised normalisation* methods in the sense that the data is first adjusted for the phenotype of interest and sources of confounding variation are then modelled in the residual variation space. Mathematically, if we denote the data matrix by X and the POI by y and we assume a functional relationship between X and y as specified by a function f, we would first perform the regression

$$X = f(y) + R \tag{3.7}$$

Surrogate variables which model the confounding factors are then constructed in the residual variation space defined by the estimated matrix of residuals R and are later incorporated as covariates in the final supervised regression model. In the case of SVA, the confounding sources of variation in the residual variation space are obtained by applying SVD/PCA to the matrix R (Leek and Storey 2007, 2008), whereas ISVA (Teschendorff et al. 2011) replaces SVD/PCA with a blind source separation algorithm (Teschendorff et al. 2014), specifically with independent component analysis (ICA) (Comon 1994). The surrogate variables are constructed in a manner which avoids overfitting. Both of these algorithms result in the construction of a number L of surrogate variables, which we can describe in terms of a $n \times L$ covariate matrix Γ. As mentioned, the last step in the SVA/ISVA procedure is then to run the model

$$X = f(y) + \lambda\Gamma + \epsilon \tag{3.8}$$

where now the error term ϵ represents approximately Gaussian white noise. The beauty of SVA or ISVA is that in principle there is no requirement for us to know the confounding factors in advance. Indeed, in principle, these algorithms should work even if confounding factors are unknown or if the true confounders are only known with error (in which case we do not wish to use them as explicit covariates in the supervised regressions).

These methods are in stark contrast to ComBat, which does not use the phenotype of interest to perform the normalisation. Thus, while ComBat returns a normalised data matrix, from which one can subsequently draw inferences about any potential phenotype of interest, SVA/ISVA performs a supervised analysis for a *given* phenotype of interest while adjusting for potential confounding factors. This is a key difference. Consequently, although SVA/ISVA returns "surrogate variables", which could be used to normalise/adjust the data matrix, it is important to remember that these surrogate variables were constructed in a residual variation space deemed orthogonal to the POI, and hence they are *dependent on the POI*. Thus, it would be incorrect to use surrogate variables inferred from running SVA/ISVA with normal/cancer status as the phenotype of interest, say, to normalise the data matrix for subsequent inferences or supervised analysis in relation to other phenotypes of interest, such as age or prognostic cancer outcome. Thus, if one desires to find features which correlate with two different phenotypes of interest, then SVA/ISVA needs to be run twice, in each case using a different POI and therefore a *different* set of surrogate variables.

Now, let us return to our head and neck cancer example. As we can see from Fig. 3.2, none of the significant components correlate with potential batch effects such as beadchip or the year in which the DNA sample was collected. Had we observed a strong beadchip effect, say, it would have been advisable to adjust the DNA methylation data for this categorical factor using the popular ComBat algorithm (Johnson et al. 2007). The same heatmap indicates, however, that the top singular vector is not only correlating with our POI, i.e. HPV status, but also with BSCE, as assessed by one of the control probes. Although in principle one could use ComBat to adjust for BSCE, this type of confounding variation (type II) may be best dealt with by using SVA or ISVA. We refer the reader who wants to learn more about SVA/ISVA to Teschendorff et al. (2014).

In what follows, we shall illustrate the application of one of these algorithms (ISVA) to the analysis of an EWAS DNA methylation data matrix subject to unknown confounding variation (type-III variation).

3.5 A Case Study: An EWAS for Smoking in Blood Tissue

We consider an Illumina 450k data set, encompassing 152 whole blood samples from women for which extensive epidemiological information is available (Anjum et al. 2014). Of particular interest is to identify CpGs whose DNA methylation level correlates with smoking pack-years, an epidemiological indicator of an individual's smoking history. This is a nice example to consider, because by now it is well established that smoking affects DNA methylation profiles in the blood and that there are highly reproducible smoking-associated differential methylation loci (Shenker et al. 2013; Zeilinger et al. 2013). After preprocessing, QC, intra-sample normalisation and dimensionality estimation using RMT, the SVD heatmap corresponding to the data matrix is shown in Fig. 3.3. As we can see, the top singular vector, accounting for over 8 % of data variation, does not correlate with any of the main biological, epidemiological or technical factors (and many factors have been suppressed in the figure for ease of visualisation). Moreover, lower-ranked components correlate most strongly with beadchip effects. Thus, if our aim is to identify features associated with smoking, we would need to be concerned about all these potentially confounding variations, which we note is also of higher magnitude (appearing in components 2–6) than the variation due to smoking (which appears in the 7th component). Indeed, using a linear regression model between smoking pack-years and DNA methylation profiles, one obtains an oddly shaped P-value histogram as shown in Fig. 3.3c. A P-value histogram where at any point the density increases in the direction of increasing P-values indicates that there is either confounding variation or that the wrong statistical test has been implemented (Leek and Storey 2008). In this case, there is no reason to believe that the oddly shaped histogram is caused by adopting a linear regression model between POI and the DNA methylation profiles. Indeed, it is much more likely that the confounding variation shown in Fig. 3.3 is causing "would-be" highly significant P-values to become less significant (i.e. P-values are being "shifted" to the right in Fig. 3.3c).

Thus, because the top component is also being driven by an unknown factor, it is ideal to approach this problem using either the SVA or ISVA framework. Here, we consider the case of ISVA. An important step in the ISVA procedure is to generate the analogue of the SVD heatmap, but now for the inferred ISVs (Fig. 3.4). These ISVs are inferred using ICA on the residual variation space, i.e. the matrix of residuals obtained after regressing out the data variance due to our phenotype of interest (here smoking). As shown in this figure, there are ISVs correlating strongly with potential technical confounders such as beadchip and BSCE. We note that there are also a few ISVs correlating, albeit very weakly so, with our POI (smoking). Because these correlations with smoking are weak, corresponding to P-values in the range $0.001 < P < 0.05$, these associations are not likely to pass multiple-testing correction, and so are likely false positives. Also note that some of the ISVs do not correlate with any factor. Given that in the original SVD heatmap the top component was driven by an unknown factor, it is very likely that one or more of these ISVs is capturing this top singular vector.

At this point, having generated the ISV heatmap, there are two modes in which ISVA can subsequently operate. In the "agnostic" mode, we use all ISVs as covariates in the final supervised regression model. In the *isva* R-package, this corresponds to using the *cf.m=NULL* option in the *DoISVA* function. This "agnostic" mode is incidentally also the mode in which SVA operates. This mode

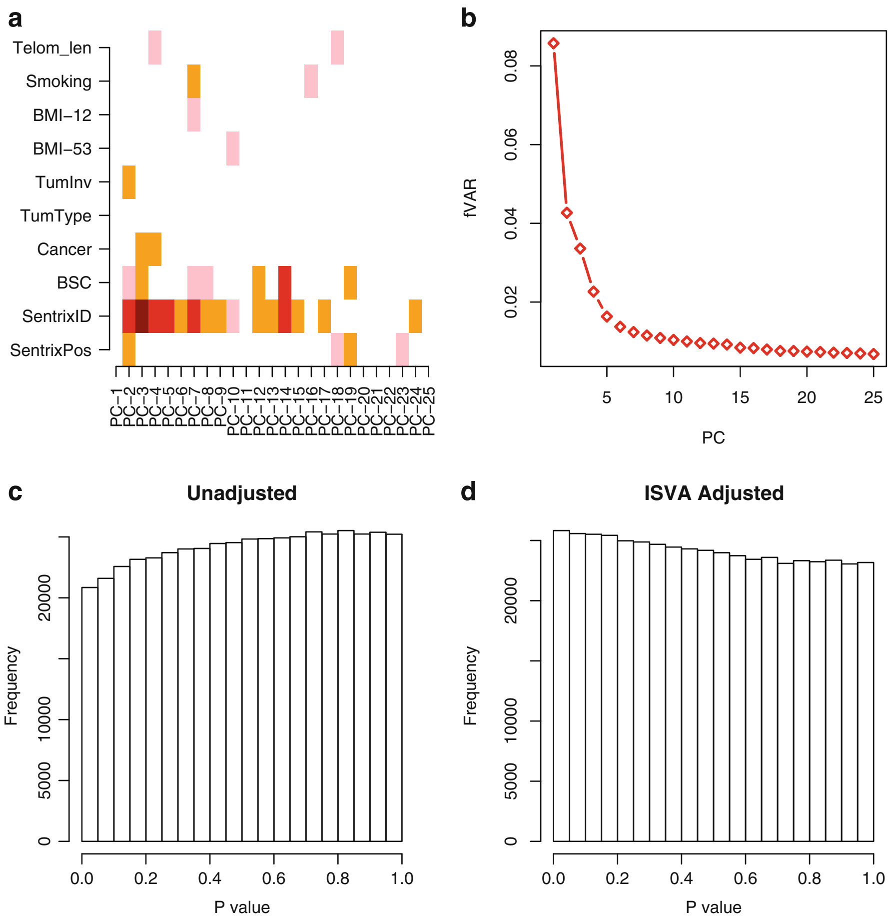


Fig. 3.3 (**a**) SVD heatmap of a 152 whole blood Illumina 450k set. RMT predicted 25 significant components of variation. Shown are a selection of biological and technical factors, including Sentrix position, beadchip (Sentrix ID), bisulfite conversion efficiency (BSC), cancer status, tumour type, invasive tumour status, body mass index at age 53 (age of sample collection), body mass index at age 12, smoking pack-years and telomere length. The colours in the heatmap represent significance levels between the singular vectors and the factors. Colour codes: *dark red* ($P < 1e-10$), *red* ($P < 1e-5$), *orange* ($P < 0.001$), *pink* ($P < 0.05$), *white* (n.s.). (**b**) Fraction of total data variation explained by each significant component. (**c**) P-value histogram of a supervised regression of DNA methylation profiles against smoking pack-years without adjustment for confounding factors. (**d**) Corresponding P-value histogram obtained by the application of ISVA

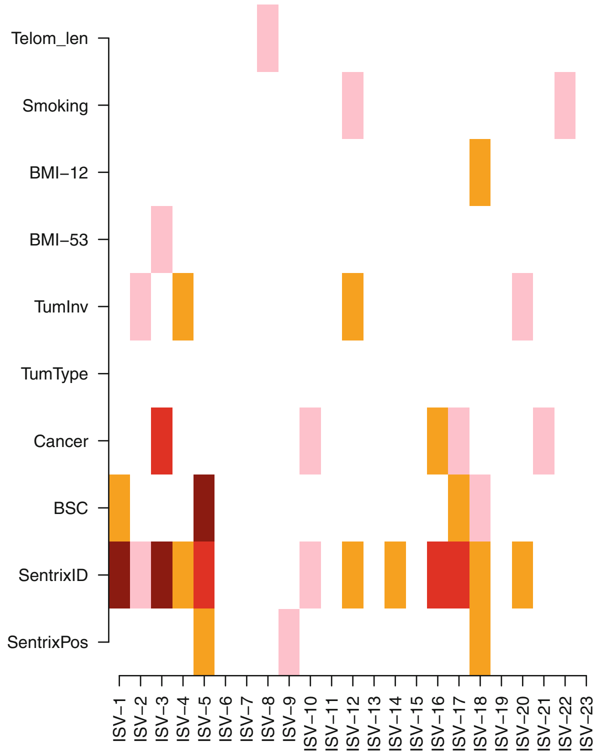


Fig. 3.4 ISV heatmap of an Illumina 450k study of 152 whole blood samples from healthy women with smoking as the POI. RMT predicted a 23-dimensional residual variation space. The x-axis of the heatmap labels the independent surrogate variables inferred from ICA in the residual variation space, while the y-axis labels some of the technical and biological factors. The colours in the heatmap represent significance levels between the singular vectors and the factors. Colour codes: *dark red* ($P < 1e-10$), *red* ($P < 1e-5$), *orange* ($P < 0.001$), *pink* ($P < 0.05$), *white* (n.s.)

works fine as long as the model used in the supervised analysis is a "good" model for describing the effect of the POI on the data. If the model is poor, for instance, if we assume a linear model between POI and the feature profiles when the dependence is clearly nonlinear, then this "agnostic mode" can break down (Teschendorff et al. 2011). Thus, when implementing this mode, caution needs to be exercised, and this is why we generated the ISV heatmap of the residual variation space to assess if this operating mode is appropriate or not. If examination of this heatmap were to reveal a residual component of variation correlating very strongly with the POI, then we might lose genuine biological signal by assuming that this residual component represents confounding variation and using it as a covariate in the final regression model (Teschendorff et al. 2014).

Alternatively, we could also run ISVA in a "non-agnostic" mode, where we declare our prior belief that specific factors may be confounders. For instance, from

Fig. 3.3a, we can see that beadchip and BSCE are potential confounders. Moreover, these same factors are also prominent in the residual variation space (Fig. 3.4). Thus, we could construct a confounding factor matrix *cf.m* of dimension 152×2 since there are 152 samples and 2 "potential" confounding factors. Running ISVA in this mode would then select specific ISVs as covariates in the final supervised regression analysis. In this case, based on Fig. 3.4, we would select ISVs 1,3,4,5,12,14,16,17,18 and 20, because these are the ones which are strongly associated with the potential confounders.

Based on the heatmaps shown in Figs. 3.3 and 3.4, which of the two modes should we adopt? In this case, it seems more reasonable to adopt an agnostic approach in which we include all ISVs as covariates in the final supervised regression analysis. This is because we know of the presence of an unknown confounder that carries most of the data variance and which will be present in the residual variation space. Using all ISVs as covariates thus ensures that we are also including this source of variation. Application of ISVA subsequently results in a P-value histogram as shown in Fig. 3.3d. Note that this P-value histogram is now exhibiting the monotonic decreasing trend, with a flattening out of the density distribution as P-values become larger, as required statistically. Thus, after ISVA adjustment, we have a statistically sensible P-value histogram, from which we can now derive more accurate estimates of genome-wide statistical significance. Indeed, we find after ISVA adjustment that there are 131 Cp Gs passing an FDR < 0.3, in contrast to only 67 Cp Gs obtained from a linear regression model without any adjustment. Thus, using ISVA we find twice as many features. Identifying more features, however, does not mean that we are identifying more true positives. Reassuringly, if we were to inspect the ISVA-derived top ranked CpGs, we would find that many map to well-known smoking-associated genes like *AHRR, CYP1A1 and F2RL3* (Shenker et al. 2013; Zeilinger et al. 2013). However, it is also especially important to inspect the list of probes obtained using ISVA which were missed in the unadjusted analysis. In this particular case, we find, for instance, a probe mapping to the shore of the CpG island region of the retinoic acid receptor alpha (*RARA*) gene, which was not observed in the unadjusted analysis. Given that independent evidence exists that smoking leads to differential methylation of the *RARA* gene (Manoli et al. 2012), this supports the view that ISVA improves the sensitivity of the assay.

3.6 Conclusions

In this chapter we have aimed to provide the reader with a very general strategy of how to critically assess inter-sample variation and how to subsequently act in terms of the inter-sample normalisation procedure. We do not provide an explicit pipeline for the following reasons. First, a pipeline assumes that all data sets have a similar study design. However, in our experience, data sets contain more often than not unique features which do not render them amenable to an inter-sample pipeline approach. Second, blind application of pipelines can lead to suboptimal or even

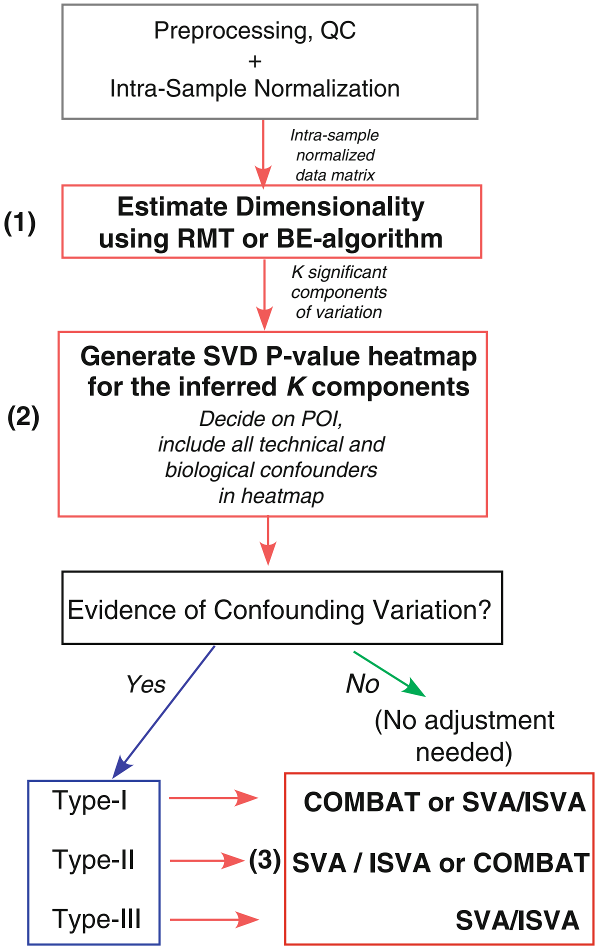


Fig. 3.5 Outline of the general strategy proposed for assessing and dealing with inter-sample variation

wrong analyses, especially when dealing with a highly complex problem. Thus, whereas pipelines for reading in data and intra-sample normalisation are extremely useful, the complexity of inter-sample variation often makes a pipeline approach unfeasible. Indeed, the nature of inter-sample variation is highly study specific and so unique features of a given study will often also require a unique approach to inter-sample normalisation.

Our general strategy and recommendation is based on a three-step approach, summarised in Fig. 3.5. Briefly, once a data set has been normalised at the intra-sample level, we recommend first estimating the dimensionality of the data matrix, i.e. to estimate the approximate number of significant components of variation. In most cases, RMT will be applicable to estimate this number. Second, we advise generating a SVD P-value heatmap to identify the nature of the significant sources of variation. This heatmap should then be used as a guide to decide if confounding variation in relation to the POI is present and, if so, how to then adjust for it. Although unsupervised batch-effect normalisation methods like ComBat are extremely useful, it is important to be aware that explicit adjustment for known batch effects is not possible if confounders are unknown or if the confounding variation is of a more complex nature. Indeed, we have shown an example of an EWAS study of smoking in blood tissue where unknown variation precluded the direct application of an algorithm such as ComBat. Thus, algorithms like SVA or ISVA, which can be applied more generally, are important methods to consider to help us obtain more reliable estimates of genome-wide statistical significance.

Acknowledgements AET is supported by the Chinese Academy of Sciences and the Max Planck Society.

References

Anjum S, Fourkala EO, Zikan M, Wong A, Gentry-Maharaj A, Jones A, Hardy R, Cibula D, Kuh D, Jacobs IJ, Teschendorff AE, Menon U, Widschwendter M. A BRCA1-mutation associated DNA methylation signature in blood cells predicts sporadic breast cancer incidence and survival. Genome Med. 2014;6(6):47.

Bell CG, Teschendorff AE, Rakyan VK, Maxwell AP, Beck S, Savage DA. Genome-wide DNA methylation analysis for diabetic nephropathy in type 1 diabetes mellitus. BMC Med Genomics. 2010;3:33.

Bishop CM. Pattern recognition and machine learning. New York: Springer; 2006.

Buja A, Eyuboglu N. Remarks on parallel analysis. Multivar Behav Res. 1992;27(4):509–40.

Comon P. Independent component analysis, a new concept? Signal Process. 1994;36(3):287–314.

de Jong S, Neeleman M, Luykx JJ, ten Berg MJ, Strengman E, den Breeijen HH, Stijvers LC, Buizer-Voskamp JE, Bakker SC, Kahn RS, Horvath S, van Solinge WW, Ophoff RA. Seasonal changes in gene expression represent cell-type composition in whole blood. Hum Mol Genet. 2014;23(10):2721–8.

Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, Cheng C, Mu XJ, Khurana E, Rozowsky J, Alexander R, Min R, Alves P, Abyzov A, Addleman N, Bhardwaj N, Boyle AP, Cayting P, Charos A, Chen DZ, Cheng Y, Clarke D, Eastman C, Euskirchen G, Frietze S, Fu Y, Gertz J, Grubert F, Harmanci A, Jain P, Kasowski M, Lacroute P, Leng J, Lian J, Monahan H, O'Geen H, Ouyang Z, Partridge EC, Patacsil D, Pauli F, Raha D, Ramirez L, Reddy TE, Reed B, Shi M, Slifer T, Wang J, Wu L, Yang X, Yip KY, Zilberman-Schapira G, Batzoglou S, Sidow A, Farnham PJ, Myers RM, Weissman SM, Snyder M. Architecture of the human regulatory network derived from encode data. Nature. 2012;489(7414):91–100.

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

Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, Wiencke JK, Kelsey KT. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinform. 2012;13:86.

Houseman EA, Molitor J, Marsit CJ. Reference-free cell mixture adjustments in analysis of DNA methylation data. Bioinformatics. 2014;30(10):1431–9.

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

Langevin SM, Houseman EA, Accomando WP, Koestler DC, Christensen BC, Nelson HH, Karagas MR, Marsit CJ, Wiencke JK, Kelsey KT. Leukocyte-adjusted epigenome-wide association studies of blood from solid tumor patients. Epigenetics. 2014;9(6):884–95.

Lechner M, Fenton T, West J, Wilson G, Feber A, Henderson S, Thirlwell C, Dibra HK, Jay A, Butcher L, Chakravarthy AR, Gratrix F, Patel N, Vaz F, O'Flynn P, Kalavrezos N, Teschendorff AE, Boshoff C, Beck S. Identification and functional validation of HPV-mediated hypermethylation in head and neck squamous cell carcinoma. Genome Med. 2013;5(2):15.

Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet. 2010;11(10):733–9.

Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. 2007;3(9):1724–35.

Leek JT, Storey JD. A general framework for multiple testing dependence. Proc Natl Acad Sci U S A. 2008;105(48):18718–23.

Manoli SE, Smith LA, Vyhlidal CA, An CH, Porrata Y, Cardoso WV, Baron RM, Haley KJ. Maternal smoking and the retinoid pathway in the developing lung. Respir Res. 2012;13:42.

Morris TJ, Butcher LM, Feber A, Teschendorff AE, Chakravarthy AR, Wojdacz TK, Beck S. ChAMP: 450k chip analysis methylation pipeline. Bioinformatics. 2014;30(3):428–30.

Philibert RA, Beach SR, Brody GH. Demethylation of the aryl hydrocarbon receptor repressor as a biomarker for nascent smokers. Epigenetics. 2012;7(11):1331–8.

Rakyan VK, Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. Nat Rev Genet. 2011;12(8):529–41.

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

Shenker NS, Polidoro S, van Veldhoven K, Sacerdote C, Ricceri F, Birrell MA, Belvisi MG, Brown R, Vineis P, Flanagan JM. Epigenome-wide association study in the European prospective investigation into cancer and nutrition (EPIC-turin) identifies novel genetic loci associated with smoking. Hum Mol Genet. 2013;22(5):843–51.

Teschendorff AE, Menon U, Gentry-Maharaj A, Ramus SJ, Gayther SA, Apostolidou S, Jones A, Lechner M, Beck S, Jacobs IJ, Widschwendter M. An epigenetic signature in peripheral blood predicts active ovarian cancer. PLoS One. 2009;4(12):e8274. doi: 10.1371/journal.pone.0008274

Teschendorff AE, Menon U, Gentry-Maharaj A, Ramus SJ, Weisenberger, DJ, Shen H, Campan M, Noushmehr H, Bell CG, Maxwell AP, Savage, DA, Mueller-Holzner E, Marth C, Kocjan G, Gayther SA, Jones A, Beck S, Wagner W, Laird PW, Jacobs IJ, Widschwendter M. Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. Genome Res. 2010;20(4):440–6.

Teschendorff AE, Renard E, Absil PA. Supervised normalisation of large-scale omic datasets using blind source separation. In: Ganesh RN, Wenwu W, editors. Blind source separation: advances in theory, algorithms and applications. Berlin: Springer; 2014.

Teschendorff AE, West J, Beck S. Age-associated epigenetic drift: implications, and a case of epigenetic thrift? Hum Mol Genet. 2013;22(NA):R7–15.

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

Zeilinger S, Kuhnel B, Klopp N, Baurecht H, Kleinschmidt A, Gieger C, Weidinger S, Lattka E, Adamski J, Peters A, Strauch K, Waldenberger M, Illig T. Tobacco smoking leads to extensive genome-wide changes in dna methylation. PLoS One. 2013;8(5):e63,812.

Chapter 4
Quantitative Comparison of ChIP-Seq Data Sets Using MAnorm

Zhen Shao and Yijing Zhang

Abstract ChIP-Seq is widely used to characterize genome-wide binding patterns of transcription factors and other chromatin-associated proteins. Although comparison of ChIP-Seq data sets is critical for understanding cell-type-dependent binding, and thus the study of cell-type-specific regulation, few quantitative approaches have been developed. This chapter describes a simple and effective method, MAnorm, for quantitative comparison of ChIP-Seq data sets. It exhibits good performance when applied to ChIP-Seq data for both epigenetic modifications and transcription factor binding site identification. The quantitative binding differences inferred by MAnorm strongly correlate with both the changes in expression of target genes and the binding of cell-type-specific regulators. Comparisons to prior methods using genome-wide signals for normalization reveal that output of MAnorm contains much lower level of bias and better reflects authentic biological changes. At the end of this chapter, an integrative pipeline of using MAnorm to identify high-confidence cell-type-specific enhancers will be presented, which can serve as a simple but illustrative example of downstream applications.

Keywords Chip-Seq • Quantitative comparison • Enhancer • Histone modification

4.1 Introduction

Chromatin immunoprecipitation followed by massively parallel DNA sequencing (ChIP-Seq) has become the preferred method to determine genome-wide binding patterns of transcription factors and other chromatin-associated proteins

Z. Shao (✉)
Key Laboratory of Computational Biology, Chinese Academy of Sciences-Max Planck Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 320 Yue Yang Road, Shanghai 200031, China
e-mail: shaozhen@picb.ac.cn

Y. Zhang
Institute of Plant Physiology and Ecology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Fenglin Road 300, Shanghai 200032, China

A.E. Teschendorff (eds.), *Computational and Statistical Epigenomics*,
Translational Bioinformatics 7, DOI 10.1007/978-94-017-9927-0_4

(Park 2009). With the rapid accumulation of ChIP-Seq data, comparison of multiple ChIP-Seq data sets becomes critical in addressing various biological questions. For example, comparison of biological replicates is commonly used to find robust binding sites, and identifying sites that are differentially bound by chromatin-associated proteins in different cellular contexts is important in elucidating underlying mechanisms of cell-type-specific regulation. However, although quite a number of methods have been proposed for finding ChIP-enriched regions in a ChIP-Seq sample compared to a suitable negative control (e.g., mock IP or nonspecific IP) (Ji et al. 2008; Zhang et al. 2008; Rozowsky et al. 2009), few methods have been proposed for comparison of ChIP-Seq samples. The simplest approach classifies peaks from each sample as either common or unique based on whether or not they overlap with peaks in other samples (Fujiwara et al. 2009; Liu et al. 2010a; Schmidt et al. 2010). Although this method can identify general relationships between peak sets from different samples, the results are highly dependent on the cutoff used in peak calling, which cannot be done in a completely objective manner. Consequently, quantitative comparison of ChIP-Seq samples based on a proper way of signal normalization, while important for extracting maximal biological information, is fraught with numerous challenges.

An intuitive and widely used approach of quantitative comparison relies on rescaling data on the basis of the total number of sequence reads. However, this method is inadequate and may introduce errors when the signal-to-background-noise (S/N) ratio varies between samples. Recently, statistical tools have been developed to discover regions that exhibit significant differences between two ChIP-Seq data sets. For example, Xu et al. (2008) proposed a Hidden Markov Model based method to detect broad chromatin domains associated with distinct levels of histone modifications between two cell types. Also, some peak-calling programs propose to identify differential binding regions between two ChIP-Seq data sets by using one data set as sample and the other as control (Zhang et al. 2008; Rozowsky et al. 2009). Since these methods also rely on the total number of reads (or background region reads) to rescale the data, they fail to circumvent problems associated with different S/N ratios. In an alternative approach, Taslim et al. proposed a nonlinear method that uses locally weighted regression (LOWESS) for ChIP-Seq data normalization (Taslim et al. 2009). The underlying assumption of this method is that the genome-wide distribution of read densities has equal mean value and variance across samples (Taslim et al. 2009). A potential problem with this approach is that global symmetry will be introduced after normalization, an assumption that is questionable when comparing biological samples with different numbers of binding sites. In addition, this method normalizes samples based on the absolute difference of read counts instead of log2ratio commonly used in traditional MA-plot methods (Smyth and Speed 2003) and thus the differences deduced by this method cannot be used directly for quantitative comparison with other observations of biological significance, such as fold changes in gene expression.

This chapter describes a simple and effective model, MAnorm, to quantitatively compare ChIP-Seq data sets (Shao et al. 2012). To circumvent the issue of differences in S/N ratio between samples, we focused on ChIP-enriched regions (peaks), and introduced a novel idea, that ChIP-Seq common peaks could serve as reference to build the rescaling model for normalization. This approach is based on the empirical assumption that if a chromatin-associated protein has a substantial number of peaks shared in two conditions, the binding at these common regions will tend to be determined by similar mechanisms, and thus should exhibit similar global binding intensities across samples. This idea is further supported by motif analysis that we present. MAnorm exhibited good performance when applied to ChIP-Seq data for both epigenetic modifications and transcription factor binding site identification. Importantly, quantitative differences inferred by MAnorm are strongly correlated with differential expression of target genes and the binding of cell-type-specific regulators. Moreover, compared to prior methods using genome-wide signals for normalization, MAnorm showed obviously lower bias in quantifying the difference of ChIP-Seq intensities and better detected authentic biological changes. Therefore, it can be used as a powerful tool in probing mechanisms of gene regulation. At the end of this chapter, we will present a schematic work flow of using the quantitative differences of enhancer-associated histone marks between two distinct developmental stages of human erythroid cells to identify high-confidence stage-specific as well as nonspecific enhancers, to serve as a simple example of downstream applications of MAnorm model.

4.2 Work Flow of MAnorm Model

Data normalization is an important step in sequencing data analysis. However, normalization of ChIP-Seq data is a difficult task due to the differential S/N ratio across samples (Xu et al. 2008; Taslim et al. 2009). These differences cannot be simply addressed by using traditional microarray data normalization methods, such as quantile normalization (Bolstad et al. 2003) and MA-plot followed by LOWESS regression (Smyth and Speed 2003). Here we borrow the idea of MA-plot and propose a method for quantitative comparison of ChIP-Seq data sets based on two empirical assumptions. First, we assume the true intensities of most common peaks should be the same between two ChIP-Seq samples. This is appropriate when these binding regions show a much higher level of co-localization between samples than that expected at random, and thus binding at the common peaks should be determined by similar mechanisms and exhibit similar global binding intensity between samples. Second, the observed differences in sequence read density in common peaks are presumed to reflect the scaling relationship of ChIP-Seq signals between two samples, which can thus be applied to all peaks. Based on these hypotheses, a novel computation model called MAnorm is proposed to quantitatively compare two ChIP-Seq data sets of the same factor but from different cell types or states based on the scaling relationship inferred from common peaks.

The work flow of MAnorm is summarized in Fig. 4.1. First, four bed files that describe the coordinates of all predefined peaks and aligned sequence reads of two ChIP-Seq samples were used as input. Second, MAnorm calculated the number of

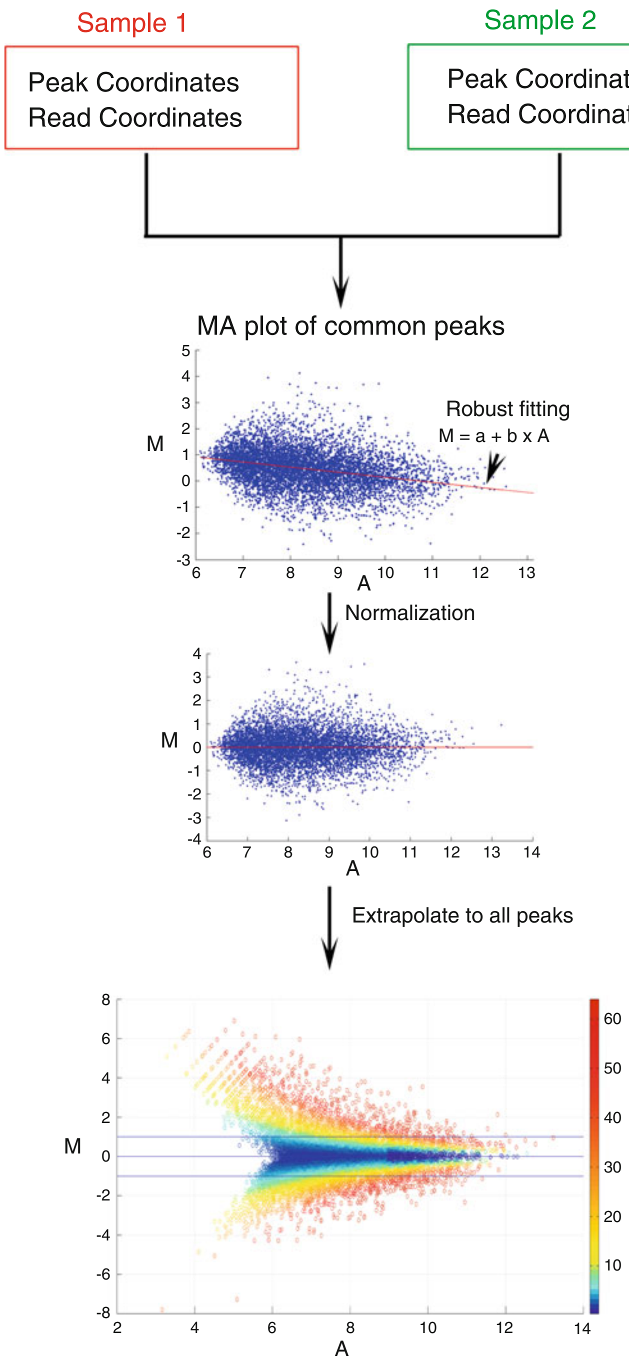


Fig. 4.1 Work flow of MAnorm model

reads in a window of the same length centered at the summit of each peak. Here the window size should be comparable to the median length of ChIP-enriched regions; we recommend 2,000 bp window size for histone modifications and 1,000 bp for transcription factor binding sites. The (*M*, *A*) value of each peak is then defined as:

$$\mathbf{M} = \mathbf{log_2}\left(\mathbf{R_1}/\mathbf{R_2}\right), \tag{4.1}$$

and

$$\mathbf{A} = \mathbf{log_2}\left(\mathbf{R_1} \times \mathbf{R_2}\right)/\mathbf{2}. \tag{4.2}$$

Here R_1 is the read density at this peak region in ChIP-Seq sample 1 and R_2 is the corresponding read density in sample 2. To avoid $\log_2 0$, we added a value of 1 to the real number of reads for all peaks. Thus, the value of *M* describes the $\log_2$ fold change of the read density at a peak region between two samples, while *A* represents the average signal intensity in terms of $\log_2$-transformed read density. To build the normalization model, each peak of the two samples being compared was further classified as a common or a unique peak, depending on whether or not it overlapped (by at least 1 nucleotide in this study) with any peak in the other sample. MAnorm model also provides a parameter for users to select common peaks based on a cutoff of peak summit-to-summit distance. By default, this value is set to 500 bp for histone modifications and 250 bp for transcription factors. In addition, when a peak overlaps with multiple peaks in the other sample, MAnorm selects the peak with the smallest summit-to-summit distance to avoid potential bias in building the normalization model. Next, robust regression, using iterative reweighted least squares with a bisquare weighting function (McKean 2004), was applied to the *M*-*A* values of common peaks and a linear model

$$\mathbf{M} = \mathbf{a} + \mathbf{b} \times \mathbf{A} \tag{4.3}$$

was derived to fit the global dependence between the (*M*, *A*) values of these peaks. To normalize the (*M*, *A*) values of all peaks, MAnorm performed coordinate transformation to make the *A* axis overlap with the linear model derived from regression. Then the corresponding (*M*, *A*) value under the new coordinate system was taken as the normalized (*M*, *A*) value of each peak. Here, the normalized *M*-value can be used as quantitative measure of differential binding in each peak region between two samples, with peak regions associated with larger absolute *M*-values exhibiting greater differences in binding. Finally, a *P*-value associated with each peak was calculated to quantify the significance of differential binding by modifying the Bayesian model developed by Audic and Claverie (1997) and applying it to the ChIP-Seq signals normalized by our method:

$$\mathbf{P}\left(\mathbf{R_2^*}\middle|\mathbf{R_1^*}\right) = \left(\mathbf{R_1^*} + \mathbf{R_2^*}\right)! \,/\, \left(\mathbf{R_1^*}! \times \mathbf{R_2^*}! \times \mathbf{2}^\wedge \left(\mathbf{R_1^*} + \mathbf{R_2^*} + \mathbf{1}\right)\right). \tag{4.4}$$

Here R^*_1 and R^*_2 are the normalized read counts of each peak in sample 1 and 2, respectively, which are calculated from the normalized (M, A) value of this peak by reversing Eqs. 4.1 and 4.2. When the read densities at most peak regions are high, most peaks associated with absolute M-values higher than 1 are associated with significant P-values. Then, the M-value can be used to rank peaks and select differential binding regions. When read densities at most peak regions are relatively low, some of the peaks associated with absolute M-values higher than 1 may still fail to obtain significant P-values. In such a case, we suggest ranking peaks by P-values and defining differential binding regions using combined cutoffs of both M-value and P-value.

The output of MAnorm includes the normalized (M, A) value and the corresponding P-value of each peak. To illustrate the normalization process, the (M, A) values of all peaks before and after normalization are plotted together with the linear model derived from common peaks. The MAnorm package will also generate three bed files presenting the genome coordinates for the nondifferential binding region and two differential binding regions based on user-specified cutoffs, together with two wig files (corresponding to the two peak lists under comparison) that can be uploaded to a genome browser for visualization of the M-value for each peak. The MAnorm packages written by different languages are available for downloading under link http://bioinfo.sibs.ac.cn/shaolab/opendata.php.

4.3 Use MAnorm to Perform Quantitative Comparison of ChIP-Seq Data Sets

4.3.1 Compare ChIP-Seq Data Sets of Histone Marks Between Different Cell Types

We applied MAnorm to analyze the differences of H3K27ac, a histone mark of active promoters as well as distal enhancers (Creyghton et al. 2010; Rada-Iglesias et al. 2011) and positively associated with gene expression, between H1 human embryonic stem (ES) cells and K562 human myeloid leukemia cell line. We downloaded corresponding ChIP-Seq data sets from the website of ENCODE project (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeBroadHistone/) and performed peak calling using MACS (Zhang et al. 2008). Peaks identified in these two cell lines showed substantial overlap, with the overlap being 17-fold greater than the overlap observed by random permutations (Fig. 4.2a). Before normalization, the MA plots exhibited an overall global dependence of M-value on A, which was closely fitted by a linear model derived by robust regression (Fig. 4.2b). A similar global dependence was evident in comparisons of biological replicates (Fig. 4.5a), indicating the dependence of M on A does not reflect biological changes but is due mainly to systemic bias and noise. After application of MAnorm to remove this dependence from the set of common peaks, the distribution of common peaks became highly symmetric with respect to

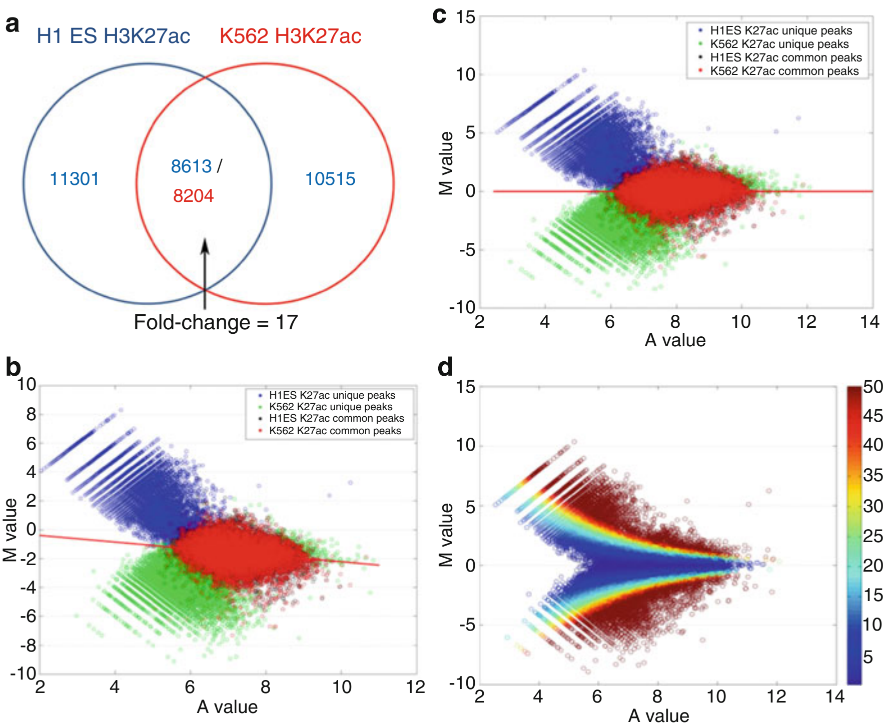


Fig. 4.2 Normalization of H3K27ac ChIP-Seq data in H1 ES cells and K562 cells. (**a**) Venn diagram representing the overlap of H3K27ac peaks between H1 ES and K562 cells. The overlap of peaks between the two cell lines was 17-fold greater than that observed for random permutations of the peak sets. (**b**, **c**) MA plots of all peaks from both samples before (**b**) and after MAnorm (**c**). *Red line* is the linear model derived from common peaks by robust regression. Blue and green circles represent unique peaks; red and black circles represent common peaks. (**d**) *P*-values associated with normalized peaks, displayed as an MA plot, with the color range representing $-\log_{10}P$-value. Most peaks associated with $|M| > 1$ have a *P*-value $< 10^{-10}$

the new *A* axis (Fig. 4.2c). These observations suggest that the ChIP-Seq signals in all peaks follow a similar scaling relationship and the extrapolation of the linear model from common peaks to all peaks is valid. The significance of differential binding in each peak region was determined using the *P*-value defined by Eq. 4.4 (Fig. 4.2d).

Next, we investigated the relationship between the *M*-value and the change in expression of peak targets between cell types. Firstly, we mapped the *M*-value of each H3K27ac peak to its target gene if this peak fell inside the promoter region of this gene, which was defined to be ranging from 8 kb upstream to 2 kb downstream of transcription start site. Secondly, we collected the gene expression data of these two cell types from the GEO database under accession numbers GSE26312 (for H1 ES cells) and GSE12056 (for K562 cells). Then, raw microarray data were processed by dChip (Li 2008), and differentially expressed genes were called by

SAM (Tusher et al. 2001) with a combined cutoff of fold change >2 and FDR <0.01. Finally, the relationship between the genes grouped by associated M-values and those genes differentially expressed between two cell types was examined by calculating the enrichment score of their mutual overlap, which was defined to be the ratio between observed number of overlapping genes and number of overlapping genes expected by chance.

In general, we found target genes associated with positive M-values, that is, peaks with higher H3K4me3 and H3K27ac read intensity in H1 ES cells, were enriched in genes more highly expressed in H1 ES cells. Conversely, target genes associated with negative M-values were enriched in genes more highly expressed in K562 cells (Fig. 4.3). These findings are consistent with the activating role of this histone mark at gene promoter regions (Lennartsson and Ekwall 2009). Notably, the enrichment score of genes more highly expressed in H1 ES cells showed strong positive correlation with the M-values, while the enrichment score of genes more highly expressed in K562 cells correlated negatively with M, suggesting that the M statistics determined by MAnorm serve as an indicator of cell-type specificity for the epigenetic marks in peak regions (Fig. 4.3). Furthermore, the target genes associated with an absolute M-value >1 were significantly enriched in genes highly expressed in the corresponding cell type among all our comparisons, implying that the absolute M-value of 1 is a suitable cutoff for defining cell-type specifically marked genes. It should be noted that many common target genes were associated with M-values far from 0, and were still highly enriched for cell-type specifically expressed genes (Fig. 4.3a), indicating that the differential epigenetic marks at these genes are also functional. On the other hand, those unique target genes with M-values near zero displayed much weaker enrichment of cell-type specifically expressed genes (Fig. 4.3b), indicating that they are not uniquely marked in one cell type. In conclusion, MAnorm quantitatively describes authentic binding differences of chromatin-associated proteins, and thus represents an improvement over arbitrary definitions of common and unique targets based on peak overlap between samples.

4.3.2 Identification of Cell-Type-Specific Regulators Associated with Differential Binding

A conventional strategy to identify cell-type-specific regulators associated with changes in epigenetic marks relies on the identification of transcription factor binding sites that are highly enriched in peak regions specific to specific cell types. However, we have shown previously that the accuracy of defining cell-type-specific regions cannot be guaranteed by simply using overlapping (Xu et al. 2012). One advantage of the continuous M-value determined by MAnorm is that it can be used to identify potential regulators driving cell-type-specific epigenetic modifications. To do so, we first downloaded the position weight matrixes (PWM) of 130 core vertebrate motifs from the JASPAR database (Sandelin et al. 2004) and performed

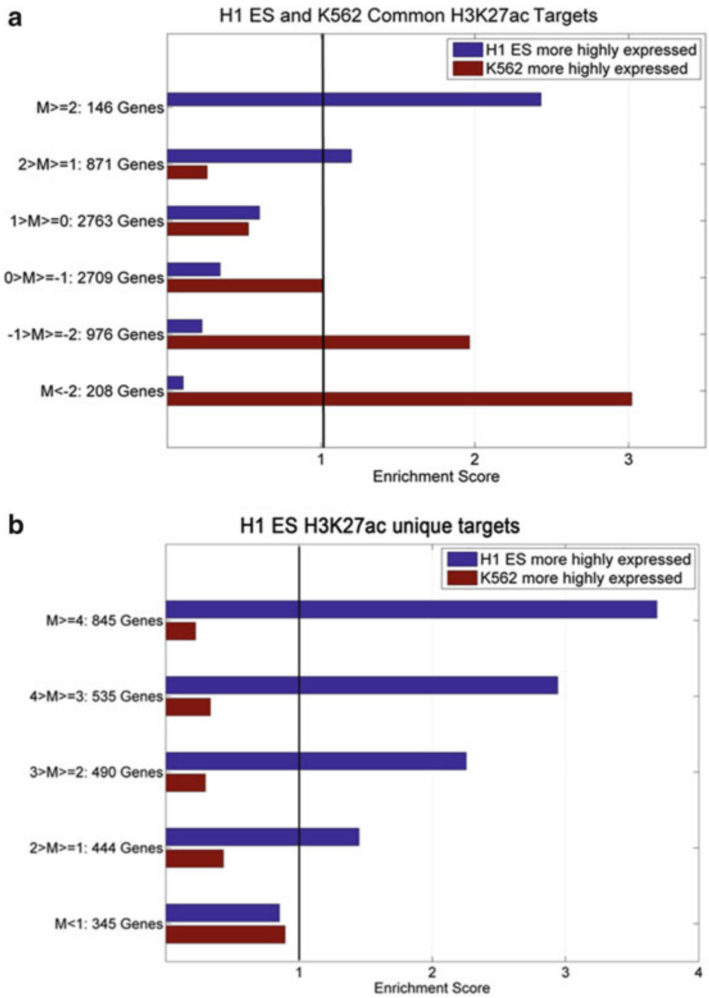


Fig. 4.3 Quantitative differences in H3K27ac marks between two cell lines are strongly correlated with cell-type-specific expression of peak target*s*. (**a**) Enrichment of the target genes of all common H3K27ac peaks in H1 ES cells and K562 cells in cell-type specifically expressed genes. Here the H3K27ac target genes were grouped by the *M*-values of H3K27ac mapped to genes. (**b**) Enrichment of the target genes of all unique H3K27ac peaks in H1 ES cells

motif scan (Liu et al. 2010b) applied to a 1,000 bp window centered at the peak summit of all H3K27ac peaks identified in H1 ES and K562 cells. For each motif *F*, the raw motif matching score at each peak *P* was calculated as

$$\max_{S \in P}\left[\log \frac{P\left(S \middle| M\right)}{P\left(S \middle| B\right)}\right],$$

in which S was sequence fragment of the same length as the motif and B was the background frequency of different nucleotides in corresponding genome. Then we searched for motifs that show strong correlation with M-values for all peaks by applying hierarchical clustering to cluster the M-value with the motif matching score of JASPAR motifs in all peaks of each cell type. We found that OCT4 (POU5F1) and SOX2 binding motifs were closely clustered with the M-value (= $\log_2$ (read density in H1 ES cells/read density in K562 cells) of H3K27ac peaks (Fig. 4.4a), suggesting the corresponding factors are closely related to the activation of ES cell-specific genes and cis-elements. In contrast, M-value (= $\log_2$ (read density in K562 cells/read density in H1 ES cells)) formed a compact module with the binding motifs for transcription factors GATA1 and SCL (also known as TAL1) (Fig. 4.4b), suggesting their roles as regulators favoring H3K27ac modification in K562 cells. These findings are consistent with the established roles of OCT4-SOX2 in ES cell self-renewal (Chambers and Smith 2004; Boyer et al. 2005) and GATA-SCL in hematopoiesis and leukemogenesis (Fujiwara et al. 2009). On the other hand, several motifs, including MYC and ETS motifs, were highly enriched in both peak sets, but showed no association with the differential binding of H3K27ac, indicating they are involved in H3K27ac modification in a non-cell-type-specific manner. This finding in turn supports the working assumption of our model that binding at most common peaks is determined by similar mechanisms. Thus, MAnorm serves as a novel and powerful tool to uncover transcription factor motifs and factors critical for cell-specific gene regulation.

4.3.3 Use MAnorm to Integrate ChIP-Seq Replicates

Integrating ChIP-Seq data from multiple biological replicates, which in some cases are generated by different labs and/or using different platforms, may be employed to reduce the false positive rate in identified binding sites. A simple approach is to define a stringent set of peaks comprised only of the common peaks shared by two or more replicates. However, this method is highly sensitive to peak-calling cutoff and may exclude peaks that have similar ChIP intensities between replicates. Moreover, some common peaks that show dramatic differences in read density are retained. Therefore, to make full use of the information in multiple replicates, a quantitative comparison of peak intensity is particularly useful. We have applied MAnorm to compare two replicates of H1 ES cell H3K27ac ChIP-Seq data generated by ENCODE project. After application of MAnorm (Fig. 4.5a, b), a large number of the unique peaks were associated with M-values close to zero, indicating that these peaks exhibit good reproducibility between replicates. On the other hand, there remained a small fraction of common peaks with M-values far from zero, representing strong signal differences between replicates. Next, we showed that the M-value between replicates is a good indicator of H3K27ac target gene expression. We grouped H3K27ac target genes by the absolute value of M statistics and calculated the expression distribution of each gene group. The genes targeted

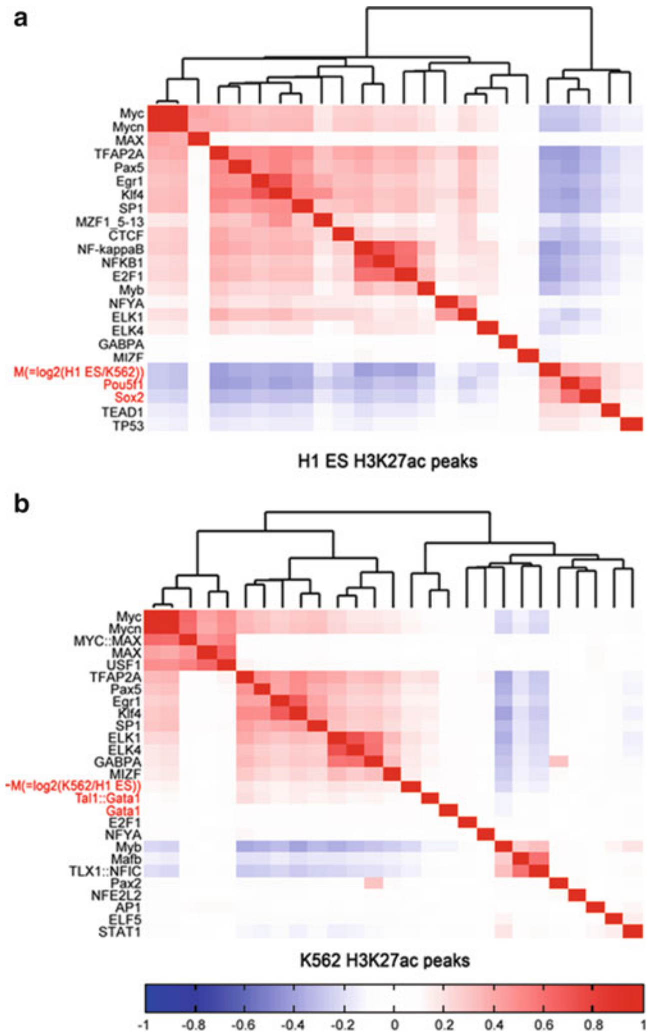


Fig. 4.4 Hierarchical clustering of the *M*/-*M* value and motif scores in all H3K27ac peaks of H1 ES cells and K562 cells. Hierarchical clustering was applied to the correlation coefficients of *M*-values (=log2 (read density in H1 ES/read density in K562)) or –*M*-values (=log2 (read density in K562/read density in H1 ES)) of all H3K27ac peaks identified in H1 ES cells (**a**) or K562 cells (**b**), with motif scores determined for 130 JASPAR vertebrate core motifs in the peak regions. Only the motifs significantly enriched in the peaks of either cell type compared to genome background are shown here

by peaks with higher absolute *M*-values, that is, peaks showing larger difference between replicates, tended to have lower expression, which is true for both common and unique peaks (Fig. 4.5c). Given that H3K27ac marks are positively associated with gene expression, peaks with low *M*-values between replicates are expected

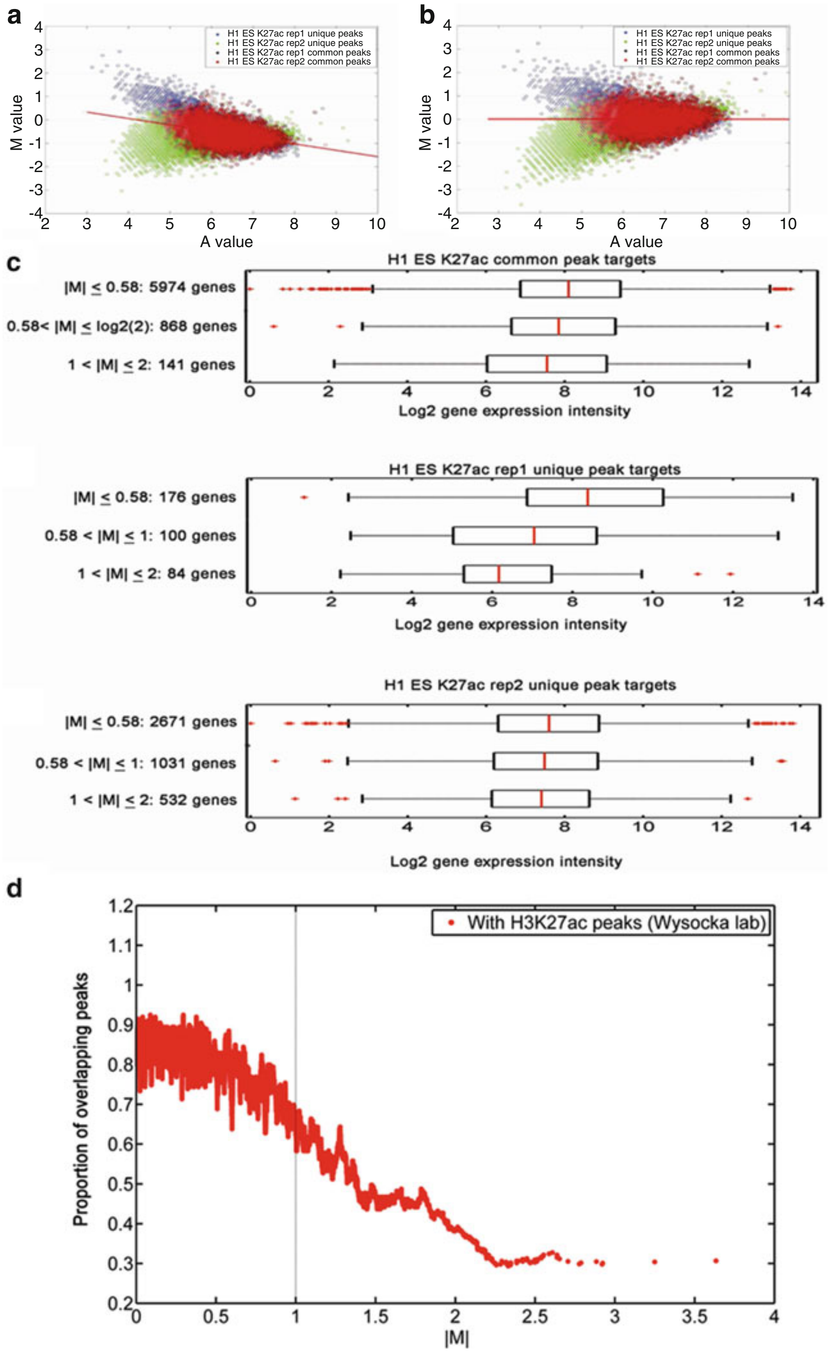


Fig. 4.5 Application of MAnorm to the integration of ChIP-Seq replicates. (**a**, **b**) MA plot comparing H3K27ac mark between two H1 ES replicates before (**a**) and after (**b**) MAnorm. (**c**) Both common and unique H3K27ac peak targets between two H1 ES replicates were divided into three groups based on absolute M-value, and the box-plot shows the log2 gene expressions for each group. (**d**) The faction of Encode H3K27ac peaks that overlap with H3K27ac peaks in H1 ES cells based on data from Rada-Iglesias et al. Here the peaks were ranked by $|M|$ values from low to high and the proportion of overlap was calculated by a moving window of 200 peaks

to be more reliable than those with high M-values. Furthermore, by overlapping these ENCODE peaks with H3K27ac peaks of H1 ES cells generated in a different laboratory (Rada-Iglesias et al. 2011), we found that a much lower proportion of the peaks with $|M| > 1$ were covered by the new peak set than those with $|M| < 1$ (Fig. 4.5d). This suggests that $|M| = 1$ can also be used as an empirical cutoff to filter unreliable peaks. Thus, MAnorm can be used both to check whether two replicates are concordant, and also to obtain high-confidence peak lists by filtering out inconsistent peaks. Compared with arbitrary removal of unique peaks, MAnorm allows for better use of replicate peak data. The MAnorm package provides the opportunity to list concordant and nonconcordant peaks between two samples based on user-specified cutoffs, with the concordant peak list corresponding to high-confidence peaks.

4.4 Performance Comparison Between MAnorm and Other Existing Methods

4.4.1 Compare the Performance in Inferring Quantitative Changes of ChIP-Seq Signals

We compared the performance of MAnorm with three widely used normalization methods that use genome-wide signals as reference, namely, normalization by total reads, quantile normalization, which assumes the genome-wide distribution of read densities to be the same across samples, and normalization using a genome-wide MA-plot followed by LOWESS regression. We used all these methods to compare H3K27ac ChIP-Seq data between H1 ES and K562 cells. For total read normalization, we divided the read intensity of each peak region by the total number of mapped sequence reads. For quantile normalization, we first divided the whole genome into nonoverlapping bins of the same size as the window used in MAnorm, that is,. 2,000 bp, and then calculated the read count in each bin. Finally, the distribution of bin read counts was normalized to be the same by matching all quantiles between samples. For normalization by genome-wide MA-plots, we first divided the whole genome into nonoverlapping 2,000 bp bins, and then calculated the (M, A) value of each bin. The dependence between M-A value was then removed by subtracting M-values with local linear model fitted by LOWESS regression from the genome-wide M-A values.

To examine which method better reflects a true biological signal, we compared M-values normalized by all four methods with the expression change of target genes. If a specific type of histone modification is closely related to gene regulation, the direction of histone modification change should be consistent with that of the change in gene expression at the target genes. By visual inspection, we found this was true for the M-values normalized by MAnorm (Fig. 4.6a). In contrast, M-values normalized by the other three methods were inconsistent with the log2-

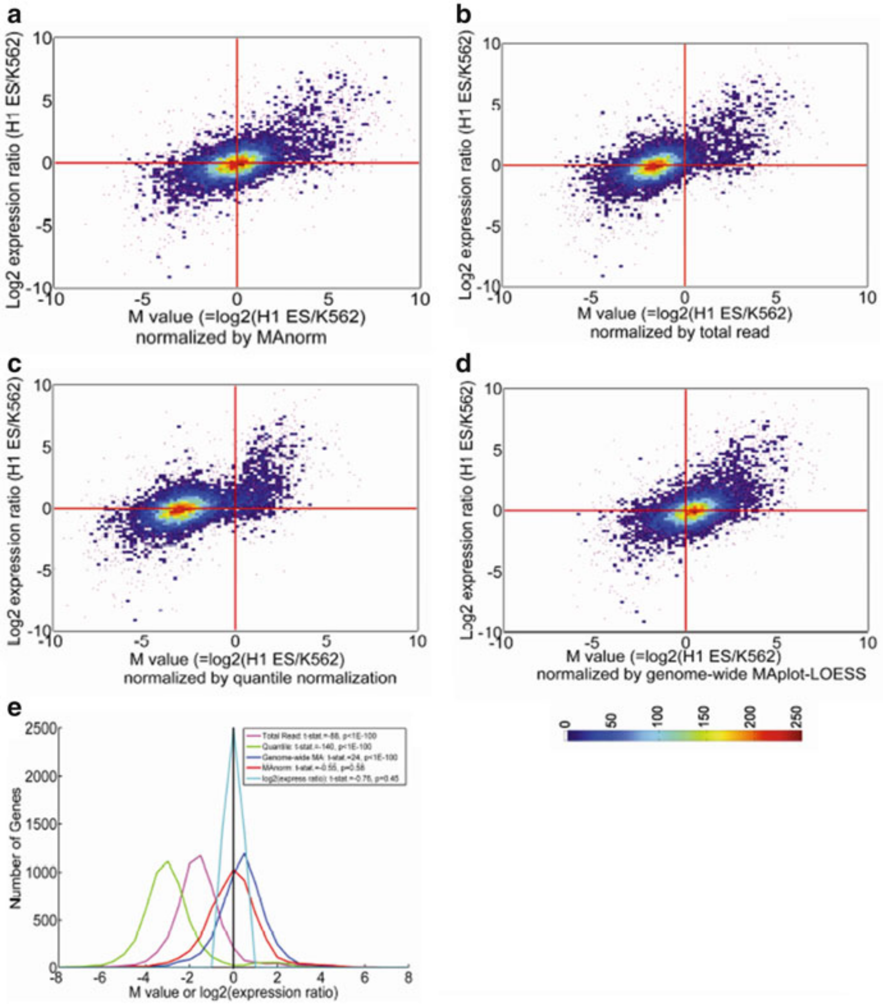


Fig. 4.6 Comparison of different normalization models. (**a–d**) Scatter plot of log2 expression ratio of target genes between H1 ES cells and K562 cells versus the *M*-values of H3K27ac normalized by MAnorm (**a**), total read (**b**), quantile normalization (**c**), and genome-wide MA-plot followed by LOWESS normalization (**d**). The color bar represents the density of dots in the scatter plot and purple dots represent the outliers separated from the others. (**e**) Distribution of *M*-values or log2 expression ratios of nondifferentially expressed target genes (fold-change <1.5). The t-statistics and *P*-value were calculated based on one sample Students' *t*-test comparing to 0

expression ratios of target genes (Fig. 4.6b–d). Specifically, most of the genes with no change in H3K27ac levels ($M = 0$) had higher (total read and quantile normalization) or lower (genome-wide MA-plot normalization) expression in H1 ES cells compared to K562 cells; while the majority of the genes expressed at similar levels in these two cell types were associated with negative (total read and quantile normalization) or positive (genome-wide MA-plot normalization) M-values, that is, they had higher (total read and quantile normalization) or lower (genome-wide MA-plot normalization) levels of H3K27ac in K562 cells. To quantitatively measure the bias of the M-values given by the above normalization methods, we first collected nondifferentially expressed genes (fold-change <1.5) between H1 ES cells and K562 cells. As shown in Fig. 4.6e, these genes are indeed not differentially expressed (t-statistics $= -0.76$ and P-value $= 0.45$ by Students' t-test in comparison to an expression ratio of 1 ($M = 0$)), indicating they are suitable for our comparison. Since H3K27ac marks are closely associated with transcriptional activation, it is reasonable to assume that these nondifferentially expressed genes should exhibit similar global H3K27ac levels. This is true only for MAnorm, where the M-values for H3K27ac of the nondifferentially expressed target genes were not significantly different from a ratio of 1 ($M = 0$; t-statistic $= -0.55$ and P-value $= 0.58$ by t-test; Fig. 4.6e). In contrast, M-values for H3K27ac obtained by the other normalization methods exhibited large deviations from $M = 0$ (t-statistic ranging from 24 to 140 and P-value $<$1e-100; Fig. 4.6e). In conclusion, MAnorm exhibits superior performance in identifying biological changes.

4.4.2 Compare the Performance in Detecting Differential Binding Regions

We also compared the performance of MAnorm in detecting differential binding regions in ChIP-Seq data sets with that of two currently used statistical methods, ChIPdiff (Xu et al. 2008) and MACS (Zhang et al. 2008), For this analysis, one data set was used as sample and the other was used as control in order to detect regions with significantly elevated ChIP-Seq signals in the first data set (Zhang et al. 2008). We applied all three methods to compare ChIP-Seq data for H3K27ac marks between H1 ES cells and K562 cells (Table 4.1). Both ChIPdiff and MACS identified four to five times more target regions associated with significantly increased ChIP-Seq signals for K562 cells compared with those found for H1 ES cells, whereas MAnorm yielded a similar number of cell-type-biased peaks in each cell line. To compare the enrichment of cell-type specifically expressed genes in the sets of target genes of the differential binding regions discovered by the three methods, we selected the same number of target genes associated with top differential binding regions identified by each method. The target genes of top differential binding regions identified by MAnorm contained similar numbers of H1 ES cell highly expressed genes but a greater number of K562 cell highly expressed genes compared to those identified by ChIPdiff and MACS (Table 4.2), suggesting MAnorm performs better in detecting differentially binding regions than the other two methods. Importantly, the fold changes of differential binding given by ChIPdiff and MACS were based on the total number of reads, which may not be appropriate, as discussed above.

Table 4.1 Enrichment of the overlap between genes more highly expressed in H1 ES cells (as compared to K562 cells) and H3K27ac H1 ES-enriched target genes, which were defined as genes whose promoter regions overlap with any H1-biased H3K27ac peaks (compared to K562 cells) identified by MAnorm, MACS, and ChIPdiff

H3K27ac H1 ES-enriched target genes	Number of genes	Overlap with H1 ES upregulated genes	Enrichment score
MAnorm ($M > 1$)	2,680	1,243	2.49
ChIPdiff (default)	1,467	884	3.24
MAnorm (top 1,467 genes; same number of genes as identified by ChIPdiff with default settings)	1,467	941	3.45
MACS (P < 1e-6)	1,589	993	3.36
MAnorm (top 1,589 genes; same number of genes as identified by MACS with $P < 1$ e-6)	1,589	987	3.34

Table 4.2 Enrichment of the overlap between genes more highly expressed in K562 ES cells (as compared to H1 ES cells) and H3K27ac K562-enriched target genes (compared to H1 ES cells) identified by MAnorm, MACS, and ChIPdiff

H3K27ac K562-enriched target genes	Number of genes	Overlap with K562 upregulated genes	Enrichment score
MAnorm ($M < -1$)	2,694	895	2.78
ChIPdiff (default)	6,733	1,402	1.74
ChIPdiff (confidence thresholds = 0.9999999999)	2,291	697	2.55
MAnorm (top 2,291 genes; same number of genes as identified by ChIPdiff with confidence threshold = 0.9999999999)	2,291	820	3.00
MACS ($P <$ 1e-6)	9,346	1,600	1.43
MACS ($P <$ 1e-150)	1,556	567	3.05
MAnorm (top 1,556 genes; same number of genes as identified by MACS with $P <$ 1e-150)	1,556	644	3.47

4.5 Define High-Confidence Cell-Type-Specific and Nonspecific Enhancers Using MAnorm

In the past decade, it has been widely found that in eukaryotic cell DNA regulatory elements are usually covered by a combination of multiple chromatin marks, which collectively define the functional status of these elements in each cell type (Creyghton et al. 2010; Rada-Iglesias et al. 2011; Lennartsson and Ekwall 2009). Thus, how to integrate the change of associated chromatin marks across different cell types becomes an inevitable problem not only for accurately classifying functional elements specific to given cell type, but also for understanding their contribution in establishing the differential gene expression programs during cell state transition. Here we use genome-wide comparison of distal enhancers between the adult and fetal stages of human erythroid cells (Xu et al. 2012) as an example of such integration. Firstly, we generated DNase-Seq and ChIP-Seq data of histone mark H3K4me1/3, H3K9ac, H3K27ac, and H3K27me3 in both human primary proerythroblast (ProE) cells at both fetal and adult stage, which were derived via an ex vivo differentiation from the corresponding hematopoietic stem/progenitor cells for 5 days. Then, we defined the putative distal enhancers of each stage as the genomic regions that

1. Are covered by H3K4me1 and H3K9ac or H3K27ac peaks
2. Contain at least one DNase I hypersensitive site and do not contain any H3K27me3 peak
3. Are located at least 2 kb away from any RefSeq annotated gene's transcription start site

Creyghton et al. (2010), Rada-Iglesias et al. (2011). Using these criteria, we identified 8,947 and 11,709 putative active distal enhancers in fetal and adult ProEs, respectively. To define stage-specific enhancers, we first merged the enhancer regions of two stages that overlap with each other and considered them as "common" enhancers (4,360 in total). The remainders were thus classified as "fetal-only" (2,594 in total) and "adult-only" (5,730 in total) (Fig. 4.7a). Then, we investigated the quantitative changes of H3K4me1, H3K9ac, and H3K27ac mark between two stages (M-values derived by MAnorm model) at these enhancers, and found in general H3K9ac and H3K27ac showed much higher changes compared to H3K4me1. In detail, we observed 31 % of total enhancers have fold change of H3K9ac or H3K27ac higher than 2, but only less than 10 % of these enhancers showed fold change of H3K4me1 higher than 2. Thus, we speculated that H3K9/27ac better reflected the functional status of active enhancers compared to H3K4me1, which is consistent with the previous findings derived from embryonic stem cells (Creyghton et al. 2010; Rada-Iglesias et al. 2011), and proceeded to define high-confidence (HC) fetal-/adult-only and common enhancers based on the differences of H3K9/27ac using the following rules:

1. HC-common enhancers are those common enhancers with $-1 <$Maximum ($M_{H3K27ac}$, M_{H3K9ac}) <1.
2. HC-fetal-only enhancers are those fetal-only enhancers with Minimum ($M_{H3K27ac}$, M_{H3K9ac}) <-1 and Maximum($M_{H3K27ac}$, M_{H3K9ac}) <0.
3. HC-adult-only enhancers are those adult-only enhancers with Maximum ($M_{H3K27ac}$, M_{H3K9ac}) >1 and Minimum($M_{H3K27ac}$, M_{H3K9ac}) >0.

Finally, we filtered out 967 and 2,024 HC-fetal/adult-only enhancers, respectively, together with 2,970 HC-common enhancers (Fig. 4.7a). To check the improvement brought by these additional rules, we systematically plotted the ChIP-Seq signals of associated histone marks in fetal and adult ProE cells at all these different groups of enhancers (Fig. 4.7b, c). It can be clearly observed

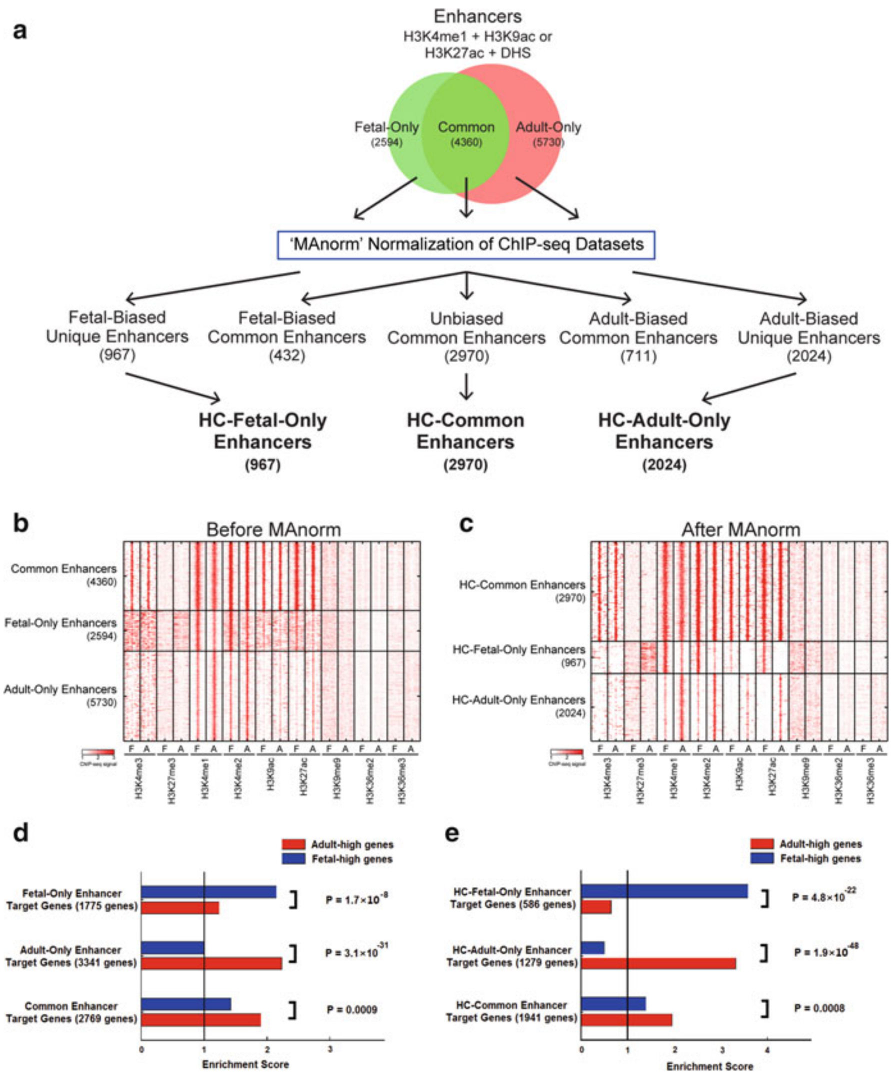


Fig. 4.7 Using MAnorm can greatly improve the accuracy in defining cell-type-specific and nonspecific regulatory elements. (**a**) Flowchart of defining different groups of enhancers. ChIP-Seq peaks for H3K4me1, H3K9ac, and H3K27ac from fetal and adult ProEs were quantitatively compared by using MAnorm. High-confidence (HC)-fetal-only and HC-adult-only enhancers were identified requiring fold change of H3K9/27ac intensities higher than 2, which represents the ratio of enhancer binding intensities between fetal and adult samples or vice versa. High-confidence common enhancers were identified requiring the corresponding fold change lower than 2 between fetal and adult samples. (**b**) ChIP-Seq read density heatmaps of the profiled histone marks within the common, fetal-only, and adult-only enhancers ("Before MAnorm"). (**c**) ChIP-Seq read density heatmaps of the profiled histone marks within the HC-common, HC-fetal-only, and HC-adult-only enhancers ("After MAnorm"). (**d**) Target genes of fetal-only, adult-only, and common enhancers were compared with genes differentially expressed between F5 and A5 ProEs. Here the *P*-values were calculated from hypergeometric distribution. (**e**) Comparing target genes of HC-fetal-only, HC-adult-only, and HC-common enhancers with genes differentially expressed between F5 and A5 ProEs

that stage-specific enhancers before filtering, especially those fetal-only ones, still contain considerable level of H3K9/27ac marks at the opposing stage (Fig. 4.7b), which clearly vanished at high-confidence stage-specific enhancers (Fig. 4.7c). Meanwhile, we also mapped each enhancer to its nearest gene within 50 kb as its target (Xu et al. 2012). Consistently, we found that compared with genes targeted by stage-specific enhancers defined by using only overlapping (Fig. 4.7d), the genes targeted by high-confidence stage-specific enhancers show much higher bias toward genes specifically expressed at the corresponding stage versus those expressed at the opposing stage (Fig. 4.7e). Taking together, these comparisons strengthened the concept that it is highly unreliable to define cell-type-specific regulatory elements by only using overlapping, and quantitative comparison of associated chromatin marks can better characterize the functional specificity of regulatory elements across different cell types.

4.6 Summary and Discussion

Normalization methods are typically based on the assumption that certain properties are invariant across samples. For example, quantile normalization in gene expression microarrays renders the distribution of expression levels of all genes constant between samples (Bolstad et al. 2003). Alternatively, normalization may be based on housekeeping genes, whose expression is presumed to remain constant across samples. The situation is quite different in ChIP-Seq studies, since the binding of most chromatin-associated proteins is highly dynamic and cell-type dependent. Thus, it is arbitrary to assume that the genome-wide distribution of ChIP-Seq signals remains constant between samples. It is also challenging to identify "hot spots" bound by a chromatin-associated protein in a non-cell-type-specific manner that can serve as a reference for normalization. Yet another difficulty underlying ChIP-Seq studies is background noise, which is often difficult to distinguish from authentic ChIP signals. Furthermore, the signal-to-noise ratio (S/N) often varies across samples. In many peak-calling models, the distribution of background signal is used to normalize sample and control data, which is reasonable when control data are comprised mainly of background signal, and the purpose is to identify sequence read-enriched regions within a sample that shows significant differences as compared to the background. However, this approach is inappropriate for sample-to-sample comparisons, especially when the S/N difference is large across samples. For example, samples relatively free of "noise" will yield a larger number of statistically significant peaks compared to samples with a higher level of background sequence reads, but these additional peaks may not be true cell-line-specific or condition-specific peaks. In MAnorm, we focused only on regions identified as significant peaks, and thus minimized the impact of S/N differences between samples. Accordingly, the output of MAnorm focuses on peak regions most likely to be of biological relevance.

MAnorm shows improved performance when compared with other methods currently used to detect differential binding regions between ChIP-Seq data sets. More importantly, MAnorm provides a quantitative measurement of binding differences, which reflects authentic biological differences. This feature is an asset for downstream analysis, including expression assays and transcription cofactor identification studies. Although the definition of peaks depends highly on the cutoff used in peak calling, MAnorm is robust to cutoff selection. Furthermore, the normalized read densities of each peak in both ChIP-Seq samples can be calculated from the (M, A) values normalized by MAnorm, and then used in other downstream analyses such as to evaluate whether the cutoffs used to define peaks were comparable between the ChIP-Seq samples being compared (Shao et al. 2012).

MAnorm relies on two working assumptions. First, MAnorm is designed for quantitative comparison of ChIP-Seq data sets that have a substantial number of peak regions in common. Second, MAnorm postulates that there are no global changes in the true ChIP signals at these common peaks. We believe these underlying hypotheses do not significantly restrict the use of MAnorm as compared to other methods. For ChIP-Seq samples for which there is not a significant overlap in peak sets, the binding of chromatin-associated proteins could be uncorrelated or even anti-correlated at a genome-wide scale and a quantitative comparison would not be important. Moreover, in cases where the binding patterns changed widely across the genome, such as following knockdown of a core subunit of a chromatin-associated protein complex (Jiang et al. 2011), more specific analysis would be required to quantitatively determine the global changes.

The pairwise approach to comparison of ChIP-Seq samples described here potentially can be extended to multiple sample comparison, as already successfully demonstrated in two-channel microarray data analysis (Smyth and Speed 2003). More specifically, multiple ChIP-Seq samples of the same factor generated from different cell types or even different individuals can be compared by statistically modeling the variation of normalized ChIP-Seq signals, which can be easily derived from the output of MAnorm program, as indicated by Pinello et al. (2014) and Kasowski et al. (2013). Furthermore, transcription factors and epigenetic modifications usually act together to modulate gene expression (Bernstein et al. 2007). Most recently, statistical models have been developed to study such combinatorial patterns in a genome-wide fashion (Pinello et al. 2014; Kasowski et al. 2013; Bernstein et al. 2007; Ji et al. 2013). However, although people have widely found changes of epigenetic marks and transcriptional factors' binding often correlate with other (Shao et al. 2012; Bernstein et al. 2007), quite few computational models were developed to dissect how they collectively define the differential expression program between different cell types (Ji et al. 2013). Here we used a systematic comparison of associated histone marks at distal active enhancers between adult and fetal stages of human erythroid cells as example, to illustrate how to design downstream applications of MAnorm for integrative analyses. However, it is still necessary to develop more statistical models in order to further understand the dynamical changes of epigenetic landscape during cell state transition.

Acknowledgments We sincerely thank Prof. Stuart H. Orkin and Prof. David J. Waxman for the great guidance during development of MAnorm model. We also thank the laboratories associated with the ENCODE project for generating and maintaining the data sets used in our analyses, as well as Drs. Jian Xu, Han Xu, and Aarathi Sugathan for useful discussions.

References

Audic S, Claverie JM. The significance of digital gene expression profiles. Genome Res. 1997;7(10):986–95.

Bernstein BE, Meissner A, Lander ES. The mammalian epigenome. Cell. 2007;128(4):669–81.

Bolstad BM, et al. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics. 2003;19(2):185–93.

Boyer LA, et al. Core transcriptional regulatory circuitry in human embryonic stem cells. Cell. 2005;122(6):947–56.

Chambers I, Smith A. Self-renewal of teratocarcinoma and embryonic stem cells. Oncogene. 2004;23(43):7150–60.

Creyghton MP, et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc Natl Acad Sci U S A. 2010;107(50):21931–6.

Fujiwara T, et al. Discovering hematopoietic mechanisms through genome-wide analysis of GATA factor chromatin occupancy. Mol Cell. 2009;36(4):667–81.

Ji H, et al. An integrated software system for analyzing ChIP-chip and ChIP-seq data. Nat Biotechnol. 2008;26(11):1293–300.

Ji H, et al. Differential principal component analysis of ChIP-seq. Proc Natl Acad Sci U S A. 2013;110(17):6789–94.

Jiang H, et al. Role for Dpy-30 in ES cell-fate specification by regulation of H3K4 methylation within bivalent domains. Cell. 2011;144(4):513–25.

Kasowski M, et al. Extensive variation in chromatin states across humans. Science. 2013;342(6159):750–2.

Lennartsson A, Ekwall K. Histone modification patterns and epigenetic codes. Biochim Biophys Acta. 2009;1790(9):863–8.

Li C. Automating dChip: toward reproducible sharing of microarray data analysis. BMC Bioinf. 2008;9:231.

Liu W, et al. PHF8 mediates histone H4 lysine 20 demethylation events involved in cell cycle progression. Nature. 2010a;466(7305):508–12.

Liu Y, Shao Z, Yuan GC. Prediction of Polycomb target genes in mouse embryonic stem cells. Genomics. 2010b;96(1):17–26.

McKean JW. Robust analysis of linear models. Stat Sci. 2004;19(4):562–70.

Park PJ. ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet. 2009;10(10):669–80.

Pinello L, et al. Analysis of chromatin-state plasticity identifies cell-type-specific regulators of H3K27me3 patterns. Proc Natl Acad Sci U S A. 2014;111(3):E344–53.

Rada-Iglesias A, et al. A unique chromatin signature uncovers early developmental enhancers in humans. Nature. 2011;470(7333):279–83.

Rozowsky J, et al. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. Nat Biotechnol. 2009;27(1):66–75.

Sandelin A, et al. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res. 2004;32(Database issue):D91–4.

Schmidt D, et al. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science. 2010;328(5981):1036–40.

Shao Z, et al. MAnorm: a robust model for quantitative comparison of ChIP-Seq data sets. Genome Biol. 2012;13(3):R16.

Smyth GK, Speed T. Normalization of cDNA microarray data. Methods. 2003;31(4):265–73.
Taslim C, et al. Comparative study on ChIP-seq data: normalization and binding pattern characterization. Bioinformatics. 2009;25(18):2334–40.
Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A. 2001;98(9):5116–21.
Xu H, et al. An HMM approach to genome-wide identification of differential histone modification sites from ChIP-seq data. Bioinformatics. 2008;24(20):2344–9.
Xu J, et al. Combinatorial assembly of developmental stage-specific enhancers controls gene expression programs during human erythropoiesis. Dev Cell. 2012;23(4):796–811.
Zhang Y, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9(9):R137.

Chapter 5
Model-Based Clustering of DNA Methylation Array Data

Devin C. Koestler and E. Andrés Houseman

Abstract Clustering refers to the "grouping" of observations into a discrete set of classes, such that observations in the same class are more similar compared to objects between classes. In the context of DNA methylation data, clustering can be used to discover novel molecular subtypes or to identify biological pathways comprised of co-methylated CpG dinucleotides, depending on whether the samples or the CpGs themselves are being clustered. In this chapter, we focus on the problem of clustering samples/subjects on the basis of their methylation profile. We begin by discussing the motivation behind clustering DNA methylation data, the nature of DNA methylation data generated from the Illumina BeadArrays, and three promising model-based clustering methods. In addition to providing a methodological overview of each of the three methods, we also demonstrate their application using a publicly available data set deposited in the Gene Expression Omnibus (GEO) database. Issues such as feature selection and comparison of clustering partitions will also be discussed.

Keywords Model-based clustering • Finite mixture models • DNA methylation • Microarray • Illumina Infinium Methylation BeadArrays

5.1 Introduction

DNA methylation has risen to the forefront as one of the most widely studied epigenetic states due to its role in regulating gene expression and gene expression potential. Although DNA methylation is a normal and essential process for human development, disrupted patterns of DNA methylation have been linked to disease

D.C. Koestler (✉)
Department of Biostatistics, University of Kansas Medical Center, Kansas City, KS 66160, USA
e-mail: dkoestler@kumc.edu

E.A. Houseman
Department of Public Health, Oregon State University, Corvallis, OR 97331, USA
e-mail: andres.houseman@oregonstate.edu

A.E. Teschendorff (eds.), *Computational and Statistical Epigenomics*,
Translational Bioinformatics 7, DOI 10.1007/978-94-017-9927-0_5

development and progression across a wide spectrum of different human diseases. Such findings have served to highlight the possibility of using DNA methylation for the purposes of diagnosis or prognosis, in which profiles of DNA methylation are used for early disease detection, risk assessment, and disease progression monitoring. Similar to analyses involving gene expression data and microarray data, unsupervised clustering analysis of DNA methylation data has emerged as one of the most widely used techniques for the identification of such profiles. While unsupervised clustering analysis can be used both to discover novel molecular subtypes (by clustering samples/subjects) and to identify biological pathways comprised of co-methylated CpG dinucleotides (by clustering CpGs), in this chapter we focus on the problem of clustering samples/subjects on the basis of their DNA methylation profile.

Numerous different unsupervised clustering methods have been applied for identifying underlying structure in DNA methylation data, including nonparametric (e.g., K-means, agglomerative hierarchical clustering, etc.) and model-based methods (Houseman et al. 2008; Kuan et al. 2010). Whereas nonparametric methods do not require an assumption about the underlying distribution of the data, model-based methods assume that the data is generated from a finite mixture of underlying probability distributions, where each mixture component corresponds to a cluster. Although there is no universal consensus on the single "best" clustering method for DNA methylation array data, existing work has demonstrated favorable performance of model-based methods over their nonparametric counterparts (Houseman et al. 2008; Kuan et al. 2010; Siegmund et al. 2004; Koestler et al. 2013). For this reason and because model-based methods allow for statistical inference on the number of mixture components (i.e., clusters) and the estimation of cluster membership probabilities, we have chosen to focus on model-based methods as the basis of this chapter. Specifically, we describe three different model-based clustering methods, mclust (Fraley and Raftery 2002), LumiWCluster (Kuan et al. 2010), and recursively partitioned mixture models (RPMM) (Houseman et al. 2008), and illustrate their application using a publicly available DNA methylation data set.

The outline for the remainder of this chapter is as follows. In Sect. 5.2 we provide an overview DNA methylation data generated from the Illumina Infinium BeadArrays, and in Sect. 5.3 we describe the DNA methylation data set that will be used throughout the chapter to illustrate the various methodologies. In Sect. 5.4 we provide an overview and application of mclust, LumiWCluster, and RPMM, and in Sect. 5.5 we discuss feature selection strategies in the context of clustering analysis. We finish with some concluding remarks and discussion points in Sect. 5.6.

Before beginning, there are a few items that deserve mentioning. Each model-based clustering method discussed in Sect. 5.4 is prefaced with a brief summary paragraph (blue boxes), which provides a high-level overview of the method/technique that follows. The purpose of the brief summary is to provide readers uninterested in the statistical details with a basic understanding of the method and its relative highlights. Also, we assume that readers have a basic understanding of statistics, i.e., probability distributions,maximum likelihood estimation, etc., and a working knowledge of the R statistical programming language. Lastly, we note that throughout this chapter we will use the terms cluster, component, and class interchangeably.

5.2 Overview of Illumina Infinium DNA Methylation Array Data

Before discussing specific methodological approaches for clustering DNA methylation data, it is critical to first begin by providing an overview of data generated from the Illumina Infinium BeadArrays. In the paragraphs that follow, we aim not to provide a comprehensive overview of this technology, as there is much existing literature along these lines, rather our goal is to provide the reader with sufficient information about the nature and characteristics of the data obtained from this technology to motivate our later discussion on model-based methods for clustering analysis.

The Illumina Infinium BeadArrays measure intensity values M and U, representing the methylated and unmethylated probe intensities, for tens to hundreds of thousand CpG dinucleotides in the genome (depending on the specific array technology used). In many applications, the methylated and unmethylated probe intensities for a given locus are combined to obtain single representative value for methylation. In the current literature, the two most commonly used measurements are the beta-value and M-value (Du et al. 2010). For a given CpG site j, the beta-value is defined as $\beta_j = \frac{M_j}{M_j+U_j}$, $j = 1, 2, \ldots, J$, where J is the total number of profiled CpG sites. In simple terms, the beta-value reflects the proportion of methylated to overall signal intensity and naturally takes on values that are bounded between 0 and 1, 0 denoting an unmethylated locus and 1 denoting a methylated locus. Note that oftentimes a small offset value is added in the denominator (i.e., 100) to regularize in situations where M_j and U_j are small.

The M-value for CpG j is defined as $m_j = \log_2\left(\frac{M_j}{U_j}\right)$, $-\infty < m_j < \infty$ and represents the $\log_2$ ratio of the methylated to unmethylated signal intensity. From this definition, it can be easily shown that the relationship between beta- and M-values is

$$m_i = \log_2\left(\frac{\beta_j}{1-\beta_j}\right) = \text{logit}_2(\beta_j) \tag{5.1}$$

Thus, M-values are equivalent to a logit transformation (log base 2 scale) applied to the methylation beta-values.

While the selection of beta- or M-values as a representative measurement for methylation is beyond the scope of this chapter, it is clear that there are advantages and disadvantages to both. As described in Saadati and Benner (2014), beta-values have the obvious advantage of interpretability; when the methylation status of a given CpG site is represented as a proportion between 0 and 1, it retains a more desirable interpretation from a biological point of view. However, beta-values exhibit severe heteroscedasticity (nonconstant variance) (Du et al. 2010) outside the middle methylation range (i.e., close to 0 or 1), raising concerns about the use of common statistical models that assume homoscedasticity. However, this characteristic could be accounted for through the use of appropriate statistical models. For example, the beta distribution has been proposed as the underlying distribution for modeling beta-values (Houseman et al. 2008; Kuan et al. 2010). The beta distribution is a family of continuous probability distributions defined on the interval (0, 1), containing two parameters, $a > 0$ and $b > 0$, which control the shape of the distribution (Fig. 5.1). The probability density function for a beta-distributed random variable is given as

$$p(x|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)+\Gamma(b)} x^{a-1}(1-x)^{b-1}, \qquad x \in (0,1) \tag{5.2}$$

where $\Gamma(.)$ is the gamma function. Here, the mean and variance are $E(x) = \frac{a}{(a+b)}$ and $\mathrm{Var}(x) = \frac{ab}{(a+b)^2(a+b+1)}$, respectively. Alternatively, the beta distribution can be reparameterized in terms of a mean μ and dispersion ϕ parameter:

$$p(x|\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} x^{\mu\phi-1}(1-x)^{(1-\mu)\phi-1}, \qquad x \in (0,1) \tag{5.3}$$

where $\mu \in (0,1)$ and $\phi > 0$. Here the mean is given by $E(x) = \mu$ and variance $\mathrm{Var}(x) = \frac{(1-\mu)\mu}{1+\phi}$. Note that since the variance is a function of the mean, the beta distribution naturally accounts for heteroscedasticity.

On the other hand, the M-value results in approximately homoscedastic values, thus permitting the use of common statistical methods such as ANOVA models

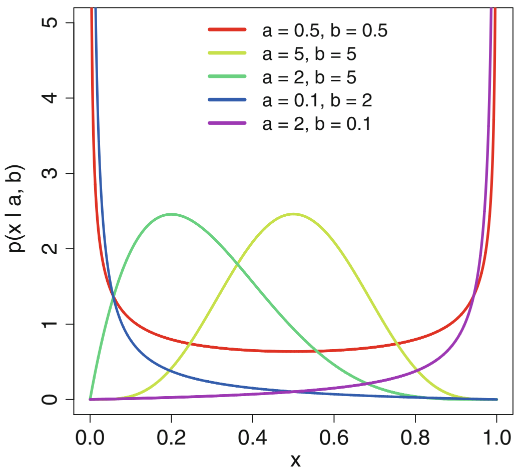


Fig. 5.1 Beta distribution for different pairs of shape parameters

and linear regression analysis. The disadvantage of M-values is that methylation is represented by a value between $(-\infty, \infty)$; this results in quantities that lack the intuitive interpretation characteristic of beta-values. In addition, it was shown in Zhuang et al. (2012) that under certain circumstances, the logit basis can lead to worse inference, as it can aggravate the effect of outliers (i.e., beta-values close to 0 or 1).

5.3 WBC DNA Methylation Data Set

Throughout this chapter, we will make use of publicly available data deposited in the Gene Expression Omnibus (accession number GSE39981) as an example data set. This data set consists of Illumina Infinium HumanMethylation27 data for different white blood cell types (WBC), including lymphoid-derived cell types (T cells ($n = 16$), natural killer cells ($n = 12$), and B cells ($n = 5$)) and myeloid-derived cell types (monocytes ($n = 5$) and granulocytes ($n = 8$)), collected from 46 different, healthy, non-diseased adults. Interested readers may refer to Houseman et al. (2012) for additional details regarding the study population, collection, and processing of these data. Given that DNA methylation is tissue specific and because several recent works have demonstrated distinct profiles of DNA methylation across WBC subtypes (Koestler et al. 2012; Reinius et al. 2012), these data represent an ideal data set on which to demonstrate clustering. We hereafter refer to this data set as the WBC DNA methylation data set. In the applications that follow, clustering of the samples was based on the top 500 most variable CpG sites (Wang et al. 2014). Justification and further discussion of this filtering/feature selection step is provided in Sect. 5.5.

5.4 Methods for Model-Based Clustering of DNA Methylation Array Data

For the remainder of this chapter, we assume that $\mathbf{X}$ is an $N \times J$ matrix of methylation data (either methylation M-values or beta-values) and $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)^T$ is a realization of this random matrix. As such, $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iJ})$ is a vector of length J representing the methylation values for subject i. Where relevant, we will be specific about whether $\mathbf{x}$ refers to the methylation M-values or beta-values. As cautionary note, several of the methods described in this chapter have R-packages whose name is identical to the function within them that performs the clustering method, i.e., the mclust package contains a function Mclust that implements the clustering algorithm. To avoid confusion we will typically refer to R-packages using **bolded** font and functions/R-objects using `courier` font.

5.4.1 Model-Based Clustering via Finite Mixture Models

Brief summary
Finite mixture models assume that the samples or observations belong to one of a fixed number of clusters and the variable denoting cluster membership is unobserved. Within each cluster, the observations are assumed to follow a prespecified distribution, often normal or Gaussian, but essentially any distribution can be assumed. For example, we might assume a Gaussian distribution for clustering samples on the basis of methylation M-values or a beta distribution for clustering samples on the basis of methylation beta-values. Regardless of the assumed distribution, the objective is the same and involves estimation of the cluster membership probabilities for each sample along with the model parameters that characterize the distribution of each cluster. In this section we discuss model-based clustering assuming a finite mixture of multivariate Gaussian distributions and demonstrate its application using the R-package **mclust**.

In model-based clustering, the data $\mathbf{x}$ are viewed as coming from a mixture density $p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k p_k(\mathbf{x})$, where $p_k(\mathbf{x})$ is the probability density function of the observations in group k and π_k is the probability that an observation comes from the kth mixture component ($0 \leq \pi_k \leq 1$, $\sum_{k=1}^{K} \pi_k = 1$). Each component $p_k(\mathbf{x})$ is usually modeled by the normal or Gaussian distribution, although theoretically any distribution, for example, the beta distribution (Eqs. 5.2 and 5.3) for methylation beta-values, can be used. Banfield and Raftery (1993) developed a general framework for clustering observations where each component density is assumed to be a multivariate normal distribution. Component distributions are characterized by the mean $\boldsymbol{\mu}_k$ and the covariance matrix $\boldsymbol{\Sigma}_k$ and have the probability density function:

$$p_k(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = (2\pi)^{-J/2}|\boldsymbol{\Sigma}_k|^{-1/2}\exp\{-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k)\} \quad (5.4)$$

Assuming a Gaussian-mixture model with K multivariate mixture components, the likelihood for data consisting of N samples is given by

$$p(\mathbf{x}|\Theta) = \mathcal{L}(\Theta|\mathbf{x}) = \prod_{i=1}^{N}\sum_{k=1}^{K} \pi_k p_k(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad (5.5)$$

where $\Theta = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K, \pi_1, \ldots, \pi_K)$ is a vector of model parameters. The problem of mixture model estimation from data $\mathbf{x}_i$, $i = 1, 2, \ldots, N$ can be formulated as to find the set of parameters Θ that gives the maximum likelihood estimate (MLE) solution:

$$\Theta^* = \underset{\Theta}{\operatorname{argmax}} \mathcal{L}(\Theta|\mathbf{x}) \tag{5.6}$$

In the common scenario where the probabilities of cluster membership π_k are unknown and need to be estimated, maximum likelihood estimates of $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ do not have closed-form analytic solutions. However, for a fixed number of components K, the model parameters Θ can be estimated by maximizing the complete data likelihood using the expectation-maximization (EM) algorithm (Dempster et al. 1977). Because each of the model-based clustering methods described in this chapter makes use of the EM algorithm for parameter estimation, we provide an intuitive overview of this technique in the section that follows. Additional details of this algorithm can be found elsewhere (Fraley and Raftery 2002) for those interested in a more comprehensive coverage of the EM algorithm in the context of finite mixture models.

5.4.1.1 The Expectation-Maximization Algorithm

The expectation-maximization (EM) algorithm is an iterative method for finding maximum likelihood estimates when there are missing values or latent variables (Dempster et al. 1977). In the mixture model context, the missing data is represented by a set of observations $\mathbf{z}$ of a discrete random variable Z, where $z_i \in \{1, \ldots, K\}$ indicates which mixture component generated the observations given in $\mathbf{x}_i$. For the time being, we shall assume that the number of mixtures K is fixed and known a priori, although treatment of this issue is provided later in this section. The likelihood of the complete data $(\mathbf{x}, \mathbf{z})$ takes the following multinomial form:

$$\begin{aligned} p(\mathbf{x}, \mathbf{z}|\Theta) = \mathcal{L}(\Theta|\mathbf{x}, \mathbf{z}) &= p(\mathbf{x}|\mathbf{z}, \Theta) p(\mathbf{z}|\Theta) \\ &= \prod_{k=1}^{K} \prod_{i=1}^{N} (\pi_k p_k(\mathbf{x}_i|\boldsymbol{\theta}_k))^{I(z_i=k)} \end{aligned}$$

where $\boldsymbol{\theta}_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k)$ is a vector of model parameters specific to class k, $I(.)$ is the indicator function, i.e., $I(z_i = k) = 1$ if $z_i = k$ and $I(z_i = k) = 0$ otherwise. The intuition behind the EM algorithm is as follows. Let Q represent the conditional expectation of the complete data $(\mathbf{x}, \mathbf{z})$, given the observed data $\mathbf{x}$ and a parameterization Θ^{t-1}:

$$\begin{aligned} Q(\Theta, \Theta^{t-1}) &= E_{\mathbf{z}}[\log(p(\mathbf{x}, \mathbf{z}|\Theta))|\mathbf{x}, \Theta^{t-1}] \\ &= \textstyle\sum_{\mathbf{z}\in\mathcal{Z}} p(\mathbf{z}|\mathbf{x}, \Theta^{t-1}) \log(p(\mathbf{x}, \mathbf{z}|\Theta)) \end{aligned} \tag{5.7}$$

where $\mathcal{Z}$ is the space of all possible values of $\mathbf{z}$. Given a parameterization Θ^t such that:

$$\Theta^t = \underset{\Theta}{\operatorname{argmax}} Q(\Theta, \Theta^{t-1}) \tag{5.8}$$

it can be shown that under regularity conditions,

$$\log(\mathcal{L}(\Theta^t|\mathbf{x})) \geq \log(\mathcal{L}(\Theta^{t-1}|\mathbf{x})) \tag{5.9}$$

This means that in maximizing Q in Eq. 5.7 with regard to a parameterization Θ^{t-1}, we obtain a parameterization Θ^t that maximizes the log-likelihood of Eq. 5.5. Based on this result, the EM algorithm proceeds by successively iterating between two steps:

1. **E-step**: In the first step, the EM algorithm involves finding the expected value of the complete likelihood given the current parameterization Θ^{t-1}.
2. **M-step**: In the second step, it searches for the set of parameters Θ^t that maximize the expectation from the E-step.

At each iteration t, the EM algorithm increases the log-likelihood and converges to a local maximum. These steps are repeated T times or until a convergence criterion is reached; for example, if there is a negligible difference in the log-likelihood between consecutive EM iterations, $\log(\mathcal{L}(\hat{\Theta}^t|\mathbf{x},\mathbf{z})) - \log(\mathcal{L}(\hat{\Theta}^{t-1}|\mathbf{x},\mathbf{z})) \approx 0$, where t signifies the tth iteration of the EM algorithm.

Before showing the components of E- and M-steps in the context of a Gaussian finite mixture model, we need to define $p(z_i = k|\mathbf{x}_i)$, the posterior probability of $z_i = k$ given $\mathbf{x}_i$. By Bayes rule this can be defined as follows:

$$\begin{aligned} p(z_i = k|\mathbf{x}_i, \Theta) &= \frac{p(z_i = k)p(\mathbf{x}_i|z_i = k, \boldsymbol{\theta}_k)}{p(\mathbf{x}_i|\Theta)} \\ &= \frac{\pi_k p_k(\mathbf{x}_i|\boldsymbol{\theta}_k)}{\sum_{k=1}^{K} \pi_k p_k(\mathbf{x}_i|\boldsymbol{\theta}_k)} \end{aligned} \tag{5.10}$$

For ease and simplicity in notation, we denote $w_{ik} = p(z_i = k|\mathbf{x}_i, \Theta)$. In simple terms, w_{ik} represents the probability that sample i belongs to cluster k, conditional on the observed data for sample i and model parameters $\boldsymbol{\theta}_k$ that define cluster k. In the case of mixture models, Eq. 5.7 can be rewritten as follows:

$$Q(\Theta, \Theta^{t-1}) = \sum_{k=1}^{K} \sum_{i=1}^{N} w_{ik} \log(\pi_k p_k(\mathbf{x}_i|\boldsymbol{\theta}_k^{t-1})) \tag{5.11}$$

E-Step The E-step amounts to finding the expected value of $\mathcal{L}(\Theta|\mathbf{x},\mathbf{z})$ given $\mathbf{x}_i$ and the current parameterization. As $\log(\mathcal{L}(\Theta|\mathbf{x},\mathbf{z}))$ is linear in $\mathbf{x}_i$, this reduces to calculating the expected value of $z_i = k$ given $\mathbf{x}_i$ and the previous parameterization Θ^{t-1},

$$\begin{aligned} E(z_i = k|\mathbf{x}_i) &= p(z_i = k|\mathbf{x}_i, \Theta^{t-1}) \\ &= w_{ik}, \text{ by Eq. 5.10} \end{aligned} \tag{5.12}$$

M-Step As previously described, the objective of the M-step is to find the set of parameters Θ^t that maximize the expectation from the E-step. This can be formally defined as follows:

$$\Theta^t = \underset{\Theta}{\operatorname{argmax}} Q(\Theta, \Theta^{t-1}) \tag{5.13}$$

To obtain the parameter estimates $\hat{\Theta}^t$, we use the maximum likelihood estimation. This involves differentiating Eq. 5.11 with respect to its parameters Θ, setting the resulting score equations equal to zero, and solving with respect to each of the parameters. In the context of a finite mixture of multivariate Gaussian distributions, it can be shown that

$$\hat{\pi}_k = \frac{\sum_{i=1}^N w_{ik}}{N}, \tag{5.14}$$

$$\hat{\boldsymbol{\mu}}_k = \frac{\sum_{i=1}^N w_{ik}\mathbf{x}_i}{\sum_{i=1}^N w_{ik}}, \text{ and} \tag{5.15}$$

$$\hat{\boldsymbol{\Sigma}}_k = \frac{\sum_{i=1}^N w_{ik}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)^T}{\sum_{i=1}^N w_{ik}} \tag{5.16}$$

One issue that deserves mentioning is the selection of the initial parameterization Θ^0 of the model, in particular, the initial selection of w_{ik}, $i = 1, \ldots, N$, and $k = 1, \ldots, K$ in Eq. 5.11. A standard way of initializing the parameters is to randomly choose w_{ik} values such that $0 \leq w_{ik} \leq 1$ and $\sum_{k=1}^K w_{ik} = 1$, followed by the estimation of parameters in the M-step. In order to deal with the effects of the random initialization, w_{ik}'s are generated several times (usually 15), and the selection that results in the highest likelihood is selected. Alternatively, Houseman et al. (2008) proposed initializing w_{ik} using a fuzzy clustering algorithm, such as the `fanny` (Kaufman and Rousseeuw 1990) algorithm in the R-package **cluster**. It is worth noting that most software for model-based clustering analysis, i.e., the `Mclust` function in the R-package **mclust**, do this step automatically, so the user does not need to supply their own set of initial values. We will demonstrate this with application of the `Mclust` function in the next section. However, it is worth noting that different initiations of w_{ik} can impact not only the final clustering solution but also the number of EM iterations for convergence and correspondingly the overall computational time. Thus, special care should be taken in the selection of the initial parameter values, Θ^0.

5.4.1.2 Parameterization of the Covariance Matrix $\boldsymbol{\Sigma}_k$

In the previous section we showed how the EM algorithm is used to obtain estimates of π_k, $\boldsymbol{\mu}_k$, and $\boldsymbol{\Sigma}_k$, when π_k is unknown, along with the estimators for these parameters in the M-step of the algorithm (Eqs. 5.14–5.16). However, in many practical applications, it might be useful to impose constraints on the geometric characteristics of $\boldsymbol{\Sigma}_k$, as the model may have more parameters than are reasonable to estimate given the available data. This is especially relevant given the high-dimensional nature of DNA methylation data (i.e., $N \ll J$). Below, we describe a framework developed by Banfield and Raftery (1993) for imposing geometric constraints on the cluster/component covariance matrices.

Data generated by mixtures of multivariate Gaussian distributions are characterized by clusters or groups of samples centered near the component means $\boldsymbol{\mu}_k$, with higher density for points nearer to the mean. The corresponding surfaces are of constant density, and geometric features (shape, volume, orientation) of the clusters – determined by the covariances $\boldsymbol{\Sigma}_k$ – may be parameterized to impose constraints across the components. As described in Fraley and Raftery (2002, 2007) and Banfield and Raftery (1993), there are a number of possible parameterizations of $\boldsymbol{\Sigma}_k$. Some common parameterizations include:

- $\boldsymbol{\Sigma}_k = \lambda\mathbf{I}$: **All components are spherical and of the same size**. In the context of methylation data $\mathbf{x}$, this amounts to assuming identical variances across CpG sites and clusters and, further, that all CpGs are uncorrelated.
- $\boldsymbol{\Sigma}_k = \lambda_k \boldsymbol{A}_k$: **All components are spherical, but need not be of the same size**. In the context of methylation data $\mathbf{x}$, this amounts to assuming a diagonal covariance structure, where the variances within a component and across components are permitted to vary. As we will later describe, this is the assumed covariance parameterization for RPMM.
- $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}$: **All components have the same geometry but need not be spherical**. In the context of methylation data $\mathbf{x}$, this amounts to assuming an identical variance-covariance across clusters. However, unlike the previous, CpGs are not constrained to be uncorrelated with one another, nor are they constrained to have identical variance.

A general framework for geometric constraints of the component covariance matrices can be obtained by parameterizing covariance matrices based on their eigenvalue decomposition (Banfield and Raftery 1993):

$$\boldsymbol{\Sigma}_k = \lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T \tag{5.17}$$

where $\mathbf{D}_k$ is an orthogonal matrix of eigenvectors, $\mathbf{A}_k$ is a diagonal matrix whose elements are proportional to the eigenvalues, and λ_k is the associated constant of proportionality. The factors $\mathbf{D}_k$, $\mathbf{A}_k$, and λ_k are treated as independent sets of parameters and are either constrained to be the same between clusters or are allowed to vary between clusters. Different assumptions about the characteristics of these factors govern the geometric properties of the component covariance

Table 5.1 Parameterizations of the multivariate Gaussian-mixture model available in the R-package **mclust**. The column labeled "Identifier" consists of a sequence of three letters that represent the geometric characteristics: volume, shape, and orientation. `E` means equal and `V` means varying across clusters. `I` refers to the identify matrix in specifying the shape and orientation of the covariance matrix. In the column labeled "# of covariance parameters," J denotes the dimension of the data (number of CpGs) and K denotes the assumed number of classes (Table abstracted from Fraley and Raftery 2007)

Indentifier	Model	# of covariance parameters	Distribution
`EII`	$\lambda\mathbf{I}$	1	Spherical
`VII`	$\lambda_k\mathbf{I}$	K	Spherical
`EEI`	$\lambda\mathbf{A}$	J	Diagonal
`VEI`	$\lambda_k\mathbf{A}$	$K + (J - 1)$	Diagonal
`EVI`	$\lambda\mathbf{A}_k$	$1 + K(J - 1)$	Diagonal
`VVI`	$\lambda_k\mathbf{A}_k$	KJ	Diagonal
`EEE`	$\lambda\mathbf{DAD}^T$	$J(J + 1)/2$	Ellipsoidal
`EEV`	$\lambda\mathbf{D}_k\mathbf{AD}_k^T$	$1 + (J - 1) + KJ(J - 1)/2$	Ellipsoidal
`VEV`	$\lambda_k\mathbf{D}_k\mathbf{AD}_k^T$	$K + (J - 1) + KJ(J - 1)/2$	Ellipsoidal
`VVV`	$\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$	$KJ(J + 1)/2$	Ellipsoidal

matrices $\mathbf{\Sigma}_k$ (Table 5.1). Specifically, $\mathbf{D}_k$ governs the orientation of the kth mixture component/cluster, $\mathbf{A}_k$ the shape, and λ_k the volume.

As noted in Table 5.1, elements within the column labeled "Identifier" consist of a sequence of three letters, `E`, `V`, and `I`, which represent the geometric characteristics of the covariance parameterization: volume, shape, and orientation. `E` means equal, `V` means varying across clusters, and `I` refers to the identify matrix in specifying the shape and orientation of the covariance matrix. For example, `VII` denotes a model in which the volumes differ between the clusters (`V`), but the shape and orientation are assumed to be the identity (`II`). Clusters in this model have diagonal covariances, where the diagonal elements (variances of the CpGs) are assumed to be equal within a cluster, but are allowed to vary between clusters.

5.4.1.3 Model Selection in Model-Based Clustering Analysis: Determining Covariance Parameterization and the Number of Clusters *K*

While the covariance parameterizations listed in Table 5.1 afford great flexibility for model-based cluster analysis via a mixture of multivariate Gaussian distributions, the question of which parameterization to select may not be immediately obvious. A "best" model can be estimated by fitting models with different parameterizations and then applying a statistical criterion for model selection. The Bayesian Information Criterion (BIC) (Schwartz 1978) is a likelihood-based criterion with a penalty for the number of parameters in the model and is the model selection criterion provided in the **mclust** package. Note that Fraley and Raftery (2002) define BIC for the model M_k as

$$\mathrm{BIC}_k = 2\log\mathcal{L}(\hat{\Theta}_k|\mathbf{x}) - \nu_k\log(N) \tag{5.18}$$

where ν_k is the number of independent parameters to be estimated in model M_k. Based on this definition, larger values of BIC_k are indicative of "better" model fit.

Up to this point, we have assumed that the number of clusters K is known. In reality, however, this is seldom the case, and the selection of K represents one of the fundamental issues in problems involving clustering (Chen 1995). Much in the same way that BIC is used for determining an appropriate covariance parameterization, the BIC can also be used for determining the number of clusters K. For example, the user begins by selecting an upper threshold for the maximum number of clusters in the data, $K_{\max}$. This could be based on any a priori knowledge about the problem/data set at hand or based on the maximum number of clusters the user is willing to accept. Models are then fit assuming $K = 1$ up to $K = K_{\max}$ clusters and the resulting BICs are used for model selection. As we will demonstrate in the next section, the `Mclust` function does this automatically, so beyond the selection of $K_{\max}$, there is no additional input required from the user.

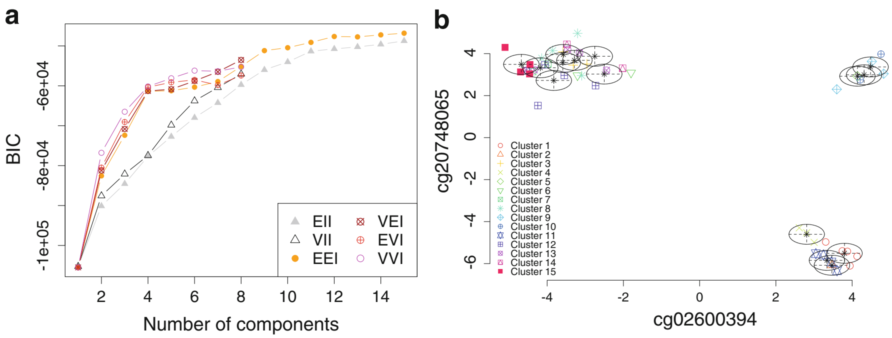
5.4.1.4 Application of Mclust in R

We illustrate model-based clustering using the WBC DNA methylation data set (Sect. 5.3), which consists DNA methylation measurements across different WBC cell types. Since `Mclust` fits Gaussian-distributed mixture models, we will cluster samples on the basis of their methylation M-values. Specifically we will cluster samples based on the top 500 most variable CpG sites. The following code computes the model using the function `Mclust` assuming a maximum of 15 clusters (i.e., $K_{\max} = 15$) and prints a summary of the model fit:

```
# BetaVals is a 46 × 500 matrix of methylation beta-values,
# consisting of the top 500 most variable CpG sites

R> library("mclust")
R> Mvals = log2(BetaVals) - log2(1-BetaVals)
R> gmm <- Mclust(Mvals, G = 1:15)
R> print(gmm)
```

This informs us that the "best model" is a model with a diagonal, equal volume, and shape covariance structure (`EEI`) with 15 components/clusters. Note that based on the above code, `Mclust` fits mixture models assuming $K = 1, 2, \ldots, 15$ and selects the optimal value $\hat{K}$ based on a comparison of the BICs calculated from each model fit. In this situation, it just happened to be the case that $\hat{K} = K_{\max}$. In such situations, we recommend refitting the model increasing $K_{\max}$, as it is likely that the optimal solution lies beyond the previous selection of $K_{\max}$. As previously described, BIC is also used as the basis for determining the optimal covariance parameterization (Table 5.1). When $N \leq J^*$, (J^* the number of features/CpGs used in clustering analysis) `Mclust` defaults to consider only spherical and diagonal covariance parameterizations (i.e., `EII`, `VII`, `EEI`, `VEI`, `EVI`, and `VVI` (Table 5.1)) because of model identifiability.



c

Cell type (columns); Cluster (**Mclust**) (rows)

Cluster	Bcell	CD4T	CD8T	Gran	Mono	NK	PanTcell
1	0	0	0	0	5	0	0
2	0	3	0	0	0	0	1
3	2	0	0	0	0	1	0
4	2	0	0	0	0	0	0
5	0	0	0	4	0	0	0
6	0	0	0	0	0	5	0
7	0	2	0	0	0	0	3
8	0	0	1	0	0	0	1
9	0	0	0	0	0	4	0
10	0	0	0	1	0	0	0
11	0	3	0	0	0	0	0
12	0	0	0	3	0	0	0
13	1	0	0	0	0	0	0
14	0	0	1	0	0	1	0
15	0	0	0	0	0	2	0

Fig. 5.2 Example output based on the model-based clustering solution obtained using the `Mclust` function. (**a**) Plot of BIC as a function of the number of assumed clusters/components K and across different covariance parameterizations (given in the plot legend). (**b**) Plot of the M-values for the first two CpGs contained in the `Mvals` object produced by the function `coordProj`. (**c**) Confusion matrix formed between the mclust cluster memberships and the cell-type classifications. *Blue* boxes indicate cell types derived from the myeloid lineage

The **mclust** R-package comes standard with several functions for visualizing the resulting model fit. The following code plots the BIC based on the resulting clustering solution (Fig. 5.2a):

```
R> plot(gmm, what = "BIC")
```

As noted from Fig. 5.2a, BIC is optimized when $K = 15$ and with `EEI` as the covariance parameterization. In addition, it might be of interest to visualize the clustering solution based on the M-values for a subset of the CpGs used in the clustering analysis. The code below extracts the cluster labels and parameters for the optimal clustering solution and generates a plot of the methylation M-values for the first two CpGs (i.e., first two columns of `Mvals`) used in the clustering analysis (Fig. 5.2b):

```
R> clust.labels = gmm$classification
R> clust.params = gmm$parameters
R> coordProj(Mvals, what = "classification", dimens = c(1,2),
     parameters = clust.params, classification = clust.labels)
```

From Fig. 5.2b we can see that the `coordProj` function also produces ellipsoids that are centered at each cluster mean with width in the x and y dimensions equal to the within-cluster estimated variance of those features. For this example, we notice that the semimajor and semiminor axes of the ellipsoids are parallel to the x and y coordinates. We also notice that the widths of the ellipsoids for the x and y dimensions are the same for each cluster, but x-dimension and y-dimension widths are different. This relates to the fact that the covariance parameterization for the optimal model was selected to be `EEI`. Referring back to Table 5.1, `EEI` corresponds to component/cluster covariances parameterized as $\mathbf{\Sigma}_k = \lambda \mathbf{A}, k = 1, 2, \ldots K$, and diagonal covariances that are constrained to be the same across components/clusters. If instead `EII` had been selected as the covariance parameterization, each of the ellipsoids in Fig. 5.2b would be replaced with circles, with equal diameters between clusters.

Now that we have the cluster labels based on the optimal clustering solution, we might next proceed by examining the relationship between cluster memberships and phenotypic, clinical, and/or demographic characteristics collected on the study samples. Since the data set considered here consists of DNA methylation data collected from different WBC subtypes, an obvious thing to examine is the distribution of WBC cell types across the 15 predicted clusters. Figure 5.2c contains the confusion matrix formed between cluster memberships and the WBC cell-type classifications. As noticed, myeloid and lymphoid cell types cluster uniquely, where cluster 1 (monocytes) and clusters 5, 10, and 12 (granulocytes) are comprised completely of myeloid-derived cell types, with the remaining clusters comprising the lymphoid-derived cell types. Thus, for this data set, `Mclust` does well in terms of identifying clusters that capture the major underlying structure of the data.

In this section we have merely covered the basics of model-based clustering via the **mclust** R-package with the intention of familiarizing users new to this technique with the fundamentals to get started. Additional examples of the functionalities of **mclust** can be found in Fraley and Raftery (2007) or in the R reference model for the **mclust** package http://cran.r-project.org/web/packages/mclust/mclust.pdf.

5.4.2 *LumiWCluster for Model-Based Clustering Analysis*

A customary preprocessing step in the analysis of Illumina BeadArray methylation data involves identification and removal of poor-quality samples and probes (Wilhelm-Benartzi et al. 2013). This is often achieved using detection P-values reported by BeadStudio, which are defined as 1-P-value computed from the background model characterizing the chance that the signal was distinguishable from negative controls. Numerous DNA methylation array analysis pipelines come standard with module options for carrying this out (see Morris and Beck 2014 for listing of such pipelines); however, the exact cutoffs for sample/probe exclusion based on detection P-values are somewhat arbitrary and conventions tend to vary among researchers (Marsit et al. 2009; Hernandez-Vargas et al. 2010; Bibikova et al. 2011). While it is crucial that quality measures pertaining to the samples and probes be taken into account, probe/sample exclusion based on hard thresholding may lead to a loss of information and unnecessary sample exclusions, further exacerbating the curse of dimensionality that is very often the case with DNA methylation array data (i.e., $N \ll J$). In an attempt to address this concern, Kuan et al. (2010) proposed a weighted model-based clustering framework for DNA methylation array data that systematically weights each observation according to the detection P-values and, in doing so, avoids discarding subsets of the data. Their method is called LumiWCluster and along with integrating information pertaining to the quality of samples/probes; LumiWCluster has a built-in procedure for selecting the most informative CpGs for clustering analysis. In what follows, we briefly describe the framework of LumiWCluster. Interested readers may refer to Kuan et al. (2010) for a more detailed account of this methodology.

Brief summary
A common quality control preprocessing step for DNA methylation data generated from the Illumina BeadArrays involves discarding samples and/or probes on the basis of their detection P-values. LumiWCluster (Il**Lumi**na **W**eighted Model-based **Cluster**ing) is a weighted model-based clustering method for Illumina BeadArray data that systematically weights each observation according to the detection P-values and, in doing so, avoids discarding subsets of the data (Kuan et al. 2010). An additional highlight of this methodology is that it has a built-in mechanism for automatically selecting informative CpGs for cluster analysis. Software for implementing the LumiWCluster methodology is freely available in the R-package **LumiWCluster**.

Let $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iJ})^T$ be a $J \times 1$ vector of methylation M-values for subject $i \in (1, 2, \ldots, N)$ and $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$. Similar to framework described in Sect. 5.4.1, assume that $\mathbf{x}$ is generated from a mixture of K multivariate Gaussian distributions. That is,

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k p_k(\mathbf{x}), \tag{5.19}$$

where $p_k(\mathbf{x})$ is given in Eq. 5.4 and π_k is the probability that an observation comes from the kth mixture component, subject to the following constraints: $0 \leq \pi_k \leq 1$,

$\sum_{k=1}^{K} \pi_k = 1$. From the above, the mixture model log-likelihood for the full data is given by

$$\ell(\Theta|\mathbf{x}) = \log(\mathcal{L}(\Theta|\mathbf{x})) = \sum_{i=1}^{N} \log \sum_{k=1}^{K} \pi_k p_k(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{5.20}$$

As before, $\Theta = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$, where $\boldsymbol{\theta}_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k)$ represent the unknown parameters for cluster $k \in (1, 2, \ldots, K)$. To avoid hard thresholding for sample exclusions based on detection P-values, Kuan et al. (2010) propose a weighted likelihood-based approach where sample weights reflect the quality of the sample. Specifically, the weighted mixture model log-likelihood function is given by

$$\ell_W(\Theta|\mathbf{x}) = \sum_{i=1}^{N} d_i \log \sum_{k=1}^{K} \pi_k p_k(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{5.21}$$

where d_i reflects the weight of sample i. Since samples with large detection P-values are indicative of lower quality, Kuan et al. (2010) suggest the following weighting scheme:

$$d_i = \frac{\text{median}_j(\log p_{ij})}{\sum_{i=1}^{N} \text{median}_j(\log p_{ij})}, 0 \leq d_i \leq 1, \tag{5.22}$$

where $\sum_{i=1}^{N} d_i = 1$ and p_{ij} denotes the detection P-value for sample i, CpG loci j. Thus, higher-quality samples (low detection P-values across the J CpG loci) are assigned larger weights and, correspondingly, have a greater influence on the estimation of the mixture model parameters. Similar to our previous discussion, estimation of Θ is achieved using the EM algorithm. As before, we define a random variable $z_i \in \{1, 2, \ldots, K\}$ that indicates which mixture component generated the observations given in $\mathbf{x}_i$. Assuming that the number of mixtures/clusters K is fixed and known, the complete weighted log-likelihood of the complete data $(\mathbf{x}, \mathbf{z})$ is given by

$$\ell_{\text{CW}}(\Theta|\mathbf{x}, \mathbf{z}) = \sum_{i=1}^{N} \sum_{k=1}^{K} d_i I(z_i = k) \left[\log \pi_k + \log p_k(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right] \tag{5.23}$$

where $I(.)$ is the indicator function. As important, if not more important as the selection of the clustering method itself, is the selection of the features (i.e., CpGs) used to cluster the samples. Up to this point, we have not considered this issue; however, it is important to note that feature selection for unsupervised clustering analysis is a difficult problem due to the absence of class labels that would guide the search for relevant features. For a given DNA methylation data set, it is reasonable to expect that only a subset $J^* \subset J$ of the CpGs vary across samples in a manner that is interesting to us, and in many scenarios this subset is likely to represent only a small fraction of all of the assayed CpGs. As the existence of many irrelevant features in clustering analysis may hinder the identification of the relevant underlying structure in the data, it is critical that feature selection be considered as a first step in the clustering analysis of DNA methylation data. With this in mind, Kuan et al. (2010) incorporate a feature selection step within their clustering framework that identifies important CpGs and clusters the N samples simultaneously. Specifically, they proposed the following penalized complete weighted likelihood:

$$\ell_{\mathrm{PCW}}(\Theta|\mathbf{x},\mathbf{z}) = \sum_{i=1}^{N}\sum_{k=1}^{K} d_i I(z_i = k)\left[\log \pi_k + \log p_k(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right] - J(\boldsymbol{\Omega}) \quad (5.24)$$

where $\boldsymbol{\Omega} = \{\mu_{kj}, k = 1,\ldots,K, j = 1,\ldots,J\}$ and $J(\boldsymbol{\Omega})$ is a penalty function. While numerous penalty functions are available, the fact that there is a natural group structure among the cluster means μ_{kj}'s (i.e., for each j, $\mu_{kj}, , k = 1,\ldots,K$ can be treated as a group since they are associated with the same CpG) and that CpG loci with large detection P-values are less reliable motivates the following penalty function:

$$J(\boldsymbol{\Omega}) = \sum_{j=1}^{J}\frac{\gamma_i}{g_i\ \mathrm{max}_k\left(|\tilde{\mu}_{kj}|^{\alpha}\right)} + \lambda\sum_{k=1}^{K}\sum_{j=1}^{J}\frac{|\delta_{kj}|}{|\tilde{\mu}_{kj}|^{\alpha}} \quad (5.25)$$

where $\mu_{kj} = \gamma_j\delta_{kj}$, $\tilde{\mu}_{kj}$'s are the unpenalized estimates of cluster means, $g_j = \frac{\mathrm{median}_i\,(\log p_{ij})}{\sum_{j=1}^{J}\mathrm{median}_i\,(\log p_{ij})}$ is the weight for locus j (larger g_j indicate more reliable probes), and α and λ are nonnegative tuning parameters. In Eq. 5.25, λ is a tuning parameter that controls the sparsity, where larger values of λ correspond to the selection of fewer CpG loci. With all other parameters fixed, we can see from Eq. 5.25 that small g_j's (indicative of a lower-quality probes) will be assigned a higher penalty and are more likely to be excluded in variable selection.

It is also worth noting that the following covariance parameterization, $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2,\ldots,\sigma_J^2)$, is assumed. That is, covariance matrices are the same across different clusters and are diagonal. This is equivalent to the `EEI` covariance parameterization under the **mclust** framework (Table 5.1). Thus, the problem of mixture model estimation can be formulated as to find the set of parameters Θ that gives the MLE solution,

$$\Theta^* = \underset{\Theta}{\mathrm{argmax}}\,\ell_{\mathrm{PCW}}(\Theta|\mathbf{x},\mathbf{z}) \quad (5.26)$$

As described previously (Sect. 5.4.1.1), this is achieved by iterating between the E- and M-steps T times or until $\ell_{\mathrm{PCW}}(\hat{\Theta}^t|\mathbf{x},\mathbf{z}) - \ell_{\mathrm{PCW}}(\hat{\Theta}^{t-1}|\mathbf{x},\mathbf{z}) \approx 0$, where t indicates the tth iteration of the EM algorithm. We refer readers to Kuan et al. (2010) for the closed-form expressions of the E- and M-steps in the context of the LumiWCluster framework.

Similar to the criterion used for selection of K in the unweighted finite mixture model approach given in Sect. 5.4.1, the BIC is used for the selection of K and λ in

LumiWCluster. To account for the weights d_i in the likelihood functions in which $\sum_{i=1}^{N} d_i = 1$, a modified version of the BIC is defined as follows:

$$\mathrm{BIC} = -2N \sum_{i=1}^{N} d_i \log\left(\sum_{k=1}^{K} \hat{\pi}_k p_k(\mathbf{x}_i|\hat{\boldsymbol{\theta}}_k)\right) + P \log(N) \tag{5.27}$$

where P is the total number of parameters in the model. Thus, the objective is to find the selection of K and λ that minimize Eq. 5.27. It is worth noting that the BIC given here differs in sign compared to the BIC expression given in the **mclust** framework. Thus, whereas we aimed to find the set of parameters that maximize the BIC in the **mclust** framework, for LumiWCluster we seek to find the set of parameters that minimize the BIC. In the section that follows, we illustrate LumiWCluster using the WBC DNA methylation data set.

5.4.2.1 Application of LumiWCluster in `R`

Before we begin with the implementation of the LumiWCluster methodology, we define the following variables:

- `DetPVals`: 46 × 500 matrix of detection P-values for the top 500 most variable CpGs
- `BetaVals`: 46 × 500 matrix of methylation beta-values for the top 500 most variable CpGs
- `CellType`: Vector of length 46 (in the same order as the rows of `BetaVals` and `DetPVals`) consisting of the cell-type classification for each of the samples

Looking at the above definitions, we notice that `BetaVals` consists of a subset of the CpGs assayed on the array (i.e., top 500 most variable CpGs), the same data as was used to illustrate the **mclust** methodology. This filtering step does not need to be done in practice as LumiWCluster has a built-in mechanism for variable selection; however, we do so here for computational convenience and for the purposes of comparison with the clustering solutions obtained from **mclust**, and later **RPMM**.

Since the `LumiWCluster` function takes as arguments the weights for the samples (i.e., d_i) and probes (i.e., g_j), we begin by computing these values using the detection P-values:

```
R> denom.d = sum(apply(DetPvals, 1, function(x) median(log(x))))
R> denom.g = sum(apply(DetPvals, 2, function(x) median(log(x))))
R> d = apply(DetPvals, 1, function(x) median(log(x)) / denom.d)
R> g = apply(DetPvals, 2, function(x) median(log(x)) / denom.g)
```

As this method assumes a weighted mixture of Gaussian distributions, we apply the `LumiWCluster` function to the methylation M-values for the WBC methylation data.

```
R> library(LumiWCluster)
R> Mvals = log2(BetaVals) - log2(1-BetaVals)
R> clusterSoln = LumiWCluster(t(Mvals), K = c(2:15), d, g)
```

Fig. 5.3 Confusion matrix constructed between the LumiWCluster cluster memberships and cell-type classification. *Blue boxes* indicate clusters uniquely comprised of myeloid-lineage cell types

	Cell Type						
Cluster (LumiWCluster)	Bcell	CD4T	CD8T	Gran	Mono	NK	PanTcell
1	0	0	0	0	5	0	0
2	0	0	1	0	0	1	0
3	0	0	0	4	0	0	0
4	2	0	0	0	0	1	0
5	0	3	0	0	0	0	0
6	0	0	0	0	0	6	0
7	0	0	0	0	0	5	0
8	0	0	1	0	0	0	1
9	0	2	0	0	0	0	3
10	0	0	0	4	0	0	0
11	3	0	0	0	0	0	0
12	0	3	0	0	0	0	1

The above function fits the LumiWCluster method assuming up to 15 clusters, consistent with our application of **mclust** in the previous section. Implementing the above code, we are informed that the optimal clustering solution consists of 12 clusters, i.e., $\hat{K} = 12$. Further, we are informed that when $K = 12$, the parameter λ, which controls the sparsity of the solution (i.e., number of informative CpGs to use in clustering), was estimated to be $\hat{\lambda} = 8.73$. Based on these estimates, we can easily extract the cluster membership for each of the samples, as well as the informative CpGs that withstood penalization:

```
R> clusterMembership = clusterSoln$ClusterID
R> InformCpGs = InformativeCpG(clusterSoln, rownames(Mvals))
```

Inspecting the `InformCpGs` object, we see that based on the estimates of λ and K, all 500 CpGs were retained in the optimal clustering model. This is not entirely surprising since the top 500 most variable CpGs were used for clustering analysis; thus, we might expect such CpGs to be informative in uncovering underlying structure in the data. Furthermore, from a confusion matrix constructed between the cluster memberships and cell-type classification, we see that similar to **mclust**, LumiWCluster was able to identify clusters that capture the major underlying structure of the data. Specifically, myeloid- and lymphoid-lineage cells cluster exclusively, though there are some clusters that are comprised of mixtures of lineage-specific cell types (i.e., clusters 4, 5, 8, 9, and 12) (Fig. 5.3).

A couple of important notes about the **LumiWCluster** R-package are discussed below. Depending on your operating system, installing and running **LumiWCluster** can be nontrivial. The `LumiWCluster` R-function passes objects to a function in C (also called LumiWCluster), which contains the code that is really the workhorse behind this method. While passing objects to C in this case substantially increases the computational efficiency of the method, this can create challenges when attempting to run the `LumiWCluster` in R. For example, after unpacking the "LumiWCluster_1.0.2.tar.gz" file, the "LumiWCluster.c code" (contained in LumiWCluster/src/) may need to be compiled in UNIX and dynamically loaded into R via `dyn.load()` before the `LumiWCluster` R-function will work. Instructions for carrying this out are provided at the following web address http://users.stat.umn.edu/~geyer/rc/.

5.4.3 Recursively Partitioned Mixture Models (RPMM)

Brief summary
The recursively partitioned mixture model (RPMM) (Houseman et al. 2008) is a model-based, binary recursive partitioning (BRP) method for clustering high-dimensional data. Like **mclust** and **LumiWCluster**, RPMM assumes that the samples arise as a finite mixture of distributions. However, unlike the previously described methods, RPMM uses recursive binary partitioning of the data to arrive at the final clustering solution. The result is an estimate of the number of clusters (without the user having to specify an upper bound), cluster membership probabilities, and, because of the hierarchical nature of this method, clusters that have a more meaningful interpretation. In addition, R-package **RPMM** comes equipped with functions for fitting both Gaussian and beta-distributed RPMMs, `glcTree` and `blcTree`, respectively. The latter is especially convenient given the inherent distribution of methylation beta-values.

In recent years binary recursive partitioning (BRP) methods have become widely popular tools for nonparametric regression and classification in many scientific fields. In the context of model-based clustering analysis, BRP methods are characterized by three crucial parts of the underlying algorithm: *partitioning* describes the fact that the algorithm arrives at a clustering solution by partitioning the samples based on a set of features (CpGs); *binary* describes the fact that, at any one step, the algorithm partitions the data into two subgroups; and *recursive* describes the fact that, within the subgroups created by partitioning the samples, the algorithm proceeds by further partitioning those subgroups based on the same or a different set of features (Merkle and Shaffer 2011). Thus, the application of BRP approaches for model-based clustering analysis represents an appealing framework for estimating the number of clusters K and, because of the hierarchical nature of BRP, results in solutions where there is an explicit structure/relationship among the clusters. For these reasons, RPMM has proved to be a formidable method and has been extensively used for clustering of DNA methylation data (Koestler et al. 2012; Marsit et al. 2011; Langevin et al. 2012; Cicek et al. 2013). In what follows, we provide a general overview of RPMM followed by an application of RPMM using the R-package **RPMM**. Interested readers may refer to Houseman et al. (2008) for further details regarding this methodology.

Much in the same way as Sects. 5.4.1 and 5.4.2, in RPMM, the data $\mathbf{x}$ are viewed as coming from a mixture density $p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k p_k(\mathbf{x})$, where $p_k(\mathbf{x})$ is the probability density function of the observations in group k and π_k is the probability that an observation comes from the kth mixture component. The likelihood contribution from sample i for a fixed number of classes K is assumed to take the following form:

$$\mathcal{L}_i(\Theta|\mathbf{x}_i) = p(\mathbf{x}_i|\Theta) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{J} p_{kj}(x_{ij}|\boldsymbol{\theta}_{kj}) \tag{5.28}$$

where $\Theta = (\boldsymbol{\theta}_{11}, \ldots, \boldsymbol{\theta}_{KJ})$ is a vector of model parameters. Although RPMM was initially described as a tool for navigating clusters in a beta-distributed mixture model (i.e., $p_{kj}(x_{ij}|\boldsymbol{\theta}_{kj}) \sim \text{Beta}(a_{kj}, b_{kj})$), theoretically any distribution for $p_{kj}(x_{ij}|\boldsymbol{\theta}_{kj})$ can be assumed. For now, we will treat $p_{kj}(x_{ij}|\boldsymbol{\theta}_{kj})$ in general terms, but note that the **RPMM** R-package comes standard with functions for fitting both beta- and Gaussian-mixture models. We also note from Eq. 5.28 that $p_k(\mathbf{x}_i|\boldsymbol{\theta}_k) = \prod_{j=1}^{J} p_{kj}(x_{ij}|\boldsymbol{\theta}_{kj})$. That is, conditional on membership in cluster k, features (CpGs) are assumed to be independent of one another. While attempts have been made to relax this assumption by modifying the covariance structure to incorporate known relationships between features, the computational efficiency of such approaches presents a major barrier and, in many scenarios, provides only marginal gains in clustering performance (Koestler et al. 2013). For a Gaussian-distributed RPMM, the assumption of cluster conditional independence of features is equivalent to assuming a `VEI` covariance parameterization (Table 5.1); i.e., a diagonal covariance structure whose elements are permitted to differ across clusters.

With methylation data observed on N subjects, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$, the full-data log-likelihood can be expressed as

$$\begin{aligned}\ell(\Theta|\mathbf{x}) = \log \mathcal{L}(\Theta|\mathbf{x}) &= \sum_{i=1}^{N} \log p(\mathbf{x}_i|\Theta) \\ &= \sum_{i=1}^{N} \log \left[\sum_{k=1}^{K} \pi_k \prod_{j=1}^{J} p_{kj}(x_{ij}|\boldsymbol{\theta}_{kj}) \right]\end{aligned} \tag{5.29}$$

At this point, the standard finite mixture model approach (Sect. 5.4.1) involves finding the set of parameters Θ^* that maximize the above equation. As previously described, MLE estimates of Θ are obtained using the EM algorithm, with the BIC typically forming the basis for model selection when mixture models with varying $K \in \{1, 2, \ldots, K_{\max}\}$ are fit. The approximate complexity for this entire operation is $NJK_{\max}^2$. To improve upon the computational complexity characteristic of the standard finite mixture model approach and to introduce a hierarchical framework that induces a natural structure between clusters, Houseman et al. (2008) proposed a recursive partitioning framework that, on average, has complexity $NJK \log K$ (K

is the true number of classes). Specifically, they proposed a weighted likelihood of the following form:

$$\ell_{\omega}^{(r)}(\Theta^{(r)}|\mathbf{x}) = \log \mathcal{L}_{\omega}^{(r)}(\Theta^{(r)}|\mathbf{x}) = \sum_{i=1}^{N} \omega_i^{(r)} \log p(\mathbf{x}_i|\Theta^{(r)}), 0 \leq \omega_i^{(r)} \leq 1 \quad (5.30)$$

When $\omega_i^{(r)} \equiv 1$ for all i, Eqs. 5.30 and 5.29 are equivalent. When $0 \leq \omega_i^{(r)} \leq 1$, sample i only partially contributes to estimation, and when $\omega_i^{(r)} = 0$, sample i is entirely excluded from estimation. While the above equation resembles the weighted likelihood equation for the LumiWCluster method (Eq. 5.21), d_i and $\omega_i^{(r)}$ are entirely different in terms of their interpretations. Whereas d_i is a weight that reflects the quality of sample i (as determined using the detection P-values), $\omega_i^{(r)}$ reflects the posterior probability of cluster membership in the parent cluster from the previous step of the recursion sequence, r. For example, if we begin by fitting a model to the data assuming $K = 2$, the result is two sets of posterior probabilities, $\omega_i^{(1)}$ and $\omega_i^{(2)}$, representing membership probabilities in clusters 1 and 2, respectively. Under the assumption that these clusters can be further split into 2 more clusters and that each sample belongs to the subsequent split with probability $\omega_i^{(1)}$ and $\omega_i^{(2)}$, the EM algorithm is applied recursively to the weighted likelihoods given by

$$\textbf{Parent cluster 1}: \ \ell_{\omega}^{(1)}(\Theta^{(1)}|\mathbf{x}) = \sum_{i=1}^{N} \omega_i^{(1)} \log p(\mathbf{x}_i|\Theta^{(1)}), 0 \leq \omega_i^{(1)} \leq 1$$

$$\textbf{Parent cluster 2}: \ \ell_{\omega}^{(2)}(\Theta^{(2)}|\mathbf{x}) = \sum_{i=1}^{N} \omega_i^{(2)} \log p(\mathbf{x}_i|\Theta^{(2)}), 0 \leq \omega_i^{(2)} \leq 1$$

to form two new clusters. The posterior probabilities for these new clusters are given as $\omega_i^{(1,1)}$ and $\omega_i^{(1,2)}$ for clusters obtained from splitting **Parent cluster 1** and $\omega_i^{(2,1)}$ and $\omega_i^{(2,2)}$ for clusters obtained from splitting **Parent cluster 2**. The above BRP process is continued until a point where splitting the data into two new clusters leads to a less parsimonious representation of the data or in situations where there is a small number of pseudo-subjects within a cluster, i.e., $\sum_{i=1}^{N} \omega_i^{(r)} \leq 5$, as the EM algorithm can become unstable when the number of pseudo-subjects is small. To address the former, Houseman et al. (2008) propose using a weighted version of the BIC for comparing model fit between successive splits of the data:

$$\text{wtdBIC}_1(r) = -2\ell_{\omega}^{(r)}(\Theta^{(r)}|\mathbf{x}) + 2J \log\left(\sum_{i=1}^{N} \omega_i^{(r)}\right) \quad (5.31)$$

$$\mathrm{wtdBIC}_2(r) = -2\ell_{\omega}^{(r)}(\Theta^{(r)*}|\mathbf{x}) + (4J + 1)\log\left(\sum_{i=1}^{N}\omega_i^{(r)}\right)$$

where the first set of parameters $\Theta^{(r)}$, defining $\mathrm{wtdBIC}_1(r)$, are obtained from the one-class model and the second set of parameters $\Theta^{(r)*}$, defining $\mathrm{wtdBIC}_2(r)$, are obtained from the two-class model. Thus, a split of the data is favored when $\mathrm{wtdBIC}_2(r) < \mathrm{wtdBIC}_1(r)$. To make these ideas more transparent, a general outline of the RPMM algorithm is given below.

RPMM Algorithm

- **Step 1**: Mixture model fit to the full data assuming $K = 1$ and $\omega_i = 1\ \forall i$.
  - Compute weighted BIC for this model, wtdBIC_1.
- **Step 2**: Mixture model fit to the full data assuming $K = 2$ and $\omega_i = 1\ \forall i$.
  - Obtain estimates of cluster membership probabilities, $\omega_i^{(1)}$ and $\omega_i^{(2)}$
  - Compute weighted BIC for this model, wtdBIC_2.
- **Step 3**: If $\mathrm{wtdBIC}_2 < \mathrm{wtdBIC}_1$ **Go on to Step 4**, Else **Stop**
- **Step 4**: Repeat **Steps 1–3** substituting ω_i with estimates $\omega_i^{(1)}$ and $\omega_i^{(2)}$ from **Step 2**

The final clustering solution consists of an estimate of the number of clusters $\hat{K}$, estimates of the model parameters $\hat{\Theta}$ defining the terminal solution, and the posterior probabilities of cluster membership for each sample across the $\hat{K}$ clusters.

5.4.3.1 Application of RPMM in R

The following R-code computes the solution for a beta-distributed RPMM based on the top 500 most variable CpGs for the WBC DNA methylation data set using the function `blcTree`:

```
# BetaVals is a 46 × 500 matrix of methylation beta-values,
# consisting of the top 500 most variable CpG sites

R> library("RPMM")
R> betaRPMM <- blcTree(BetaVals)
R> print(betaRPMM)
```

Applying the `print` statement to the object produced by the `blcTree` function provides the user with basic information about the nature of the clustering solution. In this case, the user is informed that the total number of nodes is 21 and the number of terminal nodes is 11. The total number of nodes refers to the number of nodes in the clustering dendrogram (including the root node), whereas the number of terminal

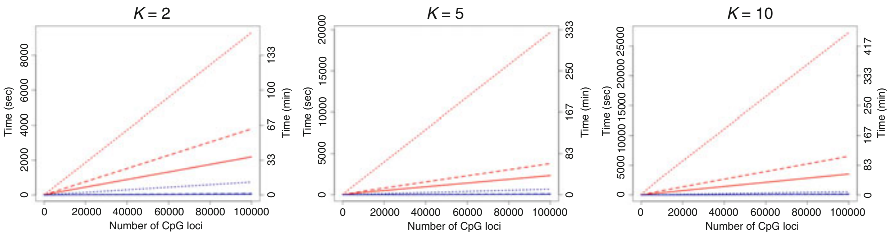


Fig. 5.4 Approximate computational time for beta (*red*)- and Gaussian (*blue*)-distributed RPMMs for varying K, number of samples N, and number of CpGs J. Line style indicates the number of samples being clustered, where *solid* ($N = 50$), *dashed* ($N = 100$), and *dotted* ($N = 500$)

nodes refers to the estimated number of clusters in the data (i.e., $\hat{K} = 11$). In lieu of or in addition to fitting a beta-distributed RPMM to the methylation beta-values, we might also be interested in fitting a Gaussian-distributed RPMM to the methylation M-values using the function `glcTree`:

```
R> Mvals = log2(BetaVals) - log2(1 - BetaVals)
R> gaussianRPMM <- glcTree(Mvals)
```

The most obvious initial difference between fitting Gaussian- and beta-distributed RPMMs is the computational time required to converge on the final clustering solution. As noted in Fig. 5.4, which shows the approximate computational time of Gaussian- and beta-distributed RPMMs when the true number of clusters K, number of samples N, and number of CpGs J are varied,[1] beta-distributed RPMMs converge at a much slower rate. This is unsurprising since the maximum likelihood estimator of beta-distribution parameters does not have a closed form and thus relies on numerical methods, which contribute to the computational burden associated with this method. While this should not be the sole determining factor in the selection of one method over another, it does deserve consideration, especially for large data sets.

Upon fitting RPMM, the next logical step is to extract the cluster memberships or membership probabilities for downstream statistical testing and/or visualize the resulting clustering solution. The `blcTree` and `glcTree` functions produce objects of the classes *blcTree* and *glcTree*, respectively, for which there exist several functions in the RPMM package for extracting information and visualizing the clustering solutions from objects of such classes. Table 5.2 provides a list of several useful functions along with a short description; however, we refer interested readers to the reference manual for RPMM (http://cran.r-project.org/web/packages/RPMM/RPMM.pdf) for additional details regarding their usage and arguments. For example, to obtain the cluster membership assignments for the objects obtained

[1]Details regarding the specification of the computing resources used for estimating computational times can be found at http://www.acf.ku.edu/wiki/.

Table 5.2 Useful functions for objects of class **blcTree**. For objects of class **glcTree**, substitute **blc** with **glc**

Function	Description
`blcTreeLeafMatrix`	Posterior probabilities of cluster membership based on clustering solution produced using `blcTree`
`blcTreeLeafClasses`	Cluster membership assignments based on highest posterior probability
`plotImage.blcTree`	Heat map based on the clustering solution
`plotTree.blcTree`	Tree dendrogram based on clustering solution
`ebayes`	Empirical Bayes predictions of the posterior probabilities of cluster membership for a new data set based on an RPMM fit

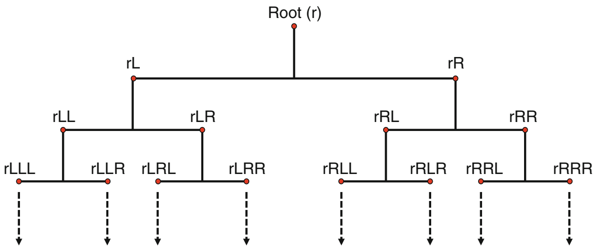


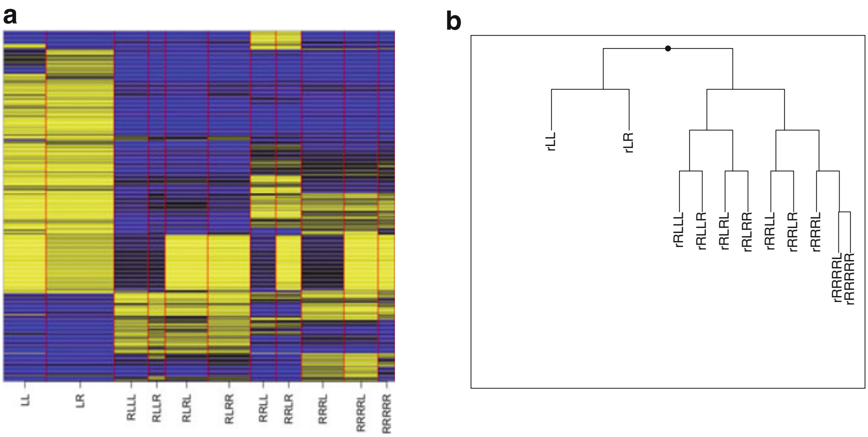
Fig. 5.5 Illustration of the RPMM cluster nomenclature. Each cluster name begins with an "r" (to denote that it derives from the root node) and is followed by a sequence of "L's" and "R's," short for left and right, to denote its location on the dendrogram tree

from fitting the `blcTree` and `glcTree` functions, we would simply do the following:

```
R> betaRPMMClasses <- blcTreeLeafClasses(betaRPMM)
R> gaussianRPMMClasses <- glcTreeLeafClasses(gaussianRPMM)
```

As noticed, clusters are named based on a sequence of "L's" and "R's," the exact sequence of which corresponds to their location on dendrogram tree (Fig. 5.5). While this may initially seem confusing, given the hierarchical nature of RPMM, this is a convenient and practical way of naming clusters in that it explicitly provides information about the relationship between clusters. For example, members of clusters "rLL" and "rLR," which are children of the same parent node, "rL," are more similar with regard to their methylation profile than are members of clusters "rLL" and "rRR." This would be missed if the cluster names followed a more generic nomenclature, i.e., cluster 1, cluster 2, etc., as is the case for **LumiWCluster** and **mclust**. In addition, it may also be of interest to produce a heat map and/or a dendrogram of the resulting clustering solution (Fig. 5.6a, b). This can be achieved using the `plotImage` and `plotTree` functions, respectively:

```
R> plotImage.blcTree(betaRPMM)
R> plotTree.blcTree(betaRPMM)
```



c

Clusters (**blcTree**) \ Clusters (**glcTree**)	rLL	rLRL	rLRR	rRLLL	rRLLR	rRLRLL	rRLRLR	rRLRR	rRRLL	rRRLR	rRRRL	rRRRRL	rRRRRR
rLL	5	0	0	0	0	0	0	0	0	0	0	0	0
rLR	0	4	4	0	0	0	0	0	0	0	0	0	0
rRLLL	0	0	0	4	0	0	0	0	0	0	0	0	0
rRLLR	0	0	0	0	2	0	0	0	0	0	0	0	0
rRLRL	0	0	0	0	0	4	1	0	0	0	0	0	0
rRLRR	0	0	0	0	0	0	3	2	0	0	0	0	0
rRRLL	0	0	0	0	0	0	0	0	3	0	0	0	0
rRRLR	0	0	0	0	0	0	0	0	0	3	0	0	0
rRRRL	0	0	0	0	0	0	0	0	0	0	5	0	0
rRRRRL	0	0	0	0	0	0	0	0	0	0	0	4	0
rRRRRR	0	0	0	0	0	0	0	0	0	0	0	0	2

Fig. 5.6 Example output based on `plotImage` and `plotTree` functions in the R-package RPMM. (**a**) Heat map of the RPMM clustering solution obtained using the `plotImage` function. Rows represent CpG loci and columns represent the samples, grouped by cluster membership. The width of each cluster is proportional to the number of samples predicted to be a member of that particular cluster. Values in the heat map reflect the within-cluster mean methylation levels for each of the CpG loci used to cluster the samples; *yellow* represents unmethylated CpG loci, and *blue* methylated CpG loci. (**b**) A dendrogram of the RPMM clustering solution produced by the `plotTree` function. (**c**) Cross-tabulation table of the clustering solutions obtained via a beta-distributed RPMM applied to the methylation beta-values (columns) and a Gaussian-distributed RPMM applied to the methylation M-values (rows)

Although RPMM automatically estimates the number of clusters K, in certain situations it might be of interest to restrict the number maximum depth that RPMM will recurse using the `maxlevel` argument in the `blcTree` and `glcTree` functions. In doing so, one imposes an upper bound on the maximum number of clusters that are considered to arrive at a final clustering solution. Since RPMM involves BRP to arrive at a final solution, setting the `maxlevel` = r corresponds to a maximum recursion depth of r branches or a maximum total of 2^r clusters. Note that setting `maxlevel` = r does not necessarily guarantee that RPMM will recurse to the maximum depth, only that RPMM will not recurse beyond this depth, as the algorithm can terminate before reaching that point if a split suggests a less parsimonious representation of the data or if there are a small number of pseudo-subjects within a cluster (Sect. 5.4.3). For example:

```
R> betaRPMMlevel3 <- blcTree(IllumBeta, maxlevel = 3)
R> gaussianRPMMlevel3 <- glcTree(Mvals, maxlevel = 3)
```

would restrict the beta- and Gaussian-distributed RPMM to recurse to a maximum of 3 branches or a maximum of 8 total clusters.

5.4.3.2 Assessing the Similarity Between Clustering Partitions

For this particular example, although the predicted number of clusters differs between the Gaussian and beta distribution ($\hat{K} = 11$ and 13, respectively), there is a high degree of agreement between the two clustering solutions (Fig. 5.6c). We can formally assess the similarity between two data clusterings, using one of several different indices. Together with the well-known Jaccard index (Jaccard 1901), the Rand Index (Rand 1971) is one of the most popular indices for assessing the correspondence between two data partitions. The Rand Index (RI) is defined as the ratio of the number of agreements $(a + d)$ and the sum of the number of agreements and disagreements $(a + d + c + b)$ between two data clusterings (Table 5.3):

$$RI = \frac{a+b}{a+b+c+d} = \frac{a+b}{\binom{n}{2}}, \qquad 0 \leq RI \leq 1 \tag{5.32}$$

where n represents the number of samples. Thus, values approaching 1 signify increasing agreement between the two data clusterings. While the Rand Index is a popular index and probably the most widely used for comparing two data partitions, it does however have some limitations, namely, that the expected value of the Rand Index does not take a constant value and the Rand statistic approaches its upper limit of unity as the number of clusters increases. Attempts to overcome these limitations include the Fowlkes-Mallows (Mallows and Fowlkes 1983) Index ($\mathrm{FM} = a/\sqrt{(a+b)/(a+c)}$) for comparing two hierarchical clusterings and the Adjusted Rand Index (ARI) proposed by Hubert and Arabie (1985).

$$\mathrm{ARI} = \frac{\binom{n}{2}(a+d) - [(a+b)(a+c) + (c+d)(b+d)]}{\binom{n}{2}^2 - [(a+b)(a+c) + (c+d)(b+d)]} \tag{5.33}$$

Table 5.3 Simplified 2 × 2 table for comparing data clusterings X and Y

	Y	
X	Pair in same group	Pair in different group
Pair in same group	a	b
Pair in different group	c	d

The ARI (Eq. 5.33) has an expected value of zero and maximum value of 1 and has been recommended as the index of choice for measuring the agreement between two partitions with a different number of clusters (Milligan and Cooper 1986). Implementation of the ARI is available as a function `adjustedRandIndex` in the R-package **mclust**. Applying `adjustedRandIndex` to the unconstrained clustering solutions given by the `betaRPMM` and `gaussianRPMM` objects,

```
R> library(mclust)
R> betaRPMMClasses <- blcTreeLeafClasses(betaRPMM)
R> gaussianRPMMClasses <- glcTreeLeafClasses(gaussianRPMM
R> adjustedRandIndex(betaRPMMClasses, gaussianRPMMClasses)
```

which yields ARI $=$ 0.80. Thus, for the example considered here, there is a strong agreement in the clustering solution produced from a beta-distributed RPMM fit to the methylation beta-values and a Gaussian-distributed RPMM fit to the methylation M-values.

5.5 Feature Selection in Clustering Analysis

While we have mentioned the importance of feature selection as an essential step of clustering analysis, this concept and existing techniques for its implementation deserve further discussion. Although treatment of feature selection is deserving of a chapter in and of itself, in the paragraphs that follow, we aim to describe several commonly used feature selection techniques in the clustering analysis DNA methylation data. Interested readers are encouraged to explore the following literature Ma and Huang (2008), Pok et al. (2010), and Wei and Billings (2007) for a more in-depth coverage of this topic.

The abundance of CpGs profiled on a typical DNA methylation microarray coupled with the small sample sizes that are characteristic of such studies renders the identification of underlying substructure in the data a difficult and daunting task. Under such circumstances, selection of the most discriminative or representative CpGs for clustering analysis inevitably becomes an important issue, and failure to do so can lead to unfavorable signal-to-noise ratios, impeding the identification of biologically interesting structure in the data regardless of the chosen clustering method. This problem is particularly acute for unsupervised clustering analysis problems due to the absence of class labels that can guide the search for relevant features. In terms of DNA methylation data, arguably the most widely used strategy for feature selection involves preselection of the top P most variable features (i.e., $P = 500$ and 1,000 are typical selections) for clustering analysis (Luo et al. 2014; Wockner et al. 2014; Milani et al. 2010; Pacheco et al. 2011). In fact, a variation of this basic approach was recommended as the first step prior to the application of RPMM (Houseman et al. 2008). Although there is a biological motivation behind using the top P most variable features for clustering analysis – that is, variable CpGs are likely to be informative with regard to underlying structure in the data – the selection of P is somewhat arbitrary and clustering solutions can be sensitive to its selection. The sensitivity of clustering solutions can however be examined by comparing the clustering memberships across a range of preselected P using, for example, the Adjusted Rand Index; yet, the final selection of P is still arbitrary. Furthermore, there is no guarantee that clustering on the top P most variable features will result in biologically or phenotypically relevant clusters. To circumvent this, and the other limitations associated with clustering using the top P most variable features, many have turned to semi-supervised clustering methods (Koestler et al. 2010; Bair and Tibshirani 2004). Semi-supervised methods involve randomly splitting the full data set into independent training and testing sets. The training set is used for preselecting features that associate with some phenotype(s) of interest (i.e., survival time, histological subtype, smoking status, etc.), along with determining the number of features for clustering analysis. Based on this information, the clustering model is then fit to the observations in the independent test set. While such approaches have demonstrated improved performance over fully unsupervised approaches in terms of identifying biologically and phenotypically relevant clusters (Bair and Tibshirani 2004), they generally require large sample sizes, which may be infeasible given the logistical and/or practical constraints of a given study.

As we previously described, the LumiWCluster method comes standard with a mechanism for feature selection. Their approach, which involves the addition of a penalty term to the log-likelihood, encourages sparse solutions and identifies important CpGs whose methylation status is used to inform the clustering solution. Along the same lines, Witten and Tibshirani (2010) proposed a novel framework for sparse clustering, in which one clusters the observations using an adaptively chosen subset of the features. Like LumiWCluster, their method uses a lasso-type penalty to select the features. However, unlike LumiWCluster where this is accomplished in a model-based framework, Witten and Tibshirani (2010) develop their framework in the context of nonparametric clustering methods, specifically, sparse K-means and hierarchical clustering.

An alternative strategy for feature selection involves clustering based on CpG loci associated with genes that have been implicated in oncogenesis or other biological processes relevant to the problem at hand. For example, a strategy that may be particularly well suited for cancer-related data sets would consist of clustering on CpGs that are associated with the genes (or subset therein) contained in the Catalogue of Somatic Mutations in Cancer (COSMIC) census gene list (Futreal et al. 2004) and/or the cancer gene list (allOnco) procured by the Bushman Laboratory (http://www.bushmanlab.org/links/genelists). Both resources contain a list of genes that are causally implicated in oncogenesis, curated from the available literature. Thus, these resources may serve as valuable tools for identifying biologically relevant CpGs for subsequent clustering analysis.

5.6 Chapter Summary and Discussion

In this chapter, we aimed to provide readers with an intuitive overview model-based clustering methods for DNA methylation array data, focusing on three different methods of which two were specifically motivated by array-based DNA methylation data. We also aimed to demonstrate a practical application of each method using their corresponding R-package and the functions within. While mclust, LumiWCluster, and RPMM all share a common framework, they differ in terms of the assumed underlying distribution for the data, flexibility to handle multiple different covariance parameterizations, ability to automatically identify informative CpGs for clustering (feature selection), and, in some cases, their computational efficiency (Fig. 5.7a). At this point, it is natural to ask, "so which method is the best and the one I should apply to my data set?" to which there is no easy answer. In the context of DNA methylation data, a comprehensive comparison of the strengths and shortcomings of each of the considered model-based methods, along with state-of-the-art nonparametric methods, remains an open research question and represents an opportunity for future work. While a comparison of the clustering solutions obtained by fitting mclust, LumiWCluster, and RPMM to the WBC DNA methylation data set showed a high degree of consistency, with the beta-RPMM and LumiWCluster demonstrating slightly better concordance with the true cell classifications (Fig. 5.7b), this is but one of the many necessary comparisons that would need to be considered in order to form a complete picture of when one method is preferred over another.

In conclusion, clustering has proved to be a valuable tool for understanding DNA methylation and represents a staple technique for the analyst toolbox. Moving

a

	Mclust	LumiWCluster	RPMM
1. Fits a mixture of Gaussian distributions	✓	✓	✓
2. Fits a mixture of beta distributions	✗	✗	✓
3. Incorporates different parameterizations of the covariance structure	✓	✗	✗
4. Performs automatic feature selection	✗	✓	✗
5. Hierarchical in nature	✗	✗	✓
6. Integrates information pertaining to the quality of methylation measurements	✗	✓	✗
7. Freely available software	✓	✓	✓

b

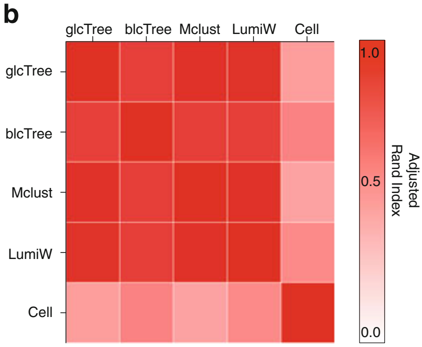


Fig. 5.7 (**a**) Side-by-side comparison of Mclust, RPMM, and LumiWCluster. *Yellow* refers to the fact that a software implementation outside of the actual R-package exists. (**b**) Image plot of the Adjusted Rand Index between clustering solutions obtained via beta- and Gaussian-distributed RPMMs (blcTree and glcTree, respectively), Mclust, LumWCluster(LumiW), and the cell classifications (Cell)

forward, it is critical that clustering methods be developed that keep pace with our evolving understanding of DNA methylation and the technologies used for its assessment.

Acknowledgements We would like to offer our deepest gratitude to Dr. Joseph Usset and Samuel Turpin for their feedback, suggestions, and comments on this chapter.

References

Houseman EA, Christensen BC, Yeh R-F, Marsit CJ, Karagas MR, Wrensch M, Nelson HH, Wiemels J, Zheng S, Wiencke JK, Kelsey KT. Model-based clustering of DNA methylation array data: a recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distributions. BMC Bioinform. 2008;9:365

Kuan PF, Wang S, Zhou X, Chu H. A statistical framework for illumina DNA methylation arrays. Bioinformatics. 2010;26:2849–55.

Siegmund KD, Laird PW, Laird-Offringa IA. A comparison of cluster analysis methods using DNA methylation data. Bioinformatics. 2004;20:1896–904.

Koestler DC, Christensen BC, Marsit CJ, Kelsey KT, Houseman EA. Recursively partitioned mixture model clustering of DNA methylation data using biologically informed correlation structures. Stat Appl Genet Mol Biol. 2013;12:225–40.

Fraley C, Raftery AE. Model-based clustering, discriminant analysis and density estimation. J Am Stat Assoc. 2002;97:611–31.

Du P, Zhang X, Huang C-C, Jafari N, Kibbe WA, Hou L, Lin SM. Comparison of beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinform. 2010;11:587

Saadati M, Benner A. Statistical challenges of high-dimensional methylation data. Stat Med. 2014;33(30):5347–57

Zhuang J, Widschwendter M, Teschendorff AE. A comparison of feature selection and classification methods in DNA methylation studies using the illumina infinium platform. BMC Bioinform. 2012;13:59

Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, Wiencke JK, Kelsey KT. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinform. 2012;13:86

Koestler DC, Marsit CJ, Christensen BC, Accomando W, Langevin SM, Houseman EA, Nelson HH, Karagas MR, Wiencke JK, Kelsey KT. Peripheral blood immune cell methylation profiles are associated with nonhematopoietic cancers. Cancer Epidemiol Biomark Prev. 2012;21:1293–302.

Reinius LE, Acevedo N, Joerink M, Pershagen G, Dahlén S-E, Greco D, Söderhäll C, Scheynius A, Kere J. Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. PLoS One. 2012;7(7):e41361.

Wang X, Laird PW, Hinoue T, Groshen S, Siegmund KD. Non-specific filtering of beta-distributed data. BMC Bioinformatics. 2014;15:199

Banfield J, Raftery A. Model-based gaussian and non-gaussian clustering. Biometrics. 1993;49:803–21.

Dempster A, Laird N, Rubin D. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser B (Methodological). 1977;39:1–38.

Kaufman L, Rousseeuw P. Finding groups in data: an introduction to cluster analysis. Hoboken, New Jersey: Wiley Interscience; 1990.

Fraley C, Raftery AE. Model-based methods of classification: using the mclust software in chemometrics. J Stat Softw. 2007;18:1–13.

Schwartz G. Estimating the dimension of a model. Ann Stat. 1978;6:461–4.

Chen J. Optimal rate of convergence for finite mixture models. Ann Stat. 1995;23:221–33.

Wilhelm-Benartzi CS, Koestler DC, Karagas MR, Flanagan JM, Christensen BC, Kelsey KT, Marsit CJ, Houseman EA, Brown R. Review of processing and analysis methods for DNA methylation array data. Br J Cancer. 2013;109:1394–402.

Morris TJ, Beck S. Analysis pipelines and packages for infinium humanmethylation450 beadchip (450k) data. Methods. 2014;72:3–8.

Marsit CJ, Christensen BC, Houseman EA, Karagas MR, Wrensch MR, Yeh R-F, Nelson HH, Wiemels JL, Zheng S, Posner MR, McClean MD, Wiencke JK, Kelsey KT. Epigenetic profiling reveals etiologically distinct patterns of DNA methylation in head and neck squamous cell carcinoma. Carcinogenesis. 2009;30:416–22.

Hernandez-Vargas H, Lambert M-P, Le Calvez-Kelm F, Gouysse G, McKay-Chopin S, Tavtigian SV, Scoazec J-Y, Herceg Z. Hepatocellular carcinoma displays distinct DNA methylation signatures with potential as clinical predictors. PLoS One. 2010;5(3):e9749.

Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, Delano D, Zhang L, Schroth GP, Gunderson KL, Fan J-B, Shen R. High density DNA methylation array with single CpG site resolution. Genomics. 2011;98:288–95.

Merkle EC, Shaffer VA. Binary recursive partitioning: background, methods, and application to psychology. Br J Math Stat Psychol. 2011;64:161–81.

Marsit CJ, Koestler DC, Christensen BC, Karagas MR, Houseman EA, Kelsey KT. DNA methylation array analysis identifies profiles of blood-derived DNA methylation associated with bladder cancer. J Clin Oncol. 2011;29:1133–9.

Langevin SM, Koestler DC, Christensen BC, Butler RA, Wiencke JK, Nelson HH, Houseman EA, Marsit CJ, Kelsey KT. Peripheral blood dna methylation profiles are indicative of head and neck squamous cell carcinoma: an epigenome-wide association study. Epigenetics. 2012;7:291–9.

Cicek MS, Koestler DC, Fridley BL, Kalli KR, Armasu SM, Larson MC, Wang C, Winham SJ, Vierkant RA, Rider DN, Block MS, Klotzle B, Konecny G, Winterhoff BJ, Hamidi H, Shridhar V, Fan J-B, Visscher DW, Olson JE, Hartmann LC, Bibikova M, Chien J, Cunningham JM, Goode EL. Epigenome-wide ovarian cancer analysis identifies a methylation profile differentiating clear-cell histology with epigenetic silencing of the HERG k+ channel. Hum Mol Genet. 2013;22:3038–47.

Jaccard P. Etude comparative de la distribution florale dans une portion des alpes et des jura. In Bull del la Soc Vaud des Sci Nat. 1901;37:547–79.

Rand W. Objective criteria for the evaluation of clustering methods. J Am Stat Assoc. 1971;66:846–50.

Mallows C, Fowlkes E. A method for comparing two hierarchical clusterings. J Am Stat Assoc. 1983;78:553–69.

Hubert L, Arabie P. Comparing partitions. J Classif. 1985;2:193–218.

Milligan G, Cooper M. A study of the comparability of external criteria for hierarchical cluster analysis. Multiv Behav Res. 1986;21:441–58.

Ma S, Huang J. Penalized feature selection and classification in bioinformatics. Brief Bioinform. 2008;9:392–403.

Pok G, Liu J-CS, Ryu KH. Effective feature selection framework for cluster analysis of microarray data. Bioinformation. 2010;4(8):385–9.

Wei H-L, Billings SA. Feature subset selection and ranking for data dimensionality reduction. IEEE Trans Pattern Anal Mach Intell. 2007;29:162–6.

Luo Y, Wong C-J, Kaz AM, Dzieciatkowski S, Carter KT, Morris SM, Wang J, Willis JE, Makar KW, Ulrich CM, Lutterbaugh JD, Shrubsole MJ, Zheng W, Markowitz SD, Grady WM. Differences in DNA methylation signatures reveal multiple pathways of progression from adenoma to colorectal cancer. Gastroenterology. 2014;147:418–29.e8.

Wockner LF, Noble EP, Lawford BR, Young RM, Morris CP, Whitehall VLJ, Voisey J. Genome-wide DNA methylation analysis of human brain tissue from schizophrenia patients. Trans Psychiatry. 2014;4:e339.

Milani L, Lundmark A, Kiialainen A, Nordlund J, Flaegstad T, Forestier E, Heyman M, Jonmundsson G, Kanerva J, Schmiegelow K, Söderhäll S, Gustafsson MG, Lönnerholm G, Syvänen A-C. DNA methylation for subtype classification and prediction of treatment outcome in patients with childhood acute lymphoblastic leukemia. Blood. 2010;115:1214–25.

Pacheco SE, Houseman EA, Christensen BC, Marsit CJ, Kelsey KT, Sigman M, Boekelheide K. Integrative DNA methylation and gene expression analyses identify DNA packaging and epigenetic regulatory genes associated with low motility sperm. PLoS One. 2011;6(6):e20280.

Koestler DC, Marsit CJ, Christensen BC, Karagas MR, Bueno R, Sugarbaker DJ, Kelsey KT, Houseman EA. Semi-supervised recursively partitioned mixture models for identifying cancer subtypes. Bioinformatics. 2010;26:2578–85.

Bair E, Tibshirani R. Semi-supervised methods to predict patient survival from gene expression data. PLoS Biol. 2004;2:E108.

Witten DM, Tibshirani R. A framework for feature selection in clustering. J Am Stat Assoc. 2010;105:713–26.

Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR. A census of human cancer genes. Nat Rev Cancer. 2004;4:177–83.

Part II
Integrative and Medical Epigenomics

Chapter 6
Integrative Epigenomics

Ming Su, Xiaoyang Dou, Hao Cheng, and Jing-Dong J. Han

Abstract In the post-genomic era, various types of functional omics data are emerging. As a result, big omics data are accumulating at an explosive rate. Epigenomics, including genome-wide DNA methylation and histone modifications, are important components of functional genomics, and play an essential role in elucidating many fundamental biological processes. Integration of epigenomic data with genomic, transcriptomic and proteomic data is increasingly valued to uncover full pictures of biological systems. Simple intersection of epigenetic features may provide interesting clues of novel patterns. Various machine learning methods are utilized to help understand chromosome segmentation and epigenetic regulation of transcription. Additionally, cluster analyses are frequently applied in cancer classifications. In this chapter, we briefly review commonly used integration methods and algorithms.

Keywords Epigenomics • Chromatin • DNA methylation • Clustering • Integration

6.1 Introduction

The great physicist Werner Heisenberg once said: "We have to remember that what we observe is not nature in itself, but nature exposed to our method of questioning". We can now obtain the genome, epigenome, transcriptome, interactome and proteome of a biological system, all of which are different parts of our subject. Only when we integrate all the parts together, we can tell what the system is.

Integration of epigenomic data, such as DNA methylation and histone modification, with other omics data has proven to be useful in answering many fundamental biological questions. How are epigenetics involved in transcription regulation and chromosome organization? What are the roles of epigenetic regulation in the

M. Su • X. Dou • H. Cheng • J.-D.J. Han (✉)
Key Laboratory of Computational Biology, Chinese Academy of Sciences-Max Planck Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 320 Yue Yang Road, Shanghai 200031, China
e-mail: jdhan@picb.ac.cn

A.E. Teschendorff (eds.), *Computational and Statistical Epigenomics*, Translational Bioinformatics 7, DOI 10.1007/978-94-017-9927-0_6

temporal and spatial control of development and ageing process? What are the relationships between diseases and genomic variations? Are there any common patterns of epigenetic variations in various cancers? How can we utilize these epigenetic patterns, combined with other molecular features, to improve molecular classification of cancers to facilitate diagnosis and therapy? The introduction and popularization of next-generation sequencing technology has led to massive omics data production and a large number of integration methods and algorithms. Recently, data generation and algorithm development were again accelerated by several collaborative projects. The NIH Roadmap Epigenomics Mapping Consortium (http://www.roadmapepigenomics.org/) aims to produce a public resource of human epigenomic data to catalyze basic biology and disease-oriented research. The Consortium has mapped DNA methylation, histone modifications, chromatin accessibility and small RNA transcripts in many cell lines and tissues. The ENCODE (http://www.genome.gov/10005107) and the modENCODE (http://www.modencode.org/) projects are dedicated to identify all functional elements in the genome of human and other model organisms by leveraging epigenomic, transcriptomic and genomic data. Meanwhile, TCGA (http://cancergenome.nih.gov/) and ICGC (https://icgc.org/) cancer genome projects aim to obtain a comprehensive description of genomic, transcriptomic and epigenomic changes in different tumour types. In addition to generating enormous amounts of omics data, all of these consortia also make great efforts on data analysis, especially on integration of multi-platform data.

In this chapter, we will introduce several typical integration methods or algorithms, which are organized based on the methodologies, namely, simple integration, machine learning–based integration and cluster-based integration.

6.2 Simple Integration

The most straightforward way to integrate omics data from different assays is intersection analysis among features extracted from these data. Omic data generated by the next-generation sequencing or microarray technology can be first simplified to “features”, which might be peaks of reads in ChIP-seq and chromatin accessibility assays, hypo-methylated or hyper-methylated regions in assays profiling DNA methylation landscapes, gene expression levels, or differentially expressed genes in transcriptome analysis assays, and genome variations called from exome or whole genome sequencing assays. Then data integration can be conveniently performed as intersecting the variety of features to obtain overlaps of special interests.

Such simple integrations are routinely used in most epigenomics studies. For example, Guttman et al. used K4–K36 domains defined by H3K4me3 and H3K36me3 ChIP-seq data and RNA-seq data to identify novel large intervening non-coding RNAs (lincRNAs) (Guttman et al. 2009). A signature of high H3K4me1 and low H3K4me3 was used to predict transcription enhancers (Heintzman et al. 2009, 2007). Liu et al. marked transcription start sites (TSSs) in the Rhesus macaque genome by H3K4me3 ChIP-seq and refined them with RNA-seq (Liu et al. 2011).

Other studies use one omic data set to identify patterns of interests and then use a different omic data set to validate them. This methodology can be considered as a simple integration in a broad sense. Hon et al. mapped methylomes in 17 adult mouse tissues and identified more than 300,000 tissue-specific differentially methylated regions (tsDMRs), which bore active marks such as H3K4me1 and H3K27ac in tissues where they were hypo-methylated but not in other tissues (Hon et al. 2013). Similarly, a set of transposable elements differentially methylated across human tissues also showed high levels of H3K4me1 modification and enhancer-associated protein P300 binding and thus were potential enhancers (Xie et al. 2013). Chromatin interaction analysis with paired-end tagging (ChIA-PET) is a novel technique to detect long distance DNA–DNA interaction. Two groups performed ChIA-PET assay separately with an antibody recognizing RNA polymerase II to explore DNAs interacting with gene promoters. They found intergenic DNAs enriched for interactions overlapped with DNase hypersensitive sites (DHSs) (Kieffer-Kwon et al. 2013) or were marked with high level of H3K4me1 modification (Zhang et al. 2013).

Simple integrations are frequently used to identify and annotate functional DNA elements such as genes and enhancers. In fact, this is also the goal of the ENCODE and modENCODE projects. Both projects released comprehensive epigenomic data as well as transcriptomic and interactomic data, and presented diverse integrative analyses (The ENCODE Project Consortium 2012; Gerstein et al. 2010; Roy et al. 2010). For the promoters, various histone modification levels were used to predict RNA expression, providing insights of epigenetic control and regulation upon transcription. More efforts were made for annotation of non-coding regions. Histone modifications were used to determine the chromatin status; DNase-seq and FAIRE-seq which reflected open chromatin helped to identify regulatory elements; DNase footprinting (Boyle et al. 2011; Hesselberth et al. 2009) and ChIP-seq for DNA binding proteins determined the transcription factors (TFs) binding regions; and finally 5C and ChIA-PET techniques revealed associations among these elements. Combining all these data sets, 80 % of the human genome and 82 % of the fruit fly genome were assigned certain functions.

When disease-associated variant data is included, convenient integration can help to elucidate how variants cause diseases. Genome-wide association studies (GWAS) have identified a huge number of disease-associated single nucleotide polymorphisms (SNPs), but few of them are located in protein-coding exons. Non-coding regulatory elements revealed by integration analyses provide an alternative solution. An example comes from Pasquali et al.'s work on type-II diabetes (T2D) (Pasquali et al. 2014). By integrating RNA-seq, FAIRE-seq, TF and histone modification ChIP-seq data, they identified genomic sequences targeted by islet-specific transcription factors. Such sequences resided in clusters of enhancers and physically associated with islet-specific gene *ISL1*, as 4C-seq data revealed. Furthermore, these sequences were enriched for SNPs found in T2D and fasting glycaemia GWASs, and one of the SNPs disrupted TF binding and islet enhancer activity. Thus, with more non-coding sequences annotated, we can explain more disease-associated variations.

Intersecting analysis requires implicit correlation among different data sets. Besides a priori knowledge, a variety of genome browsers available provide a way to discover novel correlations. The most well-known online browser is the UCSC genome browser (Kent et al. 2002), which provides plenty of public databases (e.g. Roadmap Epigenomics and ENCODE data sets) for integrative visualization and exploration. Another popular browser is the Integrative Genomics Viewer (IGV) (Robinson et al. 2011; Thorvaldsdottir et al. 2013). It is a lightweight visualization tool, allowing real-time desktop analysis of local data. Most recently, another visualization tool called Epiviz was released (Chelaru et al. 2014) with a module implemented in the commonly used statistical programming language R, which makes it easy to be called from existing analysis pipelines.

6.3 Machine Learning–Based Integrative Epigenomics

Machine learning–based integrative methods can be grouped into two categories: one is based on the combination of different epigenomic markers to classify genomic regions into different functional elements, which are also known as chromatin states, and the other is to infer the relationships among different epigenomic markers.

Hidden Markov Model assumed that the system has a series of unobserved states (hidden states), which follows a Markov process. The observed data are considered as output from the hidden states and they follow specific distributions. HMM has been widely used in engineering applications such as speech and handwriting recognition. In recent years, it has been applied to solve many problems in computational molecular biology, statistical genetics and also genome-wide studies (Choi et al. 2009; Churchill 1989; Krogh et al. 1994).

Ernst and Kellis proposed a multivariate Hidden Markov Model to capture the chromatin states in human T cells using 38 different histone markers, and H2AZ, RNA polymerase II and CTCF binding profiles (Ernst and Kellis 2010, 2012). Generally, for HMM models, two sets of parameters are considered: one is the emission probabilities and the other is the transition probabilities. In this case, the number of the hidden states is fixed. For each state, the emission parameter represents the probability that each input mark has a present call, while the transition parameter indicates the probability for this state transitioning to another chromatin state. They obtained 51 different chromatin states, such as promoter states, transcribed states, active intergenic states, repressive states and repetitive states, and so on.

Similar to Ernst and Kellis's model, Choi et al. proposed a sparsely correlated Hidden Markov Model, in which the transition probabilities depend on not only its own hidden states, but also the other related genomic regions' hidden states (Choi et al. 2013). Usually for each series of two hidden states, one can formulate N separate HMMs, if it is independent between each series. But if considering the correlation between series, a single HMM with at most 2^N hidden states can be formulated.

While this scHMM algorithm just lies between those two extreme cases, it has considered both the computational efficiency and the correlations between each series.

In 2013, Yu et al. developed an algorithm called GATE (Genomic Annotation using Temporal Epigenomic data) based on Ernst and Kellis's results (Yu et al. 2013). This model is composed of two layers: the top layer is a Finite Mixture model, which clusters genomic segments with the same epigenomic patterns; the bottom layer is based on one Hidden Markov Model representing the temporal change within each cluster. The main difference between this GATE model and the previous results is that it can not only directly annotate the epigenetic states whole genome-wide but also reflect the dynamic changes of the states across different experimental conditions.

Besides annotating the genomic regions with the combination of different epigenomic markers, many other algorithms have emerged to infer the relationships among the epigenomic markers in specific functional genomic regions, such as promoters and enhancers. To this end, we first applied Bayesian network (BN) to infer regulatory interactions among histone modifications and to de novo identify the potential causal relationships (Yu et al. 2008).

We have also developed SeqSpider, a new Bayesian network structure learning algorithm to infer regulatory or interactions between a set of biological factors using heterogeneous epigenomic data of different types (Liu et al. 2013). It can accept continuous data as well as vectored data, such as tag distribution from high-throughput sequencing data. It uses a profile-based clustering strategy for noise reduction to predict the interactions from different high-throughput epigenomic data with high accuracy and stability. One of its big advantages is that it can easily integrate heterogeneous data types, such as ChIP-seq data, BS-seq data, RNA-seq data and so on, as well as integrate data from different labs or from different batches. The regulatory networks can be inferred from various biological contexts.

Lasserre et al. proposed a sparse partial correlation network (SPCN) to infer undirected networks based on partial correlations between histone modifications (Lasserre et al. 2013). This partial correlation network focuses on direct associations of histone modifications. The algorithm is based on graphical Gaussian model (GGM), which contains the original edges and will connect the parents of a same child. Usually partial networks require normal distributions of the data. SPCNs overcome this by rank-transforming the input data. It achieved sparseness by a cross-validation scheme. Direct associations, mutual exclusivities, direct edges in a pathway and indirect edges can be revealed by SPCNs (Table 6.1).

Table 6.1 Machine learning–based methods

Methods	Tools	Citation
Chromatin states annotation	ChromHMM	Ernst and Kellis (2010, 2012)
	SCHMM	Choi et al. (2013)
	GATE	Yu et al. (2013)
Infer causal relationship	SeqSpider	Liu et al. (2013)
	SPCNs	Lasserre et al. (2013)

6.4 Application of Clustering Analyses in Integrative Cancer Epigenomics

The epigenetic state of cancer cells is profoundly altered. Human tumours undergo an overall hypomethylation but with specific hypermethylation on certain regions such as promoter of tumour-suppressor genes, which is associated with transcriptional silencing and also recognized as a key feature of cancer (Egger et al. 2004; Esteller 2005; Feinberg and Tycko 2004; Herman and Baylin 2003). In addition, these DNA methylation alternations are linked with aberrant pattern of histone modification (Ballestar et al. 2003; Fahrner et al. 2002; Fraga et al. 2005; Nguyen et al. 2001; Pruitt et al. 2006). However, the characteristic of epigenetic alternations in cancer cells are still not fully understood (Esteller 2007).

In general, cancers are classified based on pathological criteria, which rely heavily on tissue of origin. Now more and more large-scale genomic data characterizing molecular details of tumours are available. Integration of genomics and epigenomics of cancer together with relevant clinical information gathered to make a molecular-based taxonomy of cancer is possible (Hoadley et al. 2014). Such taxonomy can also give us insight into cancer prevention, diagnostics, therapeutic strategies through key genetic alternation finding, driver mutation discovery, somatic mutational signature identification, clonal evolution characteristic and epigenetic alternation (Mwenifumbo and Marra 2013).

Clustering methods can provide intuitive ways in partitioning a large data set into more easily digestible, conceptual models (Hawkins et al. 2010), and thus are widely used in integrative analyses.

6.4.1 Classical Clustering Methods

Commonly used clustering methods include hierarchical clustering and k-means with its variants. However, when performing integrative analysis using clustering methods, different types of the data sets, e.g. discrete (mutation state: mutated or not) versus continuous (gene expression, DNA methylation), often require different clustering methods. Normalization procedures before clustering and the similarity metrics should also be tailored for different data types.

Heintzman et al. integrated five histone modifications, four general transcription factors and nucleosome density at high resolution in 30 Mb of human genome to define chromatin features on enhancer and promoter regions using the k-means clustering algorithm (Heintzman et al. 2007). Lister et al., in their study of the first human methylomes, visualized DNA methylation, histone modifications and RNA-seq patterns simultaneously using hierarchical clustering (Lister et al. 2009). Shen et al. combined genetic and epigenetic alternations in 97 primary colorectal cancer patients using two kinds of clustering algorithms. First, they

Table 6.2 Clustering methods applied to integrative cancer epigenomics

Classical clustering methods	Hierarchical clustering	Lister et al. (2009)
		Shen et al. (2007)
	k-means	Heintzman et al. (2007)
		Shen et al. (2007)
k-means variants	iCluster	Shen et al. (2009)
	iCluster+	Mo et al. (2013)
		Cancer Genome Atlas Research (2012)
		Bass et al. (2014)
		Cancer Genome Atlas Research (2014)
	Adaptive super k-means clustering	Zhang et al. (2013)
	Cluster Of Cluster Assignments (COCA)	Cancer Genome Atlas (2012)
		Hoadley et al. (2014)
	SuperCluster	Cancer Genome Atlas Research et al. (2013)
		Bass et al. (2014)
		Hoadley et al. (2014)

clustered DNA methylation data using hierarchical clustering to identify three clusters corresponding to distinct genetic alternation profiles, then used k-mean clustering with k equal to three to combine the epigenetic and genetic data. These two clustering results are highly consistent (Shen et al. 2007) (Table 6.2).

6.4.2 K-Means Variants

A sequence of k-mean clustering variants emerged for large-scale data set integration especially for cancer subtyping. Shen et al. developed a jointly latent variable model called iCluster to integrate multiple data set, with the assumption that diverse molecular phenotypes can be predicted by a set of orthogonal latent variables that represent distinct molecular drivers, such as tumour subgroups with biological and clinical importance. In this model, a penalized likelihood approach with lasso penalty terms to balance the fitness and the complexity is introduced, which can be used to identify genomic features contributing the most to the biological variation and directly related to characterizing molecular subtypes. iCluster performs better in subtyping breast cancer and lung cancer compared with sample-wise hierarchical clustering (Shen et al. 2009). However, iCluster cannot handle discrete data type such as mutation state; in 2013, an enhancement of iCluster named iCluster+ was developed, which can perform pattern discovery that integrates both discrete (binary and category) and continuous variables data types by formulating a joint generalized linear regression model (Mo et al. 2013). TCGA group used iCluster+ to cluster mRNA, miRNA, methylation SCNA and mutation data of 178 lung SQCC squamous cell lung cancers (SQCC) into three distinct molecular subgroups and identified potential targetable genes or pathway alternations (Cancer Genome Atlas Research 2012). The group also clustered 295 primary gastric adenocarcinoma (Bass et al. 2014) and 230 lung adenocarcinomas (Cancer Genome Atlas Research 2014) using iCluster+ by integrating genetic and epigenetic data sets.

Different from iCluster or iCluster+, which use a joint latent variable to model the subtypes of a cancer, Zhang et al. developed an unbiased, adaptive k-means clustering approach, which can automatically determine the optimal number of clusters by maximizing a Bayesian Information Criterion (BIC) score, while introducing a free parameter lambda for tuning the penalty term BIC to achieve the flexibility in handling data with different level of noise. After optimizing k, super k-means is employed to generate the best clustering results. By applying adaptive k-mean clustering approach to integrate gene expression, DNA methylation, microRNA expression and copy number alternation profile of ovarian cancers, Zhang et al. identified seven previously unrecognized subtypes that are associated with significantly different median survival times (Zhang et al. 2013).

Both iCluster+ and adaptive super k-means clustering methods can integrate multiple data types or data sets simultaneously. Different from these two kinds of clustering methods, an integrative method called Cluster Of Cluster Assignments (COCA), developed by TCGA group, is performed in two steps: in the first step, sample-wise clusters based on each platform is performed (such as hierarchical clustering and classical k-means clustering); in the second step, subtypes defined by each platform were coded into a series of indicator variables for each subtype, then matrix 0 and 1 s representing whether samples belongs to certain subtype are clustered using ConsensusClusterPlus R package to identify modules of samples or subtypes of cancers; in other words, the input for clustering in the second step is a m*n matrix, m represents subtypes called in each data type, n represents samples. The advantage of COCA is that data across platforms are combined without the need for normalization steps prior to clustering. TCGA group integratively clustered human breast tumours on five platforms (DNA-copy number arrays, DNA methylation, exome sequencing, messenger RNA arrays, microRNA sequencing and reverse-phase protein arrays) using COCA with four subtypes defined, which are consistent with four clusters defined by gene expression previously (Cancer Genome Atlas 2012). Hoadley et al. (2014) applied this method on five genome-wide platforms and one proteomic (whole-exome DNA sequence, DNA-copy number variance, DNA methylation, genome-wide mRNA level, microRNA levels, protein levels for 131 protein and/or phosphorylated proteins) on 3,527 specimens from 12 cancer types from TCGA. They also used SuperCluster and pathway level–based clustering method PARADIGM to compare with COCA; all of the clusters derived are highly concordant.

Similar to COCA, SuperCluster is another method put forward by TCGA group. SuperCluster is also performed in two steps: in the first step, sample-wise clusters based on each platform is performed; in the second step, sample subtypes from each kind of data type are used as input for clustering again with different kinds of data type treated differently: Mutations and CNV clusters were treated as ordinal variables, others nominal. Here, the input matrix $m*n$ in the second step is different from COCA, m is data type, n is sample. In addition, SuperCluster adjusted the contribution from each data type to make their weight equal (Cancer Genome Atlas Research et al. 2013). TCGA group performed integrative clustering on gastric adenocarcinoma using SuperCluster with sample-wise unsupervised clustering followed by censuses clustering (Bass et al. 2014). Later, Hoadley et al. performed integrative analysis on 12 cancer cell types using SuperCluster, COCA and PRARDIEM; all of them are highly concordant (Hoadley et al. 2014).

6.4.3 How to Evaluate the Clustering Results

An important statistical issue for clustering is whether a cluster is real or an artefact of sampling variation. Liu et al. developed SigClust to quantify the significance of a given clustering result assuming that data within a cluster comes from a single Gaussian distribution (Liu et al. 2008). Another method called Consensus clustering (CC) (Monti et al. 2003) provides statistically stable evidence derived from repeated sampling. Later, Wilkerson et al. implemented CC in Consensus cluster plus with some extension (Wilkerson and Hayes 2010).

6.4.4 Summary

In summary, cancer subtyping in a comprehensive genomic context by epigenetic integrative analysis can better reveal molecular drivers or key genetic alternations to help understand diagnostic, prognostic and therapeutic strategies that are specific to individual patients or group of patients sharing common genetic and epigenetic features compared with clinically defined features. Moreover, molecular taxonomy defined by genetic and epigenetic integration can give more insights into subtypes, convergence of different cancer types which is different from histological classification (tissue-of-origin) and explain clinical outcomes of cancer types from same histological class.

6.5 Future Direction

We have gone through the major methodologies applied in integrative epigenomics, which are simple integration, machine learning and clustering-based methods. Such methodologies were proven to be efficient in functional genomic element annotation, chromatin segmentation, cancer sub-classification and regulatory network inference. Although the techniques are diverse, there is a principle in common, that is, how to transform such heterogeneous data into comparable status. In simple integration, one omic data is first reduced to features (e.g. peaks enriched for signals). Then we can intersect these features with those reduced from other data, or simply profile signals of other data located in these features. In clustering analysis, data normalization must be done first via data transformation to adjust distribution or via proper data discretization. In machine learning–based methods, data sets are transformed internally to certain intermediates, which are hidden states in Hidden Markov Model and discretized conditions for conditional probability calculation in Bayesian network.

However, the methods mentioned above all focus on singular genomic features such as CpG sites, genes, promoters and enhancers. The facts that all such elements function in pathways and that genetic and epigenetic alternations in cancers always converge to common pathways make pathway another layer for data integration. Tools such as Gene Ontology enrichment analysis, Gene Set Enrichment Analysis (GSEA) (Subramanian et al. 2005), Signaling Pathway Impact Analysis (SPIA) (Tarca et al. 2009) and Genomic Regions Enrichment of Annotations Tool (The Cancer Genome Atlas Research et al. 2014; McLean et al. 2010) can be used to reduce the complex data to enriched pathways for further integrative analyses. PARADIGM is a method for inferring patient-specific pathway activities from multi-dimensional cancer genomics data (Vaske et al. 2010). It first converts gene copy number and expression-level changes into pathway level with a probabilistic inference to predict the degree of perturbation of pathway activities. TCGA group have applied PARADIGM to integrate genomic and transcriptomic changes in glioblastoma multiform (GBM) (Vaske et al. 2010), breast cancer (Cancer Genome Atlas 2012) and so on. However, PARADIGM does not support any epigenomic data yet. The incorporation of epigenome data in pathway-level integration similar to PARADIGM requires further understanding of epigenomic control on pathway activities, and will be a future direction of integrative epigenetics.

References

Ballestar E, Paz MF, Valle L, Wei S, Fraga MF, Espada J, Cigudosa JC, Huang TH, Esteller M. Methyl-CpG binding proteins identify novel sites of epigenetic inactivation in human cancer. EMBO J. 2003;22:6335–45.

Bass AJ, Thorsson V, Shmulevich I, Reynolds SM, Miller M, Bernard B, Hinoue T, Laird PW, Curtis C, Shen H, et al. Comprehensive molecular characterization of gastric adenocarcinoma. Nature. 2014;513:202–9.

Boyle AP, Song L, Lee BK, London D, Keefe D, Birney E, Iyer VR, Crawford GE, Furey TS. High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. Genome Res. 2011;21:456–64.
Cancer Genome Atlas N. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490:61–70.
Cancer Genome Atlas Research N. Comprehensive genomic characterization of squamous cell lung cancers. Nature. 2012;489:519–25.
Cancer Genome Atlas Research N. Comprehensive molecular profiling of lung adenocarcinoma. Nature. 2014;511:543–50.
Cancer Genome Atlas Research N, Kandoth C, Schultz N, Cherniack AD, Akbani R, Liu Y, Shen H, Robertson AG, Pashtan I, Shen R, et al. Integrated genomic characterization of endometrial carcinoma. Nature. 2013;497:67–73.
Chelaru F, Smith L, Goldstein N, Bravo HC. Epiviz: interactive visual analytics for functional genomics data. Nat Methods. 2014;11:938–40.
Choi H, Nesvizhskii AI, Ghosh D, Qin ZS. Hierarchical hidden Markov model with application to joint analysis of ChIP-chip and ChIP-seq data. Bioinformatics. 2009;25:1715–21.
Choi H, Fermin D, Nesvizhskii AI, Ghosh D, Qin ZS. Sparsely correlated hidden Markov models with application to genome-wide location studies. Bioinformatics. 2013;29:533–41.
Churchill GA. Stochastic models for heterogeneous DNA sequences. Bull Math Biol. 1989;51:79–94.
Egger G, Liang G, Aparicio A, Jones PA. Epigenetics in human disease and prospects for epigenetic therapy. Nature. 2004;429:457–63.
Ernst J, Kellis M. Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat Biotechnol. 2010;28:817–25.
Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. Nat Methods. 2012;9:215–6.
Esteller M. Aberrant DNA methylation as a cancer-inducing mechanism. Annu Rev Pharmacol Toxicol. 2005;45:629–56.
Esteller M. Cancer epigenomics: DNA methylomes and histone-modification maps. Nat Rev Genet. 2007;8:286–98.
Fahrner JA, Eguchi S, Herman JG, Baylin SB. Dependence of histone modifications and gene expression on DNA hypermethylation in cancer. Cancer Res. 2002;62:7213–8.
Feinberg AP, Tycko B. The history of cancer epigenetics. Nat Rev Cancer. 2004;4:143–53.
Fraga MF, Ballestar E, Villar-Garea A, Boix-Chornet M, Espada J, Schotta G, Bonaldi T, Haydon C, Ropero S, Petrie K, et al. Loss of acetylation at Lys16 and trimethylation at Lys20 of histone H4 is a common hallmark of human cancer. Nat Genet. 2005;37:391–400.
Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, Yip KY, Robilotto R, Rechtsteiner A, Ikegami K, et al. Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. Science. 2010;330:1775–87.
Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP, et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature. 2009;458:223–7.
Hawkins RD, Hon GC, Ren B. Next-generation genomics: an integrative approach. Nat Rev Genet. 2010;11(7):476–86.
Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet. 2007;39:311–8.
Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, Ye Z, Lee LK, Stuart RK, Ching CW, et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature. 2009;459:108–12.
Herman JG, Baylin SB. Gene silencing in cancer in association with promoter hypermethylation. N Engl J Med. 2003;349:2042–54.

Hesselberth JR, Chen X, Zhang Z, Sabo PJ, Sandstrom R, Reynolds AP, Thurman RE, Neph S, Kuehn MS, Noble WS, et al. Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. Nat Methods. 2009;6:283–9.

Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S, Leiserson MD, Niu B, McLellan MD, Uzunangelov V, et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell. 2014;158:929–44.

Hon GC, Rajagopal N, Shen Y, McCleary DF, Yue F, Dang MD, Ren B. Epigenetic memory at embryonic enhancers identified in DNA methylation maps from adult mouse tissues. Nat Genet. 2013;45:1198–U1340.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. Genome Res. 2002;12:996–1006.

Kieffer-Kwon KR, Tang Z, Mathe E, Qian J, Sung MH, Li G, Resch W, Baek S, Pruett N, Grontved L, et al. Interactome maps of mouse gene regulatory domains reveal basic principles of transcriptional regulation. Cell. 2013;155:1507–20.

Krogh A, Brown M, Mian IS, Sjolander K, Haussler D. Hidden Markov models in computational biology. Applications to protein modeling. J Mol Biol. 1994;235:1501–31.

Lasserre J, Chung HR, Vingron M. Finding associations among histone modifications using sparse partial correlation networks. PLoS Comput Biol. 2013;9:e1003168.

Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature. 2009;462:315–22.

Liu Y, Hayes DN, Nobel A, Marron JS. Statistical significance of clustering for high-dimension, Low-sample size data. J Am Stat Assoc. 2008;103:1281–93.

Liu Y, Han D, Han Y, Yan Z, Xie B, Li J, Qiao N, Hu H, Khaitovich P, Gao Y, et al. Ab initio identification of transcription start sites in the Rhesus macaque genome by histone modification and RNA-Seq. Nucleic Acids Res. 2011;39:1408–18.

Liu Y, Qiao N, Zhu S, Su M, Sun N, Boyd-Kirkup J, Han JD. A novel Bayesian network inference algorithm for integrative analysis of heterogeneous deep sequencing data. Cell Res. 2013;23:440–3.

McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G. GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol. 2010;28:495–501.

Mo Q, Wang S, Seshan VE, Olshen AB, Schultz N, Sander C, Powers RS, Ladanyi M, Shen R. Pattern discovery and cancer gene identification in integrated cancer genomic data. Pro Natl Acad Sci U S A. 2013;110:4245–50.

Monti S, Tamayo P, Mesirov J, Golub T. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. Mach Learn. 2003;52:91–118.

Mwenifumbo JC, Marra MA. Cancer genome-sequencing study design. Nat Rev Genet. 2013;14:321–32.

Nguyen CT, Gonzales FA, Jones PA. Altered chromatin structure associated with methylation-induced gene silencing in cancer cells: correlation of accessibility, methylation, MeCP2 binding and acetylation. Nucleic Acids Res. 2001;29:4598–606.

Pasquali L, Gaulton KJ, Rodriguez-Segui SA, Mularoni L, Miguel-Escalada I, Akerman I, Tena JJ, Moran I, Gomez-Marin C, van de Bunt M, et al. Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. Nat Genet. 2014;46:136–43.

Pruitt K, Zinn RL, Ohm JE, McGarvey KM, Kang SH, Watkins DN, Herman JG, Baylin SB. Inhibition of SIRT1 reactivates silenced cancer genes without loss of promoter DNA hypermethylation. PLoS Genet. 2006;2:e40.

Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. Integrative genomics viewer. Nat Biotechnol. 2011;29:24–6.

Roy S, Ernst J, Kharchenko PV, Kheradpour P, Negre N, Eaton ML, Landolin JM, Bristow CA, Ma L, Lin MF, et al. Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science. 2010;330:1787–97.

Shen L, Toyota M, Kondo Y, Lin E, Zhang L, Guo Y, Hernandez NS, Chen X, Ahmed S, Konishi K, et al. Integrated genetic and epigenetic analysis identifies three different subclasses of colon cancer. Pro Natl Acad Sci U S A. 2007;104:18654–59.

Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics. 2009;25:2906–12.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102:15545–50.

Tarca AL, Draghici S, Khatri P, Hassan SS, Mittal P, Kim JS, Kim CJ, Kusanovic JP, Romero R. A novel signaling pathway impact analysis. Bioinformatics. 2009;25:75–82.

The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74.

Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 2013;14:178–92.

Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, Zhu J, Haussler D, Stuart JM. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics. 2010;26:i237–45.

Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics. 2010;26:1572–3.

Xie M, Hong C, Zhang B, Lowdon RF, Xing X, Li D, Zhou X, Lee HJ, Maire CL, Ligon KL, et al. DNA hypomethylation within specific transposable element families associates with tissue-specific enhancer landscape. Nat Genet. 2013;45:836–41.

Yu H, Zhu S, Zhou B, Xue H, Han JD. Inferring causal relationships among different histone modifications and gene expression. Genome Res. 2008;18:1314–24.

Yu P, Xiao S, Xin X, Song CX, Huang W, McDee D, Tanaka T, Wang T, He C, Zhong S. Spatiotemporal clustering of the epigenome reveals rules of dynamic gene regulation. Genome Res. 2013;23:352–64.

Zhang W, Liu Y, Sun N, Wang D, Boyd-Kirkup J, Dou X, Han JD. Integrating genomic, epigenomic, and transcriptomic features reveals modular signatures underlying poor prognosis in ovarian cancer. Cell Rep. 2013;4:542–53.

Chapter 7
Towards a Mechanistic Understanding of Epigenetic Dynamics

Jens Przybilla, Thimo Rohlf, and Joerg Galle

Abstract The stem cell epigenome reflects a sensitive balance of chromatin (de-)modification processes. Here, we review our recent achievements towards a mechanistic understanding of this balance.

We introduce a computational model of stem cell populations, where each cell contains an artificial genome. Transcription of the genes encoded by this genome is controlled by DNA methylation, histone modification and the action of a cis-regulatory network. Model dynamics are determined by molecular crosstalk between these different mechanisms.

The epigenetic states of the genes are subject to different types of fluctuations. We demonstrate that the timescales of these fluctuations control whether the state associated with a particular gene will undergo drifts during ongoing cell replication. In particular, our model suggests that changes in DNA methylation states are determined by histone modification dynamics. Herewith, our model provides a mechanistic understanding of the origin of tissue, age and cancer-specific DNA methylation profiles.

Keywords DNA methylation • Histone modification • Cis-regulatory networks • Mathematical modelling

---

J. Przybilla • J. Galle (✉)
Interdisciplinary Center for Bioinformatics, University Leipzig, Haertelstr. 16-18, Leipzig 04107, Germany
e-mail: przybilla@izbi.uni-leipzig.de; galle@izbi.uni-leipzig.de

T. Rohlf
Interdisciplinary Center for Bioinformatics, University Leipzig, Haertelstr. 16-18, Leipzig 04107, Germany

Max-Planck-Institute for Mathematics in the Sciences, Inselstr. 22, Leipzig 04103, Germany
e-mail: rohlf@izbi.uni-leipzig.de

A.E. Teschendorff (eds.), *Computational and Statistical Epigenomics*,
Translational Bioinformatics 7, DOI 10.1007/978-94-017-9927-0_7

7.1 Introduction

Among the plethora of known chromatin modifications, DNA methylation is probably the one that has been analyzed in most detail. Changes in DNA methylation are observed during stem cell differentiation and ageing and also in the course of many diseases (Bergman and Cedar 2013). A particular well-documented phenomenon is hyper-methylation of CpG-rich promoters during cancer development (Berdasco and Esteller 2010). This local increase of CpG methylation is often associated with a down-regulation of expression of the affected genes and thus can induce cancer phenotypes. In fact, DNA methylation patterns have been used in several tissues to classify cancer subtypes with different clinical outcomes (Hinoue et al. 2012; Sturm et al. 2012) (see Fig. 7.1).

Regardless of the enormous amount of molecular data collected so far, a mechanistic understanding of how the observed changes in DNA methylation are induced and how they impact gene expression is still largely missing. Within the last years, many experimental groups observed correlations between DNA methylation and other chromatin marks. One of the first examples here was an observation made for a certain group of genes whose promoters become hyper-methylated during ageing and cancer development. It was found that the nucleosomes associated with the promoter of these genes are often tri-methylated at lysine 4 and 27 of histone

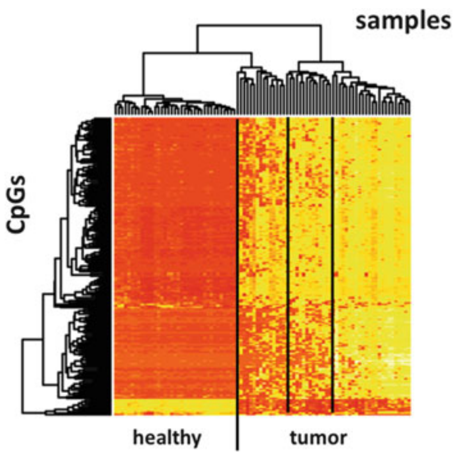


Fig. 7.1 Hierarchical clustering of DNA methylation pattern of human colorectal cancer (CRC) samples. Shown are methylation levels (*red*: low, *white*: high) that have been calculated based on 27k Illumina methylation arrays from 37 healthy and 53 tumour samples (*columns*). They have been clustered using the 200 most variant CpGs (*rows*). Data were taken from the TCGA data repository (Network 2012). Most of the CpGs become hyper-methylated in CRC, while only a few become hypo-methylated. The patterns allow to distinguish two or three different CRC methylation patterns. The origin of this epigenetic reorganization remains largely unknown

3 (Teschendorff et al. 2010; Rakyan et al. 2010). Such findings suggest that there exists a complex molecular crosstalk between the machinery of DNA methylation and that of histone modification. Some specific molecular interactions have been identified in the meanwhile (Rose and Klose 2014).

Here, we will demonstrate that mathematical modelling of this kind of molecular crosstalk can provide new insights into regulatory principles of the epigenome and can help to establish a mechanistic understanding of epigenetic reorganization, e.g. following loss of tissue homeostasis. For this purpose, in the following, we provide a brief introduction into a multi-scale model of epigenetic regulation of transcription. First, we introduce its molecular components enabling to describe DNA methylation, histone modification and cis-regulatory networks. Afterwards, we explain its extension to the cell population level. Finally, we provide some first simulation results on epigenetic changes during stem cell ageing and tissue transformation.

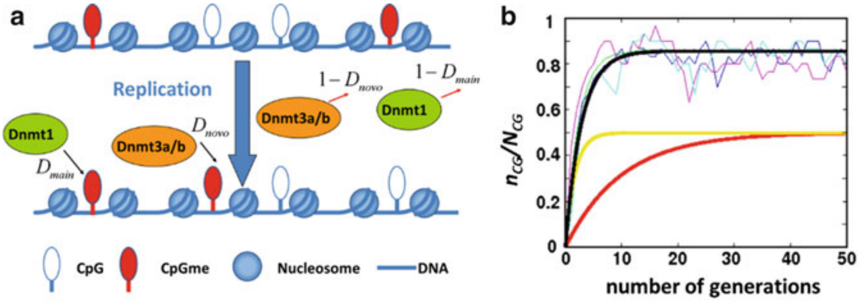
7.2 Modelling DNA Methylation Dynamics

7.2.1 Background

The role of DNA methylation (here: 5-mC methylation) in cancer development has been recognized already more than 40 years ago (Magee 1971); yet, the molecular details of the enzymatic machinery leading to establishment and maintenance of DNA methylation in normal tissue remained unknown for a long time. Despite missing knowledge about the enzymes involved, first mathematical models for DNA-methylation dynamics were proposed already in the 1990s (Otto and Walbot 1990; Pfeifer et al. 1990).

These models tried to explain the conservation of methylation states given the fact that all CpGs on the de novo synthesized DNA daughter strands are initially unmethylated. They proposed that this passive de-methylation is compensated by simultaneous action of maintenance and de novo methylation, leading to a genome-wide methylation equilibrium after a finite number of replication cycles. DNA methyltransferases (Dnmts) involved in these processes were identified experimentally some years later. Three types of Dnmts were identified in mammals, namely, Dnmt3a, Dnmt3b and Dnmt1. De novo methylation has been attributed to the action of the isoforms Dnmt3a and Dnmt3b (Okano et al. 1999), while Dnmt1 was found to be mainly responsible for maintaining the parental methylation pattern in daughter cells (Pradhan et al. 1999).

Improved models of DNA methylation, which considered these experimental findings, were introduced by Sontag et al. (2006). They introduced a linear model for independent action of Dnmt1 and Dnmt3a/b and a nonlinear model assuming cooperative dynamics between them.



c

Stochastic system:

(S1) CpGm ⟶ CpGm
with probability: D_{main}

(S2) CpGm ⟶ CpG
with probability: $1 - D_{main}$

(S3) CpG ⟶ CpGm
with probability: D_{novo}

(S4) CpG ⟶ CpG
with probability: $1 - D_{novo}$

Deterministic system:

(D1) $$\frac{dm}{dt} = (D_{main} - 1)m + D_{novo}(1-m)$$

(D2) $$m = \left(m_0 - D_{novo}/r\right)\exp(-rt) + D_{novo}/r$$
$$r = 1 - D_{main} + D_{novo}$$

Fig. 7.2 Model of DNA methylation. (**a**) Sketch of the processes considered in our model. Maintenance methylation by Dnmt1 and de novo methylation by Dnmt3a,b are either active with rate D_{main} and D_{novo}, or inactive with rate *1-*D_{main} and *1-*D_{novo}, respectively. (**b**) Different pairs (D_{novo}, D_{main}) can result in the same methylation level as seen for solutions of the deterministic system applying ($0.05/\tau$, $0.95/\tau$), red curve, and ($0.3/\tau$, $0.7/\tau$), yellow curve. Stochastic simulations for 30 CpGs show large fluctuations (*magenta, cyan, blue lines*) around the solution of the deterministic system (*black line*). (**c**) Transition probabilities for the stochastic system and the basic set of equations for the deterministic system

7.2.2 Basic Model of DNA Methylation

Since experimental support for cooperation between Dnmts is still limited, we designed a first layer of our multi-scale model adopting a simple version of the linear model by Sontag et al. (see Fig. 7.2a). We assume a single-stranded DNA molecule. Accordingly, all CpG sites of a finite DNA region (N_{CG} CpG sites) are either methylated or not. CpG methylations existing in the mother cell (n_{CG}) are restored by maintenance methylation with probability D_{main} in their daughters, while CpG sites that are unmethylated in the mother ($N_{\text{CG}} - n_{\text{CG}}$) might become de novo methylated in the daughters with probability D_{novo}. In our model, de novo methylation acts only on CpG sites that were unmethylated in the mother and not on those where methylation has not been restored in the daughters.

We assume that DNA methylation can occur only immediately after cell replication, in a short time frame compared to the cell cycle time τ. This effectively makes

DNA methylation levels a direct function of the number of successive replication events. Therefore, following Sontag et al., we implemented a probabilistic model with discrete time steps and discrete updates of CpG-methylation states. Formally, this defines a discrete Markov chain model with transition probabilities D_{novo} and D_{main}. The transition scheme for our model is shown in Fig. 7.2c (left part). The thin-lined curves in Fig. 7.2b show the results of three different simulations of the stochastic model. Shown is the fraction m (given by $n_{\mathrm{CG}}/N_{\mathrm{CG}}$) of methylated CpGs that has been observed by analyzing the methylation dynamics of 30 CpGs.

Alternatively, changes of m during ongoing replication can be analyzed using a differential equation approach (see Fig. 7.2c, right part). Such a time continuous approach is helpful to estimate methylation equilibria and convergence times. The solution of this equation yields an exponential increase or decrease to an equilibrium methylation level, depending on the initial methylation level m_0. The equilibrium methylation is given by D_{novo}/r. Examples of the temporal dynamics are shown in Fig. 7.2b. Note that D_{novo} and D_{main} represent modification rates per cell cycle time τ. Different combinations of them can lead to convergence to the same methylation level; yet, the time needed for convergence differs.

Our stochastic DNA methylation model is limited in some regards. For instance, the model alone cannot explain the coexistence of hyper- and hypo-methylated states after a large number of replications, as observed in aged tissues and during cancer development (Bergman and Cedar 2013). In their nonlinear model, Sontag et al. explained these phenomena suggesting that the efficiency of de novo methylation depends on the density of hemi-methylated sites observed after DNA replication but before maintenance and/or de novo methylation. A similar model was recently proposed by Haerter et al. (2014). Here, we use a different approach.

It is well known that recruitment of de novo Dnmts strongly depends on local histone modification states (Rose and Klose 2014). This suggests that coexisting DNA methylation states are controlled by local histone modification states. In the following, we introduce a model of cooperative histone modification dynamics which enables us to describe such regulation.

7.3 Modelling Histone Modification Dynamics

7.3.1 Background

Today, a huge number of chemical modifications on histone tails are known, including methylation, acetylation, phosphorylation, sumoylation and ubiquitination. These modifications can contribute to activation or repression of gene expression. Their combinatorial complexity is further increased by the possibility of different modification levels, e.g. mono-, di- or tri-methylation (me1, me2 or me3). Thereby, different levels of modification might have different effects on chromatin structure and transcription (Hoffman et al. 2013).

Several theoretical models of histone methylation and acetylation dynamics have been proposed so far (Dodd et al. 2007; Sedighi and Sengupta 2007). For a review of these models, see Rohlf et al. (2012). A common feature of these models is that they are based on cooperative modification dynamics. In our model of histone methylation, we enable cooperative behaviour by assuming that modified nucleosomes cooperatively recruit their own histone methyltransferase (HMT).

In the following, we first introduce a model of histone methylation for a finite number of cooperatively acting nucleosomes (Binder et al. 2013). Afterwards, we outline our strategies to integrate crosstalk between histone and DNA methylation into the model.

7.3.2 Basic Model of Histone Methylation

In our model, we only consider modification complexes that can write and read a specific modification. This is motivated by properties of polycomb group (PcG) and trithorax group (trxG) proteins (Kundu and Peterson 2009). The basic assumptions of our model regarding a reader–writer complex catalyzing histone modifications, in the following called 'interaction complex' (IC), are summarized in Fig. 7.3a. The regulatory processes are explained for an activating modification.

We assume that each IC binds to one DNA-response element (RE) which contains a variable number n_{BS} of binding sites (BS). Specifically, we identify BS with CpGs. Binding to CpGs depends on their methylation state. Adjacent REs form cooperative units (CUs) of length L_{CU}, given in units of the number of base pairs (bp) involved. Formation of CUs might occur via chromatin looping as proposed by Tiwari et al. (2008). Each CU is associated with $N_{\mathrm{H}} = L_{\mathrm{CU}}/200$ nucleosomes, where n_{HM} of them are in a modified (HM) histone state and the remaining $N_H - n_{HM}$ are in an unmodified (*H0*) histone state. We call a nucleosome modified if one of its histones is modified. In addition to the DNA BS, also the n_{HM}-modified nucleosomes within a CU facilitate IC binding. Bound ICs catalyze histone modifications, giving rise to a positive feedback loop between IC binding and histone modification. Gene transcription is activated after IC binding and repressed after IC release.

In our multi-scale model, we implemented a stochastic version of this model at the single nucleosome level. The transition probabilities for the nucleosome states per simulation time step Δt are given in Fig. 7.3c (left part). We assume that de novo modification of a histone can occur only if an IC is bound to a nearby RE. The probability of this state is quantified by the RE-occupancy Θ, which can take values between 0 and 1. Accordingly, the probability of de novo modification is given by $k_{\mathrm{M}}\ \Theta$, where k_{M} is a constant. De-modification events are assumed to occur permanently with probability k_{D}.

Using arguments from mass action kinetics, the binding process of an IC can be formalized (Binder et al. 2013). The resulting equation for the RE-occupancy Θ is given in Fig. 7.3c (Eq. D1). It is governed by the free enthalpy change Δg

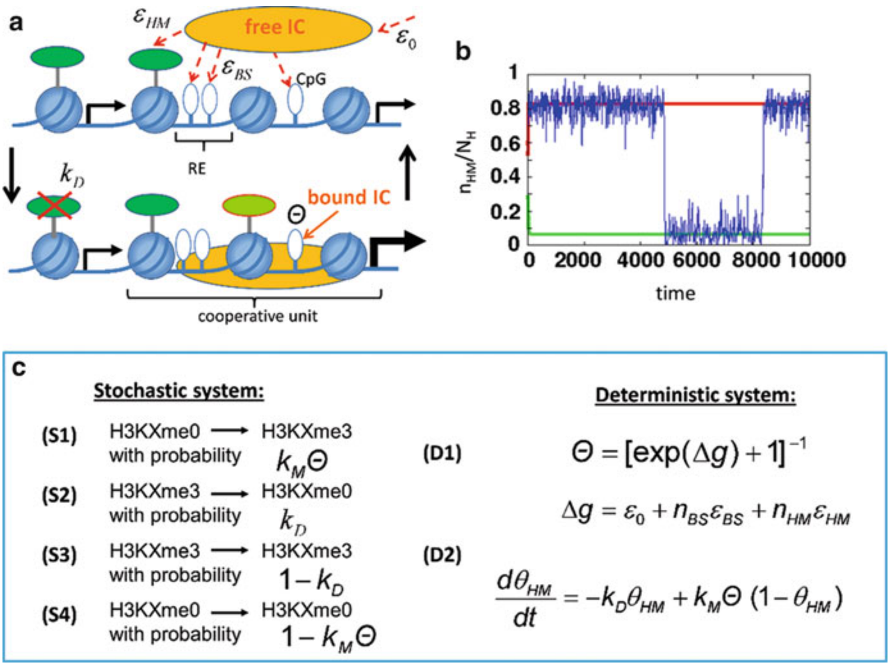


Fig. 7.3 Model of histone methylation. (**a**) Sketch of the regulatory interactions between ICs and chromatin. For an explanation, see text. (**b**) Fluctuations of the methylation level within a CU containing 48 nucleosomes, as derived from a stochastic model realization (*blue curve*). Appreciate the sudden switch from high to low modification level after about 5,000 simulation time steps and back about 3,500 steps later. The two stable solutions of the deterministic system are shown for comparison (*red and green curves*). (**c**) Basic equations of the model. In S1–S4, the X labels the lysine, e.g. 4 or 9

of IC binding, which can be decomposed into a basic repulsive term $\varepsilon_0 > 0$, and two attractive terms $n_{BS}\varepsilon_{BS} < 0$ and $n_{HM}\varepsilon_{HM} < 0$ representing the enthalpy changes according to binding of the IC to n_{BS} DNA-binding sites and to n_{HM} nucleosomes of the CU which already carry the IC-specific modification, respectively.

In Fig. 7.3b, simulation results are given for a realization of the stochastic modification process within a single CU of 48 nucleosomes. For the chosen parameter set, the system shows bistable behaviour. Stochastic fluctuations lead to switches between the two attractor states.

The temporal dynamics of the fraction θ_{HM} (given by n_{HM}/N_H) of modified histones in a CU can be described by a differential equation (Eq. D2 in Fig. 7.3c), similar to the DNA methylation process. Here, the terms k_D, $k_M\Theta$ describe rates per simulation time step Δt. Again this approach allows estimating equilibrium states of the system. It can be shown that, for relatively wide parameter ranges, the system exhibits bistable behaviour (Rohlf et al. 2012; Binder et al. 2013). Thereby, the solutions strongly depend on the number N_H of nucleosomes contained in the CU. In Fig. 7.3b, the solutions of the deterministic system (red and green lines) are compared with those of the stochastic process.

The above model can be extended easily to combinations of histone modifications. For this purpose, one has to take into account sets of transition probabilities and model parameters for every modification. Thereby, different modifications might influence each other, either directly or indirectly. In the following, we study tri-methylation of lysine 4 at histone 3 (H3K4me3) in parallel with tri-methylation of lysine 9 at histone 3 (H3K9me3). We assume that these two modifications affect each other only indirectly via their effects on DNA methylation. This kind of crosstalk is described in the following.

7.3.3 *Crosstalk Between DNA and Histone Methylation*

It is in general accepted that there is a complex crosstalk between histone modifications and DNA methylation (D'Alessio and Szyf 2006). For example, several HMTs have been demonstrated to include binding motifs either for unmethlyated (CXXC, e.g. HMTs writing H3K4me3 (Thomson et al. 2010; Fujita et al. 2003)) or for methylated (MDB, e.g. HMTs writing H3K9me3 (Fujita et al. 2003)) CpGs. So, on the one hand, local DNA methylation status impacts the recruitment of HMTs. On the other hand, the recruitment of Dnmts is affected by histone modifications. For example, H3K4me3 has been demonstrated to repel Dnmt3a (Ooi et al. 2007), whereas H3K9me3 recruits it (Feldman et al. 2006).

We model this kind of crosstalk by accounting for effects of DNA methylation on IC (i.e. HMT) recruitment. We assume that unmethylated and methylated CpGs throughout the CUs act as binding sites for the ICs catalyzing H3K4me3 and H3K9me3, respectively. In addition, we account also for effects of the histone modifications on the recruitment of Dnmts. This is achieved by modifying the de novo Dnmt probability by a factor that depends on the actual histone methylation level of the associated nucleosomes. For details, see Przybilla et al. (2013, 2014).

Simulation of the impact of DNA methylation and histone modifications on gene transcription requires model representations of the genes controlled by these epigenetic marks. For this purpose, we adopt an artificial genome model, which defines a transcription factor (TF) network that exhibits realistic gene expression properties.

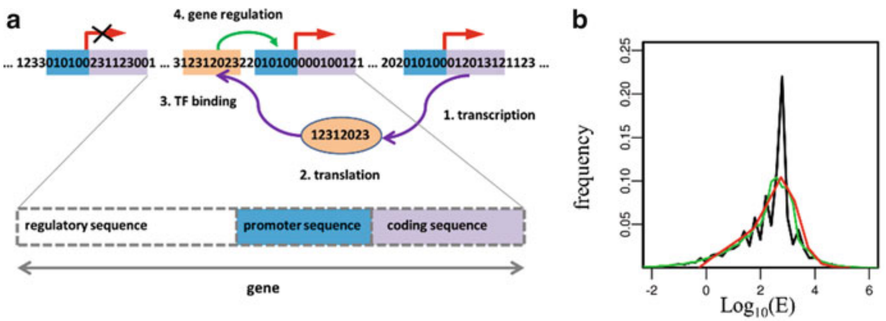
7.4 Modelling TF Networks: The Artificial Genome Approach

Artificial genomes (AGs) were originally introduced to generate gene regulatory networks that cover important biological features of real TF networks. The first idea of building an AG was published by Reil (1999). In extension, we introduced mechanisms to analyze structural evolution in AGs (Rohlf and Winkler 2009).

Moreover, we added a thermodynamic model of transcriptional regulation (Binder et al. 2010), which was adapted from Bintu et al. (2005). In its present form, the model allows calculations of gene expression based on the DNA binding and regulatory action of two types of TFs, namely, repressors and activators. Moreover, it enables straightforward integration of our DNA and histone methylation model. Thus, it represents an ideally suited backbone of our multi-scale model of transcriptional regulation.

7.4.1 Construction of an AG

According to the suggestions by Reil (1999), we generate an AG of length L_{genome} by calculating a random string composed of four different digits [0,1,2,3], where each digit denotes one DNA base [A,T,C,G] (Fig. 7.4a). We consider a single strand DNA only, neglecting all effects caused by the second DNA strand. We are using $L_{\mathrm{genome}} = 400.000$ '*bases*'. Motivated by the frequent association of promoters with TATA-boxes, we assume all short sequences [010100] that are found on the AG



c

$$\text{(D1)} \quad \frac{dT_i}{dt} = P_{\max}\, \theta_{\mathrm{Pro},i} - \delta T_i \qquad \text{(D2)} \quad \theta_{\mathrm{Proj},i} = \prod_j \frac{1 + T_{ij} \exp(\varepsilon_{TF})}{1 + T_{ij}}$$

Fig. 7.4 The artificial genome (AG). (**a**) All sequences [010100] are considered to represent base promoters. They divide the genome into genes of different length. The 8 digits downstream this sequence ([01201312] for the most right gene) define a transcript, which is translated into a protein with a specific DNA binding motif (here [12312023]). The protein can bind to DNA wherever this motif occurs. Bound proteins act as TF and regulate the nearest downstream gene (*green arrow*). (**b**) Distribution of the transcription values of AG genes for a fraction of repressors $p_a = 0.74$. Shown is an average over 100 AGs (*black line*) applying the data set described in the text (*green line*: smoothed version). The result is compared with a distribution measured for colon tissue by RNA-seq (Network 2012) (*red line*). (**c**) Basic equations enabling calculation of expression T_i of gene i. The product in D2 runs over all TFs j regulating gene i

to represent base promoters of genes. The next $L_{cod} = 8$ bases downstream of the promoters are assumed to represent their coding regions. All bases upstream of a gene up to the end of the coding region of the preceding gene define the regulatory region L_{reg} of the gene. Its length can be different for every gene. According to these assumptions, each gene of the AG is divided into three regions, a regulatory, a base promoter and a coding region. All genes potentially encode TFs.

All TFs together form a TF network. This network is constructed by the following rules: Each coding region of a gene defines a transcript. The transcript is translated into a protein with a specific binding motif. This motif is calculated by applying a simple transition rule: each digit of the coding region is updated by adding 1, if the sum is 4 it is replaced by 0. Accordingly, a coding sequence (01201312) will be translated to binding motif (12312023). The protein can bind to identical DNA motifs in the regulatory regions of all genes and act as a TF. Bound TFs regulate the next downstream gene.

According to these building rules, each AG has an intrinsic TF-network structure, which is completely defined by the promoter length, the length of the coding region, and the length of the genome. Using the parameters given above, each TF can regulate on average 6 different genes and each gene can be regulated on average by 6 TFs. For statistical properties of such kind of TF networks, we refer to Binder et al. (2010).

Whether a TF binds to DNA or not depends on the binding energies ε_{TF} and on the concentration of the TF which is set to be equal to the expression level T of its transcript. Bound TFs regulate the occupancy θ_{Pro} of the nearest downstream promoter by RNA-polymerase II by changing the polymerase binding energy. Activators increase it and repressors decrease it. Whether a TF is an activating or a repressing one is chosen with probability p_a and $(1\text{-}p_a)$, respectively. The promoter occupancy $\theta_{Pro,i}$ of the base promoter of gene i is assumed to be proportional to the transcription of the gene. A typical distribution of the expression values of an AG is shown in Fig. 7.4b. It agrees very well with experimentally measured distributions.

The equations describing how the transcription T_i of an individual gene i is calculated for our AG are given in Fig. 7.4c. Here, δ is the degradation constant of the transcript and P_{max} is the maximum promoter activity. Both parameters are assumed to be identical for all genes of the AG.

7.4.2 Crosstalk Between Chromatin Modifiers and Polymerase II Binding

Our AG model describes regulation of gene expression by a TF network. However, experimental results indicate that there is in addition a complex interplay between the expression of genes and their epigenetic status (Cui et al. 2009). In the following, we summarize our assumptions regarding this kind of crosstalk. A sketch describing the interactions considered in the model is given in Fig. 7.5.

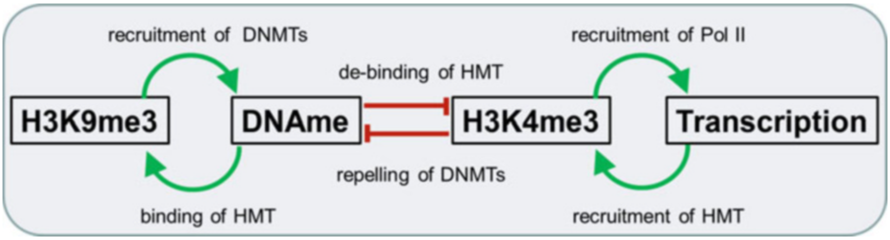


Fig. 7.5 Epigenetic crosstalk. The epigenetic regulation is described by two positive and one negative feedback loop. The sketch denotes the mechanisms that are considered to create these loops

First of all, implementing such crosstalk requires linking the AG model to our model of epigenetic regulation of transcription. For that one has to specify the nucleosomes whose status can contribute to the regulation of a particular gene. Given the structure of our AG, we assume that the modification of all nucleosomes associated with the regulatory region of gene *i* affect the transcription of this gene. This means that we identify the regulatory regions of the genes of our AG with the CUs of our histone model. The modification state of a particular gene is thus the modification state of the nucleosomes associated with the regulatory region of this gene.

As already pointed out in Sect. 7.3, histone modifications can activate as well as repress gene expression. The H3K4me3 mark has been suggested to contribute to gene activation via an improved recruitment of polymerase II in presence of this modification (Vermeulen et al. 2007). In our model we assume that the transcription level of gene *i* is proportional to the occupancy $\Theta_i^{H3K4me3}$ of the REs of this gene by H3K4me3-specific ICs; i.e. by H3K4-HMTs. Accordingly, we replace $P_{\max}$ in Eq. D1 in Fig. 7.4 by $P_{\max}\ \Theta_i^{H3K4me3}$. H3K9me3 is frequently associated with silenced genes. However, we assume that it affects transcription only indirectly via its activity in recruiting de novo DNA methylation.

Recently, it has been shown that not only transcription depends on the histone modification state of the gene but that, vice versa, also the stability of the histone modification states is affected by the transcriptional activity of the associated genes. Experimental findings suggest that the C-terminal domain of the RNA polymerase II subunit Rpb1 undergoes dynamic phosphorylation and that this process helps recruiting the H3K4-HMT complex during early elongation (Buratowski and Kim 2010). In our model, we assume that the recruitment of H3K4me3-specific ICs is enforced at transcribed promoters. Actually, we assume that the basic repulsive term ε_0 of the free enthalpy of binding of H3K4me3-specific ICs is equal to ε_1-$ln(T\delta/P_{\max})$, where ε_1 is a constant. This leads to a positive feedback, stabilizing H3K4me3 at transcribed regions. Binding of H3K9me3-specific ICs is assumed to be not affected by transcription. Accordingly, we assume $\varepsilon_0 = \varepsilon_2$ to be constant for this modification.

The assumed crosstalk determines the dynamics of our multi-scale model. In particular, it controls the stability of its regulatory states. In the next section, we describe how these states can be inherited through iterative replication cycles and how regulatory states evolve on the population scale.

7.5 Modelling Cell Population Behaviour

7.5.1 Model of Cell Replication

In our model, gene expression depends on the histone modification states of the genes. These states are coupled to the local DNA methylation status, which changes during cell division. These changes are different between the daughter cells. Consequently, long-term drifts of transcription, DNA and histone methylation states of the cells can be captured by explicit simulation of cell replication only, and their analysis requires the simulation of large cell populations. We model cell replication assuming that each cell undergoes stochastic growth steps with rate R and divides after N_R successful growth steps.

The changes in DNA methylation during cell division (after initial equilibration) due to limited maintenance and de novo methylation are rather moderate perturbations of the regulatory state of the cells. Much stronger changes can result in parallel from processes of nucleosome re-assembly. During cell division, the core nucleosomes of the mother cell are distributed onto the daughter cells and there become complemented by de novo synthesized, unmodified nucleosomes. This leads to a strong dilution of the modified nucleosomes in the daughter cells. These changes can be different in each daughter due to an unequal distribution of modified nucleosomes of the mother cell onto its daughters (Margueron and Reinberg 2010). In accordance with experimental results (Xu et al. 2010), we assume a random distribution (see Fig. 7.6). In parallel, we assume that de novo synthesis of unmodified nucleosomes guarantees that in the daughter cells the same number of nucleosomes is established as in the mother cell. Thus, we neglect any variance in this property.

The histone modification states immediately after cell division are non-equilibrium states and drift until stable states are reached. If the modification state of a particular gene is bistable, the strong dilution of modified histones after cell division can lead to spontaneous de-modification of the histones. As a result, an asymmetry of the daughters with respect to the modification state of these genes will become manifest. These changes might induce subsequent transcriptional changes and those even further regulatory changes. The regulatory states approached after cell division can be different across daughter cells and thus can induce a strong heterogeneity in an expanding cell clone on all levels of regulation. In order to cover this heterogeneity, we model cell populations, where in each cell the same regulatory network is active but undergoes independent development of its regulatory states.

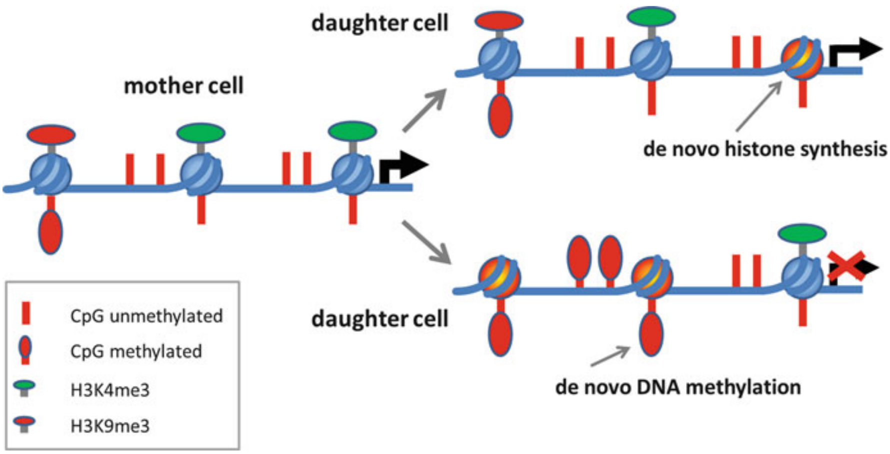


Fig. 7.6 Epigenetic changes during cell replication. During cell replication, the mother nucleosomes are randomly distributed onto the daughter strands and there become complemented with de novo synthesized, unmodified nucleosomes. Thus, the number of modified nucleosomes on each daughter strand becomes diluted. In parallel, DNA methylation state becomes updated separately on each strand. Both processes can induce an asymmetric phenotype in the daughter cells

7.5.2 Regulatory States at the Population Level

Transcriptional states (Wu et al. 2014) and in part DNA methylation states (Guo et al. 2013) of individual cells can be measured experimentally. In contrast, histone modification states are currently accessible on the population level only. In our model, all these states are calculated on the single cell level. Thus, to compare model results with available experimental data, we have to average the regulatory states of the individual genes over a population of cells.

An example of such a calculation is given in Fig. 7.7a showing simulated epigenetic drifts in a proliferative population. The parameter set used in this simulation is given in Table 7.1. In this example, two histone modification states, the H3K4me3 and H3K9me3 state, have been considered. Initially all nucleosomes were marked by both modifications. Due to the dynamics described above, H3K4me3 as well as H3K9me3 is lost within the regulatory region of many genes over time. Loss of H3K4me3 induces decreasing transcription and enables, as de novo DNA methylation is no longer blocked, stabilization of the repressed state by DNA methylation. Thereby, the stability of the modification depends on the number of cooperative nucleosomes. In order to visualize this phenomenon, we have sorted the genes from bottom to top by the increasing number N_H of cooperatively acting nucleosomes associated with the gene. For low numbers, loss of H3K4me3 occurs fast, while for higher numbers it slows down and modified states become stable. Similar effects are observed for H3K9me3. Here, conservation of the initial modification state is associated with DNA methylation.

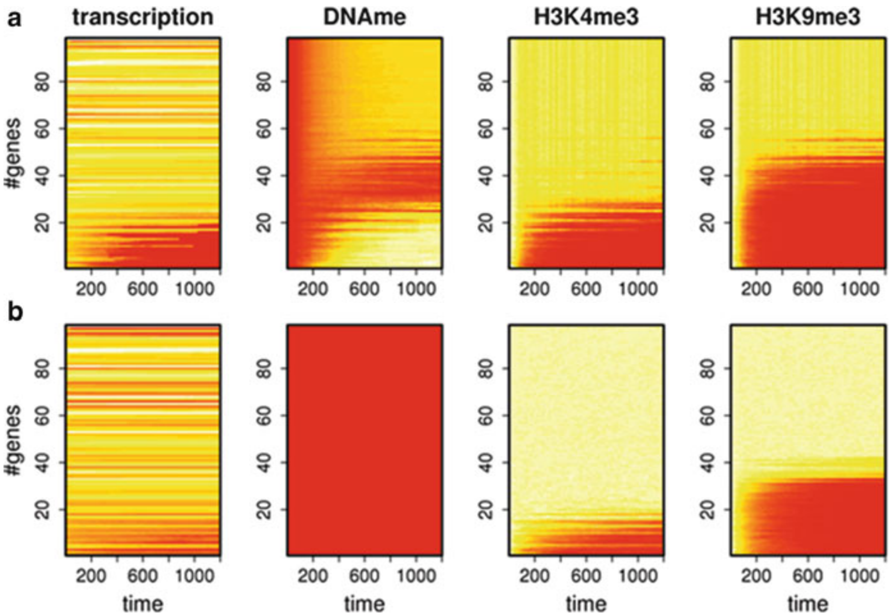


Fig. 7.7 Simulated drifts on the population level. Shown is the development of regulatory states for a proliferative active (**a**) and a quiescent (**b**) cell population. Regulatory states of individual genes are characterized by their transcription level ($\text{Log}_{10}(T)$), their DNA methylation level and by the level of H3K4me3 and H3K9me3 modification of the associated nucleosomes. In their initial state, the expression of all genes was set to the equilibrium state of the isolated AG; DNA methylation level was set to 0 (*red*) and the histone modification levels to 1 (*white*). The genes have been sorted from bottom to top by the increasing number N_H of nucleosomes associated with them. Drifts of the regulatory states are seen for both compartments. In the quiescent population, no changes of the DNA methylation status can occur

The effect of cell division on the regulatory states can be analyzed comparing the dynamics of the regulatory states of proliferative active (Fig. 7.7a) with that of quiescent cells (Fig. 7.7b). In a quiescent cell, the number of cooperating nucleosomes that is required to ensure stable modification decreases and more genes remain stably modified. This is due to inactive DNA methylation, allowing de-modified nucleosomes to become modified again following fluctuations in their histone modification status.

Recently, we have shown that, according to this mechanism, age-dependent drifts in histone modification states are partly reversible if the cells become located in a quiescent niche (Przybilla et al. 2014). In this study, we simulated hematopoietic stem cells in their niche. In order to cope with experimental data (Dykstra et al. 2011; Verovskaya et al. 2013), we enabled the cells to switch between a compartment where they are proliferative active and one where they are quiescent. We assumed the transition rates per simulation time step between the compartments to depend on the number of cells in the compartment they leave. This enables a

Table 7.1 Model parameters. Typical model parameters used in the simulation of Fig. 7.7. Energy terms are scaled by the Boltzmann unit. Rates are given per simulation time step Δt. The parameters of the AG are set as described in the text; those of the TF network were chosen as in Binder et al. (2010)

Symbol	Value	Meaning
P_{max}	1,000	Maximum promoter activity
δ	0.1	Degradation rate of transcripts
D_{main}	0.8	DNA maintenance methylation probability
D_{novo}	0.3	DNA de novo methylation probability
k_D	0.005	De-modification rate for H3K4me3 and H3K9me3
k_M	0.05	Modification rate for H3K4me3 and H3K9me3
ε_{K4}	6	Interaction energy between DNMT and HMT: H3K4me3
ε_{K9}	6	Interaction energy between DNMT and HMT: H3K9me3
ε_{HM}	−1.5	Free enthalpy change of HMT binding to H3K4me3, H3K9me3
ε_{BS}	−5.5	Free enthalpy change of HMT binding to unmethylated (H3K4me3) or methylated (H3K9me3) CpGs
ε_1	7	Ground enthalpy per bound HMT: H3K4me3
ε_2	10	Ground enthalpy per bound HMT: H3K9me3
R	0.1	Growth rate
N_R	10	Number of growth steps towards cell division

stabilization of the number of cells in each compartment. Alternative assumptions are described by Glauche et al. (2009).

7.6 Application of the Model: DNA Methylation Profiles in Tumours

Simulation of stem cell ageing and tissue transformation were major objectives guiding the development of our multi-scale model. In such simulations, we derive hypotheses about the mechanisms underlying the associated changes of the epigenome. So far, we have linked ageing to the limited inheritance of histone modification states (Przybilla et al. 2014) and suggested that epigenetic drifts during tissue transformation originate in an accelerated ageing process, which is often paralleled by drifts induced by mutation of epigenetic pathways (Przybilla et al. 2013).

Figure 7.8 summarizes some of our simulation results. Shown are results of a hierarchical clustering of DNA methylation pattern. These results have been obtained analyzing the consequences of changing activity and of mutations of chromatin modifiers. It can be see that changes of the modifiers can induce both DNA hyper- and hypo-methylation phenotypes. As expected, hypo-methylation is seen for inefficient DNA maintenance methylation ($D_{main} = 0.5$). However, similar patterns are induced also by a knock-out of the H3K4 histone demethylase (HDM) activity (90 % reduction). Hyper-methylation is induced by a knock-out of the

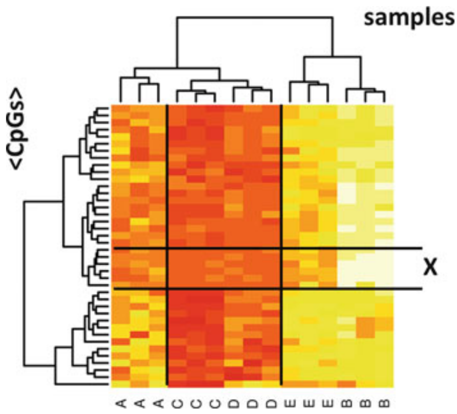


Fig. 7.8 Hierarchical clustering of simulated DNA methylation pattern. Shown are results for five different simulation scenarios, each of them in three replicates. The colour code quantifies the average methylation of CpGs within the regulatory region of genes at a defined time point (*red*: low, *white*: high). We have selected the 40 most variant genes. The characters denote different regulatory conditions, *A* normal ageing, *B* slow histone modification velocity, *C* inefficient maintenance DNA methylation, *D* knock-out of the H3K4 HDM activity, *E* knock-out of the H3K9 HDM activity. Replicates cluster together on the first level. At the second level, conditions C and D and F and B cluster, due to their similar hypo- and hyper-methylation pattern, respectively. Hyper-methylation pattern F and B can be distinguished by the methylation of the CpG subset X

H3K9 HDM activity (90 % reduction) and also in case of decelerated histone modification dynamics (90 % reduction). The latter two patterns are distinguished by the methylation level of only a few genes (compare: cluster X). This suggests that even small groups of CpGs could be very important markers for specific kinds of deregulation.

Overall these results demonstrate that our model is capable of explaining complex changes in DNA methylation pattern by changes in individual chromatin modification pathways. The model suggests that changes in DNA methylation pattern are governed by histone modification dynamics.

7.7 Discussion

Transcriptional changes during stem cell differentiation and also during tissue transformation are commonly thought to be induced by changes in cis-regulatory networks. Chromatin modifications appear to function in stabilization of these changes (Wutz 2013). However, chromatin reorganization can neither establish completely stable nor perfectly inheritable transcriptional states because cell replication induces strong perturbations of the regulatory states. As a consequence, continuous replication results in epigenetic drifts that contribute in controlling the emergence of age-related phenotypes.

We here have introduced a multi-scale model of transcriptional regulation that combines models of DNA and histone methylation with a model of cis-regulatory networks. The combined model enables to analyze the temporal changes of global regulatory states and their dependence on the activity of the individual regulatory layers. Moreover, it allows to generate substantial hypotheses about the interrelations between the different layers of transcriptional regulation and about the potential changes following loss or gain of function in chromatin modification. We have shown how this model can be extended to simulate regulatory phenomena in proliferative active cell populations and that proliferation does strongly feedback on the states of the epigenome.

Our multi-scale model clearly contains various simplifications. For instance: (1) our cis-regulatory model is based on a single strand AG that does neglect evolutionary developed non-random structures, (2) our model of DNA methylation does not consider active DNA de-methylation, (3) the model of histone modification describes only a specific kind of potentially inheritable modifications, namely those set by reader-writer complexes, and focuses on modifications of the cis-regulatory regions only.

Regardless of these simplifications our model provides new insights into transcriptional regulation, e.g. by pointing to the importance of the timescale ratio between proliferation and histone modification for the stability of regulatory states (Przybilla et al. 2013, 2014). Moreover, our model adds new arguments to the histone code debate, suggesting that chromatin computation acts on a very restricted state space, because only a few of the possible combinatorial states are stable.

Although covering several time and length scales of transcriptional regulation, our model still might lack some important regulatory processes. As an example we like to highlight 3D chromatin organization. Changes in the 3D organization potentially affect the cooperative behaviour of the histone modification process, and thus might substantially impact the regulatory states. Actually, we and others observed a dramatic change in the length distribution of specifically modified chromatin during stem cell differentiation processes (Steiner et al. 2012).

In the model, simulations presented here we largely neglected extrinsic regulation of the epigenome. In fact, environmental effects have been considered only by assuming compartment-specific signals that support or block proliferation. However, there are many more environmental signals affecting DNA and histone methylation (Burgess et al. 2014). These signals often depend on the spatial position of the cells in the tissue. For example, stem cells in spatially structured niches, e.g. intestinal stem cells, have been shown to receive local signals that trigger their phenotype. The associated regulatory changes involve also epigenetic changes (Sheaffer et al. 2014). As a first example of a spatially structured model, we plan simulating an intestinal crypt where an artificial genome is transcribed in each of the cells.

Models of the transcriptional regulation by epigenetic processes, as mathematical models in general, will never be comprehensive. However, even at the current state of the art, they allow to generate experimentally testable hypotheses about the mechanisms driving global re-organization of the stem cell epigenome. Thus, computational model approaches, as that presented here, are well on the way to support a better understanding of epigenetic dynamics during differentiation, ageing and tissue transformation.

Acknowledgements This work was supported by the Federal Ministry of Education and Research (BMBF): HNPCC-Sys (grant No. 031 6065), MAGE (grant No. 50500541) and INDRA (grant No. 031A312). We thank Peter Buske for computing data sets used in Fig. 7.4b.

References

Berdasco M, Esteller M. Aberrant epigenetic landscape in cancer: how cellular identity goes awry. Dev Cell. 2010;19:698–711.

Bergman Y, Cedar H. DNA methylation dynamics in health and disease. Nat Struct Mol Biol. 2013;20:274–81.

Binder H, Wirth H, Galle J. Gene expression density profiles characterize modes of genomic regulation: theory and experiment. J Biotechnol. 2010;149:98–114.

Binder H, Steiner L, Przybilla J, Rohlf T, Prohaska S, Galle J. Transcriptional regulation by histone modifications: towards a theory of chromatin re-organization during stem cell differentiation. Phys Biol. 2013;10:026006.

Bintu L, Buchler NE, Garcia HG, Gerland U, Hwa T, Kondev J, Kuhlman T, Phillips R. Transcriptional regulation by the numbers: applications. Curr Opin Genet Dev. 2005;15:125–35.

Buratowski S, Kim T. The role of cotranscriptional histone methylations. Cold Spring Harb Symp Quant Biol. 2010;75:95–102.

Burgess RJ, Agathocleous M, Morrison SJ. Metabolic regulation of stem cell function. J Intern Med. 2014;276:12–24.

Cui K, Zang C, Roh T-Y, Schones DE, Childs RW, Peng W, Zhao K. Chromatin signatures in multipotent human hematopoietic stem cells indicate the fate of bivalent genes during differentiation. Cell Stem Cell. 2009;4:80–93.

D'Alessio AC, Szyf M. Epigenetic tête-à-tête: the bilateral relationship between chromatin modifications and DNA methylation. Biochem Cell Biol. 2006;84:463–76.

Dodd IB, Micheelsen MA, Sneppen K, Thon G. Theoretical analysis of epigenetic cell memory by nucleosome modification. Cell. 2007;129:813–22.

Dykstra B, Olthof S, Schreuder J, Ritsema M, DE Haan G. Clonal analysis reveals multiple functional defects of aged murine hematopoietic stem cells. J Exp Med. 2011;208:2691–703.

Feldman N, Gerson A, Fang J, Li E, Zhang Y, Shinkai Y, Cedar H, Bergman Y. G9a-mediated irreversible epigenetic inactivation of Oct-3/4 during early embryogenesis. Nat Cell Biol. 2006;8:188–94.

Fujita N, Watanabe S, Ichimura T, Tsuruzoe S, Shinkai Y, Tachibana M, Chiba T, Nakao M. Methyl-CpG binding domain 1 (MBD1) interacts with the Suv39h1-HP1 heterochromatic complex for DNA methylation-based transcriptional repression. J Biol Chem. 2003;278:24132–8.

Glauche I, Moore K, Thielecke L, Horn K, Loeffler M, Roeder I. Stem cell proliferation and quiescence–two sides of the same coin. PLoS Comput Biol. 2009;5:e1000447.

Guo H, Zhu P, Wu X, Li X, Wen L, Tang F. Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. Genome Res. 2013;23:2126–35.

Haerter JO, Loevkvist C, Dodd IB, Sneppen K. Collaboration between CpG sites is needed for stable somatic inheritance of DNA methylation states. Nucleic Acids Res. 2014;42:2235–44.

Hinoue T, Weisenberger DJ, Lange CP, Shen H, Byun HM, van den Berg D, Malik S, Pan F, Noushmehr H, van Dijk CM, Tollenaar RA, Laird PW. Genome-scale analysis of aberrant DNA methylation in colorectal cancer. Genome Res. 2012;22:271–82.

Hoffman MM, Ernst J, Wilder SP, Kundaje A, Harris RS, Libbrecht M, Giardine B, Ellenbogen PM, Bilmes JA, Birney E, Hardison RC, Dunham I, Kellis M, Noble WS. Integrative annotation of chromatin elements from ENCODE data. Nucleic Acids Res. 2013;41:827–41.
Kundu S, Peterson CL. Role of chromatin states in transcriptional memory. Biochim Biophys Acta. 2009;1790:445–55.
Magee PN. The possible role of nucleic acid methylases in the induction of cancer. Cancer Res. 1971;31:599–604.
Margueron R, Reinberg D. Chromatin structure and the inheritance of epigenetic information. Nat Rev Genet. 2010;11:285–96.
Network CGA. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012;487:330–7.
Okano M, Bell DW, Haber DA, Li E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell. 1999;99(3):247–57.
Ooi SKT, Qiu C, Bernstein E, Li K, Jia D, Yang Z, Erdjument-Bromage H, Tempst P, Lin S-P, Allis CD, Cheng X, Bestor TH. DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. Nature. 2007;448:714–7.
Otto SP, Walbot V. DNA methylation in eukaryotes: kinetics of demethylation and de novo methylation during the life cycle. Genetics. 1990;124:429–37.
Pfeifer GP, Steigerwald SD, Hansen RS, Gartler SM, Riggs AD. Polymerase chain reaction-aided genomic sequencing of an X chromosome-linked CpG island: methylation patterns suggest clonal inheritance, CpG site autonomy, and an explanation of activity state stability. Proc Natl Acad Sci U S A. 1990;87:8252–6.
Pradhan S, Bacolla A, Wells RD, Roberts RJ. Recombinant human DNA (cytosine-5) methyltransferase. I. Expression, purification, and comparison of de novo and maintenance methylation. J Biol Chem. 1999;274:33002–10.
Przybilla J, Buske P, Binder H, Galle J. Histone modifications control DNA methylation profiles during ageing and tumour expansion. Front Life Sci. 2013;7:31–43.
Przybilla J, Rohlf T, Loeffler M, Galle J. Understanding epigenetic changes in aging stem cells–a computational model approach. Aging Cell. 2014;13:320–8.
Rakyan VK, Down TA, Maslau S, Andrew T, Yang T-P, Beyan H, Whittaker P, Mccann OT, Finer S, Valdes AM, Leslie RD, Deloukas P, Spector TD. Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains. Genome Res. 2010;20:434–9.
Reil T. Dynamics of gene expression in an artificial genome – implications for biological and artificial ontogeny. Adv Artif Life Proc. 1999;1674:457–66.
Rohlf T, Winkler CR. Emergent network structure, evolvable robustness, and nonlinear effects of point mutations in an artificial genome model. Adv Complex Syst. 2009;12:293–310.
Rohlf T, Steiner L, Przybilla J, Prohaska S, Binder H, Galle J. Modeling the dynamic epigenome: from histone modifications towards self-organizing chromatin. Epigenomics. 2012;4:205–19.
Rose NR, Klose RJ. Understanding the relationship between DNA methylation and histone lysine methylation. Biochim Biophys Acta. 2014;1839:1362–72.
Sedighi M, Sengupta AM. Epigenetic chromatin silencing: bistability and front propagation. Phys Biol. 2007;4:246.
Sheaffer KL, Kim R, Aoki R, Elliott EN, Schug J, Burger L, Schubeler D, Kaestner KH. DNA methylation is required for the control of stem cell differentiation in the small intestine. Genes Dev. 2014;28:652–64.
Sontag LB, Lorincz MC, Georg Luebeck E. Dynamics, stability and inheritance of somatic DNA methylation imprints. J Theor Biol. 2006;242:890–9.
Steiner L, Hopp L, Wirth H, Galle J, Binder H, Prohaska SJ, Rohlf T. A global genome segmentation method for exploration of epigenetic patterns. PLoS One. 2012;7:e46811.
Sturm D, Witt H, Hovestadt V, Al E. Hotspot mutations in H3F3A and IDH1 define distinct epigenetic and biological subgroups of glioblastoma. Cancer Cell. 2012;22:425–37.
Teschendorff AE, Menon U, Gentry-Maharaj A, Ramus SJ, Weisenberger DJ, Shen H, Campan M, Noushmehr H, Bell CG, Maxwell AP, Savage DA, Mueller-Holzner E, Marth C, Kocjan

G, Gayther SA, Jones A, Beck S, Wagner W, Laird PW, Jacobs IJ, Widschwendter M. Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. Genome Res. 2010;20:440–6.
Thozmson JP, Skene PJ, Selfridge J, Clouaire T, Guy J, Webb S, Kerr AR, Deaton A, Andrews R, James KD, Turner DJ, Illingworth R, Bird A. CpG islands influence chromatin structure via the CpG-binding protein Cfp1. Nature. 2010;464:1082–6.
Tiwari VK, Mcgarvey KM, Licchesi JD, Ohm JE, Herman JG, Schubeler D, BAYLIN SB. PcG proteins, DNA methylation, and gene repression by chromatin looping. PLoS Biol. 2008;6:2911–27.
Vermeulen M, Mulder KW, Denissov S, Pijnappel WW, van Schaik FM, Varier RA, Baltissen MP, Stunnenberg HG, Mann M, Timmers HT. Selective anchoring of TFIID to nucleosomes by trimethylation of histone H3 lysine 4. Cell. 2007;131:58–69.
Verovskaya E, Broekhuis MJ, Zwart E, Ritsema M, van Os R, de Haan G, Bystrykh LV. Heterogeneity of young and aged murine hematopoietic stem cells revealed by quantitative clonal analysis using cellular barcoding. Blood. 2013;211:487–97.
Wu AR, Neff NF, Kalisky T, Dalerba P, Treutlein B, Rothenberg ME, Mburu FM, Mantalas GL, Sim S, Clarke MF, Quake SR. Quantitative assessment of single-cell RNA-sequencing methods. Nat Methods. 2014;11:41–6.
Wutz A. Epigenetic regulation of stem cells : the role of chromatin in cell differentiation. Adv Exp Med Biol. 2013;786:307–28.
Xu M, Long C, Chen X, Huang C, Chen S, Zhu B. Partitioning of histone H3-H4 tetramers during DNA replication-dependent chromatin assembly. Science. 2010;328:94–8.

Chapter 8
Systems Epigenomics and Applications to Ageing and Cancer

Andrew E. Teschendorff

Abstract One way to view epigenomics is in terms of representing the software of living cells. It is increasingly recognised that complex diseases like cancer are not only driven by defects in the genetic machinery (i.e. the underlying hardware) but also by defects in the epigenome. However, to improve our understanding of how epigenomic aberrations may contribute to the causal development of diseases like cancer will require a systems-level epigenomics approach which integrates different omic data types together. In this chapter, we describe three systems-level statistical methods which have been successful in identifying novel biomarkers for ageing, for cancer risk and for early detection of cancer. In addition, these systems-level methods have provided us with substantial novel insights into systems-level aspects of carcinogenesis, which we also describe.

Keywords DNA methylation • Network biology • ChIP-Seq • Integrative epigenomics • Cancer • Risk prediction • Early detection

8.1 Systems-Level Integration of DNA Methylation and Gene Expression: The Functional Epigenetic Modules (FEM) Algorithm

8.1.1 Integration of DNA Methylation and Gene Expression Using Illumina 450k Arrays

DNA methylation is an epigenetic mark which is well known to correlate with gene expression (Deaton and Bird 2011; Tate and Bird 1993). In cells of normal physiology, promoter CpG islands, which are usually unmethylated, are associated with a transcriptionally permissive chromatin state. In contrast, the DNA methylation of

A.E. Teschendorff (✉)
CAS Key Laboratory of Computational Biology, Chinese Academy of Sciences and Max-Planck Gesellschaft Partner Institute for Computational Biology, Shanghai, China

UCL Cancer Institute, University College London, London, UK
e-mail: andrew@picb.ac.cn; a.teschendorff@ucl.ac.uk

A.E. Teschendorff (eds.), *Computational and Statistical Epigenomics*, Translational Bioinformatics 7, DOI 10.1007/978-94-017-9927-0_8

promoter CpG islands, as is often observed in cancer, is associated with a closed chromatin configuration and hence with gene repression (Deaton and Bird 2011; Feinberg et al. 2006). Importantly, however, the absence of DNA methylation in a promoter CpG island (CGI) is not always associated with transcriptional activity. For instance, in human embryonic stem cells (hESCs), a large class of genes, which are bivalently marked by the active H3K4me3 and the repressive H3K27me3 marks, is associated with unmethylated promoters, yet most of these "bivalent" genes are also not expressed in hESCs (Bernstein et al. 2006; Lee et al. 2006). Thus, the relation between DNA methylation and gene expression is distinctively nonlinear: plotting DNA methylation on the x-axis and gene expression on the y-axis, the relation is described fairly accurately by an "L-type" shape, with low methylation of promoter CGIs associated with either high or low expression, but with methylated promoters generally associated with gene silencing (Fig. 8.1).

Integrative analysis of DNA methylation and gene expression is of considerable interest, especially in a disease context, as this can pinpoint genes causally implicated in disease aetiology or disease progression. In the same way that integration of copy-number and gene expression data has been a promising strategy to identify novel cancer drivers (see, e.g. Chin et al. 2006, 2007; Curtis et al. 2012), the expectation would be that analogous integration of DNA methylation and gene expression could unravel other key cancer drivers. Although DNA methylation levels are limited as predictors of a gene's expression level and more accurate predictors require consideration of histone modification marks (Budden et al. 2014; Karlic et al. 2010), DNA methylation constitutes the only epigenetic mark which

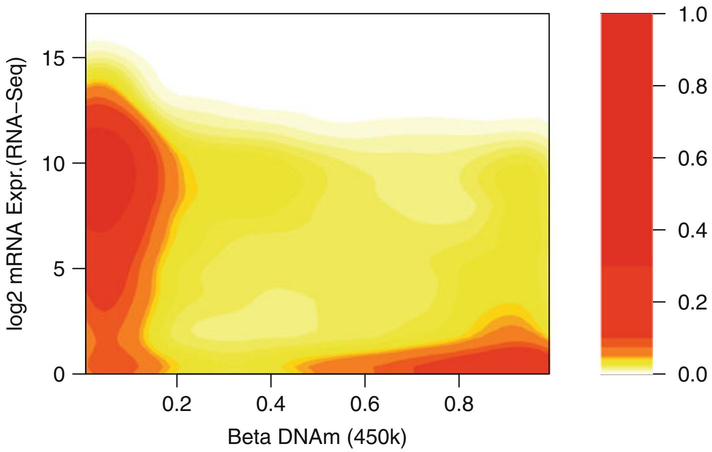


Fig. 8.1 Density scatterplot of promoter DNA methylation values (x-axis) against log2 normalised RNA-Seq values (y-axis) for 15,899 genes. For each gene, DNA methylation values were averaged over 17 normal endometrial samples, and similarly for RNA-Seq. *Red* indicate regions of high point density

can be comprehensively integrated with gene expression in a disease context. This is because, unlike histone modification profiles, genome-wide DNA methylation can be reliably measured in a large number of samples, including limited DNA specimens as often required in a clinical context.

In performing an integrative analysis of DNA methylation and gene expression, a key consideration becomes the specific region to use as a DNA methylation predictor of a gene's expression level. While a number of studies have indicated that DNA methylation of the CGI shores of a gene appears to be most predictive of gene expression variation (e.g. Irizarry et al. 2009), this is still a matter of debate (Deaton and Bird 2011), with some recent studies proposing predictors which go beyond measures of average DNA methylation (e.g. Vanderkraats et al. 2013).

Another key consideration when deciding how best to integrate DNA methylation and gene expression is the technology used for generating the DNA methylation data. Indeed, which gene region to use to build a DNA methylation-based predictor will also largely depend on the underlying technology being used, since not all regions may be equally represented. Currently, one of the most popular technologies for performing genome-wide DNA methylation analysis is the Illumina Human Methylation 450k beadchip, which allows the DNA methylation level of over 450,000 CpGs in the human genome to be measured (Sandoval et al. 2011). The Illumina Infinium 450k chip is in fact still the technology of choice for studies from The Cancer Genome Atlas (TCGA) (see, e.g. Kandoth et al. 2013) and epigenome-wide association studies (EWAS) (Beck 2010; Rakyan et al. 2011). For this reason, it has become important to assess which gene region represented on the 450k array is, at the DNAm level, most informative of gene expression. In doing so, another issue that arises is whether it is best to use single CpG site levels or DNAm levels averaged over neighbouring probes. Since DNAm is generally well correlated on length scales up to 500 bp and in some instances up to 1 kb (Eckhardt et al. 2006), it makes sense to use an average over probes on these length scales, yet this is also a matter of debate (see, e.g. Vanderkraats et al. 2013). A recent study (Jiao et al. 2014) averaged the DNAm levels of probes falling within different gene regions and, using high-quality matched Illumina 450k and RNA-Seq data of many normal tissue samples, concluded that in all samples analysed the region 200 bp upstream of the TSS (TSS200) was the most informative of gene expression, followed by DNAm levels in the 1st exon, and finally DNAm levels located up to 1,500 bp upstream of the TSS (TSS1500). Specifically, for the TSS200, 1st exon and TSS1500 regions, DNAm levels in these regions generally exhibited an anti-correlation to gene expression, in line with the usual paradigm (Deaton and Bird 2011; Tate and Bird 1993). Given that these results were obtained in individual normal samples, using high-quality matched data, and that these results were congruent across so many independent samples, it does indeed support the view that for the 450k array probes, the TSS200 and 1st exon regions are the most predictive of gene expression (Jiao et al. 2014).

8.1.2 Systems-Level Integration of DNA Methylation and Gene Expression Using FEM

Once a DNA methylation value has been assigned to any given gene, integration of the two data types can proceed in both a univariate or multivariate fashion. The standard univariate approach would be to use regressions between gene expression and DNA methylation, one for each gene, to identify putative drivers. Alternatively, one may find the overlap of genes exhibiting both differential methylation and differential expression in cancer. The resulting gene list could then be used as input to integrative clustering algorithms such as iCluster (Shen et al. 2012, 2009), jNMF (Wang et al. 2015) or JIVE (Lock et al. 2013), to identify tumour subgroups characterised by covariation of DNAm and gene expression, although we point out that some of these algorithms (e.g. JIVE) would not require prior selection of correlated genes, since it is able to dissect data-type-specific variation from the common variation across data types. An alternative possibility would be to use canonical correlation analysis (CCA) and specifically their penalised sparse and/or semi-supervised versions (Witten et al. 2009; Witten and Tibshirani 2009).

A list of genes exhibiting simultaneous differential methylation and differential expression could also be used for performing a Gene Set Enrichment Analysis (GSEA) (Subramanian et al. 2005). In this regard, it is worth pointing out that a list of differentially methylated genes may be enriched for important biological terms; for instance, a DNA methylation study of breast cancer identified a strong immune cell component among prognostic CpGs (Dedeurwaerder et al. 2011), in line with corresponding studies done previously at the gene expression level (see, e.g. Teschendorff et al. 2010, 2007). For this reason, it may also be fruitful to perform functional supervised analyses, by direct integration of promoter DNA methylation levels with a functional gene network, encoding functional relationships between genes, such as that provided by a protein-protein interaction (PPI) network (Cerami et al. 2011; Prasad et al. 2009; Rolland et al. 2014). It is worth noting that integration of DNA methylation changes with a PPI network has been performed in other contexts (Liu et al. 2011; West et al. 2013) and appears well justified on biological grounds (Timp et al. 2009). Functional network-supervised analysis methods have been shown to be very fruitful in the gene expression context (Chuang et al. 2007; Mitra et al. 2013) and subsequently later also in the DNA methylation context (West et al. 2013). Indeed, quite remarkably, one can identify interactome hotspots of differential methylation associated with ageing, which are validated in many independent data sets and which pinpointed several stem-cell differentiation pathways (including the WNT signalling pathway) as significantly altered in ageing (Teschendorff et al. 2013; West et al. 2013).

Further justifying the integration of DNA methylation data with a PPI network, we observed that, at the level of the PPI network, gene promoter DNA methylation is characterised by a “correlation modularity” (Teschendorff and Widschwendter 2014; West et al. 2013), defined as the propensity of neighbouring proteins in the network to share a more similar gene promoter DNA methylation level than a randomly picked protein pair in the network. This correlation modularity was demonstrated within a phenotype, and specifically across normal samples of a given tissue type (Teschendorff and Widschwendter 2014; West et al. 2013), and is driven mainly by a more similar promoter CpG density level of interacting proteins (Teschendorff and Widschwendter 2014). It should be clear, however, that this correlation modularity is not a requirement for pursuing differential DNA methylation analysis between two phenotypes in a PPI network context. For instance, copy-number aberrations in a given cancer sample tend to occur in a mutually exclusive fashion within a specific signalling pathway (Ciriello et al. 2012), so a similar pattern of mutual exclusivity may be expected for DNA methylation levels. Mutual exclusivity would mean that DNAm correlation modularity, as assessed across two phenotypes, would be absent, yet a given pathway module may still represent a hotspot of differential methylation since different genes may be aberrantly methylated in different subgroups of cancers (West et al. 2013).

Given that supervised network analysis of DNA methylation data is a promising approach (West et al. 2013), it is therefore natural to also consider a three-way integration of DNA methylation and gene expression with a PPI network, as done by the FEM (Functional Epigenetic Modules) algorithm (Jiao et al. 2014; Jones et al. 2013) (Fig. 8.2). Doing so in the context of cancer may not only reveal functional epigenetic drivers of cancer but may also shed light on specific signalling pathways or mechanisms which contribute to carcinogenesis. Indeed, a clear example of how such integration can pinpoint a cancer driver was a study performed in endometrial tumours, which using FEM identified a gene called *HAND2* as causally implicated in the genesis of this cancer (Jiao et al. 2014; Jones et al. 2013; Teschendorff and Widschwendter 2014). The FEM algorithm is freely available from www.bioconductor.org.

Briefly, the FEM algorithm integrates the statistics of differential DNA methylation and differential mRNA expression with a comprehensive PPI network, as provided, for instance, by the PathwayCommon resource (Cerami et al. 2011). Because the integration is performed at the level of statistics, there is no requirement for the DNA methylation and gene expression data to be matched (i.e. to come from the same samples). Of course, in the unmatched setting, for the integration of the statistics to make sense, one does require that the two cohorts used for the DNA methylation and gene expression profiling are in some sense similar (e.g. similar types of breast cancer). In what follows we don't make a distinction as to whether we have a matched or unmatched setting, since at the level of statistics, the integration proceeds in an identical fashion. Thus, for each gene represented in the PPI network, one has a statistic of differential methylation and another for differential expression, which are obtained by comparing two phenotypes (here we consider normal vs cancer tissues) using one of many possible statistical tests (e.g. moderated t-statistics (Smyth et al. 2003)). Since the expected association between promoter DNA methylation and gene expression is that of an anti-correlation (as mentioned earlier), one can assign an overall statistic to each node/gene as

$$t_g = \left\{ H\left(-t_g^{(D)}\right) H\left(t_g^{(R)}\right) + H\left(t_g^{(D)}\right) H\left(-t_g^{(R)}\right) \right\} \left| t_g^{(D)} - t_g^{(R)} \right| \qquad (8.1)$$

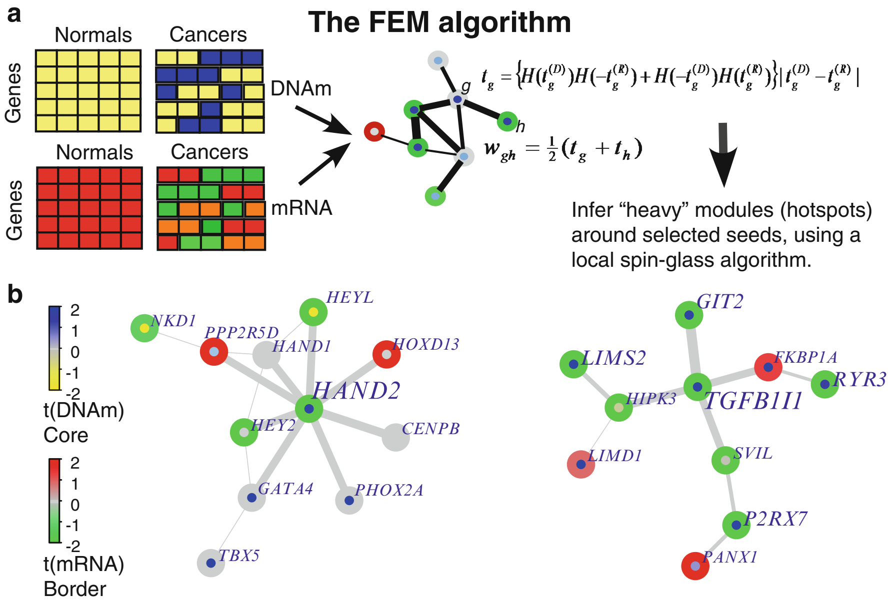


Fig. 8.2 (**a**) Schematic of the FEM algorithm. Unmatched or matched DNA methylation and gene expression data from normals and cancers are used to derive the statistics of differential DNA methylation and differential expression for genes represented in a PPI network. Looking specifically for anticorrelated patterns between differential DNAm and differential expression (using Heaviside functions to impose the anti-correlation) leads to an overall semi-positive statistic t_g for each gene in the network. Edge weights are then constructed as averages of these statistics, and a spin-glass algorithm is subsequently used in a local greedy fashion to identify hotspots of simultaneous differential DNAm and mRNA expression. FEM is available as a Bioconductor package (www.bioconductor.org). (**b**) Two examples of hotspots inferred from the TCGA endometrial cancer data implicated two genes with roles in the tumour suppressor progesterone receptor pathway. *HAND2* is the clear target of one hotspot and mediates the tumour-suppressive effects of progesterone (not shown in the diagram), while *TGFB1I1/HIC5* is one possible target of the other hotspot and is a co-activator of the progesterone receptor. Both *HAND2* and *HIC5* are silenced in endometrial cancer through promoter DNA hypermethylation, as shown

where $H(t)$ is the Heaviside function, which means that genes which are hypomethylated and underexpressed (or hypermethylated and overexpressed) are assigned a statistic of 0. This anti-correlation assumption is by no means necessary, but is a procedure which can help focus on the more likely true positive associations, given that, globally, there is an anti-correlation (Jiao et al. 2014). The FEM algorithm then proceeds by weighting the PPI network with the weight of the edge connecting genes g and h, defined by

$$w_{gh} = \frac{1}{2}(t_g + t_h) \tag{8.2}$$

Encoding the associations at the DNA methylation and gene expression level in the weights of the network has the advantage that hotspots of differential methylation and expression can then be identified using module detection algorithms which aim to maximise the weight density of subnetworks. Although computationally very intensive, it is possible to identify the global maxima (Mitra et al. 2013). However, the robustness of such global maxima is likely to be low, and so for this reason, greedy approaches to module detection, which are also much more scalable, have proved extremely popular (Mitra et al. 2013). Indeed, in a greedy local approach to module detection, one would perform the search for heavy subnetworks (i.e. subnetworks of large weight density) in a local fashion, starting from a prespecified set of seed genes. The FEM algorithm implements such a local greedy search by choosing as seeds those genes which were top-ranked in the combined supervised analysis, i.e. those genes with the maximal t_g values. The specific algorithm used to do the search is based on a spin-glass model, which formulates the modularity (i.e. weight density) to be maximised in terms of the negative of an energy function (Reichardt and Bornholdt 2006). Assuming that seeds are not too close to each other in the resulting network, such a greedy local implementation can nevertheless search most of the network, and although identification of true global maxima is not guaranteed, the identified modules are likely to represent fairly robust features. The fact that a simpler version of FEM, the EpiMod algorithm, which implements the same module detection algorithm in the context of only DNA methylation data, was so successful in retrieving age-associated differential methylation hotspots which could subsequently be validated in many independent data sets (West et al. 2013), attests to the robustness and validity of the procedure.

When implementing FEM, there are three important issues to keep in mind. First, since the weights are derived from statistics which are obtained by combining those from DNA methylation and those from mRNA expression, it is important that the two sets of statistics are comparable. This requires scaling the statistics of one data type (say DNA methylation) by a constant factor to ensure that the variances of the two statistics distributions are the same. This will help remove unwanted bias of resulting modules towards one of the two data types. Second, not all seeds will lead to a module/hotspot. This only reflects the chance that a highly ranked seed represents an isolated node of association, with none of its neighbours exhibiting an association with the phenotype at the DNAm or mRNA level. Third, a key issue in module detection within biological networks is *module size*. Modules which are too large will be difficult to interpret and are unlikely to validate in independent data (West et al. 2013). On the other hand, modules which are too small, containing only a handful of genes, are likely to arise by chance and to represent false positives. Thus, when detecting modules within biological networks, there is an optimal module size range, which according to previous analyses appears to be around 10–100 genes (West et al. 2013). In this size range, gene modules are more likely to validate in independent data (West et al. 2013). However, this size range is also motivated by GSEA approaches in the gene expression context which often identify enriched biological terms encompassing this number of genes. This size range also represents roughly the number of genes found within specific signalling pathways. Thus, it makes sense if the module detection algorithm is tuned to identifying modules in this particular size range. Fortunately, the spin-glass algorithm implemented in FEM is characterised by a single free parameter (called γ), which controls the average size of the inferred modules. Thus, this parameter can be tuned on a training set, say, to achieve modules in the desired size range, and then applied on the data of interest. Such an approach was used in West et al. (2013), identifying $\gamma \sim 0.5$–0.6 as an optimal parameter range and which was found to be largely robust/independent of data type and data set (Jiao et al. 2014; West et al. 2013).

In Jiao et al. (2014), the FEM algorithm was applied to matched Illumina 450k and RNA-Seq data from the endometrial cancer TCGA data set (Kandoth et al. 2013), identifying a number of FEM modules, representing differential DNA methylation and gene expression hotspots in this cancer type. The list of FEM modules included one centred around *HAND2*, which was one of the top-ranked seeds and the clear target of the module (Fig. 8.2). Indeed, while other key transcription factors like *GATA4* (which interact with *HAND2*) were also hypermethylated in cancer, these were generally not silenced. Importantly, *HAND2* is a target of the progesterone receptor, mediating the effects of this tumour suppressor on the endometrial epithelium (Jones et al. 2013). As shown in Jones et al., methylation-induced silencing of *HAND2* is causally implicated in endometrial carcinogenesis, predicts non-response to progesterone treatment and could be used to detect the earliest stages of the disease by noninvasive vaginal swab collections (Jones et al. 2013). The importance of FEM is that it allowed the identification of a module implicated in the progesterone receptor pathway, the key tumour suppressor pathway in endometrial cancer. Indeed, another FEM module implicated in the same tumour suppressor pathway centred around *TGFB1I1/HIC5* (Fig. 8.2), which is a well-known co-activator of the progesterone receptor. In the case of the *HIC5* module, however, there were several other genes silenced in cancer through promoter DNA methylation, indicating potential tumour-suppressive roles for these other genes. This example highlights the power of such integrative network analyses to identify novel potentially causal mechanisms underlying carcinogenesis.

8.2 Systems-Level Integration of DNA Methylation and ChIP-Seq Data

Another type of systems-level integration one could consider is between DNA methylation and ChIP-Seq data. Although the interplay between DNA methylation and transcription factor (TF) binding is undoubtedly complex, with many different proposed models (Blattler and Farnham 2013; Hu et al. 2013; Vermeulen 2013), one prevailing model is that of transcription factor binding acting to block de novo DNA methylation (Blattler and Farnham 2013). Indeed, binding of site-specific transcription factors like *CTCF* or *SP1* can protect regulatory regions from gaining methylation (Blattler and Farnham 2013). Moreover, active transcription of GC-rich promoter regions is thought to protect the promoter-proximal transcribed region from the action of DNA methyltransferases (DNMTs) (Blattler and Farnham 2013). Conversely, the presence of DNA methylation in high-CpG-density promoter regions may act to block TF binding, leading to gene repression (Deaton and Bird 2011) (Fig. 8.3a).

Given the complexity of the crosstalk between DNA methylation, TF binding and the action of epigenetic enzymes which mediate the crosstalk (Ruscio et al. 2013), it is clear that matched genome-wide DNA methylation and ChIP-Seq data for many cell types are required to elucidate the underlying rules. Generating such matched data on a large scale and with clinical samples is at present, however, not possible. Thus, from an integrative analysis viewpoint, one feasible approach is to use ChIP-Seq profiles generated in model cellular systems (Bernstein et al. 2012) as surrogates, to subsequently integrate with sample-specific DNA methylation. One of the first studies to conduct such an integration did this in the context of cellular development (Ziller et al. 2013). This landmark study integrated whole-genome bisulfite sequencing data (WGBS) of hESCs and many differentiated cell types with genome-wide ChIP-Seq profiles, as generated by the ENCODE consortium (Bernstein et al. 2012; Gerstein et al. 2012; Thurman et al. 2012). A key insight of this study was the identification of dynamic differentially methylated regions (DMRs) (i.e. regions whose DNAm levels change during cellular development) which were highly enriched for regulatory elements, including transcription factor binding sites derived from ChIP-Seq experiments performed in cells which were relevant to the comparisons of interest. Fundamentally, this study showed how DMRs between hESCs and differentiated cells of a given tissue type (say liver) could be used to identify tissue-specific transcription factors. Specifically, in the case of liver, it was observed how binding sites of liver-specific transcription factors like *HNF4A* become significantly hypomethylated in liver samples (Ziller et al. 2013). Likewise, the binding sites of well-known pluripotency factors, such as *NANOG* or *OCT4*, were found to be invariably hypomethylated in the hESC state compared to differentiated cells. Thus, DNA methylation can act as a mark distinguishing active from inactive TF binding and thus help identify regulatory factors which are disrupted in complex phenotypes and diseases. Once gain, this is particularly noteworthy given that DNA methylation can be reliably measured in a high-throughput manner and on a large scale, including clinical specimens (Beck 2010).

Importantly, the results obtained by Ziller et al. (2013) have recently been validated in the context of Illumina 450k data, thus opening up integrative analyses of DNA methylation and ChIP-Seq data to existing and upcoming 450k EWAS and TCGA studies (Tian et al. 2015). Indeed, it was possible to use DMRs inferred from comparing Illumina 450k data of hESCs to that of differentiated cell types, to identify known pluripotency and tissue-specific factors (Tian et al. 2015) (Fig. 8.3b, c). In these enrichment analyses, it is possible to incorporate combinatorial information, by recasting them as multivariate regressions between

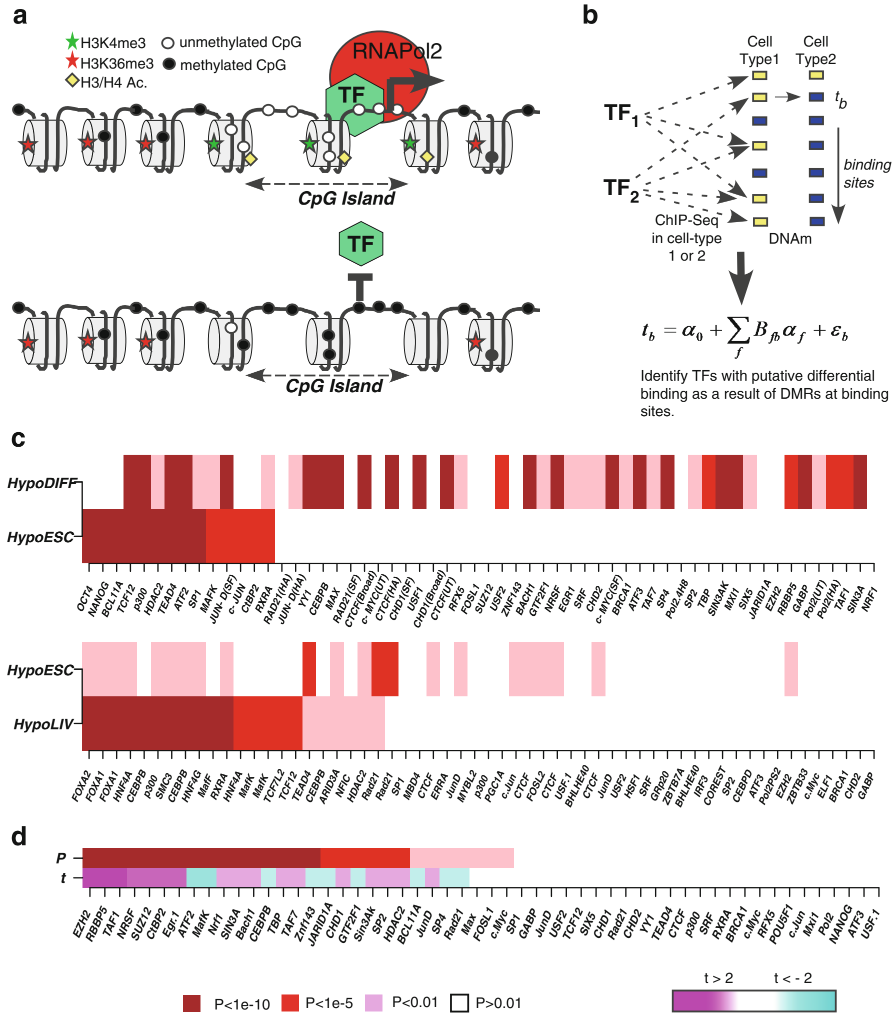


Fig. 8.3 (**a**) One prevailing model of the complex crosstalk between DNA methylation and TF binding implicates promoter DNA methylation at CpG islands as a mechanism which blocks site-specific TFs from binding, thus leading to gene repression. (**b**) Integration of ChIP-Seq data obtained in a relevant cell-line model with the differentially methylated regions between two cell types of interest. The statistics of differential methylation can be regressed in a multivariate model against TF binding profiles to identify those TFs, whose binding sites are enriched within DMRs independently of other TFs. (**c**) *Upper panel*: ranking of TFs according to the enrichment of their binding sites (as assessed in a hESC line) among DMRs between hESCs and differentiated cell types. *Lower panel*: as upper panel, but now for ChIP-Seq data of TFs obtained in a liver cancer cell line and for DMRs between liver cells and hESCs. (**d**) As (**c**), but now for ChIP-Seq data in a hESC line and with DMRs selected as age-associated DMRs derived from over 600 whole blood samples

the statistics of differential DNAm and TF binding profiles (Fig. 8.3b), thus allowing in principle the more fundamental TFs to be identified.

Application of this strategy in the context of ageing (Tian et al. 2015), using a large whole blood 450k data set (Hannum et al. 2013), confirmed known regulatory factors associated with ageing (e.g. polycomb components) (Maegawa et al. (2010); Rakyan et al. (2010); Teschendorff et al. (2010)), but also many novel ones (Fig. 8.3d). For instance, many age DMRs mapped to the binding sites of *NRSF/REST* (Tian et al. 2015), a transcription factor which, although not directly linked to ageing, is nevertheless implicated in suppressing genes that promote Alzheimers (Lu et al. 2014). Given that many age-associated DNA methylation changes appear in common between blood and brain tissues (Horvath 2013; Horvath et al. 2012), it is plausible that a similar pattern of DNA methylation change affecting *REST* binding sites is indeed present in aged brain tissue. Age-associated acquisition of DNA methylation at these specific binding sites could thus compromise the suppression of these key genes, resulting in an increased risk of Alzheimer's, consistent with age being the major risk factor.

The integration of age DMRs and ChIP-Seq data also identified epigenetic regulatory factors, such as *CTCF* and *RBBP5* (Tian et al. 2015), which also have not been directly linked to ageing. In the case of *RBBP5* (retinoblastoma binding protein-5), this protein is part of the MLL1/MLL complex, whose role is to methylate/di-methylate lysine 4 of histone H3, which in turn is a tag for epigenetic transcriptional activation. Although the methylation of H3K4 has been widely observed to vary with age (see, e.g. Lui et al. 2014; Walter et al. 2014), *RBBP5*'s potential role in ageing was only implied from homology (JenAge AgeFactDB database) (Huehne et al. 2014). Thus, the new data suggests that the age-induced impairment of *RBBP5* binding could lead to the functional disruption of the MLL complex and hence to the loss of H3K4 methylation, a well-known ageing effect (Lui et al. 2014; Walter et al. 2014).

In summary, integrative analysis of genome-wide ChIP-Seq and DNA methylation data will be a fruitful approach to identifying regulatory networks disrupted in complex diseases and phenotypes and may also provide more refined DNAm signatures to be used in prognosis and prediction. However, generation of ChIP-Seq profiles for relevant transcription factors in relevant cell types will be key to make the most of these analyses.

8.3 A System's Epigenomics Approach to Cancer Risk Prediction

8.3.1 Introduction

The emphasis in cancer genomics is slowly shifting from prognosis and treatment to early detection and risk prediction (Anjum et al. 2014; Dumanski et al. 2015;

Forsberg et al. 2014; Genovese et al. 2014; Jacobs et al. 2012; Laurie et al. 2012; Teschendorff et al. 2012; Xie et al. 2014; Xu et al. 2013). This shift is happening for various reasons. First, individual cancers have now been shown to be remarkably heterogeneous entities (Gerlinger et al. 2012), which endows them with a remarkable aptitude to evolve drug-resistant clones, either through selection of pre-existing resistant clones or through new driver mutations (Crystal et al. 2014; Pisco et al. 2013). Thus, although some authors suggest that it may be possible to turn cancer into a chronic disease by continuous monitoring of new driver mutations in circulating tumour cells (CTCs) and by subsequent administration of corresponding targeted therapies (Beck and Ng 2014), it is unclear whether such a strategy could ever provide a long-lasting cure of advanced diagnosed cancers. Detecting a tumour at an early stage, when it is likely to be less heterogeneous, is key to a good clinical outcome (Maley et al. 2006). Even better would be to be able to predict the risk of neoplastic transformation, since such a putative risk index could be used to assign at-risk individuals to especially designed screening programmes (Kitchener et al. 2009). Close monitoring of at-risk individuals would thus help detect more tumours at an earlier stage, or possibly even prevent them (Hood and Friend 2011). However, predicting the risk of a given cancer requires, in principle, access to a relevant tissue, the most relevant one being the normal cell of origin that gives rise to the cancer. In the case of blood-borne cancers, this is, in principle, feasible, and indeed recent studies indicate that prospective risk prediction of haematological cancers may be possible (Anjum et al. 2014; Dumanski et al. 2015; Forsberg et al. 2014; Genovese et al. 2014; Jacobs et al. 2012; Laurie et al. 2012; Xie et al. 2014; Xu et al. 2013).

However, the case of epithelial cancers is far more challenging given that the normal cell of origin is usually not readily accessible. Ideally, one would be able to acquire such cells in a noninvasive manner as part of routine screening programmes, which could have been set up to detect other common diseases. One of the few cancers where access to the cell of origin is possible is cervical cancer, since cervical smear samples are routinely collected for screens against cytological abnormalities and/or HPV infection, which is the major risk factor (Kitchener et al. 2009). In the context of cervical cancer, it is possible to collect a large number of samples at different disease stages and in particular to follow up healthy women in order to see who progress to a high-grade intra-epithelial cervical neoplasia (CIN2+), a condition which normally precedes an invasive cervical cancer (Kitchener et al. 2009). The ability to measure molecular profiles (e.g. DNA methylation) in normal cells, well before they become neoplastic, may thus not only offer novel insights into carcinogenesis but possibly also novel risk biomarkers.

A recent study tried to correlate DNA methylation patterns measured in cytologically normal cervical samples from 152 women to the prospective risk of CIN2+ (75 of the 152 women developed a CIN2+ within 3 years of the 1st round of sample collection) (Teschendorff et al. 2012). However, genome-wide significance levels were poor (minimum FDR ~ 0.6), indicating that DNAm-based risk prediction from normal cells may require larger sample sets (Teschendorff et al. 2012). Interestingly, the same authors realised that a more fundamental reason for the poor FDR could be that the homogeneity assumption underlying t-tests and their nonparametric equivalents may not be valid in the stages prior to carcinogenesis. Indeed, the same authors proposed a different statistical model to select risk predicting CpGs ("risk CpGs"), one based on the paradigm of differential variability (Teschendorff et al. 2012; Teschendorff and Widschwendter 2012). Specifically, the authors posited that although the mean levels of DNA methylation of risk CpGs may not differ significantly between the normal samples which remain normal and the normal samples which become CIN2+, nevertheless, a small subset of the prospective CIN2+ cases may show significant "outlier" changes in DNAm. Thus, while t-tests, or their nonparametric equivalents, would fail to select these CpGs, a test for differential variability would rank them at top, since the variance is far more sensitive to outliers. Differential variability would thus identify CpGs with outlier methylation in a subgroup of samples within one phenotype but with highly stable methylation across all samples of the other phenotype. In Fig. 8.4, we illustrate and clarify the key distinction between differential variability (as assessed, e.g. by the Bartlett or Levene test) and differential means (as assessed, e.g. by the t-test or its nonparametric counterparts) and how these different tests may select for very different types of profiles/CpGs. Importantly, what this figure also demonstrates is that t-tests, and generally speaking also their nonparametric equivalents, will select

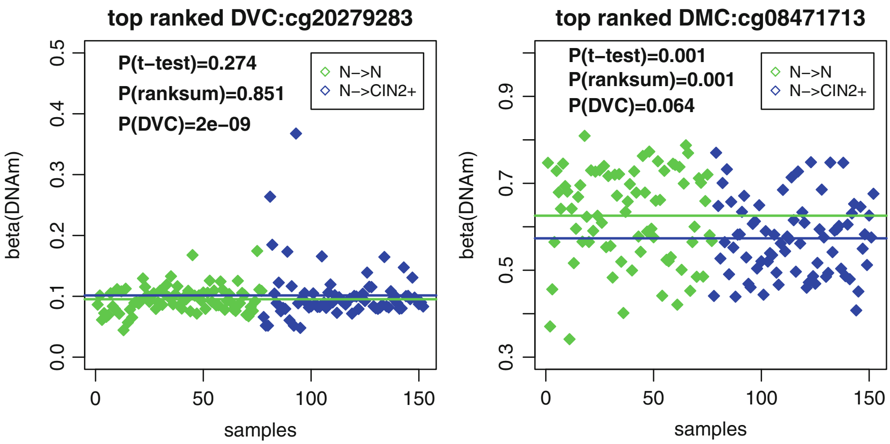


Fig. 8.4 Examples of differentially variable (DVC) and differentially methylated (DMC) CpGs. Shown are the DNAm beta values of a top-ranked DVC and top-ranked DMC across 152 cervical smear samples (all of normal cytology). The women who 3 years later developed a CIN2+ are indicated in *blue*. We provide P-values of a t-test, a Wilcoxon rank sum test and that of a Bartlett test (which tests for differential variability). In this particular data set, the top-ranked DMCs show little separability with a difference in the mean methylation level of the two groups of only about 5 %. For this reason, no risk prediction is possible by selecting DMCs. However, using a test for differential variability identifies a different set of CpGs, which remain significant after multiple testing and which allow the future risk of CIN2+ to be predicted using the EVORA risk prediction algorithm (Teschendorff et al. 2012). Note how a t-test or its nonparametric equivalent would not be able to identify this type of DVC, as the corresponding P-values are nonsignificant

for CpGs which show *homogeneous* changes in DNA methylation between the two phenotypes. These same tests are not able to identify features with *heterogenous* outlier profiles, where samples do not show any variation in DNAm levels, except for a relatively small number of outliers which are present exclusively in one phenotype.

Using a novel prediction algorithm, called EVORA (Epigenetic Variable Outliers for Risk prediction Analysis), designed to identify differentially variable CpGs, the authors demonstrated that the prospective risk of CIN2+ could be predicted, albeit with a low AUC (AUC = 0.66 with 95 %CI of (0.58–0.75)). The key novel insight of this study, however, was the presence of specific promoter regions, targeting bivalent domains in hESCs (Bernstein et al. 2006; Lee et al. 2006), which were prone to hypervariable (i.e. outlier) DNA methylation, specifically in the subset of women who later developed a CIN2+. These same loci were stably unmethylated in all the women who did not progress to the CIN2+ stage. For one of these given risk loci, the number of women exhibiting abnormally high DNA methylation in their cytologically normal samples was a relatively small fraction (typically 5–10 %) of all women developing CIN2+ 3 years later. Using 140 of these risk loci, however, EVORA was able to predict the prospective risk of CIN2+ by constructing a risk index, obtained by counting the number of loci which exhibit aberrant hypermethylation in a given sample (Teschendorff et al. 2012). Importantly, this risk index progressed in value when computed in CIN2+ samples, as well as in invasive cervical cancers, providing a diagnostic test of very high sensitivity and specificity for both CIN2+ and invasive cervical cancer (AUC for CIN2+ = 0.93, AUC for cervical cancer = 1) (Teschendorff et al. 2012).

8.3.2 Adopting a Systems View: The Dynamics of DNAm Covariation During Carcinogenesis

As demonstrated in Teschendorff et al. (2012) and Teschendorff and Widschwendter (2012) and as explained again above, differential variability can be key for identifying DNA methylation changes in normal epithelial cells which may indicate the future risk of neoplastic transformation. Furthermore, we have seen how a test for differential variability can pick out heterogeneous outlier DNAm profiles, with outliers happening exclusively in the normal samples who later progress to CIN2+. As Fig. 8.4 demonstrates, the outliers are defined by relatively “big jumps” in methylation, on the order of 10–30 %. Biologically, this means that in these at-risk normal samples, $\sim$30 % of the cells have the specific risk CpG site methylated. Although a number of risk CpG sites may exhibit such outlier DNAm values in the same sample, this does not mean that these are happening in the same subclone. Outlier methylation occurring at in-phase or spatially close CpGs is much more likely to represent the same subclone. Risk CpG sites may also exhibit outlier methylation in different sets of samples or exhibit overlap for a subset of samples. These considerations are important in view of a hypothesis put forward by a number of authors (Feinberg and Irizarry 2010; Issa 2011) and which posits that environmental risk factors (e.g. smoking, sunlight exposure, inflammation, viral infection) mediate cancer risk by increasing intra-sample epigenetic (and genetic) heterogeneity. Although evidence is mounting that genetic and epigenetic heterogeneity both play a role (Anjum et al. 2014; Dumanski et al. 2015; Feinberg and Irizarry 2010; Forsberg et al. 2014; Genovese et al. 2014; Issa 2011; Jacobs et al. 2012; Laurie et al. 2012; Maley et al. 2006; Teschendorff et al. 2012; Xie et al. 2014), clonal mosaicism generally may increase the risk of cancer, since a more heterogeneous cell population is more likely to give rise to a future clone with neoplastic properties. We posited that at the tipping point of the emergence of such a neoplastic clone, the intra-sample molecular heterogeneity would be specially high and this may also drive a high inter-sample variability, as observed between unrelated individuals who are all at the same prior disease stage (Teschendorff et al. 2014). We furthermore proposed that by analysing the genome-wide covariation in DNA methylation between CpGs in subsequent disease stages, we might be able to detect a subset of CpGs, whose covariation is maximal in a disease stage immediately prior to the onset of cervical cancer. Cervical cancer itself would be characterised by the emergence of a dominant malignant clone, with increased selection pressure, and thus possibly by a small reduction in the intra-sample clonal mosaicism, which would manifest itself as a reduction in the covariation patterns as assessed over independent samples. This model is illustrated with real data in Fig. 8.5, which shows the progressive changes in the DNAm of two risk CpGs between three disease stages (normal, CIN2+ and cervical cancer) (Fig. 8.5a).

As we can see, the CIN2+ stage is characterised by a striking bimodality, with some samples showing similar DNAm levels as in the normal state but with other samples exhibiting much higher fractions of DNAm. A key point to appreciate here is that in the subsequent cancer stage, this bimodality disappears as effectively most cancers now demonstrate some level of DNAm at these CpG sites. From the perspective of the covariation strength of these two CpGs, we can see in Fig. 8.5b how the covariation (represented there by the R^2 value) is maximal in the CIN2+ stage, *and not in the cancer stage*. Figure 8.5c shows the covariation heatmaps of the 91 risk CpGs identified in Teschendorff et al. (2014) across the three main disease stages, which clearly demonstrates that the covariation of this *particular set of CpGs* is higher in the CIN2+ stage compared to cervical cancer, although cervical cancer also exhibits stronger covariation patterns than the normal state.

The covariation strength of a given set or module of CpGs (labelled below by m) can be quantified in terms of the following heuristic score:

$$S_m = \mathrm{SD}_m \frac{\mathrm{PCC}_m}{\mathrm{PCC}_{o,m}} \tag{8.3}$$

where SD_m is the average standard deviation of the DNA methylation profiles of the CpGs making up the module m, as computed over independent samples (all within the same clinical disease stage), PCC_m denotes their average pairwise Pearson

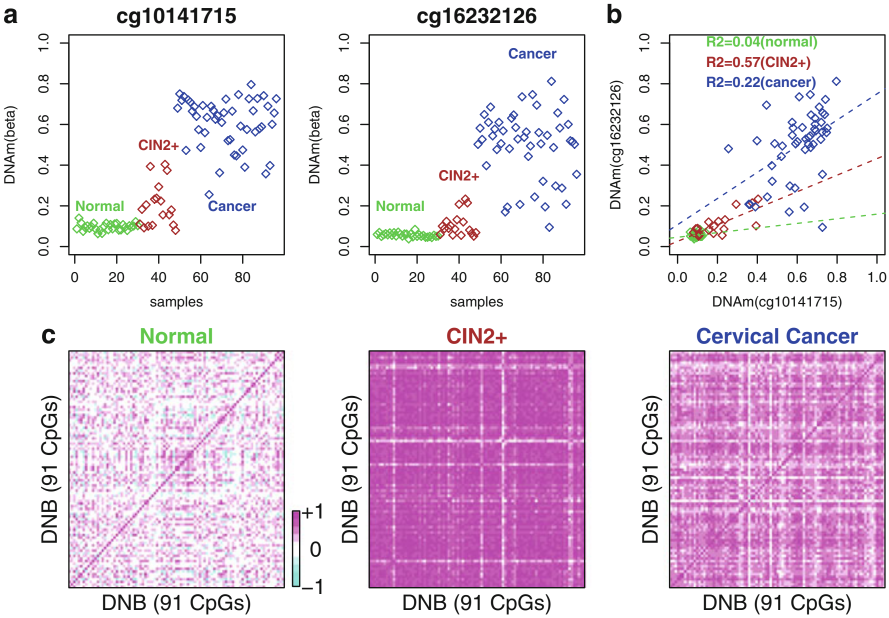


Fig. 8.5 (**a**) DNAm profiles of two risk CpG sites across three successive disease stages as indicated. Note the bimodality in the CIN2+ stage, in the sense that some CIN2+ samples show same methylation levels as the normals but others show increased levels. Also note how the DNAm levels of these CpGs progress to even higher levels in cervical cancer and that effectively all cancers show some level of increased methylation relative to the normal state. (**b**) Scatterplot of the DNAm levels of the same two CpGs depicted in (**a**) with samples coloured by disease stage. R^2 values of a linear model fitted to the samples of each disease stage is shown. This shows that the correlation is strongest for the CIN2+ stage, followed by cervical cancer and lowest in the normal state. (**c**) Covariation/correlation heatmaps of the 91 CpGs that make up the DNRB/DNB across the three successive disease stages as indicated, confirming that the covariation is highest in the CIN2+ stage, i.e. a stage prior to cervical cancer

correlation coefficient (as estimated across the same samples) and $\mathrm{PCC}_{o,m}$ denotes the average Pearson correlation between the module CpGs and their complement, i.e. all other CpGs not in the module m. By studying the "dynamic" changes of this score across successive disease stages and doing so for different candidate modules m, we posited that it would be possible to pinpoint a module m exhibiting a maximum in the disease stage immediately prior to cancer, thus allowing us to identify a set of "risk CpGs".

One immediate question is how to find the modules m? The algorithm is described in detail in Teschendorff et al. (2014). Briefly, one first uses differential methylation analysis to identify CpGs which are differentially methylated between a given disease stage and the baseline reference stage, i.e. the stage of normal physiology. A clustering algorithm is then applied on these selected CpGs to identify the main clusters, which then constitute our candidate modules of interest. Thus,

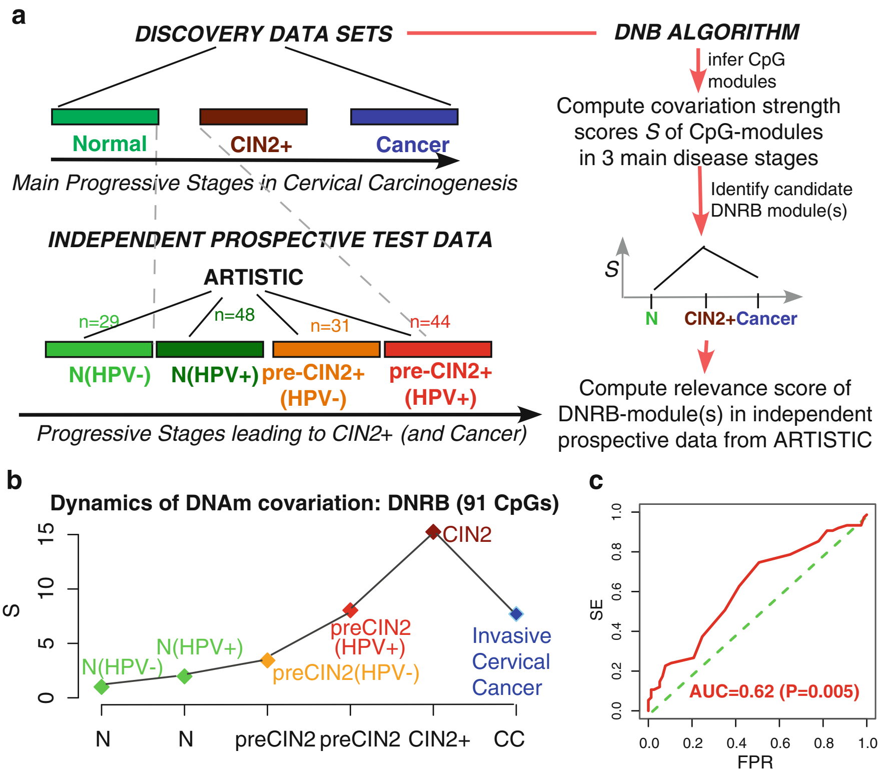


Fig. 8.6 (**a**) Flowchart of the strategy underlying the DNB algorithm. DNAm data sets representing the three major stages of carcinogenesis (here cervical carcinogenesis) are used to identify a set of candidate modules. Inspection of the dynamic change of the covariation strength of all modules may identify one or more modules with a maximum in the stage prior to cancer. These candidate risk predicting DNBs are then assessed in independent data, which should include prospective data to test whether the DNB can predict the risk of neoplastic transformation. (**b**) The dynamic changes in the covariation strength of the identified 91 DNRB CpGs, including the independent prospective data from ARTISTIC. Note how the intermediate disease stages prior to CIN2+ take on values which are intermediate between the normal and CIN2+ states, as required by consistency with the model. (**c**) Risk prediction ROC curve of the 91 Cp Gs of the DNB in the ARTISTIC data, confirming its ability to predict the prospective risk of CIN2+

for a given disease stage (i.e. CIN2+), we would obtain a number of modules, and similarly for cervical cancer, we would obtain another set of modules. For all of these modules, the covariation strength (as given by the above formula) can then be computed across all disease stages. Modules exhibiting nonlinear dynamics, exhibiting a maximum at a disease stage immediately prior to cervical cancer, represent candidate risk-indicating module(s), which we here call Dynamic Network Risk Biomarkers (DNRBs) (or also DNBs) (Fig. 8.6a).

It is important to note that because of the way modules are inferred, it is more likely that a module inferred from comparing normal to CIN2+ samples, say, will exhibit a maximum in the CIN2+ stage. It is therefore important to validate the observed covariation strengths of the modules in independent data representing the same disease stages (Teschendorff et al. 2014). However, even more importantly of course is to test the original hypothesis that a DNRB can indeed predict the risk of neoplastic transformation. This requires access to independently prospectively collected data. In the case of cervical cancer, such data is available from the ARTISTIC trial (Kitchener et al. 2009). Cervical smear samples, all of normal cytology, were obtained from a nested case-control subset consisting of 152 women, with 75 of these developing a CIN2+ at a 2nd round of screening performed 3 years later. The 152 cytologically normal samples can thus be divided into four "stage" groups: normal cytology at the 2nd round and HPV-free in the 1st round (N(HPV−)), normal cytology in the 2nd round but HPV+ in the 1st round (N(HPV+)), CIN2+ in the 2nd round but HPV− in the 1st round (preCIN2+(HPV−)) and CIN2+ in the 2nd round and HPV+ in the 1st round (preCIN2+(HPV+)) (Fig. 8.6a). Thus, for the candidate DNRB identified from the discovery study (which involved normal HPV−, CIN2+ samples and cervical cancer samples), we can compute the covariation strength in the four groups of the ARTISTIC data. The expectation would be that the covariation strength increases monotonically from the normal HPV− cells to the normal HPV+ cells and increases further in the normal cells which become CIN2+ 3 years later, which is indeed what is observed (Fig. 8.6b). Moreover, for the CpGs making up the DNRB, a risk index can be estimated, for instance, using an adaptive index algorithm as implemented in the EVORA model or by averaging their DNAm levels across the DNRB CpGs (Teschendorff et al. 2014). Doing so confirmed that the DNRB CpGs could discriminate the two groups of women (those remaining healthy and those developing CIN2+), that is, it predicts the prospective risk of CIN2+, with an AUC of 0.62 ($P = 0.005$) (Fig. 8.6c), similar to the AUC obtained using the EVORA algorithm (Teschendorff et al. 2012).

8.3.3 *Relation of the DNB Risk Prediction Framework to EVORA*

Although the DNB risk prediction model does not outperform EVORA, it is important to note that the DNB model allowed the identification of risk CpGs from analysing DNAm data of only three main disease stages (normal, CIN2+ and cancer), without requiring prospectively collected data, although the latter is of course needed to validate the risk prediction model. In fact, the 91 DNRB CpGs, which were derived from comparing established CIN2+ samples to normals from one cohort, overlapped strongly with the 140 risk CpGs derived using EVORA from the prospectively collected ARTISTIC data (Teschendorff et al. 2014). Although it remains to be seen whether the DNB model applies to other epithelial cancer types, one key advantage of the DNB framework is that it only relies on the covariation of DNAm between CpGs, and so, in principle, it does not rely on the existence of heterogeneous outlier profiles (as required by EVORA). This is an important point given that the epithelial cell of origin of most cancers is not easily accessible and that surrogate tissues, e.g. blood or buccal cells, are being considered for building risk prediction models (Anjum et al. 2014). In a surrogate tissue, it is less likely that heterogeneous outlier profiles would be present, especially in a complex tissue such as blood, which is composed of many different cell types and where changes in a specific cell subtype may be harder to detect. However, changes in the covariation patterns could be more easily seen.

8.4 Conclusions

In summary, in this chapter we have presented three systems-epigenomics strategies/models for addressing a variety of different challenges in the epigenomics of cancer and ageing. We presented the FEM algorithm, which was successful in identifying a causal epigenetic driver event (DNA methylation-induced silencing of the *HAND2* gene) in endometrial cancer, measurement of which in vaginal DNA swabs could provide an early detection test of high sensitivity and specificity (Jiao et al. 2014; Jones et al. 2013). The identification of *HAND2* and other tumour suppressors like *HIC5* was made possible by performing the integration of DNAm with gene expression data in the context of a human PPI network. It will be of great interest to apply such supervised functional network analyses in the context of other cancers and complex diseases. Other types of functional networks may also provide more powerful ways of integrating DNAm and gene expression data together, for instance, networks which start to incorporate low or even high-dimensional chromatin state information (Dixon et al. 2012; Ernst and Kellis 2012, 2013; Ernst et al. 2011).

We also described an approach to integrate genome-wide DNAm with ChIP-Seq data, allowing putative differential binding of TFs in development and complex diseases to be identified. Given that generation of high-quality ChIP-Seq data in a large number of clinical specimens is still challenging, the ability to detect disrupted regulatory networks in disease from analysing genome-wide DNA methylation data represents an attractive alternative. Of particular interest will be to extend these integrative models to include RNA-Seq data as well chromatin state and TF ChIP-Seq data generated in more relevant cell types.

Finally, we described a model for risk prediction of cancer, which was based on performing a systems-level analysis of dynamic DNAm changes happening during carcinogenesis. Studying the dynamics of covariation of DNAm patterns during cervical carcinogenesis allowed the identification of genomic loci, whose aberrant DNAm patterns in normal cells could predict their risk of future neoplastic transformation. This systems-level approach, focusing on genome-wide covariation in DNAm patterns, also supports the view that the disease stages prior to invasive cervical cancer are characterised by a high level of uncertainty/unpredictability in the DNAm levels one may observe across individual samples, which nevertheless also means that the covariation of specific CpG modules may be maximal in this disease stage. Further in-depth study of the genetic and epigenetic mosaicism in normal samples at stages prior to disease onset will be important to further elucidate the connections between intra-sample and inter-sample heterogeneity and how this varies across disease stages.

Acknowledgements AET is supported by the Chinese Academy of Sciences and the Max Planck Society.

References

Anjum S, Fourkala EO, Zikan M, Wong A, Gentry-Maharaj A, Jones A, Hardy R, Cibula D, Kuh D, Jacobs IJ, Teschendorff AE, Menon U, Widschwendter M. A BRCA1-mutation associated DNA methylation signature in blood cells predicts sporadic breast cancer incidence and survival. Genome Med. 2014;6(6):47.

Beck S. Taking the measure of the methylome. Nat Biotechnol. 2010;28(10):1026–8.

Beck S, Ng T. C2C: turning cancer into chronic disease. Genome Med. 2014;6(5):38.

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

Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K, Jaenisch R, Wagschal A, Feil R, Schreiber SL, Lander ES. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell. 2006;125(2):315–26.

Blattler A, Farnham PJ. Cross-talk between site-specific transcription factors and DNA methylation states. J Biol Chem. 2013;288(48):34,287–94.

Budden DM, Hurley DG, Cursons J, Markham JF, Davis MJ, Crampin EJ. Predicting expression: the complementary power of histone modification and transcription factor binding data. Epigenetics Chromatin. 2014;7(1):36.

Cerami EG, Gross BE, Demir E, Rodchenkov I, Babur O, Anwar N, Schultz N, Bader GD, Sander C. Pathway commons, a web resource for biological pathway data. Nucleic Acids Res. 2011;39(Database):D685–90.

Chin K, DeVries S, Fridlyand J, Spellman PT, Roydasgupta R, Kuo WL, Lapuk A, Neve RM, Qian Z, Ryder T, Chen F, Feiler H, Tokuyasu T, Kingsley C, Dairkee S, Meng Z, Chew K, Pinkel D, Jain A, Ljung BM, Esserman L, Albertson DG, Waldman FM, Gray, J.W. Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. Cancer Cell. 2006;10(6):529–41.

Chin SF, Teschendorff AE, Marioni JC, Wang Y, Barbosa-Morais NL, Thorne NP, Costa JL, Pinder SE, van de Wiel MA, Green AR, Ellis IO, Porter PL, Tavaré S, Brenton JD, Ylstra B, Caldas C. High-resolution aCGH and expression profiling identifies a novel genomic subtype of ER negative breast cancer. Genome Biol. 2007;8(10):R215.

Chuang HY, Lee E, Liu YT, Lee D, Ideker T. Network-based classification of breast cancer metastasis. Mol Syst Biol. 2007;3:140.

Ciriello G, Cerami E, Sander C, Schultz N. Mutual exclusivity analysis identifies oncogenic network modules. Genome Res. 2012;22(2):398–406.

Crystal AS, Shaw AT, Sequist LV, Friboulet L, Niederst MJ, Lockerman EL, Frias RL, Gainor JF, Amzallag A, Greninger P, Lee D, Kalsy A, Gomez-Caraballo M, Elamine L, Howe E, Hur W, Lifshits E, Robinson HE, Katayama R, Faber AC, Awad MM, Ramaswamy S, Mino-Kenudson

M, Iafrate AJ, Benes CH, Engelman JA. Patient-derived models of acquired resistance can identify effective drug combinations for cancer. Science. 2014;346(6216):1480–6.
Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y, Gräf S, Ha G, Haffari G, Bashashati A, Russell R, McKinney S, Langerod A, Green A, Provenzano E, Wishart G, Pinder S, Watson P, Markowetz F, Murphy L, Ellis I, Purushotham A, Borresen-Dale AL, Brenton JD, Tavare S, Caldas C, Aparicio S. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012;486(7403):346–52.
Deaton AM, Bird A. CpG islands and the regulation of transcription. Genes Dev. 2011;25(10):1010–22.
Dedeurwaerder S, Desmedt C, Calonne E, Singhal SK, Haibe-Kains B, Defrance M, Michiels S, Volkmar M, Deplus R, Luciani J, Lallemand F, Larsimont D, Toussaint J, Haussy S, Rothe F, Rouas G, Metzger O, Majjaj S, Saini K, Putmans P, Hames G, van Baren N, Coulie PG, Piccart M, Sotiriou C, Fuks F. DNA methylation profiling reveals a predominant immune component in breast cancers. EMBO Mol Med. 2011;3(12):726–41.
Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 2012;485(7398):376–80.
Dumanski JP, Rasi C, Lonn M, Davies H, Ingelsson M, Giedraitis V, Lannfelt L, Magnusson PK, Lindgren CM, Morris AP, Cesarini D, Johannesson M, Tiensuu-Janson E, Lind L, Pedersen NL, Ingelsson E, Forsberg LA. Mutagenesis smoking is associated with mosaic loss of chromosome Y. Science. 2015;347(6217):81–3.
Eckhardt F, Lewin J, Cortese R, Rakyan VK, Attwood J, Burger M, Burton J, Cox TV, Davies R, Down TA, Haefliger C, Horton R, Howe K, Jackson DK, Kunde J, Koenig C, Liddle J, Niblett D, Otto T, Pettett R, Seemann S, Thompson C, West T, Rogers J, Olek A, Berlin K, Beck S. DNA methylation profiling of human chromosomes 6, 20 and 22. Nat Genet. 2006;38(12):1378–85.
Ernst J, Kellis M. Chromhmm: automating chromatin-state discovery and characterization. Nat Methods. 2012;9(3):215–6.
Ernst J, Kellis M. Interplay between chromatin state, regulator binding, and regulatory motifs in six human cell types. Genome Res. 2013;23(7):1142–54.
Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, Ku M, Durham T, Kellis M, Bernstein BE. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature. 2011;473(7345):43–9.
Feinberg AP, Irizarry RA. Evolution in health and medicine sackler colloquium: stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. Proc Natl Acad Sci U S A. 2010;107:1757–64.
Feinberg AP, Ohlsson R, Henikoff S. The epigenetic progenitor origin of human cancer. Nat Rev Genet. 2006;7(1):21–33.
Forsberg LA, Rasi C, Malmqvist N, Davies H, Pasupulati S, Pakalapati G, Sandgren J, de Diaz S, Zaghlool A, Giedraitis V, Lannfelt L, Score J, Cross NC, Absher D, Janson ET, Lindgren CM, Morris AP, Ingelsson E, Lind L, Dumanski JP. Mosaic loss of chromosome Y in peripheral blood is associated with shorter survival and higher risk of cancer. Nat Genet. 2014;46(6):624–8.
Genovese G, Kahler AK, Handsaker RE, Lindberg J, Rose SA, Bakhoum SF, Chambert K, Mick E, Neale BM, Fromer M, Purcell SM, Svantesson O, Landen M, Hoglund M, Lehmann S, Gabriel SB, Moran JL, Lander ES, Sullivan PF, Sklar P, Gronberg H, Hultman CM, McCarroll SA. Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. N Engl J Med. 2014;371(26):2477–87.
Gerlinger M, Rowan AJ, Horswell S, Larkin J, Endesfelder D, Gronroos E, Martinez P, Matthews N, Stewart A, Tarpey P, Varela I, Phillimore B, Begum S, McDonald NQ, Butler A, Jones D, Raine K, Latimer C, Santos CR, Nohadani M, Eklund AC, Spencer-Dene B, Clark G,

Pickering L, Stamp G, Gore M, Szallasi Z, Downward J, Futreal PA, Swanton C. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N Engl J Med. 2012;366(10):83–92.

Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, Cheng C, Mu XJ, Khurana E, Rozowsky J, Alexander R, Min R, Alves P, Abyzov A, Addleman N, Bhardwaj N, Boyle AP, Cayting P, Charos A, Chen DZ, Cheng Y, Clarke D, Eastman C, Euskirchen G, Frietze S, Fu Y, Gertz J, Grubert F, Harmanci A, Jain P, Kasowski M, Lacroute P, Leng J, Lian J, Monahan H, O'Geen H, Ouyang Z, Partridge EC, Patacsil D, Pauli F, Raha D, Ramirez L, Reddy TE, Reed B, Shi M, Slifer T, Wang J, Wu L, Yang X, Yip KY, Zilberman-Schapira G, Batzoglou S, Sidow A, Farnham PJ, Myers RM, Weissman SM, Snyder M. Architecture of the human regulatory network derived from encode data. Nature. 2012;489(7414):91–100.

Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sadda S, Klotzle B, Bibikova M, Fan JB, Gao Y, Deconde R, Chen M, Rajapakse I, Friend S, Ideker T, Zhang K. Genome-wide methylation profiles reveal quantitative views of human aging rates. Mol Cell. 2013;49(2):359–67.

Hood L, Friend SH. Predictive, personalized, preventive, participatory (P4) cancer medicine. Nat Rev Clin Oncol. 2011;8(3):184–7.

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

Horvath S, Zhang Y, Langfelder P, Kahn RS, Boks MP, van Eijk K, van den Berg LH, Ophoff RA. Aging effects on DNA methylation modules in human brain and blood tissue. Genome Biol. 2012;13(10):R97.

Hu S, Wan J, Su Y, Song Q, Zeng Y, Nguyen HN, Shin J, Cox E, Rho HS, Woodard C, Xia S, Liu S, Lyu H, Ming GL, Wade H, Song H, Qian J, Zhu H. DNA methylation presents distinct binding sites for human transcription factors. Elife. 2013;2:e00,726.

Huehne R, Thalheim T, Suehnel J. Agefactdb–the jenage ageing factor database–towards data integration in ageing research. Nucleic Acids Res. 2014;42:D892–6.

Irizarry RA, Ladd-Acosta C, Wen B, Wu Z, Montano C, Onyango P, Cui H, Gabo K, Rongione M, Webster M, Ji H, Potash JB, Sabunciyan S, Feinberg AP. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nat Genet. 2009;41(2):178–86.

Issa JP. Epigenetic variation and cellular darwinism. Nat Genet. 2011;43(8):724–26.

Jacobs KB, Yeager M, Zhou W, Wacholder S, Wang Z, Rodriguez-Santiago B, Hutchinson A, Deng X, Liu C, Horner MJ, Cullen M, Epstein CG, Burdett L, Dean MC, Chatterjee N, Sampson J, Chung CC, Kovaks J, Gapstur SM, Stevens VL, Teras LT, Gaudet MM, Albanes D, Weinstein SJ, Virtamo J, Taylor PR, Freedman ND, Abnet CC, Goldstein AM, Hu N, Yu K, Yuan JM, Liao L, Ding T, Qiao YL, Gao YT, Koh WP, Xiang YB, Tang ZZ, Fan JH, Aldrich MC, Amos C, Blot WJ, Bock CH, Gillanders EM, Harris CC, Haiman CA, Henderson BE, Kolonel LN, Le Marchand L, McNeill LH, Rybicki BA, Schwartz AG, Signorello LB, Spitz MR, Wiencke JK, Wrensch M, Wu X, Zanetti KA, Ziegler RG, Figueroa JD, Garcia-Closas M, Malats N, Marenne G, Prokunina-Olsson L, Baris D, Schwenn M, Johnson A, Landi MT, Goldin L, Consonni D, Bertazzi PA, Rotunno M, Rajaraman P, Andersson U, Beane FLE, Berg CD, Buring JE, Butler MA, Carreon T, Feychting M, Ahlbom A, Gaziano JM, Giles GG, Hallmans G, Hankinson SE, Hartge P, Henriksson R, Inskip PD, Johansen C, Landgren A, McKean-Cowdin R, Michaud DS, Melin BS, Peters U, Ruder AM, Sesso HD, Severi G, Shu XO, Visvanathan K, White E, Wolk A, Zeleniuch-Jacquotte A, Zheng W, Silverman DT, Kogevinas M, Gonzalez JR, Villa O, Li D, Duell EJ, Risch HA, Olson SH, Kooperberg C, Wolpin BM, Jiao L, Hassan M, Wheeler W, Arslan AA, de Mesquita HBB, Fuchs CS, Gallinger S, Gross MD, Holly EA, Klein AP, LaCroix A, Mandelson MT, Petersen G, Boutron-Ruault MC, Bracci PM, Canzian F, Chang K, Cotterchio M, Giovannucci EL, Goggins M, Hoffman BJA, Jenab M, Khaw KT, Krogh V, Kurtz RC, McWilliams RR, Mendelsohn JB, Rabe KG, Riboli E, Tjonneland A, Tobias GS, Trichopoulos D, Elena JW, Yu H, Amundadottir L, Stolzenberg-Solomon RZ, Kraft P, Schumacher F, Stram D, Savage SA, Mirabello L, Andrulis IL, Wunder JS, Patino GA, Sierrasesumaga L, Barkauskas DA, Gorlick RG, Purdue M, Chow WH, Moore LE, Schwartz KL, Davis FG, Hsing AW, Berndt SI, Black A, Wentzensen N, Brinton LA, Lissowska J, Peplonska B, McGlynn KA, Cook MB, Graubard BI, Kratz CP, Greene MH, Erickson RL, Hunter DJ, Thomas G, Hoover RN, Real FX, Fraumeni JF, Caporaso NE, Tucker M, Rothman N, Perez-Jurado LA, Chanock SJ. Detectable clonal mosaicism and its relationship to aging and cancer. Nat Genet. 2012;44(6):651–8.

Jiao Y, Widschwendter M, Teschendorff AE. A systems-level integrative framework for genome-wide DNA methylation and gene expression data identifies differential gene expression modules under epigenetic control. Bioinformatics. 2014;30(16):2360–6.

Jones A, Teschendorff AE, Li Q, Hayward JD, Kannan A, Mould T, West J, Zikan M, Cibula D, Fiegl H, Lee SH, Wik E, Hadwin R, Arora R, Lemech C, Turunen H, Pakarinen P, Jacobs IJ, Salvesen HB, Bagchi MK, Bagchi IC, Widschwendter M. Role of DNA methylation and epigenetic silencing of HAND2 in endometrial cancer development. PLoS Med. 2013;10:e1001,551.

Kandoth C, Schultz N, Cherniack AD, Akbani R, Liu Y, Shen H, Robertson AG, Pashtan I, Shen R, Benz CC, Yau C, Laird PW, Ding L, Zhang W, Mills GB, Kucherlapati R, Mardis ER, Levine DA. Integrated genomic characterization of endometrial carcinoma. Nature. 2013;497(7447):67–73.

Karlic R, Chung HR, Lasserre J, Vlahovicek K, Vingron M. Histone modification levels are predictive for gene expression. Proc Natl Acad Sci U S A. 2010;107(7):2926–31.

Kitchener HC, Almonte M, Thomson C, Wheeler P, Sargent A, Stoykova B, Gilham C, Baysson H, Roberts C, Dowie R, Desai M, Mather J, Bailey A, Turner A, Moss S, Peto J. HPV testing in combination with liquid-based cytology in primary cervical screening (artistic): a randomised controlled trial. Lancet Oncol. 2009;10(7):672–82.

Laurie CC, Laurie CA, Rice K, Doheny KF, Zelnick LR, McHugh CP, Ling H, Hetrick KN, Pugh EW, Amos C, Wei Q, Wang LE, Lee JE, Barnes KC, Hansel NN, Mathias R, Daley D, Beaty TH, Scott AF, Ruczinski I, Scharpf RB, Bierut LJ, Hartz SM, Landi MT, Freedman ND, Goldin LR, Ginsburg D, Li J, Desch KC, Strom SS, Blot WJ, Signorello LB, Ingles SA, Chanock SJ, Berndt SI, Marchand LL, Henderson BE, Monroe KR, Heit JA, de Andrade M, Armasu SM, Regnier C, Lowe WL, Hayes MG, Marazita ML, Feingold E, Murray JC, Melbye M, Feenstra B, Kang JH, Wiggs JL, Jarvik GP, McDavid AN, Seshan VE, Mirel DB, Crenshaw A, Sharopova N, Wise A, Shen J, Crosslin DR, Levine DM, Zheng X, Udren JI, Bennett S, Nelson SC, Gogarten SM, Conomos MP, Heagerty P, Manolio T, Pasquale LR, Haiman CA, Caporaso N, Weir BS. Detectable clonal mosaicism from birth to old age and its relationship to cancer. Nat Genet. 2012;44(6):642–50.

Lee TI, Jenner RG, Boyer LA, Guenther MG, Levine SS, Kumar RM, Chevalier B, Johnstone SE, Cole MF, Isono K, Koseki H, Fuchikami T, Abe K, Murray HL, Zucker JP, Yuan B, Bell GW, Herbolsheimer E, Hannett NM, Sun K, Odom DT, Otte AP, Volkert TL, Bartel DP, Melton DA, Gifford DK, Jaenisch R, Young RA. Control of developmental regulators by polycomb in human embryonic stem cells. Cell. 2006;125(2):301–13.

Liu H, Su J, Li J, Liu H, Lv J, Li B, Qiao H, Zhang Y. Prioritizing cancer-related genes with aberrant methylation based on a weighted protein-protein interaction network. BMC Syst Biol. 2011;5:158.

Lock EF, Hoadley KA, Marron JS, Nobel AB. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. Ann Appl Stat. 2013;7(1):523–42.

Lu T, Aron L, Zullo J, Pan Y, Kim H, Chen Y, Yang TH, Kim HM, Drake D, Liu XS, Bennett DA, Colaiacovo MP, Yankner BA. Rest and stress resistance in ageing and Alzheimer's disease. Nature. 2014;507(7493):448–54.

Lui JC, Chen W, Cheung CS, Baron J. Broad shifts in gene expression during early postnatal life are associated with shifts in histone methylation patterns. PLoS One. 2014;9(1):e86,957.

Maegawa S, Hinkal G, Kim HS, Shen L, Zhang L, Zhang J, Zhang N, Liang S, Donehower LA, Issa JP. Widespread and tissue specific age-related DNA methylation changes in mice. Genome Res. 2010;20(3):332–40.

Maley CC, Galipeau PC, Finley JC, Wongsurawat VJ, Li X, Sanchez CA, Paulson TG, Blount PL, Risques RA, Rabinovitch PS, Reid BJ. Genetic clonal diversity predicts progression to esophageal adenocarcinoma. Nat Genet. 2006;38(4):468–73.

Mitra K, Carvunis AR, Ramesh SK, Ideker T. Integrative approaches for finding modular structure in biological networks. Nat Rev Genet. 2013;14(10):719–32.

Pisco AO, Brock A, Zhou J, Moor A, Mojtahedi M, Jackson D, Huang, S.: Non-darwinian dynamics in therapy-induced cancer drug resistance. Nat Commun. 2013;4:2467.

Prasad TS, Kandasamy K, Pandey A. Human protein reference database and human proteinpedia as discovery tools for systems biology. Methods Mol Biol. 2009;577:67–79.

Rakyan VK, Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. Nat Rev Genet. 2011;12(8):529–41.

Rakyan VK, Down TA, Maslau S, Andrew T, Yang TP, Beyan H, Whittaker P, McCann OT, Finer S, Valdes AM, Leslie RD, Deloukas P, Spector TD. Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains. Genome Res. 2010;20(4):434–9.

Reichardt J, Bornholdt S. Statistical mechanics of community detection. Phys Rev E. 2006;74:016,110.

Rolland T, Tasan M, Charloteaux B, Pevzner SJ, Zhong Q, Sahni N, Yi S, Lemmens I, Fontanillo C, Mosca R, Kamburov A, Ghiassian SD, Yang X, Ghamsari L, Balcha D, Begg BE, Braun P, Brehme M, Broly MP, Carvunis AR, Convery-Zupan D, Corominas R, Coulombe-Huntington J, Dann E, Dreze M, Dricot A, Fan C, Franzosa E, Gebreab F, Gutierrez BJ, Hardy MF, Jin M, Kang S, Kiros R, Lin GN, Luck K, MacWilliams A, Menche J, Murray RR, Palagi A, Poulin MM, Rambout X, Rasla J, Reichert P, Romero V, Ruyssinck E, Sahalie JM, Scholz A, Shah AA, Sharma A, Shen Y, Spirohn K, Tam S, Tejeda AO, Trigg SA, Twizere JC, Vega K, Walsh J, Cusick ME, Xia Y, Barabasi AL, Iakoucheva LM, Aloy P, De Las RJ, Tavernier J, Calderwood MA, Hill DE, Hao T, Roth FP, Vidal M. A proteome-scale map of the human interactome network. Cell. 2014;159(5):1212–26.

Ruscio AD, Ebralidze AK, Benoukraf T, Amabile G, Goff LA, Terragni J, Figueroa ME, de Figueiredo Pontes LL, Alberich-Jorda M, Zhang P, Wu M, D'Alo F, Melnick A, Leone G, Ebralidze KK, Pradhan S, Rinn JL, Tenen DG. DNMT1-interacting RNAs block gene-specific DNA methylation. Nature. 2013;503(7476):371–6.

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

Shen R, Mo Q, Schultz N, Seshan VE, Olshen AB, Huse J, Ladanyi M, Sander C. Integrative subtype discovery in glioblastoma using icluster. PLoS One. 2012;7(4):e35,236.

Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics. 2009;25(22):2906–12.

Smyth GK, Yang YH, Speed TP. Statistical issues in microarray data analysis. In: Brownstein MJ, Khodursky AB. editors. Functional genomics: methods and protocols. Methods in molecular biology, vol 224. Totowa: Humana Press; 2003. p. 111–36.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102(43):15,545–50.

Tate PH, Bird AP. Effects of DNA methylation on DNA-binding proteins and gene expression. Curr Opin Genet Dev. 1993;3(2):226–31.

Teschendorff AE, Gomez S, Arenas A, El-Ashry D, Schmidt M, Gehrmann M, Caldas C. Improved prognostic classification of breast cancer defined by antagonistic activation patterns of immune response pathway modules. BMC Cancer. 2010;10:604.

Teschendorff AE, Jones A, Fiegl H, Sargent A, Zhuang JJ, Kitchener HC, Widschwendter M. Epigenetic variability in cells of normal cytology is associated with the risk of future morphological transformation. Genome Med. 2012;4(3):24.

Teschendorff AE, Liu X, Caren H, Pollard SM, Beck S, Widschwendter M, Chen L. The dynamics of DNA methylation covariation patterns in carcinogenesis. PLoS Comput Biol. 2014;10(7):e1003,709.

Teschendorff AE, Menon U, Gentry-Maharaj A, Ramus SJ, Weisenberger DJ, Shen H, Campan M, Noushmehr H, Bell CG, Maxwell AP, Savage DA, Mueller-Holzner E, Marth C, Kocjan

G, Gayther SA, Jones A, Beck S, Wagner W, Laird PW, Jacobs IJ, Widschwendter M. Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. Genome Res. 2010;20(4):440–6.

Teschendorff AE, Miremadi A, Pinder SE, Ellis IO, Caldas C. An immune response gene expression module identifies a good prognosis subtype in estrogen receptor negative breast cancer. Genome Biol. 2007;8(8):R157.

Teschendorff AE, West J, Beck S. Age-associated epigenetic drift: implications, and a case of epigenetic thrift? Hum Mol Genet. 2013;22(NA):R7–15.

Teschendorff AE, Widschwendter M. Differential variability improves the identification of cancer risk markers in DNA methylation studies profiling precursor cancer lesions. Bioinformatics. 2012;28(11):1487–94.

Teschendorff AE, Widschwendter M. A network systems approach to identify functional epigenetic drivers in cancer. In: Shen B. editor. Bioinformatics for diagnosis, prognosis and treatment of complex diseases. Beijing: Springer; 2014.

Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, Garg K, John S, Sandstrom R, Bates D, Boatman L, Canfield TK, Diegel M, Dunn D, Ebersol AK, Frum T, Giste E, Johnson AK, Johnson EM, Kutyavin T, Lajoie B, Lee BK, Lee K, London D, Lotakis D, Neph S, Neri F, Nguyen ED, Qu H, Reynolds AP, Roach V, Safi A, Sanchez ME, Sanyal A, Shafer A, Simon JM, Song L, Vong S, Weaver M, Yan Y, Zhang Z, Zhang Z, Lenhard B, Tewari M, Dorschner MO, Hansen RS, Navas PA, Stamatoyannopoulos G, Iyer VR, Lieb JD, Sunyaev SR, Akey JM, Sabo PJ, Kaul R, Furey TS, Dekker J, Crawford GE, Stamatoyannopoulos JA. The accessible chromatin landscape of the human genome. Nature. 2012;489(7414):75–82.

Tian Y, Jiao Y, de Jong S, Ophoff RA, Beck S, Teschendorff AE. An integrative multi-scale analysis of the dynamic DNA methylation landscape in aging. PLoS Genet. 2015;11(2):e1004996. doi: 10.1371/journal.pgen.1004996. eCollection 2015 Feb.

Timp W, Levchenko A, Feinberg AP. A new link between epigenetic progenitor lesions in cancer and the dynamics of signal transduction. Cell Cycle. 2009;8(3):383–90.

Vanderkraats ND, Hiken JF, Decker KF, Edwards JR. Discovering high-resolution patterns of differential DNA methylation that correlate with gene expression changes. Nucleic Acids Res. 2013;41(14):6816–27.

Vermeulen M. Making the most of methylation. Elife. 2013;2:e01,387.

Walter D, Matter A, Fahrenkrog B. Loss of histone H3 methylation at lysine 4 triggers apoptosis in saccharomyces cerevisiae. PLoS Genet. 2014;10(1):e1004,095.

Wang HQ, Zheng CH, Zhao XM. jNMFMA: a joint non-negative matrix factorization meta-analysis of transcriptomics data. Bioinformatics. 2015;31(4):572–80. doi: 10.1093/bioinformatics/btu679. Epub 2014 Oct 16.

West J, Beck S, Wang X, Teschendorff AE. An integrative network algorithm identifies age-associated differential methylation interactome hotspots targeting stem-cell differentiation pathways. Sci Rep. 2013;3:1630.

West J, Widschwendter M, Teschendorff AE. Distinctive topology of age-associated epigenetic drift in the human interactome. Proc Natl Acad Sci U S A. 2013;110(35):14,138–43.

Witten DM, Tibshirani R, Hastie T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics. 2009;10(3):515–34.

Witten DM, Tibshirani RJ. Extensions of sparse canonical correlation analysis with applications to genomic data. Stat Appl Genet Mol Biol. 2009;8:Article28.

Xie M, Lu C, Wang J, McLellan MD, Johnson KJ, Wendl MC, McMichael JF, Schmidt HK, Yellapantula V, Miller CA, Ozenberger BA, Welch JS, Link DC, Walter MJ, Mardis ER, Dipersio JF, Chen F, Wilson RK, Ley TJ, Ding L. Age-related mutations associated with clonal hematopoietic expansion and malignancies. Nat Med. 2014;20(12):1472–8.

Xu Z, Bolick SC, Deroo LA, Weinberg CR, Sandler DP, Taylor JA. Epigenome-wide association study of breast cancer using prospectively collected sister study samples. J Natl Cancer Inst. 2013;105(10):694–700

Ziller MJ, Gu H, Muller F, Donaghey J, Tsai LT, Kohlbacher O, Jager PD, Rosen ED, Bennett DA, Bernstein BE, Gnirke A, Meissner A. Charting a dynamic DNA methylation landscape of the human genome. Nature. 2013;500(7463):477–81.

Chapter 9
Epigenomic Biomarkers for the Advance of Personalized Medicine

Jesus Mendez-Gonzalez and Juan Sandoval

Abstract Epigenetic factors (DNA methylation, histone modifications, or ncRNAs) are involved in gene expression regulation. Thus, determining abnormal epigenetic changes is a suitable approach to extract meaningful information about human diseases. An altered pattern of epigenetic modifications has been firstly defined as a hallmark for cancer, although it is also a key element to many common human diseases, such as cardiovascular, metabolic, and neurological pathologies. During the last decade, the advent of genome-scale analysis techniques applied to epigenetics has provided a massive amount of data, enabling an important advance in the molecular mechanisms underlying disease initiation, progression, and expansion. Disease-specific epigenomic signatures, mainly based on DNA methylation analysis, have been studied for several clinical purposes including prognostics and diagnostics, as well as disease-specific chemotherapy response. Using noninvasive specimens, epigenetic profiling holds the promise of being of clinical value in the management of patients, even at the early stages of disease. Additionally, epigenetic marks have also been catalogued as targets for pharmacological drugs. The upgrade of epigenetic research to epigenomics together with other –omics would tackle the many unanswered questions in the field, paving the path to achieve a more precise personalized medicine.

Keywords Epigenetics • DNA methylation • Lung cancer • Biomarkers • Personalized medicine

---

J. Mendez-Gonzalez
Cancer Epigenetics and Biology Program (PEBC), Biomedical Research Institute of Bellvitge (IDIBELL), Hospital Duran i Reynals Av. Gran Via 199, 08907 – L'Hospitalet de Llobregat, Barcelona, Spain

J. Sandoval, PhD (✉)
Laboratory of Personalized Medicine, Genomic Unit, Instituto de Investigacion Sanitaria La Fe Torre A, Sótano. Avda. Fernando Abril Martorell, 106, 46026 – Valencia, Spain
e-mail: juan.sandoval@uv.es

A.E. Teschendorff (eds.), *Computational and Statistical Epigenomics*,
Translational Bioinformatics 7, DOI 10.1007/978-94-017-9927-0_9

9.1 Introduction

Nowadays, the high grade of complexity of the biology and the influence of environment in the etiology and development of the majority of diseases are well known. However, despite extended research focused on different aspects, including genetic defects, a contributing factor has been missed for many years. In this sense, epigenetics has been hailed as a molecular transductor of environmental exposures or genetics and many common diseases (Cortessis et al. 2012). Thus, lifestyle, stress, drugs, physiopathological situations, and pharmacological interventions have a great impact on the epigenetic code of the cells by altering their epigenetic factors. Epigenetics has emerged as a promising field in recent years adding a new layer of complexity. It is one of the most rapidly developing fields of biological research and has become a key factor to complete the whole picture of biology complexity.

Despite constant effort in research, there is an urgent necessity to find biomarkers for improving diagnosis and therapeutic intervention for several diseases. Research on genetic and epigenetic biomarkers are fundamental tools in human healthcare with numerous supporting studies. Epigenetic research for molecular biomarkers encourages the translation of this field from the bench to the clinical practice. In this sense, uncovering and deciphering the locations and timing of intricate epigenetic changes in several molecular processes, mainly human malignancies, is a new challenge for near future.

Breakthroughs in technologies have enabled the possibility of genome-wide epigenetic research and revolutionized the concept of analyzing in an unbiased manner. This recent groundbreaking technological innovation has enormously increased the data available for assessing epigenetic features of human biology and disease. High numbers of novel epigenetic biomarkers are and will be discovered in present and near future. Therefore, a titanic effort of credentialing and independently validating these biomarkers for using with patients should be also carried out. The incorporation of this patient level predictor information will help to increase healthcare efficiency by implementing effectively individualized treatment strategies.

In this chapter, we shortly describe the high throughput technologies currently available for epigenomic studies and present epigenetic features with potential biomarker value. In addition, we describe recent and most promising epigenomic biomarkers in some of the most prevalent types of tumors such as lung, colorectal, and breast cancer. Finally, we describe the most recent advances in DNA methylation profiling in other entities, such as neurological and cardiovascular diseases.

9.2 Personalized Medicine

Personalized medicine has been a long-awaited promise that currently is starting to become a reality. This new concept will bring along changes in the practice of medicine and the dynamics of drug development, leading to new healthcare economic models. Hopefully, it will facilitate medical prevention and optimal therapy selection, improving patient's quality of life and reducing overall cost of public healthcare. However, further progress will require a team effort from different players of healthcare members.

Personalized medicine determines the unique molecular characteristics of a patient for diagnosing more finely a disease or predicting an individual's susceptibility before clinical signs and symptoms appear. Besides, personalized medicine has also been defined as therapy decisions tailored to individual patients, aiming to improve therapeutic efficiencies and to minimize side effects. The current clinical practice includes limited targeted therapies for stratified populations of patients who have a greater likelihood of responding based on molecular biomarker information. In many areas, the clinical interventions can be life saving. The advent of high throughput screening technologies has enabled more comprehensive identification strategies and suggests a plethora of new valuable biomarkers and druggable molecules for future clinical applications (Heyn et al. 2013a). In this sense, epigenetic biomarkers have emerged into the field as promising valuable entities, especially in cancer context.

9.3 Epigenetics Definition and Factors

A first definition of epigenetics was proposed by Conrad Waddington in 1942 (Waddington 1942) as the study of how genotypes give rise to phenotypes through programmed changes during development. New concepts were subsequently added to this original definition. Then, nowadays, epigenetics may be defined as the study of heritable changes in gene expression that are not due to changes in the primary DNA sequence, being essential for gene transcription, development, and differentiation of cells and organisms. There are basically three types of mechanistic layers in the field of epigenetics: post-translational modifications of histone proteins, DNA methylation, and noncoding RNAs. During the last two decades the best studied epigenetic process has been DNA methylation, although the rest are an emerging field that promises auspicious results in a near future.

9.3.1 DNA Methylation

DNA methylation is a covalent chemical modification, resulting in the addition of a methyl group at the carbon 5 position of the cytosine ring in the DNA that is essential for the correct development of the organisms. The adding of methyl groups from S-adenosylmethionine is catalyzed by specific enzymes called DNA methyltransferases. The deregulation of specific methylated genes and aberrant methylation profiles has been associated with several diseases, including cancer (Esteller 2008). Interestingly, CpG sites are underrepresented and unequally distributed across the human genome, giving rise to vast low-density CpG regions interspersed with CpG clusters located mainly in denominated CpG islands (Sandoval and Esteller 2012). From a clinical point of view, the application of DNA methylation for diagnostic purposes entails several advantages when compared with other type of biomarkers, such as genetic mutations, or gene expression profiles. In contrast to gene expression-based biomarkers, alterations in DNA methylation are mainly found in exact regions (CpG islands) and can be detected by a wide range of sensitive and cost-efficient techniques (Mikeska et al. 2012). Besides, DNA methylation is a stable mark that is not easily altered and therefore does not vary in response to external stimuli, unlike gene expression patterns, where larger cohorts are needed (Kratz et al. 2012).

9.3.2 Histone Modifications

Histones and 146 bp of wrapped DNA around the core of histones form the tridimensional basic particle that generates the nucleosome. Histones can undergo multiple post-translational covalent modifications leading to either activation or repression, depending upon which amino acids are modified, and the type and number of the modifications presents (Sharma et al. 2010). Evidence of the implication of histone modifications in some diseases, and the identification of histone variants has increased the enthusiasm for investigating the use of histones as biomarkers.

9.3.3 MicroRNAs

MicroRNAs (miRNAs) are a large family of short noncoding RNAs involved in many biological processes, such as cellular development, differentiation, apoptosis, and proliferation (He and Hannon 2004), and also in tumor evolution, metastatic dissemination, and resistance/sensitivity to therapy (Garzon et al. 2010). Since miRNAs play diverse roles in a broad spectrum of biological processes, present enhanced stability, and that specific miRNA profiles have been identified in several malignancies, these small molecules have been proposed as potential biomarkers. The identification of these miRNA biomarkers will facilitate the development of new diagnostic, prognostic, and therapeutic applications (Gargalionis and Basdra 2013).

9.3.4 Interplay Between Epigenetic Factors

The entire epigenetic machinery acts together and interconnected to ensure the correct chromatin conformation and levels of accessibility so that normal levels of gene expression are eventually achieved. MicroRNA expression can be affected by other epigenetic factors (DNA methylation and histone modifications), while a subgroup of small noncodifying RNA molecules (called epi-miRNAs) can also directly or indirectly regulate the expression of components of the epigenetic machinery, such as DNA methyltransferases (DNMTs) or histone deacetylases (HDACs), creating a highly controlled feedback mechanism (Iorio et al. 2010). Genetic mechanisms also affect epigenetic effectors and vice versa. Therefore, an altered balance between the key epigenetic players and genetic factors is associated with human malignancies.

9.4 Epigenomics (from Single to Genome-Wide Strategies)

In the last century, before genome-scale analysis techniques were possible, epigenetic biomarkers associated with diseases were sought using a candidate-gene approach. The appearance of bisulfite conversion was crucial for the development of DNA methylation research. Basically, 5-methyl cytosines are protected to deamination by sodium bisulfite in contrast to cytosines that are converted to uracil. Consequently, the product of this reaction was coupled to PCR (methyl-specific PCR (MSP)). Nowadays, there are different secondary strategies: such as sequencing-based technologies (bisulfite genomic sequencing and pyrosequencing) or MALDI-TOF MassArray spectrometry. The advent of chromatin immunoprecipitation (ChIP) technique was a fundamental contribution to address the study of the rest of epigenetic factors, mainly in histone modifications. ChIP is a powerful technique for analyzing targeted proteins that bind to particular sequences of DNA. From that moment on, many ChIP-grade antibodies that recognized most of the histone modifications and chromatin modifying players were produced, increasing exponentially the knowledge of the relationship between epigenetic players and control of gene expression (for an extended review of these techniques see (Heyn and Esteller 2012; García-Giménez et al. 2012; Sandoval et al. 2013a)). However, following the genomics trail, the application of genome-scale analysis techniques such as sequencing and array platforms, has opened a new era in the epigenetic field: leading to epigenomics. DNA methylation platforms have proved to be useful for addressing genome-wide DNA methylation profiling in large cohorts of patients due to its reproducibility, rapidness and reasonable low price per sample (Sandoval et al. 2011). The latest version from Illumina, the Infinium Human methylation 450K beadchip assay, permits the analysis of more than 480,000 CpGs covering 99 % of all referenced gene sequences. On the other hand, whole-genome bisulfite sequencing permits the rapid unbiased analysis of the total DNA methylome at a single-base resolution of an organism (Laird 2010). Regarding histone modifications, in 2007, the laboratory of Wold and Myers contributed to the progression of global genomic-scale analysis by combining chromatin immunoprecipitation and massively parallel sequencing (ChIP-seq) to identify mammalian DNA sequences bound by transcription factors in vivo (Johnson et al. 2007). Soon after, different laboratories used ChIP-seq for large-scale profiling of histone modifications and chromatin modifying complexes (Barski et al. 2007; Robertson et al. 2008). Several platforms measuring miRNA expression profiles (miRNome) have been rapidly developed (such as Affymetrix, Agilent, and Illumina microarray platforms) and next-generation sequencing (NGS) miRNA-Seq technologies are also available (Vaz et al. 2010).

9.5 Clinical Applications of Biomarkers

9.5.1 Epi-biomarkers in Cancer

The improvement of current clinical outcomes in cancer patients depends on factors such as an early stage diagnosis of the disease and the use of prognostic and predictive biomarkers coupled with appropriate and effective treatments. The accomplishment of these objectives would be translated into a better disease outcome, survival, and quality of life.

For early diagnosis, population screening on high-risk patients could result in early detection and led to avoid radical surgical procedure and fewer side effects of chemotherapy. The most important aspect for early diagnostics is to identify markers associated to cancer using noninvasive or minimally invasive methods for sample collection. The number of sources of DNA is very large. It can be obtained from nipple aspirate fluid, exfoliated cells, urine, plasma, saliva, stool, and material from bronchial brushes and bronchoalveolar lavages (BALs). Regarding prognostic and predictive biomarkers, they are necessary for tailoring therapies, guiding treatment decisions in order to maximize efficacy and diminish unnecessary toxic effects.

There is increasing evidence that epigenetic biomarkers can be one of the most prevalent molecular markers for human cancers. Although histone modifications and miRNAs are promising, DNA methylation, and specifically hypermethylation of CpG island promoters, has been by far the most valuable tool to date and the list of methylated genes identified in a broad range of cancer types is still growing and prone to be translated into a clinical setting. Here we will summarize the most advanced studies to date in the search of diagnostic and prognostic epigenetic biomarkers focusing in three of the most prevalent tumor types, such as lung, colorectal, and breast cancer. Subsequently, we will highlight the main advances in epigenetic biomarkers as therapeutic predictors.

9.5.1.1 Diagnostic and Prognostic Predictors

Lung Cancer

Lung cancer (LC) is the leading cause of cancer-related death worldwide with 1.3 million deaths annually, accounting for approximately a third of all cancer deaths, following data from the World Health Organization (WHO) in 2011. Lung cancer has been associated for a long time with chronic exposure to tobacco carcinogens. It is estimated that about 90 % of lung cancer deaths are due to smoking (Jemal et al. 2007). LC is one of the most aggressive cancers and its survival after diagnosis is very poor. A major factor in the high mortality of lung cancer patients is a late diagnosis and consequently its late stage presentation with metastases occurring in approximately two thirds of patients at the time of diagnosis. At that time the options for effective therapeutic intervention are limited. It has been estimated that detection of lung cancer at earlier stages could potentially increase survival rates by 10- to 50-fold. However, less than 25 % of patients are diagnosed at clinical stage I (Henschke et al. 2006). For this reason, there is an urgent necessity of finding and validating new early stages epigenetic biomarkers to initiate rapid therapeutic strategies in patients that in turn may increase the survival rate.

Diagnosis

Before the advent of omics technologies, the application of DNA methylation targeted approaches for diagnostic purposes has achieved an "acceptable" success. Hypermethylation of driver genes involved in critical cellular processes has been found in lung tumor tissues at early stages of tumorigenesis. For example, p16INK4a gene was reported to be silenced by promoter hypermethylation not only in lung cancer tissues but also in early stages of lung cancer development (Belinsky 2005). Methylation aberrant changes have also been used for distinguishing the major histological lung cancer types. As an example, the frequencies of methylation of a selected eight genes panel not only differentiate between non-small, small cell lung cancer, and neuroendocrine tumors, but also between non-small cell lung cancer (NSCLC) subtypes: adenocarcinomas and squamous (Toyooka et al. 2001). Interestingly, a progression of methylation levels in a panel of seven genes from normal lung to adenomatous hyperplasia and synchronous lung cancer (defined as two separate lung tumors in a same individual) has also been reported (Licchesi et al. 2008). Regarding histone modifications, several studies report on global changes in expression of specific histone modifications in lung cancer. Hypoacetylation of H4K12 and H4K16, hyperacetylation of H4K5 and H4K8 and decreased levels of H4K20me3 has been identified in lung adenocarcinoma and squamous cell carcinoma compared with normal lung parenchyma (Van Den Broeck et al. 2008). MiRNAs have also been extensively studied in the carcinogenesis of the lung. One of the most interesting examples is the downregulation of mir-let7 which correlates with higher expression of K-Ras (one of the main mir-let7 targets) in lung tumors compared to normal tissues (Johnson et al. 2005).

Nowadays, the challenge is to find new and robust epigenetic diagnostic biomarkers in samples obtained by less invasive methods, such as blood, sputum, or bronchoalveolar lavages (BALs). Nowadays, there is no epigenetic biomarker for clinical practice in LC. Interestingly, preneoplasic lung cancer diseases already show an aberrant epigenetic landscape, and several studies have described gaining of methylation of genes in plasma, serum, or sputums of LC patients (Esteller et al. 1999a; Usadel et al. 2002; Wang et al. 2007; Belinsky et al. 2005, 2007). In this sense, SHOX2 hypermethylation biomarker has recently been certificated by the "Conformité Européenne In vitro Diagnostic" in bronchial aspirates (Dietrich et al. 2012). Besides, several miRNAs biomarkers with potential diagnostic value have been described. A recent study identified 11 selected miRNAs that distinguish lung cancer form normal by qRT-PCR in plasma samples (Sanfiorenzo et al. 2013).

Recent development of high throughput technologies is generating high number of epigenetic data that help to characterize the role of epigenetics in lung cancer disease and is leading to the discovery of new potential epigenetic biomarkers. Different genome-wide methylation analyses in lung cancer have been performed using different methodologies (for an extended review, see Heyn and Esteller 2012; Bock 2012). Studies have been carried out using restriction landmark genomics scanning (RLGS) and Methylated-CpG island recovery assay (MIRA) in lung cancer and adjacent normal lung tissue: For example, Park et al. identified 21 new hypermethylated genes not reported previously and the silencing of *SLC5A8* (Park et al. 2005, 2013). Moreover, Rauch et al. identified a potential tumor-specific methylation signature with diagnostic methylation using MIRA technology in stage I squamous samples (Rauch et al. 2008). Nowadays the combination of bisulfate conversion of DNA with microarray technology has become the gold standard for epigenomic analysis. The first analysis was performed by Bibikova et al. using the Golden gate Beadarray platform where 371 lung cancer-related genes were identified in adenocarcinoma lung cancer samples compared to normal lung tissue (Bibikova et al. 2006). Recently, more advanced platforms have been delivered to the market, such as the 450K beadchip array from Illumina. In these 4 years, several genome-wide analyses have been carried using non-small/small cell lung cancer and normal lung tissues identifying new aberrant epigenetic changes with potential diagnostic value (Nelson et al. 2012; Heller et al. 2013; Morán et al. 2012; Son et al. 2011). Interestingly, a study comparing 169 adenocarcinomas and 72 squamous cell carcinomas revealed divergent genomic and epigenomic landscapes in non-small lung cancer subtypes (Lockwood et al. 2012). Finally, a recent manuscript validated a previously identified three gene methylation signature (*CDO1*, *HOXA9*, and *TAC1*) with diagnostic value (100 % specificity and 83–99 % sensitivity depending on the cohort) in lung cancer using TCGA 450K array data (Wrangle et al. 2014).

Prognosis

Since the beginning of the twenty-first century, a worse prognostic impact of DNA hypermethylation, especially in non-small cell lung cancer (NSCLC) primary tumors, has been described for several genes. For instance, *DAPK* methylation was shown to be strongly associated with survival in stage I NSCLC patients (Tang et al. 2000), finding that was confirmed later (Lu et al. 2004). Additionally, several meta-analysis have shown *P16* (Xing et al. 2013a) and *RASSF1A* (Liu et al. 2013) hypermethylation as potential independent prognostic factors for poor survival in surgically treated NSCLC. Brock et al. (2008) showed that methylation of *P16*, *CDH13*, *RASSF1A*, and *APC* in tumors as well as in histological tumor-negative lymph nodes—probably indicating undetectable micrometastasis—was associated with disease recurrence.

In the recent years, high throughput epigenetic analyses with prognostic relevance have begun to appear. For instance, genome-wide search for methylated CpG islands patients by combining methylated DNA immunoprecipitation and microarray analysis in 101 stages I–III NSCLC found 477 tumor-specifically methylated genes and, importantly, showed that *HOXA2* and *HOXA10* may be of prognostic relevance in squamous cell carcinoma (Heller et al. 2013). In a more powerful approach, another study (Sandoval et al. 2013b) used the Infinium 450k array to study tumoral DNA obtained from 444 patients with NSCLC, including 237 stage I tumors. Unsupervised clustering identified patients with high-risk stage I NSCLC who had shorter relapse-free survival. The analysis in an independent validation cohort of the most significant methylated sites found that hypermethylation of five genes was significantly associated with shorter RFS in stage I NSCLC: *HIST1H4F*, *PCDHGB6*, *NPBWR1*, *ALX1*, and *HOXA9*. These data led to a signature based on the number of hypermethylated events able to distinguish patients with high- and low-risk stage I NSCLC.

Histone modifications have been also studied in terms of prognosis. Although they need to be validated, global histone H3 and H4 modification patterns were shown to be potential markers of tumor recurrence and disease-free survival in a cohort of 408 NSCLC patients (Song et al. 2012). By using miRNA arrays, miRNA expression signatures have been also proposed as prognostic tools in NSCLC. A five-miRNA signature (*miR-137*, *miR-372*, *miR-182*, *miR-221*, and *miR-let7a*) correlated with disease-free survival in a cohort of 122 NSCLC patients (Yu et al. 2008). Raponi et al. (2009) identified several miRNAs, including *miR-155* and *miR-let7*, which had previously been shown to have prognostic value in adenocarcinoma, as influencing prognosis in squamous cell carcinoma, and proposed *miR-146b* as the strongest individual predictor. Patnaik et al. (2010) also defined a miRNA signature that predicted postoperative recurrence of stage I NSCLCs. More recently, a microRNA signature in plasma showed predictive, diagnostic, and prognostic value in a screening context (Sozzi et al. 2014). However, as is the case with gene expression profiling, miRNA signatures suggested by different groups are almost nonoverlapping. More detailed studies are needed to clarify these issues.

Colon Cancer

Colorectal cancer (CRC) is the third leading cause of cancer mortality worldwide accounting for over 600,000 deaths annually, and rise higher positions in developed countries (Jemal et al. 2011). Since epigenetics plays a major role in the initiation and progression of colorectal cancers, aberrant epigenetic changes including locus specific and global DNA methylation, histone modifications, and miRNAs are becoming a useful alternative to standard methods for early detection of colorectal cancer.

Diagnosis

A variety of tumor suppressors genes such as *RB*, *CDKN2A*, *MGMT*, and *ARF* have been described as aberrantly methylated in CRC (Grady and Carethers 2008; Schweiger et al. 2013). A study from 2009 analyzed the aberrant promoter methylation of two DNA repair *hMLH1* and *MGMT* in different segments of the colon from healthy donors from a colonoscopy screening. Specifically, they showed the prevalence of *hMLH1* and *MGMT* methylation increased significantly with age, particularly in the right colon. Concomitant methylation of both promoters was also significantly more common in the right colon of women, suggesting that epigenetically altered and silenced genes may be important in the early carcinogenic process (Menigatti et al. 2009). During tumorigenesis, global and specific DNA hypomethylation undergoes with genomic instability and activation of proto-oncogenes. In this sense, an increased risk of CRC is associated with loss of imprinting of *IGF2* gene by DNA hypomethylation (Timp et al. 2009) and global DNA hypomethylation of *LINE-1* has also been reported (Estécio et al. 2007). A subset (approximately 20 %) of CRCs shows a widespread GpG island hypermethylation, named as the CpG Island Methylator Phenotype (CIMP). High degree CIMP colorectal cancers are associated with several features such as gender, age, tumor location, *B-RAF* and *TP53* mutation status, methylation levels of *LINE-1* and promoter *MLH1*, and chromosome stability. While it is generally well accepted that etiologically and clinically distinct subgroups exist in this disease, a precise definition of CIMP remains to be established (Hughes et al. 2012). Another important epigenetic mechanism related to CRC tumorigenesis is histone modification. In this sense Nakazawa et al., analyzing endoscopically resected specimens of colorectal tumors, in which the authors observed high levels of H3K9me2 compared with normal colon mucosa, suggested that these post-translational histone modifications occur during CRC (Nakazawa et al. 2012). Another epigenetic change that has been described in early stages of CRC is the aberrant methylation of microRNAs. It has been suggested that *miR-137* acts as a tumor suppressor in the colon, since its gain of promoter methylation is correlated with gene expression silencing (Balaguer et al. 2010).

The ideal DNA methylation biomarker for early detection of CRC should be present in high frequency and detectable in bodily fluids or secretions. Studies have reported the existence of specific hypermethylated genes in blood and stool of patients with CRC (Herbst et al. 2009; Ahlquist et al. 2012a). In this context, the analysis of *SEPT9* promoter DNA methylation levels in peripheral blood has demonstrated to be a promising biomarker for detection of colorectal cancer at all stages and locations (Warren et al. 2011; Tóth et al. 2012). Interestingly, the recently commercially available *SEPT9* methylation in plasma showed high sensitivity and specificity for detection of colorectal cancer and could improve the performance of current fecal occult blood and DNA testing (deVos et al. 2009). On other hand, *SFRP2* was the first reported DNA methylation marker in stool and the most sensitive biomarker for CRC, showing a sensitivity of 77–90 % and specificity of 77 % (Müller et al. 2004). A follow-up study found that *SFRP2* methylation was detectable in the stool of almost half of all patients with hyperplasic polyps or colorectal adenomas (Oberwalder et al. 2008). All these results further support its potential use in the detection of early CRC lesions (Tang et al. 2011). Finally, it has also been reported that a DNA stool test detects methylated *BMP3*, *NDRG4*, *VIMENTIN*, and *TFPI2* combined with mutant KRAS gene that detects adenomas larger than 2 cm with 82 % of sensitivity and CRC with 91 and 93 % sensitivity and specificity, respectively (Ahlquist et al. 2012b). Regarding histone modifications in noninvasive samples, interesting results were obtained by Gezer et al. demonstrating lower levels of H3K9me3 and H4K20me3 in circulating blood nucleosomes of colorectal cancer patients (Gezer et al. 2013). Many studies have investigated the potential use of miRNAs as biomarkers in the early diagnosis of CRC. It has been recently described a three-miRNA plasma panel (*miR-409-3p*, *miR-7*, and *miR-93*) and a multimarker panel tested on plasma samples that are able to accurately discriminate CRC patients from healthy controls, highlighting its promising clinical value in the early CRC detection (Wang et al. 2013; Luo et al. 2013; for an extended revision, see Kim and Reitmair (2013) and Gall et al. (2013).

In vitro genome-wide analyses have been useful strategies for the discovery of new epigenetic biomarkers in CRC. Researchers often compare profiles of CRC cell lines with normal colorectal cells and then compile a list of candidate biomarkers for further study. In this sense, Khamas et al. conducted a genome-wide screen of 15 CRC cell lines and 23 paired tumor and normal samples from CRC patients to identify a set of methylation-silenced genes in CRC, combining gene expression arrays with treatment with demethylating agents. From the 54,613 genes analyzed, they reported 139 genes epigenetically regulated in CRC. Interestingly, derived from this study the *THSD1* methylation appeared to have the potential for diagnostic use (Khamas et al. 2012). The combination of bisulfite treatment with array platform was a groundbreaking point for the characterization of colorectal cancer phenotypes. Using the first version of the array platforms, termed Golden Gate, several studies analyzed the association between CIMP+ CRCs and mutations, including BRAF (Hinoue et al. 2009) or the characterization of CRC subgroups according to CIMP levels (Ang et al. 2010). However, a recent manuscript has described discrepancies in low and intermediate CIMP subgroups (Karpinski et al. 2013) and using more advanced DNA methylation platforms more subgroups with different CIMP levels have been described (Hinoue et al. 2012), even with paraffin embedded samples (Dumenil et al. 2014). Several studies have been performed using the next-generation platforms (27K and 450K beadchip arrays) comparing CRC cohorts and their corresponding adjacent normal colorectal tissues. These analyses derived in the identification of different signatures and specific locus with potential diagnostic value (Kibriya et al. 2011; Kim et al. 2011). Interestingly, the combination of early stages (adenomas) with CRC cohorts suggested that alteration in DNA methylation occur during early stages of the transition from adenomas to CRC and permits the identification of epigenetic biomarker for early detection of CRC (Oster et al. 2011; Luo et al. 2014), including nonfrequent hypomethylated biomarkers (Oster et al. 2013). Finally, the discovery of early biomarkers using genome-wide approaches and latter verification in noninvasive samples, such as blood or stool, will be of great clinical value. An example for this approach has been carried out by Lange et al., with the identified methylation of *THBD-M* as a promising clinical biomarker in plasma and serum (Lange et al. 2012).

Prognosis

It has been shown that DNA methylation of *P14*, *RASSF1A*, and *APC* genes is associated with a poor prognosis subset of CRC patients, independently of tumor stage and differentiation (Nilsson et al. 2013). A meta-analysis of 11 studies concluded that *P16* hypermethylation might be a predictive factor for unfavorable prognosis of colorectal cancer patients (Xing et al. 2013b), and *IGF2* hypomethylation has been also shown as a potential biomarker for worse prognosis (Baba et al. 2010). On the contrary, hypermethylation of *hMLH1* has been related to better survival (Jensen et al. 2013). Some miRNAs specifically associated with patient survival have also been described (Wu et al. 2011). For instance, patients with higher miR-200c expression have a shorter survival time compared with patients with lower expression, and expression levels of *miR-320* and *miR-498* are also correlated with the probability of recurrence-free survival in stage II colon cancer patients.

The CIMP phenotype has been independently associated with significantly worse prognosis in CRC patients (Juo et al. 2014). Genome-wide epigenomic analyses helped to identify that methylation of several genes involved in extracellular matrix components are associated with worse survival in CRC patients, suggesting that methylation of this pathway might represent a prognostic signature in CRC. Specifically, methylation of *IGFBP3* and *EVL* genes was validated as an independent prognostic marker in an independent cohort (Yi et al. 2011). Another comprehensive study that analyzed the methylation status of around 14,000 genes (using Infinium 27k from Illumina) in 144 CRC samples was able to identify subgroups of tumors correlating with prognostic markers such as *hMLH1* hypermethylation, *KRAS* or p53 mutations, and *BRAF*(V600E) mutation, suggesting that a combination of epigenetic and genetic analysis might improve the accuracy of prognosis (Hinoue et al. 2012). The analysis of the expression 315 human miRNAs in 10 normal mucosa samples and 49 stage II colon cancers could predict recurrence of disease with an overall performance accuracy of 81 %, indicating a potential role of miRNAs in determining tumor aggressiveness (Schepeler et al. 2008).

Breast Cancer

Breast cancer is the most common cancer and is the second leading cause of death among women. Although it is well established that inherited and acquired mutations in genetic material are known to be principal contributors to the initiation and development of breast cancer, epigenetic changes are also important factors. The incorporation to the field of cancer research of this novel layer of complexity will help to improve clinical challenges of breast cancer.

Diagnosis

Breast cancer genomes usually contain thousands of genetic changes, of which only a small subset might actually drive development of the disease such as *BRCA1* or *2* (Stratton et al. 2009). CpG island promoter hypermethylation of DNA repair genes (*BRCA1*, *PALB2*, *ATM*) and other tumor suppressor genes, including *CDKN2A*, *FZR1*, *RARB2*, and *GSTP1*, as a mechanism of gene inactivation, has been found in breast cancers arising in familial and sporadic cases (Esteller et al. 2001; Potapova et al. 2008; Vo et al. 2004). Other methylated genes in breast cancer include those important for evasion of apoptosis, invasion, and metastasis (Widschwendter and Jones 2002; Berman et al. 2005). Additionally, global and locus specific DNA methylation including juxtacentromeric satellite DNA and proto-oncogenes have been found to be associated with breast tumorigenesis (Lo and Sukumar 2008; Jackson et al. 2004). DNA methylation changes in gene promoter are usually associated with histone modifications changes. In this sense, a study analyzed histone lysine acetylation (H3K9ac, H3K18ac, H4K12ac, and H4K16ac), lysine methylation (H3K4me2 and H4K20me3), and arginine methylation (H4R3me2) marks in a cohort of 880 human breast carcinoma. Their analyses revealed low or absent H4K16ac and relatively high levels of H3K18ac and H420me3 in the majority of breast cancer cases suggesting that these alterations may represent an early sign of breast cancer. Clustering analysis identified three groups of histone displaying distinct pattern in breast cancer, which have distinct relationships to known prognostic factors and clinical outcome (Elsheikh et al. 2009). Specific miRNAS have been identified with aberrant expression in breast cancer (Iorio et al. 2005; Veeck and Esteller 2010). Besides, a study from the consortia of Cancer Genome Atlas Network using a cohort of 522 tumors and 22 matched normal tissues have identified deregulated miRNAs and seven miRNA subtypes (Cancer Genome Atlas Network 2012).

Epigenetic analysis in noninvasive breast samples is an ideal model system for studying early breast tumor development. Several studies in ductal lavage fluids or periareolar fine needle aspiration samples have reported hypermethylation of *P16INK4α* combined with other hypermethylated promoter genes (Bean et al. 2007; Locke et al. 2007; Vasilatos et al. 2009). These results suggested that epigenetic biomarkers analysis in ductal lavage fluids could be clinically valuable for breast cancer detection and risk prediction. Moreover, hypermethylation of promoter genes in serum or blood DNA have been tested in breast cancer patients (Brooks et al.

2009; Hu et al. 2003). Interestingly, promoter hypermethylation of three selected genes (*APC*, *RASSF1*, and *DAPK*) was detected by methylation-specific PCR analysis in serum DNA from patients with preinvasive and early-stage breast cancer (Dulaimi et al. 2004).

Genome-wide epigenetic profiling has revealed new alterations in breast carcinogenesis. A study using MeDIP-ChIP, a technique that combines inmunoprecipitation of methylated DNA with hybridization in microarrays, evaluated specific methylation patterns in familial breast cancers. They also integrated methylation data with expression and SNP CGH arrays and concluded that methylation profiles for familial breast cancers are defined by the mutation status and are distinct from the intrinsic subtypes (Flanagan et al. 2010). Several groups have used the 27K Ilumina array for characterizing the diversity of the disease and finding potential biomarkers for improving the diagnosis of breast cancer. In this sense, Dedeurwaerder et al. showed that DNA methylation profiling identified the existence of unrecognized breast cancer groups characterized by different cell type composition of the tumor microenvironment and immune components (Dedeurwaerder et al. 2011). Other groups have identified or characterized patients groups with different CIMP features (Fang et al. 2011) or hormone receptor status (Li et al. 2010; Fackler et al. 2011; Hill et al. 2011). It is well established that DNA methylation profiling in blood is an ideal approach for identifying potential epigenetic markers of cancer risk. Two recent studies have analyzed blood samples. On one hand Xu et al. identified 250 altered CpGs with potential cancer detection value using the 27K array in a cohort of 298 women (Xu et al. 2013). On the other hand, our group using 15 twin pairs discordant for breast cancer (to avoid genetic variation) and high resolution 450K DNA methylation array identified and validated *DOK7* as a potential early biomarker for breast cancer (Heyn et al. 2013b). Recent advances in detection of novel epigenetic biomarkers in circulating tumor DNA will offer robust and convenient approaches for noninvasive breast cancer detection (Fackler et al. 2014).

Prognosis

As in other tumor types, hypermethylation of the *P16* tumor suppressor gene in breast cancer has been associated with higher mortality (Xu et al. 2010). Additionally, *GSTP1*, *TWIST*, and *RARβ* promoter methylation has also been significantly correlated with an increase in breast cancer-specific mortality (Cho et al. 2012). In another study, methylation-specific PCR of six known tumor suppressor genes was used to generate a hypermethylation profile of primary breast tumors, and the methylation states of different genes—including *GSTP1*—were found to be significantly associated with several known prognostic factors (Shinozaki et al. 2005). The expression of several microRNAs has been related to metastasis and worse prognosis and differential expression of histone deacetylases *HDAC1*, *2*, and *3*—overexpression of *HDAC2* and *HDAC3*—has been associated with clinicopathological indicators of disease progression (Müller et al. 2013).

Epigenetic profiling of tumor DNAs by using the 27k array was used to show that a hierarchical clustering based on methylation levels was able to divide the specimens in three distinct groups with significant differences in relapse-free survival and lymph node metastasis. Additionally, six individual methylated genes were associated with worse tumor recurrence (Hill et al. 2011). Fang et al. also used the 27k array to first discover a "methylator" phenotype, as previously delineated in CRC. In this case, the coordinated methylation of a large group of CpG island in groups of tumors was termed B-CIMP (breast CpG island methylator phenotype). Opposite to CRC, this B-CIMP was associated with lower risk of breast cancer metastasis and improved rates of survival independently of other known breast cancer prognostic markers, such as ER+ status (Fang et al. 2011). This study, together with others (Pakneshan et al. 2004; Campbell et al. 2004) led to the hypothesis that DNA hypomethylation is a driving force in breast cancer metastasis, and could have therapeutic implications and concerns in guiding the use of epigenetic drugs, at least in breast cancer therapy (Szyf 2009). Microarrays of microRNAs expression have also been useful for determining the tumors' aggressiveness, for example, in node-negative ER+ tumors (Foekens et al. 2008) or in demonstrating their crucial role in the metastatic process, leading the scientific world coin the definition "metastomiRs" (Iorio et al. 2011).

9.5.1.2 Epigenetic Biomarkers as Therapeutic Predictors

Despite advances in pharmacology, it is evident that not all patients respond favorably to particular drugs. Thus, the identification of biomarkers predicting response to drug treatments is one of the main requirements to enable personalized cancer therapies.

Stratification according to individual tumor characteristics is increasingly allowing to identify those patients more prone to benefit from specific therapies, thus improving responses and avoiding side effects from unnecessary approaches. Single-gene biomarkers are already guiding treatment decisions for several types of cancer. Illustrative examples are those represented by genetic alterations of *EGFR*, *HER2*, or *BCR-ABL*, which support treatments with gefitinib/erlotinib/afatinib in non-small cell lung cancer (NSCLC), trastuzumab in breast cancer and imatinib in chronic myeloid leukemia, respectively. Even more recently, it has been approved the use of vemurafenib in patients with *BRAF*-mutated melanomas or crizotinib in those NSCLC patients harboring *EMT-ALK* translocations. Although validated biomarkers are mostly related to targeted therapies (especially kinase inhibitors), efforts are also directed to establishing predictive biomarkers for broader chemotherapeutic agents. Promising candidates are currently being evaluated. This is the case of *BRCA* mutations, which result in defects in DNA repair and seem to confer sensitivity to chemotherapeutic drugs such as cisplatin. Consistently, high levels of *BRCA1* expression have been related to cisplatin resistance. Similarly, other DNA repair related enzymes have also been associated, as high expression of *ERCC1*, *RRM1*, *TS*, or *MSH2* have been related to cisplatin resistance, especially in lung cancer (Felip and Martinez 2012). However, despite great efforts, the number of examples valuable for clinical use is still very limited.

While some genetic-based biomarkers are currently being identified and/or evaluated in clinical trials, epigenetic alterations—especially DNA methylation—represents a potentially wide biomarker for clinical application in drug prediction. Nowadays, hypermethylation of *MGMT* is the most advanced example in predicting treatment success, specifically in gliomas treated with alkylating agents such as carmustine and temozolomide. However, although there is a plethora of additionally potential epigenetic biomarkers with predictive potential, the value of epigenetic modifications in personalized medicine is yet poorly understood. Here, we will first highlight some of the best studied single-gene candidate biomarkers from hypothesis-driven approaches to date. Secondly, we will introduce some epigenomic procedures that have been already developed.

Hypothesis-Driven Approaches for Chemotherapy Response

The best example today of gene promoter hypermethylation linked to drug resistance is displayed by the DNA repair gene O6-methylguanine-DNA methyltransferase (*MGMT*). The *MGMT* gene encodes a DNA repair protein that removes alkyl groups from the O6 position of guanine, an important site of DNA alkylation. This process protects the cells form genetic modification caused by carcinogens such as nitrosamides. However, and somewhat paradoxically, high levels of MGMT activity in cancer cells create a resistant phenotype by blunting the therapeutic effect of alkylating agents such as temozolomide or carmustine, and have been shown to be an important determinant of treatment failure in gliomas (Hegi et al. 2005).

MGMT gene is not commonly mutated or deleted. However, methylation of the CpG island in the *MGMT* gene prevents transcription of the gene and was shown (Esteller et al. 2000) and confirmed (Hegi et al. 2005) to be a useful predictor of the responsiveness of gliomas to alkylating agents. Additionally, despite focusing on glioma therapy stratification, *MGMT* testing could be applicable for different tumor types with frequent hypermethylation events (Esteller et al. 1999b).

Epigenetic inactivation of other DNA repair genes has been associated to drug response. Consistently, not only *BRCA1* mutations have been linked to sensitivity to cisplatin: the frequent inactivation of the gene by hypermethylation in breast and ovarian cancers has been recently related to cisplatin response (Stefansson et al. 2012). Additionally, epigenetic inactivation of *BRCA1* is currently being studied as a biomarker for sensitivity to *PARP* inhibitors. PARP is DNA repair protein that works by base excision: thus, targeting PARP function in *BRCA1* mutant or deficient cells leads to an overload of DNA damage and cell death. Currently, genetic lesions are being tested in clinical trials to evaluate their role in therapy decision for those *BRCA1* mutation carriers (Fong et al. 2009). As this is a rather infrequent event, it is possible that *BRCA1* hypermethylation, which occur in almost 20 % of these patients, could represent a broader biomarker for the efficacy of the drug in this particular context (Veeck et al. 2010).

Enzymes involved in the process of xenobiotic detoxification have also been evaluated as chemotherapy response biomarkers. This is the case of GSTP1, a protein whose action of catalyzing detoxification of xenobiotic and carcinogens is favorable for the healthy cell, but that minimizes therapeutic actions in tumor cells, as it also removes therapeutic drugs. Hypermethylation of the *GSTP1* gene promoter is present in a high frequency of prostate tumors and is being proposed as a candidate complement for the discussed PSA testing (Van Neste et al. 2012). In terms of drug prediction, hypermethylation and gene inactivation of *GSTP1* has been associated to prolonged survival in breast cancer patients treated with doxorubicin (Dejeux et al. 2010). Consistently, promoter hypermethylation of *ABCB1*—a membrane-bound transporter that actively effluxes a wide range of compounds from cells, including chemotherapy drugs—was also related to doxorubicin response (Chekhun et al. 2006), thus reinforcing the role of this particular mechanism in the treatment of breast cancer patients with doxorubicin.

There are several other epigenetic biomarkers that have been related to drug response, mainly other enzymes related to DNA repair and detoxification, but also involved in programmed cell death or signal transduction (for a more extensive review, see Heyn et al. 2013a). They have established the basis that epigenetics is a promising avenue in advancing to a more personalized medicine. However, only a few of them are close to be used in a clinical setting. Better and more powerful biomarkers will most likely come from unbiased epigenomic strategies able to identify previously unsuspected associations.

Data-Driven Epigenomic Biomarkers for Chemotherapy Response

Single-gene approaches are hypothesis-driven strategies. They are based on preliminary data supporting a possible role of particular genes in specific cellular functions or pathways. On the contrary, whole-genome profiling consists on an unbiased data-driven research, with the potential of identifying previously unknown, more powerful and somewhat unexpected associations. The use of massive data to address questions such as drug sensitivity is specially promising but has not been particularly harnessed yet, as breakthroughs in sequencing and array technologies have been recently incorporated to research.

One of the pioneering studies addressing the role of methylation profiling in drug prediction was performed taking advantage of the NCI-60 Anticancer Drug Screening panel. This was an in vitro drug-discovery tool composed of 60 cell lines from 9 different types of tumors that were tested for thousands of different antitumor compounds. The extensive genetic and transcriptomic analyses led, for example, to the identification of P-glycoproteins targeting compounds or to discover inhibitors of mutant BRAF signaling (Shoemaker 2006). In terms of epigenetics, although only 32 promoter-associated CpG islands were studied, this initial profiling led to the identification of the hypermethylation of the p53 homolog DNA damage sensor *TP73* as predictor for alkylating agent response, including the alkylating-like cisplatin (Shen et al. 2007).

A more comprehensive epigenomics approach recently analyzed 82 NSCLC cell lines with the Illumina Infinium450 BeadArray (Walter et al. 2012). In this study, DNA methylation patterns were able to divide NSCLCs into epithelial-like and mesenchymal-like subsets. Five hundred and forty-nine DMRs were identified and incorporated into an epithelial–mesenchymal transition (EMT) classifier. Importantly, and consistent with previous findings (Dave et al. 2012), nearly all epithelial-like lines according to the classifier were associated with sensitivity to erlotinib —an *EGFR* targeted drug—whereas nearly all mesenchymal-like lines were resistant. Additionally, this approach served to validate seven single biomarkers as predictive or erlotinib sensitivity in vitro.

Another recent approach (Wrangle et al. 2013) used the Illumina Infinium 450 BeadArray to study the DNA methylation (together with expression changes) in NSCLC cell lines upon an AZA demethylating treatment, which has shown to develop a "priming" effect to subsequent standard therapies, including also immunotherapy. In most cells a multifaceted upregulation, involving hundreds of genes of the immune profiles was observed, including the target of immune checkpoint therapy, the tumor ligand PD-L1. The authors used the TCGA information to identify a significant proportion of lung tumors with this particular "immune evasion" signature. This subgroup of patients is hypothesized to benefit from AZA priming together with immune checkpoint therapy, outlining a signature that may identify successful predictive biomarkers.

In some cases, prognostic signatures have the potential of becoming predictive, too. Although adjuvant platinum-based chemotherapy is beneficial in NCSCL advanced resected disease, it has failed to show a survival benefit for patients at stage I (Pignon et al. 2008). One potential explanation for these negative data in the early stages is the lack of biologic factors predicting their recurrence and the fact that, in the absence of useful biomarkers, all stage I NSCLCs are pooled, making it more difficult to draw meaningful clinical conclusions. If high-risk patients can be identified, they probably could benefit for standard therapies, as patients with more advanced tumors do. In this sense, a recent study (Sandoval et al. 2013b) showed that unsupervised clustering of the 10,000 most variable DNA methylation sites stage I NSCLC who had shorter relapse-free survival (RFS). The study in a validation cohort of the most significant methylated sites found that hypermethylation of five genes was significantly associated with shorter RFS. Further studies are needed to confirm if these particular patients could benefit from adjuvant therapies.

Comprehensive studies involving other epigenetic factors have might also be valuable in drug prediction. For instance, a two miRNA signature (*miR-149* and *miR-375*) was found predictive for response in NSCLC patients treated with cisplatin and vinorelbine (Berghmans et al. 2013) and a six-miRNA-based classifier—extracted from miRNAs microarrays—was shown to be a reliable prognostic and predictive tool for disease recurrence in patients with stage II colon cancer, being able to predict which patients benefit from adjuvant chemotherapy (Zhang et al. 2013).

9.5.2 Other Diseases

Although, the first and currently the majority of epigenetic research conducted to date has been developed in cancer, an increasing interest in the scientific community is directed towards other type of pathologies. In the following section, we will shortly summarize recent findings about the role of epigenetics in a neurodegenerative disease, such as Alzheimer's, and in cardiovascular diseases.

9.5.2.1 Neurodegenerative: Alzheimer's Disease

Epigenetic marks are essential for the development of the highly specialized structure of central nervous system, which requires fine-tuning of gene expression. Thus, altered epigenetic has been associated with a range of neurological disorders (Jakovcevski and Akbarian 2012). Alzheimer's disease (AD) is a progressive neurodegenerative disorder, with over 35 million cases worldwide. Alzheimer's disease is the most common type of dementia in the elderly. It is characterized by the deposition of two forms of aggregates within the brain, the amyloid β plaques and tau neurofibrillary tangles. The majority of AD cases are sporadic suggesting that epigenetics may play an important role in the pathology. Only about 5 % of cases are familial or early onset AD which is associated with rare mutations on the *APP*, *PSEN1*, or *PSEN2* genes (Goate et al. 1991; Sherrington et al. 1995; Tanzi 2012). Currently, no pharmacological agent is approved for the treatment of AD (Selkoe 2012).

A seminal study using monozygotic twins discordant for AD showed less methylation in the cortex of the AD twin compared to the non-AD twin (Mastroeni et al. 2009). The same group and others have reported abnormal DNA methylation in AD patients (Mastroeni et al. 2010; Rao et al. 2012). Later, a genome-wide DNA methylation studies showed that more than 900 CpG sites representing 918 unique genes might be associated with late onset AD. The best candidate gene turned out to be *TMEM59*, whose promoter was hypomethylated in AD (Bakulski et al. 2012). Finally, a recent study from our group has analyzed 12 distinct mouse brain regions according to their genome-wide DNA methylation patterns and characterized their unique epigenetic landscapes. Using these methylome DNA methylation-associated silencing of three targets genes has been identified: *TBXA2R*, *SORBS3*, and *SPTBN4* (Sanchez-Mut et al. 2013). More recently, using 27K platform DNA methylation profiles of human hippocampus from controls and AD patients, they have also identified the functional role of promoter hypermethylation silencing in AD of *DUSP22* gene (Sanchez-Mut et al. 2014). Further studies will be needed to cross-validate and extrapolate these biomarkers to minimally invasive specimens, such as blood, for the long-awaited early detection of AD.

9.5.2.2 Cardiovascular Disease

Despite advances in the prevention and management of cardiovascular disease (CVD), this group of multifactorial disorders remains a leading cause of mortality worldwide. Today, although CVD has been associated with multiple genetic and environmental risk factors, these can only explain a limited part of the variability in CVD risk. In this context, the role of epigenetics in the pathogenesis of CVD—although considerably less explored than in cancer—is a promising field.

One of the first links between epigenetics and CVD was homocysteine (Hcy). This particular amino acid is metabolized to methionine after activation to S-adenosylmethionine, which is known to act as the main methyl donor in human body. Interestingly, elevated plasma total Hcy is associated with increased risk for vascular disease and decreased global methylation (Jamaluddin et al. 2007). However, the evidence linking global methylation patterns—usually measured in blood cells—with cardiovascular outcomes remains conflicting (Aslibekyan et al. 2015).

Candidate gene methylation studies have identified a number of promising epigenetic targets in CVD. For instance, a large-scale prospective cohort study linked methylation of *F2RL3*, a known locus linked with tobacco use, with CVD risk factors such as C-reactive protein as well as with overall mortality (Breitling et al. 2012). In a similar approach, the methylation of *IGF2* was found to be associated with a greater triglyceride-to-HDL cholesterol ratio, and predicted development of obesity (Perkins et al. 2012). Epigenomic studies have also started to be developed. One example of this approach used the Infinium 27k array to identify new loci associated with HDL cholesterol in the setting of familial hypercholesterolemia (Guay et al. 2012), and epigenome-wide studies of cardiomyopathy found distinct patterns of DNA methylation and histone 3 lysine-36 trimethylation in left ventricle tissues between patients and controls (Movassagh et al. 2011).

Importantly, global DNA hypomethylation has been described as a landmark of human advanced atherosclerotic lesions (Lund and Zaina 2011), whereas *ERα* promoter has been detected in atheromas as well as in the phenotypic switch from quiescent SMCs to a proliferative state (Turunen et al. 2009). Additionally, epigenetic modifications induced by environmental factors are emerging as important modulators of diabetes (Paneni et al. 2013). However, although most of these studies have proved to be useful in elucidating underlying pathophysiological mechanisms, their clinical relevance still remains unclear. Future studies are needed to fulfill the promises that cardiovascular epigenetics have raised.

9.6 Concluding Remarks

Long-awaited personalized medicine is currently becoming a reality. This emerging medical practice takes advantage of individual's molecular profiles to guide physician decisions for prevention, diagnosis, and disease treatment. Breakthrough of genome-scale analysis techniques, including the recently developed next-generation sequencing, has enabled an invaluable advance from hypothesis-driven single-gene approaches to whole-genome profiling and unbiased data-driven strategies. The combination and integration of epigenomic and other –omics studies is essential to fully understand the biology of diseases and is paving the way to identify unexpected biomarkers candidates. Currently, a plethora of newly identified biomarkers for diverse diseases are being described, but their clinical adaptation is hardly achieved yet due to lack of independent validation with different and large cohort of patients. Therefore, while pursuing best and more powerful biomarkers, scientific community should make an effort for cross-validating those already identified as most promising, thus closing the gap between bench and bedside.

Bibliography

Ahlquist DA, Zou H, Domanico M, Mahoney DW, Yab TC, Taylor WR, Butz ML, Thibodeau SN, Rabeneck L, Paszat LF, Kinzler KW, Vogelstein B, Bjerregaard NC, Laurberg S, Sørensen HT, Berger BM, Lidgard GP. Next-generation stool DNA test accurately detects colorectal cancer and large adenomas. Gastroenterology. 2012a;142(2):248–56.

Ahlquist DA, Taylor WR, Mahoney DW, Zou H, Domanico M, Thibodeau SN, Boardman LA, Berger BM, Lidgard GP. The stool DNA test is more accurate than the plasma septin 9 test in detecting colorectal neoplasia. Clin Gastroenterol Hepatol. 2012b;10(3):272–7.e1.

Ang PW, Loh M, Liem N, Lim PL, Grieu F, Vaithilingam A, Platell C, Yong WP, Iacopetta B, Soong R. Comprehensive profiling of DNA methylation in colorectal cancer reveals subgroups with distinct clinicopathological and molecular features. BMC Cancer. 2010;10:227.

Aslibekyan S, Claas SA, Arnett DK. Clinical applications of epigenetics in cardiovascular disease: the long road ahead. Transl Res. 2015;165(1):143–53. pii: S1931-5244(14)00128-5.

Baba Y, Nosho K, Shima K, Huttenhower C, Tanaka N, Hazra A, Giovannucci EL, Fuchs CS, Ogino S. Hypomethylation of the IGF2 DMR in colorectal tumors, detected by bisulfite pyrosequencing, is associated with poor prognosis. Gastroenterology. 2010;139(6):1855–64.

Bakulski KM, Dolinoy DC, Sartor MA, Paulson HL, Konen JR, Lieberman AP, Albin RL, Hu H, Rozek LS. Genome-wide DNA methylation differences between late-onset Alzheimer's disease and cognitively normal controls in human frontal cortex. J Alzheimers Dis. 2012;29(3):571–88.

Balaguer F, Link A, Lozano JJ, Cuatrecasas M, Nagasaka T, Boland CR, Goel A. Epigenetic silencing of miR-137 is an early event in colorectal carcinogenesis. Cancer Res. 2010;70(16): 6609–18.

Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K. High-resolution profiling of histone methylations in the human genome. Cell. 2007;129(4):823–37.

Bean GR, Bryson AD, Pilie PG, Goldenberg V, Baker Jr JC, Ibarra C, Brander DM, Paisie C, Case NR, Gauthier M, Reynolds PA, Dietze E, Ostrander J, Scott V, Wilke LG, Yee L, Kimler BF, Fabian CJ, Zalles CM, Broadwater G, Tlsty TD, Seewaldt VL. Morphologically normal-appearing mammary epithelial cells obtained from high-risk women exhibit methylation silencing of INK4a/ARF. Clin Cancer Res. 2007;13(22 Pt 1):6834–41.

Belinsky SA. Silencing of genes by promoter hypermethylation: key event in rodent and human lung cancer. Carcinogenesis. 2005;26(9):1481–7.

Belinsky SA, Klinge DM, Dekker JD, Smith MW, Bocklage TJ, Gilliland FD, Crowell RE, Karp DD, Stidley CA, Picchi MA. Gene promoter methylation in plasma and sputum increases with lung cancer risk. Clin Cancer Res. 2005;11(18):6505–11.

Belinsky SA, Grimes MJ, Casas E, Stidley CA, Franklin WA, Bocklage TJ, Johnson DH, Schiller JH. Predicting gene promoter methylation in non-small cell lung cancer by evaluating sputum and serum. Br J Cancer. 2007;96(8):1278–83.

Berghmans T, Ameye L, Willems L, Paesmans M, Mascaux C, Lafitte JJ, Meert AP, Scherpereel A, Cortot AB, Cstoth I, Dernies T, Toussaint L, Leclercq N, Sculier JP, European Lung Cancer Working Party. Identification of microRNA-based signatures for response and survival for non-small cell lung cancer treated with cisplatin-vinorelbine A ELCWP prospective study. Lung Cancer. 2013;82(2):340–5.

Berman H, Zhang J, Crawford YG, Gauthier ML, Fordyce CA, McDermott KM, Sigaroudinia M, Kozakiewicz K, Tlsty TD. Genetic and epigenetic changes in mammary epithelial cells identify a subpopulation of cells involved in early carcinogenesis. Cold Spring Harb Symp Quant Biol. 2005;70:317–27.

Bibikova M, Lin Z, Zhou L, Chudin E, Garcia EW, Wu B, Doucet D, Thomas NJ, Wang Y, Vollmer E, Goldmann T, Seifart C, Jiang W, Barker DL, Chee MS, Floros J, Fan JB. High-throughput DNA methylation profiling using universal bead arrays. Genome Res. 2006;16(3):383–93.

Bock C. Analysing and interpreting DNA methylation data. Nat Rev Genet. 2012;13(10):705–19.

Breitling LP, Salzmann K, Rothenbacher D, Burwinkel B, Brenner H. Smoking, F2RL3 methylation, and prognosis in stable coronary heart disease. Eur Heart J. 2012;33(22):2841–8.

Brock MV, Hooker CM, Ota-Machida E, Han Y, Guo M, Ames S, Glöckner S, Piantadosi S, Gabrielson E, Pridham G, Pelosky K, Belinsky SA, Yang SC, Baylin SB, Herman JG. DNA methylation markers and early recurrence in stage I lung cancer. N Engl J Med. 2008;358(11):1118–28.

Brooks J, Cairns P, Zeleniuch-Jacquotte A. Promoter methylation and the detection of breast cancer. Cancer Causes Control. 2009;20(9):1539–50.

Campbell PM, Bovenzi V, Szyf M. Methylated DNA-binding protein 2 antisense inhibitors suppress tumourigenesis of human cancer cell lines in vitro and in vivo. Carcinogenesis. 2004;25(4):499–507.

Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490(7418):61–70.

Chekhun VF, Kulik GI, Yurchenko OV, Tryndyak VP, Todor IN, Luniv LS, Tregubova NA, Pryzimirska TV, Montgomery B, Rusetskaya NV, Pogribny IP. Role of DNA hypomethylation in the development of the resistance to doxorubicin in human MCF-7 breast adenocarcinoma cells. Cancer Lett. 2006;231(1):87–93.

Cho YH, Shen J, Gammon MD, Zhang YJ, Wang Q, Gonzalez K, Xu X, Bradshaw PT, Teitelbaum SL, Garbowski G, Hibshoosh H, Neugut AI, Chen J, Santella RM. Prognostic significance of gene-specific promoter hypermethylation in breast cancer patients. Breast Cancer Res Treat. 2012;131(1):197–205.

Cortessis VK, Thomas DC, Levine AJ, Breton CV, Mack TM, Siegmund KD, Haile RW, Laird PW. Environmental epigenetics: prospects for studying epigenetic mediation of exposure-response relationships. Hum Genet. 2012;131(10):1565–89.

Dave B, Mittal V, Tan NM, Chang JC. Epithelial-mesenchymal transition, cancer stem cells and treatment resistance. Breast Cancer Res. 2012;14(1):202.

Dedeurwaerder S, Desmedt C, Calonne E, Singhal SK, Haibe-Kains B, Defrance M, Michiels S, Volkmar M, Deplus R, Luciani J, Lallemand F, Larsimont D, Toussaint J, Haussy S, Rothé F, Rouas G, Metzger O, Majjaj S, Saini K, Putmans P, Hames G, van Baren N, Coulie PG, Piccart M, Sotiriou C, Fuks F. DNA methylation profiling reveals a predominant immune component in breast cancers. EMBO Mol Med. 2011;3(12):726–41.

Dejeux E, Rønneberg JA, Solvang H, Bukholm I, Geisler S, Aas T, Gut IG, Børresen-Dale AL, Lønning PE, Kristensen VN, Tost J. DNA methylation profiling in doxorubicin treated primary locally advanced breast tumours identifies novel genes associated with survival and treatment response. Mol Cancer. 2010;9:68.

deVos T, Tetzner R, Model F, Weiss G, Schuster M, Distler J, Steiger KV, Grützmann R, Pilarsky C, Habermann JK, Fleshner PR, Oubre BM, Day R, Sledziewski AZ, Lofton-Day

C. Circulating methylated SEPT9 DNA in plasma is a biomarker for colorectal cancer. Clin Chem. 2009;55(7):1337–46.

Dietrich D, Kneip C, Raji O, Liloglou T, Seegebarth A, Schlegel T, Flemming N, Rausch S, Distler J, Fleischhacker M, Schmidt B, Giles T, Walshaw M, Warburton C, Liebenberg V, Field JK. Performance evaluation of the DNA methylation biomarker SHOX2 for the aid in diagnosis of lung cancer based on the analysis of bronchial aspirates. Int J Oncol. 2012;40(3):825–32.

Dulaimi E, Hillinck J, Ibanez de Caceres I, Al-Saleem T, Cairns P. Tumor suppressor gene promoter hypermethylation in serum of breast cancer patients. Clin Cancer Res. 2004;10(18 Pt 1):6189–93.

Dumenil TD, Wockner LF, Bettington M, McKeone DM, Klein K, Bowdler LM, Montgomery GW, Leggett BA, Whitehall VL. Genome-wide DNA methylation analysis of formalin-fixed paraffin embedded colorectal cancer tissue. Genes Chromosomes Cancer. 2014;53(7):537–48.

Elsheikh SE, Green AR, Rakha EA, Powe DG, Ahmed RA, Collins HM, Soria D, Garibaldi JM, Paish CE, Ammar AA, Grainge MJ, Ball GR, Abdelghany MK, Martinez-Pomares L, Heery DM, Ellis IO. Global histone modifications in breast cancer correlate with tumor phenotypes, prognostic factors, and patient outcome. Cancer Res. 2009;69(9):3802–9.

Estécio MR, Gharibyan V, Shen L, Ibrahim AE, Doshi K, He R, Jelinek J, Yang AS, Yan PS, Huang TH, Tajara EH, Issa JP. LINE-1 hypomethylation in cancer is highly variable and inversely correlated with microsatellite instability. PLoS One. 2007;2(5):e399.

Esteller M. Epigenetics in cancer. N Engl J Med. 2008;358(11):1148–59.

Esteller M, Sanchez-Cespedes M, Rosell R, Sidransky D, Baylin SB, Herman JG. Detection of aberrant promoter hypermethylation of tumor suppressor genes in serum DNA from non-small cell lung cancer patients. Cancer Res. 1999a;59(1):67–70.

Esteller M, Hamilton SR, Burger PC, Baylin SB, Herman JG. Inactivation of the DNA repair gene O6-methylguanine-DNA methyltransferase by promoter hypermethylation is a common event in primary human neoplasia. Cancer Res. 1999b;59(4):793–7.

Esteller M, Garcia-Foncillas J, Andion E, Goodman SN, Hidalgo OF, Vanaclocha V, Baylin SB, Herman JG. Inactivation of the DNA-repair gene MGMT and the clinical response of gliomas to alkylating agents. N Engl J Med. 2000;343(19):1350–4.

Esteller M, Fraga MF, Guo M, Garcia-Foncillas J, Hedenfalk I, Godwin AK, Trojan J, Vaurs-Barrière C, Bignon YJ, Ramus S, Benitez J, Caldes T, Akiyama Y, Yuasa Y, Launonen V, Canal MJ, Rodriguez R, Capella G, Peinado MA, Borg A, Aaltonen LA, Ponder BA, Baylin SB, Herman JG. DNA methylation patterns in hereditary human cancers mimic sporadic tumorigenesis. Hum Mol Genet. 2001;10(26):3001–7.

Fackler MJ, Umbricht CB, Williams D, Argani P, Cruz LA, Merino VF, Teo WW, Zhang Z, Huang P, Visvananthan K, Marks J, Ethier S, Gray JW, Wolff AC, Cope LM, Sukumar S. Genome-wide methylation analysis identifies genes specific to breast cancer hormone receptor status and risk of recurrence. Cancer Res. 2011;71(19):6195–207.

Fackler MJ, Lopez Bujanda Z, Umbricht C, Teo WW, Cho S, Zhang Z, Visvanathan K, Jeter S, Argani P, Wang C, Lyman JP, de Brot M, Ingle JN, Boughey J, McGuire K, King TA, Carey LA, Cope L, Wolff AC, Sukumar S. Novel methylated biomarkers and a robust assay to detect circulating tumor DNA in metastatic breast cancer. Cancer Res. 2014;74(8):2160–70.

Fang F, Turcan S, Rimner A, Kaufman A, Giri D, Morris LG, Shen R, Seshan V, Mo Q, Heguy A, Baylin SB, Ahuja N, Viale A, Massague J, Norton L, Vahdat LT, Moynahan ME, Chan TA. Breast cancer methylomes establish an epigenomic foundation for metastasis. Sci Transl Med. 2011;3(75):75ra25.

Felip E, Martinez P. Can sensitivity to cytotoxic chemotherapy be predicted by biomarkers? Ann Oncol. 2012;23 Suppl 10:x189–92.

Flanagan JM, Cocciardi S, Waddell N, Johnstone CN, Marsh A, Henderson S, Simpson P, da Silva L; kConFab Investigators, Khanna K, Lakhani S, Boshoff C, Chenevix-Trench G. DNA methylome of familial breast cancer identifies distinct profiles defined by mutation status. Am J Hum Genet. 2010;86(3):420–33.

Foekens JA, Sieuwerts AM, Smid M, Look MP, de Weerd V, Boersma AW, Klijn JG, Wiemer EA, Martens JW. Four miRNAs associated with aggressiveness of lymph node-negative, estrogen receptor-positive human breast cancer. Proc Natl Acad Sci U S A. 2008;105(35):13021–6.

Fong PC, Boss DS, Yap TA, Tutt A, Wu P, Mergui-Roelvink M, Mortimer P, Swaisland H, Lau A, O'Connor MJ, Ashworth A, Carmichael J, Kaye SB, Schellens JH, de Bono JS. Inhibition of poly(ADP-ribose) polymerase in tumors from BRCA mutation carriers. N Engl J Med. 2009;361(2):123–34.

Gall TM, Frampton AE, Krell J, Castellano L, Stebbing J, Jiao LR. Blood-based miRNAs as noninvasive diagnostic and surrogative biomarkers in colorectal cancer. Expert Rev Mol Diagn. 2013;13(2):141–5.

García-Giménez JL, Sanchis-Gomar F, Lippi G, Mena S, Ivars D, Gomez-Cabrera MC, Viña J, Pallardó FV. Epigenetic biomarkers: a new perspective in laboratory diagnostics. Clin Chim Acta. 2012;413(19–20):1576–82.

Gargalionis AN, Basdra EK. Insights in microRNAs biology. Curr Top Med Chem. 2013;13(13):1493–502.

Garzon R, Marcucci G, Croce CM. Targeting microRNAs in cancer: rationale, strategies, and challenges. Nat Rev Drug Discov. 2010;9(10):775–89.

Gezer U, Ustek D, Yörüker EE, Cakiris A, Abaci N, Leszinski G, Dalay N, Holdenrieder S. Characterization of H3K9me3- and H4K20me3-associated circulating nucleosomal DNA by high-throughput sequencing in colorectal cancer. Tumour Biol. 2013;34(1):329–36.

Goate A, Chartier-Harlin MC, Mullan M, Brown J, Crawford F, Fidani L, Giuffra L, Haynes A, Irving N, James L. Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease. Nature. 1991;349(6311):704–6.

Grady WM, Carethers JM. Genomic and epigenetic instability in colorectal cancer pathogenesis. Gastroenterology. 2008;135(4):1079–99.

Guay SP, Voisin G, Brisson D, Munger J, Lamarche B, Gaudet D, Bouchard L. Epigenome-wide analysis in familial hypercholesterolemia identified new loci associated with high-density lipoprotein cholesterol concentration. Epigenomics. 2012;4(6):623–39.

He L, Hannon GJ. MicroRNAs: small RNAs with a big role in gene regulation. Nat Rev Genet. 2004;5(7):522–31.

Hegi ME, Diserens AC, Gorlia T, Hamou MF, de Tribolet N, Weller M, Kros JM, Hainfellner JA, Mason W, Mariani L, Bromberg JE, Hau P, Mirimanoff RO, Cairncross JG, Janzer RC, Stupp R. MGMT gene silencing and benefit from temozolomide in glioblastoma. N Engl J Med. 2005;352(10):997–1003.

Heller G, Babinsky VN, Ziegler B, Weinzierl M, Noll C, Altenberger C, Müllauer L, Dekan G, Grin Y, Lang G, End-Pfützenreuter A, Steiner I, Zehetmayer S, Döme B, Arns BM, Fong KM, Wright CM, Yang IA, Klepetko W, Posch M, Zielinski CC, Zöchbauer-Müller S. Genome-wide CpG island methylation analyses in non-small cell lung cancer patients. Carcinogenesis. 2013;34(3):513–21.

Henschke CI, Yankelevitz DF, Libby DM, Pasmantier MW, Smith JP, Miettinen OS. Survival of patients with stage I lung cancer detected on CT screening. N Engl J Med. 2006;355(17): 1763–71.

Herbst A, Wallner M, Rahmig K, Stieber P, Crispin A, Lamerz R, Kolligs FT. Methylation of helicase-like transcription factor in serum of patients with colorectal cancer is an independent predictor of disease recurrence. Eur J Gastroenterol Hepatol. 2009;21(5):565–9.

Heyn H, Esteller M. DNA methylation profiling in the clinic: applications and challenges. Nat Rev Genet. 2012;13(10):679–92.

Heyn H, Méndez-González J, Esteller M. Epigenetic profiling joins personalized cancer medicine. Expert Rev Mol Diagn. 2013a;13(5):473–9.

Heyn H, Carmona FJ, Gomez A, Ferreira HJ, Bell JT, Sayols S, Ward K, Stefansson OA, Moran S, Sandoval J, Eyfjord JE, Spector TD, Esteller M. DNA methylation profiling in breast cancer discordant identical twins identifies DOK7 as novel epigenetic biomarker. Carcinogenesis. 2013b;34(1):102–8.

Hill VK, Ricketts C, Bieche I, Vacher S, Gentle D, Lewis C, Maher ER, Latif F. Genome-wide DNA methylation profiling of CpG islands in breast cancer identifies novel genes associated with tumorigenicity. Cancer Res. 2011;71(8):2988–99.

Hinoue T, Weisenberger DJ, Pan F, Campan M, Kim M, Young J, Whitehall VL, Leggett BA, Laird PW. Analysis of the association between CIMP and BRAF in colorectal cancer by DNA methylation profiling. PLoS One. 2009;4(12):e8357.

Hinoue T, Weisenberger DJ, Lange CP, Shen H, Byun HM, Van Den Berg D, Malik S, Pan F, Noushmehr H, van Dijk CM, Tollenaar RA, Laird PW. Genome-scale analysis of aberrant DNA methylation in colorectal cancer. Genome Res. 2012;22(2):271–82.

Hu XC, Wong IH, Chow LW. Tumor-derived aberrant methylation in plasma of invasive ductal breast cancer patients: clinical implications. Oncol Rep. 2003;10(6):1811–5.

Hughes LA, Khalid-de Bakker CA, Smits KM, van den Brandt PA, Jonkers D, Ahuja N, Herman JG, Weijenberg MP, van Engeland M. The CpG island methylator phenotype in colorectal cancer: progress and problems. Biochim Biophys Acta. 2012;1825(1):77–85.

Iorio MV, Ferracin M, Liu CG, Veronese A, Spizzo R, Sabbioni S, Magri E, Pedriali M, Fabbri M, Campiglio M, Ménard S, Palazzo JP, Rosenberg A, Musiani P, Volinia S, Nenci I, Calin GA, Querzoli P, Negrini M, Croce CM. MicroRNA gene expression deregulation in human breast cancer. Cancer Res. 2005;65(16):7065–70.

Iorio MV, Piovan C, Croce CM. Interplay between microRNAs and the epigenetic machinery: an intricate network. Biochim Biophys Acta. 2010;1799(10–12):694–701.

Iorio MV, Casalini P, Piovan C, Braccioli L, Tagliabue E. Breast cancer and microRNAs: therapeutic impact. Breast. 2011;20 Suppl 3:S63–70.

Jackson K, Yu MC, Arakawa K, Fiala E, Youn B, Fiegl H, Müller-Holzner E, Widschwendter M, Ehrlich M. DNA hypomethylation is prevalent even in low-grade breast cancers. Cancer Biol Ther. 2004;3(12):1225–31.

Jakovcevski M, Akbarian S. Epigenetic mechanisms in neurological disease. Nat Med. 2012;18(8):1194–204. doi:10.1038/nm.2828.

Jamaluddin MS, Yang X, Wang H. Hyperhomocysteinemia, DNA methylation and vascular disease. Clin Chem Lab Med. 2007;45(12):1660–6.

Jemal A, Siegel R, Ward E, Murray T, Xu J, Thun MJ. Cancer statistics, 2007. CA Cancer J Clin. 2007;57(1):43–66.

Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. CA Cancer J Clin. 2011;61(2):69–90.

Jensen LH, Rasmussen AA, Byriel L, Kuramochi H, Crüger DG, Lindebjerg J, Danenberg PV, Jakobsen A, Danenberg K. Regulation of MLH1 mRNA and protein expression by promoter methylation in primary colorectal cancer: a descriptive and prognostic cancer marker study. Cell Oncol (Dordr). 2013;36(5):411–9.

Johnson SM, Grosshans H, Shingara J, Byrom M, Jarvis R, Cheng A, Labourier E, Reinert KL, Brown D, Slack FJ. RAS is regulated by the let-7 microRNA family. Cell. 2005;120(5):635–47.

Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-DNA interactions. Science. 2007;316(5830):1497–502.

Juo YY, Johnston FM, Zhang DY, Juo HH, Wang H, Pappou EP, Yu T, Easwaran H, Baylin S, Engeland M, Ahuja N. Prognostic value of CpG island methylator phenotype among colorectal cancer patients: a systematic review and meta-analysis. Ann Oncol. 2014;25(12):2314–27. pii: mdu149.

Karpinski P, Walter M, Szmida E, Ramsey D, Misiak B, Kozlowska J, Bebenek M, Grzebieniak Z, Blin N, Laczmanski L, Sasiadek MM. Intermediate- and low-methylation epigenotypes do not correspond to CpG island methylator phenotype (low and -zero) in colorectal cancer. Cancer Epidemiol Biomarkers Prev. 2013;22(2):201–8.

Khamas A, Ishikawa T, Shimokawa K, Mogushi K, Iida S, Ishiguro M, Mizushima H, Tanaka H, Uetake H, Sugihara K. Screening for epigenetically masked genes in colorectal cancer using 5-Aza-2′-deoxycytidine, microarray, and gene expression profile. Cancer Genomics Proteomics. 2012;9(2):67–75.

Kibriya MG, Raza M, Jasmine F, Roy S, Paul-Brutus R, Rahaman R, Dodsworth C, Rakibuz-Zaman M, Kamal M, Ahsan H. A genome-wide DNA methylation study in colorectal carcinoma. BMC Med Genomics. 2011;4:50.

Kim T, Reitmair A. Non-coding RNAs: functional aspects and diagnostic utility in oncology. Int J Mol Sci. 2013;14(3):4934–68.

Kim YH, Lee HC, Kim SY, Yeom YI, Ryu KJ, Min BH, Kim DH, Son HJ, Rhee PL, Kim JJ, Rhee JC, Kim HC, Chun HK, Grady WM, Kim YS. Epigenomic analysis of aberrantly methylated genes in colorectal cancer identifies genes commonly affected by epigenetic alterations. Ann Surg Oncol. 2011;18(8):2338–47.

Kratz JR, He J, Van Den Eeden SK, Zhu ZH, Gao W, Pham PT, Mulvihill MS, Ziaei F, Zhang H, Su B, Zhi X, Quesenberry CP, Habel LA, Deng Q, Wang Z, Zhou J, Li H, Huang MC, Yeh CC, Segal MR, Ray MR, Jones KD, Raz DJ, Xu Z, Jahan TM, Berryman D, He B, Mann MJ, Jablons DM. A practical molecular assay to predict survival in resected non-squamous, non-small cell lung cancer: development and international validation studies. Lancet. 2012;379(9818):823–32.

Laird PW. Principles and challenges of genomewide DNA methylation analysis. Nat Rev Genet. 2010;11(3):191–203.

Lange CP, Campan M, Hinoue T, Schmitz RF, van der Meulen-de Jong AE, Slingerland H, Kok PJ, van Dijk CM, Weisenberger DJ, Shen H, Tollenaar RA, Laird PW. Genome-scale discovery of DNA-methylation biomarkers for blood-based detection of colorectal cancer. PLoS One. 2012;7(11):e50266.

Li L, Lee KM, Han W, Choi JY, Lee JY, Kang GH, Park SK, Noh DY, Yoo KY, Kang D. Estrogen and progesterone receptor status affect genome-wide DNA methylation profile in breast cancer. Hum Mol Genet. 2010;19(21):4273–7.

Licchesi JD, Westra WH, Hooker CM, Herman JG. Promoter hypermethylation of hallmark cancer genes in atypical adenomatous hyperplasia of the lung. Clin Cancer Res. 2008;14(9):2570–8.

Liu WJ, Tan XH, Guo BP, Ke Q, Sun J, Cen H. Associations between RASSF1A promoter methylation and NSCLC: a meta-analysis of published data. Asian Pac J Cancer Prev. 2013;14(6):3719–24.

Lo PK, Sukumar S. Epigenomics and breast cancer. Pharmacogenomics. 2008;9(12):1879–902.

Locke I, Kote-Jarai Z, Fackler MJ, Bancroft E, Osin P, Nerurkar A, Izatt L, Pichert G, Gui GP, Eeles RA. Gene promoter hypermethylation in ductal lavage fluid from healthy BRCA gene mutation carriers and mutation-negative controls. Breast Cancer Res. 2007;9(1):R20.

Lockwood WW, Wilson IM, Coe BP, Chari R, Pikor LA, Thu KL, Solis LM, Nunez MI, Behrens C, Yee J, English J, Murray N, Tsao MS, Minna JD, Gazdar AF, Wistuba II, MacAulay CE, Lam S, Lam WL. Divergent genomic and epigenomic landscapes of lung cancer subtypes underscore the selection of different oncogenic pathways during tumor development. PLoS One. 2012;7(5):e37775.

Lu C, Soria JC, Tang X, Xu XC, Wang L, Mao L, Lotan R, Kemp B, Bekele BN, Feng L, Hong WK, Khuri FR. Prognostic factors in resected stage I non-small cell lung cancer: a multivariate analysis of six molecular markers. J Clin Oncol. 2004;22(22):4575–83.

Lund G, Zaina S. Atherosclerosis: an epigenetic balancing act that goes wrong. Curr Atheroscler Rep. 2011;13(3):208–14.

Luo X, Stock C, Burwinkel B, Brenner H. Identification and evaluation of plasma microRNAs for early detection of colorectal cancer. PLoS One. 2013;8(5):e62880.

Luo Y, Wong CJ, Kaz AM, Dzieciatkowski S, Carter KT, Morris SM, Wang J, Willis JE, Makar KW, Ulrich CM, Lutterbaugh JD, Shrubsole MJ, Zheng W, Markowitz SD, Grady WM. Differences in DNA methylation signatures reveal multiple pathways of progression from adenoma to colorectal cancer. Gastroenterology. 2014;147(2):418–29.e8. pii: S0016-5085(14)00595-2.

Mastroeni D, McKee A, Grover A, Rogers J, Coleman PD. Epigenetic differences in cortical neurons from a pair of monozygotic twins discordant for Alzheimer's disease. PLoS One. 2009;4(8):e6617.

Mastroeni D, Grover A, Delvaux E, Whiteside C, Coleman PD, Rogers J. Epigenetic changes in Alzheimer's disease: decrements in DNA methylation. Neurobiol Aging. 2010;31(12): 2025–37.

Menigatti M, Truninger K, Gebbers JO, Marbet U, Marra G, Schär P. Normal colorectal mucosa exhibits sex- and segment-specific susceptibility to DNA methylation at the hMLH1 and MGMT promoters. Oncogene. 2009;28(6):899–909.

Mikeska T, Bock C, Do H, Dobrovic A. DNA methylation biomarkers in cancer: progress towards clinical implementation. Expert Rev Mol Diagn. 2012;12(5):473–87.

Morán A, Fernández-Marcelo T, Carro J, De Juan C, Pascua I, Head J, Gómez A, Hernando F, Torres AJ, Benito M, Iniesta P. Methylation profiling in non-small cell lung cancer: clinical implications. Int J Oncol. 2012;40(3):739–46.

Movassagh M, Choy MK, Knowles DA, Cordeddu L, Haider S, Down T, Siggens L, Vujic A, Simeoni I, Penkett C, Goddard M, Lio P, Bennett MR, Foo RS. Distinct epigenomic features in end-stage failing human hearts. Circulation. 2011;124(22):2411–22.

Müller HM, Oberwalder M, Fiegl H, Morandell M, Goebel G, Zitt M, Mühlthaler M, Ofner D, Margreiter R, Widschwendter M. Methylation changes in faecal DNA: a marker for colorectal cancer screening? Lancet. 2004;363(9417):1283–5.

Müller BM, Jana L, Kasajima A, Lehmann A, Prinzler J, Budczies J, Winzer KJ, Dietel M, Weichert W, Denkert C. Differential expression of histone deacetylases HDAC1, 2, and 3 in human breast cancer–overexpression of HDAC2 and HDAC3 is associated with clinicopathological indicators of disease progression. BMC Cancer. 2013;13:215.

Nakazawa T, Kondo T, Ma D, Niu D, Mochizuki K, Kawasaki T, Yamane T, Iino H, Fujii H, Katoh R. Global histone modification of histone H3 in colorectal cancer and its precursor lesions. Hum Pathol. 2012;43(6):834–42.

Nelson HH, Marsit CJ, Christensen BC, Houseman EA, Kontic M, Wiemels JL, Karagas MR, Wrensch MR, Zheng S, Wiencke JK, Kelsey KT. Key epigenetic changes associated with lung cancer development: results from dense methylation array profiling. Epigenetics. 2012;7(6):559–66.

Nilsson TK, Löf-Öhlin ZM, Sun XF. DNA methylation of the p14ARF, RASSF1A, and APC1A genes as an independent prognostic factor in colorectal cancer patients. Int J Oncol. 2013;42(1):127–33.

Oberwalder M, Zitt M, Wöntner C, Fiegl H, Goebel G, Zitt M, Köhle O, Mühlmann G, Ofner D, Margreiter R, Müller HM. SFRP2 methylation in fecal DNA–a marker for colorectal polyps. Int J Colorectal Dis. 2008;23(1):15–9.

Oster B, Thorsen K, Lamy P, Wojdacz TK, Hansen LL, Birkenkamp-Demtröder K, Sørensen KD, Laurberg S, Orntoft TF, Andersen CL. Identification and validation of highly frequent CpG island hypermethylation in colorectal adenomas and carcinomas. Int J Cancer. 2011;129(12):2855–66.

Oster B, Linnet L, Christensen LL, Thorsen K, Ongen H, Dermitzakis ET, Sandoval J, Moran S, Esteller M, Hansen TF, Lamy P; COLOFOL steering group, Laurberg S, Ørntoft TF, Andersen CL. Non-CpG island promoter hypomethylation and miR-149 regulate the expression of SRPX2 in colorectal cancer. Int J Cancer. 2013;132(10):2303–15.

Pakneshan P, Szyf M, Farias-Eisner R, Rabbani SA. Reversal of the hypomethylation status of urokinase (uPA) promoter blocks breast cancer growth and metastasis. J Biol Chem. 2004;279(30):31735–44.

Paneni F, Costantino S, Volpe M, Lüscher TF, Cosentino F. Epigenetic signatures and vascular risk in type 2 diabetes: a clinical perspective. Atherosclerosis. 2013;230(2):191–7.

Park J, Brena RM, Gruidl M, Zhou J, Huang T, Plass C, Tockman MS. CpG island hypermethylation profiling of lung cancer using restriction landmark genomic scanning (RLGS) analysis. Cancer Biomark. 2005;1(2–3):193–200.

Park JY, Kim D, Yang M, Park HY, Lee SH, Rincon M, Kreahling J, Plass C, Smiraglia DJ, Tockman MS, Kim SJ. Gene silencing of SLC5A8 identified by genome-wide methylation profiling in lung cancer. Lung Cancer. 2013;79(3):198–204.

Patnaik SK, Kannisto E, Knudsen S, Yendamuri S. Evaluation of microRNA expression profiles that may predict recurrence of localized stage I non-small cell lung cancer after surgical resection. Cancer Res. 2010;70:36–45.

Perkins E, Murphy SK, Murtha AP, Schildkraut J, Jirtle RL, Demark-Wahnefried W, Forman MR, Kurtzberg J, Overcash F, Huang Z, Hoyo C. Insulin-like growth factor 2/H19 methylation at birth and risk of overweight and obesity in children. J Pediatr. 2012;161(1):31–9.

Pignon JP, Tribodet H, Scagliotti GV, Douillard JY, Shepherd FA, Stephens RJ, Dunant A, Torri V, Rosell R, Seymour L, Spiro SG, Rolland E, Fossati R, Aubert D, Ding K, Waller D, Le Chevalier T, LACE Collaborative Group. Lung adjuvant cisplatin evaluation: a pooled analysis by the LACE Collaborative Group. J Clin Oncol. 2008;26(21):3552–9.

Potapova A, Hoffman AM, Godwin AK, Al-Saleem T, Cairns P. Promoter hypermethylation of the PALB2 susceptibility gene in inherited and sporadic breast and ovarian cancer. Cancer Res. 2008;68(4):998–1002.

Rao JS, Keleshian VL, Klein S, Rapoport SI. Epigenetic modifications in frontal cortex from Alzheimer's disease and bipolar disorder patients. Transl Psychiatry. 2012;2:e132.

Raponi M, Dossey L, Jatkoe T, Wu X, Chen G, Fan H, Beer DG. MicroRNA classifiers for predicting prognosis of squamous cell lung cancer. Cancer Res. 2009;69:5776–6783.

Rauch TA, Zhong X, Wu X, Wang M, Kernstine KH, Wang Z, Riggs AD, Pfeifer GP. High-resolution mapping of DNA hypermethylation and hypomethylation in lung cancer. Proc Natl Acad Sci U S A. 2008;105(1):252–7.

Robertson AG, Bilenky M, Tam A, Zhao Y, Zeng T, Thiessen N, Cezard T, Fejes AP, Wederell ED, Cullum R, Euskirchen G, Krzywinski M, Birol I, Snyder M, Hoodless PA, Hirst M, Marra MA, Jones SJ. Genome-wide relationship between histone H3 lysine 4 mono- and tri-methylation and transcription factor binding. Genome Res. 2008;18(12):1906–17.

Sanchez-Mut JV, Aso E, Panayotis N, Lott I, Dierssen M, Rabano A, Urdinguio RG, Fernandez AF, Astudillo A, Martin-Subero JI, Balint B, Fraga MF, Gomez A, Gurnot C, Roux JC, Avila J, Hensch TK, Ferrer I, Esteller M. DNA methylation map of mouse and human brain identifies target genes in Alzheimer's disease. Brain. 2013;136(Pt 10):3018–27.

Sanchez-Mut JV, Aso E, Heyn H, Matsuda T, Bock C, Ferrer I, Esteller M. Promoter hypermethylation of the phosphatase DUSP22 mediates PKA-dependent TAU phosphorylation and CREB activation in Alzheimer's disease. Hippocampus. 2014;24(4):363–8.

Sandoval J, Esteller M. Cancer epigenomics: beyond genomics. Curr Opin Genet Dev. 2012;22(1):50–5.

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

Sandoval J, Peiró-Chova L, Pallardó FV, García-Giménez JL. Epigenetic biomarkers in laboratory diagnostics: emerging approaches and opportunities. Expert Rev Mol Diagn. 2013a;13(5):457–71.

Sandoval J, Mendez-Gonzalez J, Nadal E, Chen G, Carmona FJ, Sayols S, Moran S, Heyn H, Vizoso M, Gomez A, Sanchez-Cespedes M, Assenov Y, Müller F, Bock C, Taron M, Mora J, Muscarella LA, Liloglou T, Davies M, Pollan M, Pajares MJ, Torre W, Montuenga LM, Brambilla E, Field JK, Roz L, Lo Iacono M, Scagliotti GV, Rosell R, Beer DG, Esteller M. A prognostic DNA methylation signature for stage I non-small cell lung cancer. J Clin Oncol. 2013b;31(32):4140–7.

Sanfiorenzo C, Ilie MI, Belaid A, Barlési F, Mouroux J, Marquette CH, Brest P, Hofman P. Two panels of plasma microRNAs as non-invasive biomarkers for prediction of recurrence in resectable NSCLC. PLoS One. 2013;8(1):e54596.

Schepeler T, Reinert JT, Ostenfeld MS, Christensen LL, Silahtaroglu AN, Dyrskjøt L, Wiuf C, Sørensen FJ, Kruhøffer M, Laurberg S, Kauppinen S, Ørntoft TF, Andersen CL. Diagnostic and prognostic microRNAs in stage II colon cancer. Cancer Res. 2008;68(15):6416–24.

Schweiger MR, Hussong M, Röhr C, Lehrach H. Genomics and epigenomics of colorectal cancer. Wiley Interdiscip Rev Syst Biol Med. 2013;5(2):205–19.

Selkoe DJ. Preventing Alzheimer's disease. Science. 2012;337(6101):1488–92.

Sharma S, Kelly TK, Jones PA. Epigenetics in cancer. Carcinogenesis. 2010;31(1):27–36.

Shen L, Kondo Y, Ahmed S, Boumber Y, Konishi K, Guo Y, Chen X, Vilaythong JN, Issa JP. Drug sensitivity prediction by CpG island methylation profile in the NCI-60 cancer cell line panel. Cancer Res. 2007;67(23):11335–43.

Sherrington R, Rogaev EI, Liang Y, Rogaeva EA, Levesque G, Ikeda M, Chi H, Lin C, Li G, Holman K, Tsuda T, Mar L, Foncin JF, Bruni AC, Montesi MP, Sorbi S, Rainero I, Pinessi L, Nee L, Chumakov I, Pollen D, Brookes A, Sanseau P, Polinsky RJ, Wasco W, Da Silva HA, Haines JL, Perkicak-Vance MA, Tanzi RE, Roses AD, Fraser PE, Rommens JM, St George-Hyslop PH. Cloning of a gene bearing missense mutations in early-onset familial Alzheimer's disease. Nature. 1995;375(6534):754–60.

Shinozaki M, Hoon DS, Giuliano AE, Hansen NM, Wang HJ, Turner R, Taback B. Distinct hypermethylation profile of primary breast cancer is associated with sentinel lymph node metastasis. Clin Cancer Res. 2005;11(6):2156–62.

Shoemaker RH. The NCI60 human tumour cell line anticancer drug screen. Nat Rev Cancer. 2006;6(10):813–23.

Son JW, Jeong KJ, Jean WS, Park SY, Jheon S, Cho HM, Park CG, Lee HY, Kang J. Genome-wide combination profiling of DNA copy number and methylation for deciphering biomarkers in non-small cell lung cancer patients. Cancer Lett. 2011;311(1):29–37.

Song JS, Kim YS, Kim DK, Park SI, Jang SJ. Global histone modification pattern associated with recurrence and disease-free survival in non-small cell lung cancer patients. Pathol Int. 2012;62(3):182–90.

Sozzi G, Boeri M, Rossi M, Verri C, Suatoni P, Bravi F, Roz L, Conte D, Grassi M, Sverzellati N, Marchiano A, Negri E, La Vecchia C, Pastorino U. Clinical utility of a plasma-based miRNA signature classifier within computed tomography lung cancer screening: a correlative MILD trial study. J Clin Oncol. 2014;32(8):768–73.

Stefansson OA, Villanueva A, Vidal A, Martí L, Esteller M. BRCA1 epigenetic inactivation predicts sensitivity to platinum-based chemotherapy in breast and ovarian cancer. Epigenetics. 2012;7(11):1225–9.

Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature. 2009;458(7239):719–24.

Szyf M. Epigenetics, DNA methylation, and chromatin modifying drugs. Annu Rev Pharmacol Toxicol. 2009;49:243–63.

Tang X, Khuri FR, Lee JJ, Kemp BL, Liu D, Hong WK, Mao L. Hypermethylation of the death-associated protein (DAP) kinase promoter and aggressiveness in stage I non-small cell lung cancer. J Natl Cancer Inst. 2000;92(18):1511–6.

Tang D, Liu J, Wang DR, Yu HF, Li YK, Zhang JQ. Diagnostic and prognostic value of the methylation status of secreted frizzled-related protein 2 in colorectal cancer. Clin Invest Med. 2011;34(2):E88–95.

Tanzi RE. The genetics of Alzheimer's disease. Cold Spring Harb Perspect Med. 2012;2(10). pii: a006296.

Timp W, Levchenko A, Feinberg AP. A new link between epigenetic progenitor lesions in cancer and the dynamics of signal transduction. Cell Cycle. 2009;8(3):383–90.

Tóth K, Sipos F, Kalmár A, Patai AV, Wichmann B, Stoehr R, Golcher H, Schellerer V, Tulassay Z, Molnár B. Detection of methylated SEPT9 in plasma is a reliable screening method for both left- and right-sided colon cancers. PLoS One. 2012;7(9):e46000.

Toyooka S, Toyooka KO, Maruyama R, Virmani AK, Girard L, Miyajima K, Harada K, Ariyoshi Y, Takahashi T, Sugio K, Brambilla E, Gilcrease M, Minna JD, Gazdar AF. DNA methylation profiles of lung tumors. Mol Cancer Ther. 2001;1(1):61–7.

Turunen MP, Aavik E, Ylä-Herttuala S. Epigenetics and atherosclerosis. Biochim Biophys Acta. 2009;1790(9):886–91.

Usadel H, Brabender J, Danenberg KD, Jerónimo C, Harden S, Engles J, Danenberg PV, Yang S, Sidransky D. Quantitative adenomatous polyposis coli promoter methylation analysis in tumor tissue, serum, and plasma DNA of patients with lung cancer. Cancer Res. 2002;62(2):371–5.

Van Den Broeck A, Brambilla E, Moro-Sibilot D, Lantuejoul S, Brambilla C, Eymin B, Khochbin S, Gazzeri S. Loss of histone H4K20 trimethylation occurs in preneoplasia and influences prognosis of non-small cell lung cancer. Clin Cancer Res. 2008;14(22):7237–45.

Van Neste L, Herman JG, Otto G, Bigley JW, Epstein JI, Van Criekinge W. The epigenetic promise for prostate cancer diagnosis. Prostate. 2012;72(11):1248–61.

Vasilatos SN, Broadwater G, Barry WT, Baker Jr JC, Lem S, Dietze EC, Bean GR, Bryson AD, Pilie PG, Goldenberg V, Skaar D, Paisie C, Torres-Hernandez A, Grant TL, Wilke LG, Ibarra-Drendall C, Ostrander JH, D'Amato NC, Zalles C, Jirtle R, Weaver VM, Seewaldt VL. CpG island tumor suppressor promoter methylation in non-BRCA-associated early mammary carcinogenesis. Cancer Epidemiol Biomarkers Prev. 2009;18(3):901–14.

Vaz C, Ahmad HM, Sharma P, Gupta R, Kumar L, Kulshreshtha R, Bhattacharya A. Analysis of microRNA transcriptome by deep sequencing of small RNA libraries of peripheral blood. BMC Genomics. 2010;11:288.

Veeck J, Esteller M. Breast cancer epigenetics: from DNA methylation to microRNAs. J Mammary Gland Biol Neoplasia. 2010;15(1):5–17.

Veeck J, Ropero S, Setien F, Gonzalez-Suarez E, Osorio A, Benitez J, Herman JG, Esteller M. BRCA1 CpG island hypermethylation predicts sensitivity to poly(adenosine diphosphate)-ribose polymerase inhibitors. J Clin Oncol. 2010;28(29):e563–4.

Vo QN, Kim WJ, Cvitanovic L, Boudreau DA, Ginzinger DG, Brown KD. The ATM gene is a target for epigenetic silencing in locally advanced breast cancer. Oncogene. 2004;23(58):9432–7.

Waddington C. The epigenotype. Endeavour. 1942;1:18–20.

Walter K, Holcomb T, Januario T, Du P, Evangelista M, Kartha N, Iniguez L, Soriano R, Huw L, Stern H, Modrusan Z, Seshagiri S, Hampton GM, Amler LC, Bourgon R, Yauch RL, Shames DS. DNA methylation profiling defines clinically relevant biological subsets of non-small cell lung cancer. Clin Cancer Res. 2012;18(8):2360–73.

Wang Y, Yu Z, Wang T, Zhang J, Hong L, Chen L. Identification of epigenetic aberrant promoter methylation of RASSF1A in serum DNA and its clinicopathological significance in lung cancer. Lung Cancer. 2007;56(2):289–94.

Wang S, Xiang J, Li Z, Lu S, Hu J, Gao X, Yu L, Wang L, Wang J, Wu Y, Chen Z, Zhu H. A plasma microRNA panel for early detection of colorectal cancer. Int J Cancer. 2013;136(1):152–61.

Warren JD, Xiong W, Bunker AM, Vaughn CP, Furtado LV, Roberts WL, Fang JC, Samowitz WS, Heichman KA. Septin 9 methylated DNA is a sensitive and specific blood test for colorectal cancer. BMC Med. 2011;9:133.

Widschwendter M, Jones PA. DNA methylation and breast carcinogenesis. Oncogene. 2002;21(35):5462–82.

World Health Organization. World health statistics report. 2011. http://www.who.int/whosis/whostat/2011/en/.

Wrangle J, Wang W, Koch A, Easwaran H, Mohammad HP, Vendetti F, Vancriekinge W, Demeyer T, Du Z, Parsana P, Rodgers K, Yen RW, Zahnow CA, Taube JM, Brahmer JR, Tykodi SS, Easton K, Carvajal RD, Jones PA, Laird PW, Weisenberger DJ, Tsai S, Juergens RA, Topalian SL, Rudin CM, Brock MV, Pardoll D, Baylin SB. Alterations of immune response of non-small cell lung cancer with azacytidine. Oncotarget. 2013;4(11):2067–79.

Wrangle J, Machida EO, Danilova L, Hulbert A, Franco N, Zhang W, Glöckner SC, Tessema M, Van Neste L, Easwaran H, Schuebel KE, Licchesi J, Hooker CM, Ahuja N, Amano J, Belinsky SA, Baylin SB, Herman JG, Brock MV. Functional identification of cancer-specific methylation of CDO1, HOXA9, and TAC1 for the diagnosis of lung cancer. Clin Cancer Res. 2014;20(7):1856–64.

Wu WK, Law PT, Lee CW, Cho CH, Fan D, Wu K, Yu J, Sung JJ. MicroRNA in colorectal cancer: from benchtop to bedside. Carcinogenesis. 2011;32(3):247–53.

Xing X, Cai W, Shi H, Wang Y, Li M, Jiao J, Chen M. The prognostic value of CDKN2A hypermethylation in colorectal cancer: a meta-analysis. Br J Cancer. 2013a;108(12):2542–8.

Xing XB, Cai WB, Luo L, Liu LS, Shi HJ, Chen MH. The prognostic value of p16 hypermethylation in cancer: a meta-analysis. PLoS One. 2013b;8(6):e66587.

Xu X, Gammon MD, Zhang Y, Cho YH, Wetmur JG, Bradshaw PT, Garbowski G, Hibshoosh H, Teitelbaum SL, Neugut AI, Santella RM, Chen J. Gene promoter methylation is associated with increased mortality among women with breast cancer. Breast Cancer Res Treat. 2010;121(3):685–92.

Xu Z, Bolick SC, DeRoo LA, Weinberg CR, Sandler DP, Taylor JA. Epigenome-wide association study of breast cancer using prospectively collected sister study samples. J Natl Cancer Inst. 2013;105(10):694–700.

Yi JM, Dhir M, Van Neste L, Downing SR, Jeschke J, Glöckner SC, de Freitas Calmon M, Hooker CM, Funes JM, Boshoff C, Smits KM, van Engeland M, Weijenberg MP, Iacobuzio-Donahue CA, Herman JG, Schuebel KE, Baylin SB, Ahuja N. Genomic and epigenomic integration identifies a prognostic signature in colon cancer. Clin Cancer Res. 2011;17(6):1535–45.

Yu SL, Chen HY, Chang GC, Chen CY, Chen HW, Singh S, Cheng CL, Yu CJ, Lee YC, Chen HS, Su TJ, Chiang CC, Li HN, Hong QS, Su HY, Chen CC, Chen WJ, Liu CC, Chan WK, Chen WJ, Li KC, Chen JJ, Yang PC. MicroRNA signature predicts survival and relapse in lung cancer. Cancer Cell. 2008;13:48–57.

Zhang JX, Song W, Chen ZH, Wei JH, Liao YJ, Lei J, Hu M, Chen GZ, Liao B, Lu J, Zhao HW, Chen W, He YL, Wang HY, Xie D, Luo JH. Prognostic and predictive value of a microRNA signature in stage II colon cancer: a microRNA expression analysis. Lancet Oncol. 2013;14(13):1295–306.