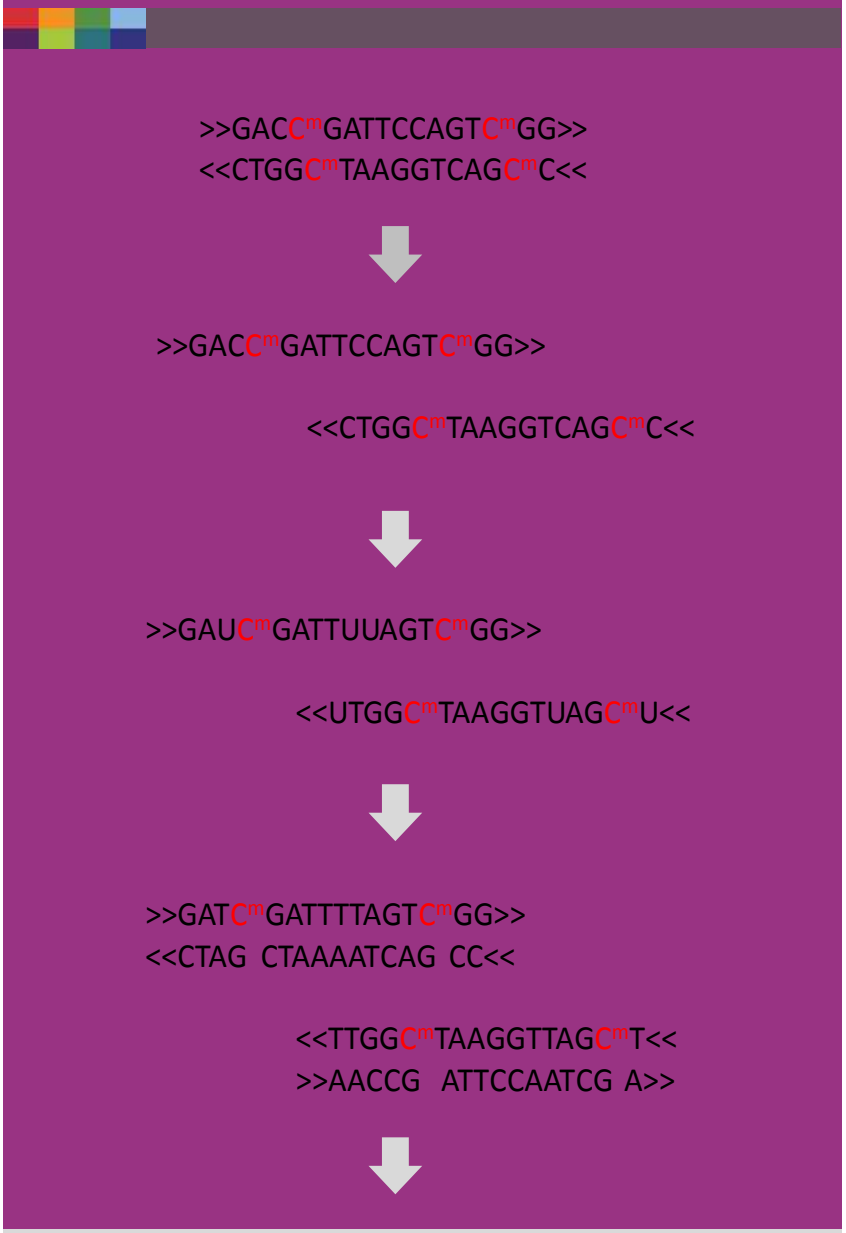


Development of computational approaches for the analysis of bisulfite next-generation sequencing data

Dissertation zur Erlangung des Doktorgrades der Mathematisch-Naturwissenschaftlichen Fakultät der Christian-Albrechts-Universität zu Kiel

vorgelegt von

Benjamin Kreck

Kiel, 2012

Development of computational approaches for the analysis of bisulfite next-generation sequencing data

Dissertation zur Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Christian-Albrechts-Universität zu Kiel

vorgelegt von Benjamin Kreck

Kiel, 2012

Referent/in:	Prof. Dr. Andre Franke
Koreferent/in:	Prof. Dr. Hinrich Schulenburg
Tag der mündlichen Prüfung:	06.08.2012
Zum Druck genehmigt:	Prof. Dr. Andre Franke
Dekan:	Prof. Dr. Wolfgang Duschl

„Probleme kann man niemals mit der selben Denkweise lösen, durch die sie entstanden sind.“

(Albert Einstein)

Für Nick, Kristina und meine Familie

Acknowledgments

Prof. Dr. Andre Franke	for providing the PhD project, giving me the opportunity to get insight into such a fascinating research field and all his support
Prof. Dr. Stefan Schreiber	for providing the research environment and the excellent working conditions
Prof. Dr. Reiner Siebert	for constructive discussions and the interesting DAUDI project
Dr. Felix Krüger	for helpful discussions about bisulfite sequencing
Eva Ellinghaus, Tobias Balschun, Georg Hemmrich-Stanisak	for an enriching office community
Next-generation sequencing lab	for generating bisulfite sequencing libraries and helpful discussions
All of the members of the IKMB	for a good working atmosphere

Declaration

Herewith I confirm that this thesis is completely the result of my own work. Apart from the advice of my supervisors, all sources are listed in the references. This thesis has not been submitted elsewhere. It has been carried out in strict accordance with the rules of good scientific practice of the *Deutsche Forschungsgesellschaft*.

Ich bestätige hiermit, dass diese Dissertation das Resultat meiner eigenen Arbeit ist. Abgesehen vom Rat meiner Betreuer, sind sämtliche Quellen in den Referenzen aufgeführt. Diese Arbeit lag und liegt nirgends sonst im Rahmen eines Promotionsverfahrens vor. Die Arbeit wurde unter Einhaltung der Regeln guter wissenschaftlicher Praxis der *Deutschen Forschungsgesellschaft* verfasst.

_____ _____

Signature Date

Table of Contents

List of Figures

List of Tables

Abbreviations

ASM	Allele-specific DNA methylation
A^{vy}	The agouti viable yellow allele
BRCA1	Breast-cancer susceptibility gene 1
BS-seq	Bisulfite sequencing
BWS	Beckwith-Wiedemann syndrome
CpG dinucleotides	Cytosine and guanine nucleotides separated by a phosphate
DMRs	Differentially methylated regions
DNA	Deoxyribonucleic acid
DNMT	DNA methyltransferase
DRD4	The dopamine receptor 4 gene
EBV	Epstein-Barr virus
ES cells	Embryonic stem cells
EWASs	Epigenome-wide association studies
Gb	Gigabases
GRNs	Gene regulatory networks
GWASs	Genome-wide association studies
hMLH1	A homologue of MutL *Escherichia coli*
HSM	haplotype-specific DNA methylation
IHGSC	International Human Genome Sequencing Consortium
InDels	Sequence insertion or deletion
iPSCs	Induced pluripotent stem cells
MAOA	Monoamine oxidase A gene
MBD-seq	Methylated DNA binding domain sequencing
MeDIP-seq	Methylated DNA immunoprecipitation sequencing
NGS	Next-generation sequencing
PBMCs	Human peripheral blood mononuclear cells
PCR	Polymerase chain reaction
Rb	Retinoblastoma gene
RNA-seq	RNA sequencing
RRBS	Reduced Representation Bisulfite Sequencing
S100A4	S100 calcium binding protein A4 gene
SERT	The serotonin transporter gene
siRNAs	Small interfering RNAs
SMRT	Single-molecule real-time sequencing
SNP	Single nucleotide polymorphisms
TET family	Ten-eleven translocation family
TSS	Transcription start site
VMRs	Variably methylated regions

List of Publications

Published Works by the Author Incorporated into the Thesis

"DNA methylome analysis using short bisulfite sequencing data."

Krueger F[*], **Kreck B**[*], Franke A, Andrews SR

Nature Methods (2012), Jan 30;9(2):145-51

Citation format for paper – Added as Appendix B.

[*]Authors that contributed equally to the manuscript.

"B-SOLANA: An approach for the analysis of two-base encoding bisulfite sequencing data"

Kreck B, Marnellos G, Richter J, Krueger F, Siebert R, Franke A

Bioinformatics (2011), 2012 Feb 1;28(3):428-9. Epub 2011 Dec 6

Citation format for paper – Added as Appendix A.

Submitted Works by the Author Incorporated into the Thesis

"Analysis of a Base-Pair Resolution DNA Methylome from an Endemic Burkitt Lymphoma"

Kreck B, Richter J, Ammerpohl O, Barann M, Esser D, Pedersen BS, Vater I, Murga Penas EM, Chung CA, Seisenberger S, Boyd VL, Smallwood S, Drexler HG, MacLeod RAF, Hummel M, Krueger F, Häsler R, Schreiber S, Rosenstiel P, Franke A, Siebert R

Submitted to **Leukemia** (August 2012)

Citation format for paper – Added as Appendix C.

Co-authorships of Works not Incorporated into the Thesis

"A tissue-specific landscape of sense/antisense transcription in the mouse intestine."
Klostermeier UC, Barann M, Wittig M, Haesler R, Franke A, Gavrilova O, **Kreck B**, Sina C, Schilhabel MB, Schreiber S, Rosenstiel P
BMC Genomics (2011), Jun 10;12:305

Co-authorships of Works Under Revision not Incorporated into the Thesis

"From next-generation sequencing alignments to accurate comparison and validation of single nucleotide variants: the pibase software"
Forster M, Forster P, Elsharawy A, Hemmrich G, **Kreck B**, Wittig M, Thomsen I, Stade B, Barann M, Ellinghaus D, Petersen BS, May S, Melum E, Schilhabel MB, Keller A, Schreiber S, Rosenstiel P, Franke A
Under revision, **Nucleic Acids Research** (2012)

Chapter 1 Introduction

1.1 Epigenetics

Epigenetic mechanisms were originally described by the British embryologist C.H. Waddington in 1939 as "the causal interactions between genes and their products, which bring the phenotype into being" (Waddington, 1939). Today, epigenetics is defined as the study of mechanisms that involve changes in gene expression which are not accompanied by changes of the DNA sequence (Holliday, 1987). The fact that classical genetics alone cannot explain the development of an organism's phenotype, was exemplified by Fraga *et al.* and Humpherys *et al.*, who showed that monozygotic twins or cloned animals can have different phenotypes and different susceptibilities to a disease although they exhibit identical DNA sequences (Mario F Fraga et al., 2005; Humpherys et al., 2001). Epigenetic mechanisms offer a partial explanation for these phenomena and by now, a range of different chemical modifications to deoxyribonucleic acid (DNA) and histones are known to be associated with changes in gene expression (see Figure 1.1).

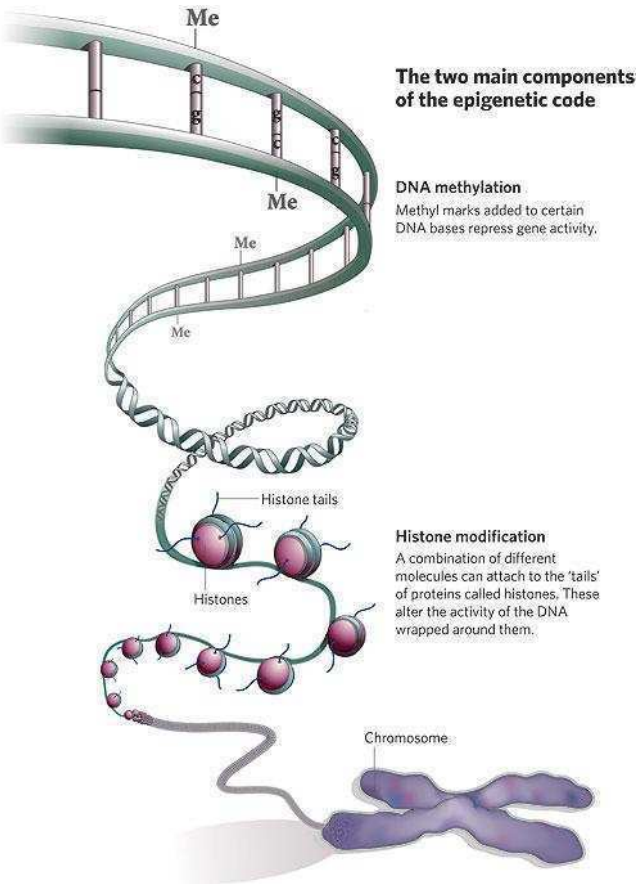


Figure 1.1: Epigenetic mechanisms: DNA methylation and histone modifications.

Figure from (Qiu, 2006)

The most-studied epigenetic changes are DNA methylation and histone modifications. DNA methylation involves the addition of a methyl group from S-adenosylmethionin to the 5 position of the cytosine pyrimidine ring or the number 6 nitrogen of the adenine purine ring (Holliday and Pugh, 1975). Histone modifications are post-translational enzymatic modifications of the histones by acetylation (addition of an acetyl group), methylation (addition of methyl group), phosphorylation (the addition of a phospholyration group) and ubiquitination (addition of an ubiquinitation protein) (Bártová et al., 2008). These epigenetic mechanisms are necessary for the development of higher eukaryotes and are particularly important in several key physiological processes, including regulation of gene expression, X-chromosome inactivation, imprinting, as well as silencing of germline-specific genes and repetitive elements (Robertson, 2005). Detailed information about DNA methylation is provided in section 1.2.

Inheritance is defined as the transmission of information between generations of organisms. Thus, the epigenetic property to be replicated during mitotic cell division should rather be considered as mitotic stability (Skinner, 2011). Mitotic stability of DNA methylation is comprehensively described in section 1.2. The ability of the epigenome (the overall epigenetic state of a cell) to be replicated between generations of species is called epigenetic meiotic inheritance (Bock and Lengauer, 2008). The pluripotency of cells of the early embryo is ensured by a reset mechanism of the epigenetic information after fertilization (Wolf Reik, 2007). This reprogramming enables embryonic stem cells (ES cells) to differentiate down to any pathway. Epigenetic meiotic inheritance occurs by an incomplete reprogramming in the early embryo (Morgan et al., 1999). Consequently, epigenetic patterns might be carried from parent to offspring. Meiotic inheritance was initially identified in plants (Bender and Fink, 1995). Bender and Fink described meiotic inheritance of DNA methylation patterns and their associated phenotypes of a gene family of the Wassilewskaija strain of *Arabidopsis*. They were able to show that methylated and silenced loci can be meiotically transmitted through self-pollination and outcrosses (Bender and Fink, 1995). Further research showed that DNA methylation patterns can be replicated over more than one generation (Johannes et al., 2009) (see section 1.5). However, the underlying functional mechanisms remain to be clarified. Epigenetic inheritance and stability and its impact on the cell population or associated physiology is part of recent epigenetic

research. Reik provided a fundamental insight into heritable aspects of epigenetic gene regulation in mammalian development (Wolf Reik, 2007).

Interestingly, environmental factors are able to modify the epigenome of somatic cells (Skinner, 2011). The following section highlights environmental-based influences on epigenetic modifications. The replication of environmentally influenced epigenetic patterns by mitotic stability modifies the somatic cell differentiation and function throughout the development of an organism and its susceptibility to a disease (Jirtle and Skinner, 2007). Honeybees pose an interesting example for such an environmental-based alteration of a phenotype regulated by DNA methylation (Kucharski et al., 2008). Larvae predominantly fed royal jelly become more likely queens due to a specific signal cascade (Maleszka). Kucharski *et al.* pointed out that down-regulation of the DNA methyltransferase during larval development leads to an increased number of queens not fed royal jelly (Kucharski et al., 2008). This study exemplifies a potential influence of nutrition on DNA methylation.

Recent biological research shows the importance of studying epigenetic influences on complex diseases and aging (Flintoft, 2010) (see Table 1). Especially several potential links between epigenetic mechanisms and cancer have been identified so far (Manel Esteller, 2008). A study in the field of aging research showed that age-dependent methylation patterns have an impact on neurologic disorders, autoimmunity, and the development of cancer in elderly people (Richardson, 2003). A tissue-specific loss of DNA methylation has been identified, which may lead to chromosomal instability and neoplasia and Richardson *et al.* stated that global DNA hypermethylation increases the risk of colon cancer with advancing age (Richardson, 2003).

Table 1: Associations between epigenetic modifications and human diseases.

Table modified from (Rodenhiser and Mann, 2006)

Disease/condition	*Gene*	*Epigenetic process*	*References*
Beckwith-Weidemann syndrome	11p15	Imprinting	(Weksberg et al., 2003)
Breast cancer	BRCA1	Hypermethylation	(Mancini et al., 1998)
Colon cancer	Multiple genes	Hypermethylation	(M Esteller, Corn, et al., 2001)
Leukemia	p15	Hypermethylation	(M Esteller, Corn, et al., 2001)
Lung cancer	p16, p73	Hypermethylation	(M Esteller, Corn, et al., 2001)
Prader-Willi syndrome or Angelman syndrome	15q11-q13	Imprinting	(Nicholls and Knepper, 2001)
Shizophrenia	RELN	Hypermethylation	(Sharma, 2005; Costa et al., 2002)
Stomach cancer	Cyclin D2	Hypomethylation	(Oshimo et al., 2003)

In summary, epigenetics involves the study of alterations in gene expression caused by heritable and non-heritable biochemical mechanisms, other than changes in the underlying DNA sequence. Recent research focuses on DNA methylation and histone modifications and their impact on transcriptional control. These mechanisms play a determining role in the development and growth of cells. It has been shown that epigenetic abnormalities pose a riskfactor for complex disease.

1.2 DNA methylation

Nowadays several research projects, such as the Human Genome Project, have sequenced various genomes (IHGSC, 2004). However the DNA sequence of an organism is insufficient to describe its phenotype. It is important to know when and where a specific gene will be transcribed. DNA methylation, as a heritable epigenetic modification, is able to control gene expression. It is the only known epigenetic mechanism that directly concerns the DNA without changing the underlying DNA sequence.

DNA methylation involves the addition of a methyl group from S-adenosylmethionin to the 5 position of the cytosine pyrimidine ring or to the number 6 nitrogen of the adenine purine ring (Holliday and Pugh, 1975). Eukaryotes solely make use of cytosine methylation, whereas in prokaryotes both cytosine and adenine can be methylated (Jeltsch, 2002). DNA

methylation in prokaryotes controls DNA replication and gene expression and has a protective function in terms of distinction of self and non-self DNA. This distinction has been associated with defense against bacteriophages (Arber and Linn, 1969). Such protective mechanism also exists in eukaryotes: transgenes, introduced into humans or mice, can be detected and silenced by DNA methylation (Kisseljova et al., 1998; Sasaki et al., 1993). Thus, this distinguishing mechanism seems to be conserved in evolution.

During mitotic cell division, both genetic and epigenetic information must be replicated to daughter cells. This phenomenon results in differentiated states of cells and enables a normal development process. The enzyme DNA methyltransferase (DNMT) identifies DNA methylation patterns of the parental cell during the replication of the DNA (synthesis phase) and methylates the replicated strand of the daughter cell (Goll and T. H. Bestor, 2005). Goll and Bestor showed that especially repetitive DNA sequences and RNA-DNA interaction mediate the establishment and maintenance of DNA methylation by DNMT. Originally, the cytosine methylatransferase was identified in 1988 (T. Bestor et al., 1988) and so far several homologues of this enzyme are known. They are split into three groups categorized based on their C-terminal catalytic domains: DNMT1 family, DNMT3 family, and chromomethylase family (Goll and T. H. Bestor, 2005; Goll et al., 2006). Originally, DNMT2 was also identified as a DNA methyltransferase (Goll and T. H. Bestor, 2005), however, Goll *et al.* showed that DNMT2 acts as RNA specific methyltransferase (Goll et al., 2006). Meanwhile, DNMT2 is known as TRDMT1, which is the only identified RNA methyltransferase so far (Squires et al., 2012).

Human DNA methylation establishment and maintenance is regulated by DNMT1 and DNMT3 (T. Bestor et al., 1988). It has been shown that hemimethylated DNA (only one strand of a double-stranded DNA sequence is methylated) is much faster methylated than completely unmethylated DNA sequences (Stein et al., 1982). This maintaining mechanism is carried out by DNMT1, which involves a 5-30 times faster DNA methylation mechanism compared to *de novo* DNA methylation (Yoder et al., 1997). *De novo* DNA methylation in humans is mediated mainly by DNMT1 but can also be mediated by members of the DNMT3 family, which consists of DNMT3A, DNMT3B, and DNMT3L (Okano et al., 1998; Goll and T. H. Bestor, 2005). In cancer, DNMT1 is used for *de novo* and maintaining DNA methylation of tumor suppressor genes (Jair et al., 2006; Ting et al., 2006). The DNMT3 family involves two different types of regulatory mechanisms. DNMT3A and DNMT3B mediate *de novo* and

maintaining DNA methylation of CpG dinucleotides (cytosine and guanine nucleotides separated by a phosphate). In contrast, DNMT3L does not involve any methyltransferase activity (Goll and T. H. Bestor, 2005). It especially regulates mechanisms in germ cells and is important for the establishment of maternal genomic imprinting patterns (Goll and T. H. Bestor, 2005). Goll and Bestor stated that DNMT3A and DNMT3B do not involve any sequence specificity beyond CpG dinucleotides (Goll and T. H. Bestor, 2005).

In prokaryotes, DNA methylation regulates the mitotic replication process of the DNA sequence. After the replication, the synthesized strand is not directly methylated thereby allowing the mismatch repair system to differentiate between the template and nascent strands (Cooper et al., 1993). This self-adjusting error correction does not exist in eukaryotic organisms. Araujo *et al.* investigated whether there is also a lag between DNA replication and DNA methylation patterns by methyltransferase in mammalian cells (Araujo et al., 1998). They stated that DNA methylation at CpG dinucleotides simultaneously occurs with the replication of the underlying DNA sequence. This tight coordination of genetic and epigenetic replication is characteristic for mammalian cells (Araujo et al., 1998). Methylation occurs independently of the genomic distance to the origin of the replication and prior to ligation of Okazaki fragments (short molecules of single-stranded DNA that are formed on the lagging strand during DNA replication).

DNA methylation usually takes place at CpG dinucleotides (Pelizzola and Ecker, 2010). Almost a fifth (~19%) of all bases in the human reference (hg19/build37) DNA sequence are Cs and another 19% are Gs, whilst only ~1.8% of all dinucleotides are CpGs. The frequency of CpG dinucleotides is therefore much lower than expected based on the GC content (Bird, 1980), which is due to the inherent mutability of methylated cytosines (Venter et al., 2001). Deamination of cytosine includes the hydrolysis reaction of cytosine into uracil, whereas spontaneous deamination of methylated cytosine results in thymine (Venter et al., 2001). Because uracils are a component of the RNA but not DNA, a mechanism exists, which recognizes and repairs deaminated cytosines, but deaminated methylated cytosines, namely thymines, remain unmodified (Singal and Ginder, 1999; Bird, 1980; Duncan and Miller, 1980). As a consequence, methylated CpG sites are more likely to get lost during cell differentiation.

Methylated CpG sites in the genome are not equally distributed. Regions characterized by a high G+C content and a high frequency of CpG dinucleotides, are called CpG islands (Bird,

1986). There are several numerical definitions of CpG islands but the most commonly used definition is: a CpG island is a genomic region with a G+C content of greater than 60% and a ratio of CpGs to GpCs of at least 0.6 (S B Baylin et al., 1998). Most of CpG islands are unmethylated and are often located within upstream regions of genes (Pelizzola and Ecker, 2010). Hence, they have a strong impact on transcriptional gene regulation (Koga et al., 2009). CpG sites within CpG islands usually involve weak DNA methylation levels in the germ line, which involves a protective mechanism regarding deamination of methylated cytosines during mitotic cell division (Fazzari and Greally, 2004).

Methylated cytosines also occur outside of CpG dinulceotides. These sites are called non-CpG methylation sites, whereas CHG and CHH (H being A, C or T) methylation sites involve a regulatory effect on transcriptional levels (Lister et al., 2009; Pelizzola and Ecker, 2010). Plants often exhibit an enrichment of methylated non-CpG sites, which mediate dynamical interaction with small interfering RNAs (siRNAs) (Koga et al., 2009). Variying distribution of CpG and non-CpG methylation frequencies can be found in different eukaryotic organisms, as exemplified in the following (see Table 2). A notably high amount of non-CpG methylation can be found in Physcomitrella patens, a moss plant, which exhibits 29.7% of methylated CHG and 23.2% of methylated CHH sites (Pelizzola and Ecker, 2010). Until today, non-CpG methylation in human genomes has predominantly been found in ES cells (Lister et al., 2009). Almost 25% of all methylation sites in ES cells occur within non-CpG dinucleotides, which implicates further, so far unknown, regulatory mechanisms of methylation (Lister et al., 2011). Most often gene bodies show an enrichment of non-CpG methylation sites, whereby upstream regions of genes and protein binding sites are depleted (Lister et al., 2009). Furthermore, non-CpG methylation disappears during cell differentiation and is re-established in induced pluripotent stem cells (Lister et al., 2009) (see sections 1.2.1).

Table 2: Frequencies of CpG and non-CpG methylation in 23 eukaryotes.

Table modified from (Pelizzola and Ecker, 2010)

Eukaryotic organisms	*Genome size (Mb)*	*mCpG*	*mCHG*	*mCHH*
Nematosella vectensis	297	9.4	0.16	0.15
Mus musculus	2716	74.2	0.30	0.29
Homo sapies	3080	67-82	0.09-2.8	0.04-0.9
Tetraodon nigrovirdis	302	65.5	0.25	0.34
Danio rerio	1563	80.3	1.22	0.91
Ciona intestinalis	141	21.6	0.28	0.28
Apis melifera	231	0.51	0.11	0.16
Drosophila melanogaster	162	0.12	0.11	0.11
Bombyx mori	431	0.71	0.08	0.09
Tribolium castaneum	151	0.11	0.12	0.12
Uncinocarpus reesii	22	0.67	-	-
7Coprinopsis cinerea	36	12.2	-	-
Phycomyces blakesleeanus	51	4.9	-	-
Selaginella moellendorffii	101	12.5	9.0	0.92
Oryza sativa	372	50.0	27.4	5.2
Arabidopsis thaliana	120	22.3	5.92	1.51
Populus trichocarpa	485	41.9	20.9	3.25
Physcomitrella patens	454	29.5	29.7	23.2
Cholella sp. NC64A	42	80.5	2.2	0.25
Volvox carteri	126	2.6	0.08	0.08
Chlamydomonas reinhardtii	120	5.4	2.59	2.49

A modification of DNA methylation is the oxidation of 5-methylcytosine to 5-hydroxymethylated cytosine (Ficz et al., 2011). Most of the methods, which are used to detect DNA methylation are inapplicable for the determination of hydroxymethylation patterns (Harris et al., 2010). Corresponding approaches make use of thin-layer chromotography (Jin et al., 2010). Hydroxymethylation is catalyzed by proteins of the ten-eleven translocation (TET) family that are highly expressed in ES cells (Tahiliani et al., 2009). Recent studies showed that 5-hydroxymethylated cytosines are enriched in euchromatic parts of the genome including regulatory regions as for instance promoters and CpG islands (Ficz et al., 2011). Ficz *et al.* found out that frequent appearance of 5-hydroxymethylation

correlates with high expression level of the underlying DNA sequence. However, the underlying functional mechanism remains to be clarified.

1.2.1 Human Methylomes

Recent biotechnological developments in the field of next-generation sequencing enable cost-effective analysis of methylomes for multiple-gigabase genomes, like the human genome. Lister *et al.* provided the first genome-wide insight into DNA methylation at single-base resolution in mammalian genomes in 2009 (Lister et al., 2009). They generated maps of methylated cytosines of human ES cells and fetal fibroblasts. Constitutive differences of DNA methylation patterns were identified for these two cell types (Lister et al., 2009). This primordial map of DNA methylation patterns can be used as a reference for methylomes of differentiated cell types. By now, additional human methylomes were analyzed, as for instance of human colon cancer cells and of human peripheral blood mononuclear cells (Hansen et al., 2011; Li et al., 2010).

Typically, differences based on DNA methylation are associated with cell type specific patterns of CpG methylation. Lister *et al.* additionally detected substantial distinctions based on non-CpG methylation (Lister et al., 2009). ES cells exhibit an enrichment of methylated non-CpG sites compared to differentiated cells (Lister et al., 2009). Nearly 25% of all methylation sites in ES cells were within non-CpG dinucleotides, exhibiting an enrichment of CAH and CAG trinucleotides. They further reported the following sequence specificity of DNA methylation. CpG and non-CpG sites with a spacing of 8–10 bases were more likely methylated, whereas only for methylated CHG sites a periodicity of two pairs of 8-base separated cytosines with a spacing of 13 bases could be identified (Lister et al., 2009). Lister *et al.* found out that especially genes, involving a high transcriptional level, are affected by non-CpG methylation (Lister et al., 2009). More precisely, methylated CHH sites on the antisense strand in gene bodies were enriched, whereas CHG sites on the sense strand were almost exclusively unmethylated. Thus, non-CpG methylation might involve a stimulating effect on gene expression, in contrast to CpG methylation, which has a silencing effect. Lister *et al.* additionally identified specific CpG methylation patterns at exon-intron boundaries. These findings indicate the key role of non-CpG methylation in the origin and maintenance of ES cell as a pluripotent cell type.

A further publication of Lister *et al.* pursued the question whether similar DNA methylation properties, as they were observed in ES cells, also underlie induced pluripotent stem cells (iPSCs). They pointed out that iPSCs are subjected to a reprogramming mechanism for CpG and non-CpG methylation. Altogether, methylomes of iPSCs are very similar to those of ES cells (Lister et al., 2011). Nevertheless there are genomic regions, which exhibit differences in CpG DNA methylation patterns. Most of them are located in CpG islands and gene regions (Lister et al., 2011). These differentially methylated regions (DMRs) might be traced to transmission of incomplete reprogrammed somatic cell DNA methylation patterns or they might be even specific for iPSCs (Lister et al., 2011). Most of these DMRs were analyzed by analyzing independent iPSC lines. However, there seems to be unique DMRs for each iPSC line, which results in an interclone variability (Lister et al., 2011). Mega-scale genomic regions have been additionally indetified, which are resistant to reprogramming of non-CpG methylation patterns. These regions are also associated with histone modifications, such as H3K9me3, and transcriptional activity (Lister et al., 2011). In summary, CpG and non-CpG DMRs, varying histone modifications and expression patterns can be used as markers for incomplete reprogramming of iPSCs (Lister et al., 2011).

1.2.2 DNA Methylation in Cancer

Already in 1983 a first comparison of DNA hypomethylation (decrease of DNA methylation) in tumors and corresponding control tissue was published by Feinberg and Vogelstein (A P Feinberg and Vogelstein, 1983). Several articles have been published describing DNA hypomethylation in cancer (Bedford and van Helden, 1987; Cadieux et al., 2006). The opposite relative DNA methylation state, DNA hypermethylation, is especially associated with promoter regions of tumor suppressor genes (Graff et al., 1995; Melki et al., 1999; Costello et al., 2000). However, these gene-specific hypermethylated tumors additionally consist of genomic regions exhibiting predominant low DNA methylation (Manel Esteller, 2008). Consequently, tumor methylomes cannot be categorized as DNA hypomethylated or DNA hypermethylated; they rather should be distinguished regarding their underlying genomic region. A large proportion of DNA hypomethylation is located within repetitive genomic regions, which are usually predominantly methylated (Hoffmann and Schulz, 2005). For example, satellites within pericentromeric heterochromatin of chromosome 1 are

predominantly unmethylated in many human cancers (N. Wong et al., 2001). There are several genes, which are affected by DNA hypomethylation in tumor cell lines, such as the unmethylated melanoma antigen family, which encode tumor antigens (De Smet et al., 1999), the unmethylated S100 calcium binding protein A4 gene (*S100A4*) in colon cancer (Nakamura and Takenaga, 1998), the unmethylated serine protease inhibitor gene *SERPINB5* in gastric cancer (Akiyama et al., 2003), and the unmethylated oncogene γ-synuclein (*SNCG*) in breast and ovarian cancer (Gupta et al., 2003). In summary, DNA hypomethylation can be separated into three different mechanisms. It affects genomic instability early in tumorigenesis, reactivates transposable elements, and can lead to loss of imprinting (Manel Esteller, 2008).

DNA hypermethylation of CpG islands, especially for those overlapping promoter regions of tumor-suppressor genes, is also often associated with cancer cells (Peter A Jones and Stephen B Baylin, 2007). Several DNA hypermethylated CpG islands overlapping promoter regions of cancer associated genes could be identified: the tumor-suppressor genes retinoblastoma *Rb* (Greger et al., 1989; Sakai et al., 1991), *VHL* (associated with von Hippel-Lindau disease) and $p16^{INK4a}$ (J G Herman et al., 1994; Merlo et al., 1995; J G Herman et al., 1995; Gonzalez-Zulueta et al., 1995), *hMLH1* (a homologue of MutL *Escherichia coli*) (James G Herman and Stephen B Baylin, 2003), and *BRCA1* (breast-cancer susceptibility gene 1) (James G Herman and Stephen B Baylin, 2003; M Esteller et al., 2000). The Knudson two-hit hypothesis states that all copies of a tumor-suppressor gene have to be impaired to obtain a complete loss of function (Knudson, 2001). In genetics, such a phenomenon might be given by a germ-line (in familial cancers) or a somatic (in non-inherited tumors) mutation within the coding region of one of two copies of a tumor-suppressor gene. The second copy of the tumor-suppressor gene is usually affected by a somatic mutation, which ultimately leads to a malignant transformation of the cell (Knudson, 2001). DNA methylation of promoter regions of tumor-suppressor genes might act as a disruption, which has a similar effect as mutations within the coding region of the underlying gene (M Esteller, M F Fraga, et al., 2001). In familial cancer types, DNA methylation usually interferes with the activity of a tumor-suppressor gene on the second level of the Knudson two-hit hypothesis i.e. the first copy is affected by an inherited mutation (M Esteller, M F Fraga, et al., 2001). Esteller *et al.* additionally verified several types of nonfamilial tumors, which involve two fully methylated promoters of an associated tumor-suppressor gene. Altogether, it can be concluded that

there are more cancer-related genes influenced by DNA methylation than by mutations (Peter A Jones and Stephen B Baylin, 2002; Merlo et al., 1995).

Several epigenetic research projects identified hundreds of cancer genes associated with a DNA hypermethylated promoter region. Even an individual tumor type might include a variety of DNA hypermethylated loci. Recent cancer biology assumes that the stem cell state is an integral component of cancer development (Manel Esteller, 2008). Baylin and Ohm reviewed contributions of epigenetically silenced groups of genes at the chromatin level, which control the maintenance of cells in a stem cell state (Stephen B Baylin and Ohm, 2006). The concept of the "cancer stem cell" involves the cell population which is responsible for perpetuating the tumor (Bjerkvig et al., 2005). Cancer stem cells might originate from tissue-specific and bone marrow stem cells. Furthermore they might be derived from somatic cells that undergo transdifferentiation processes, or they are the result of cell fusion or horizontal gene-transfer processes (Bjerkvig et al., 2005). An accurate distinction of cancer stem cells and normal stem cells remains to be clarified.

The strong association of DNA methylation and cancer raises the request of epigenetic therapeutics. Especially drugs involving a reduction of DNA methylation is of great interest, since it could reverse gene silencing of tumor-suppressor genes. Constantinides initially described the impact of azanucleoside drugs on the expression level of cells in 1977 (Constantinides et al., 1977). In the meantime, it has been proven that treatment of 5-azacytidine has an inhibiting effect on DNA methylation (Santi et al., 1983). Further DNA methylation influencing drugs, such as fluoro-2'-deoxycytidine (P A Jones and Taylor, 1980) and zebularine (Cheng et al., 2004), are currently in development. A disadvantage of such nucleosides is their need to be incorporated directly into the DNA to exploit their full capability. Consequently, there have been several alternative attempts to develop drugs acting without any direct incorporation into DNA. Procainamide (Cornacchia et al., 1988) and tea polyphenols (Fang et al., 2003) include a DNA methylation inhibiting effect. However they solely have a weak effect on living cells (Peter A Jones and Stephen B Baylin, 2007).

1.3 Genomic Imprinting

Genomic imprinting is a epigenetic mechanism, that regulates gene expression in a parental-origin-specific manner (Ferguson-Smith, 2011). Either the maternal or the paternal allele is transcribed while the respectively other allele is silenced by an epigenetic mechanism. Thus, imprinting does not follow Mendelian rules. Originally, this phenomenon was discovered in experiments with mouse embryos that contain only one of the two sets of parental chromosomes (uniparental embryos) and those that inherit solely specific chromosomes from one parent (uniparental disomy) (Surani et al.; McGrath and Solter, 1984). Both experiments showed that genes have different properties and effects depending on their parental origin.

In 1991, the first imprinted genes could be identified, which are expressed in a parental-specific manner (DeChiara et al., 1991; Barlow et al., 1991; Bartolomei et al., 1991). Today, it is known that genomic imprinting takes place in several mammalian organisms (W Reik and Walter, 2001). Today, especially DNA methylation is associated with genomic imprinting. Parental-specific incomplete reprogrammed DNA methylation patterns in the embryo result in varying gene expression levels of differentiated cells. Figure 1.2 exemplifies a hypothetical pedigree of familial inheritance of an imprinted disorder through five generations. In this case, the mutation takes place in a paternally imprinted gene. Hence a paternal mutation cannot affect the offspring, since it is repressed.

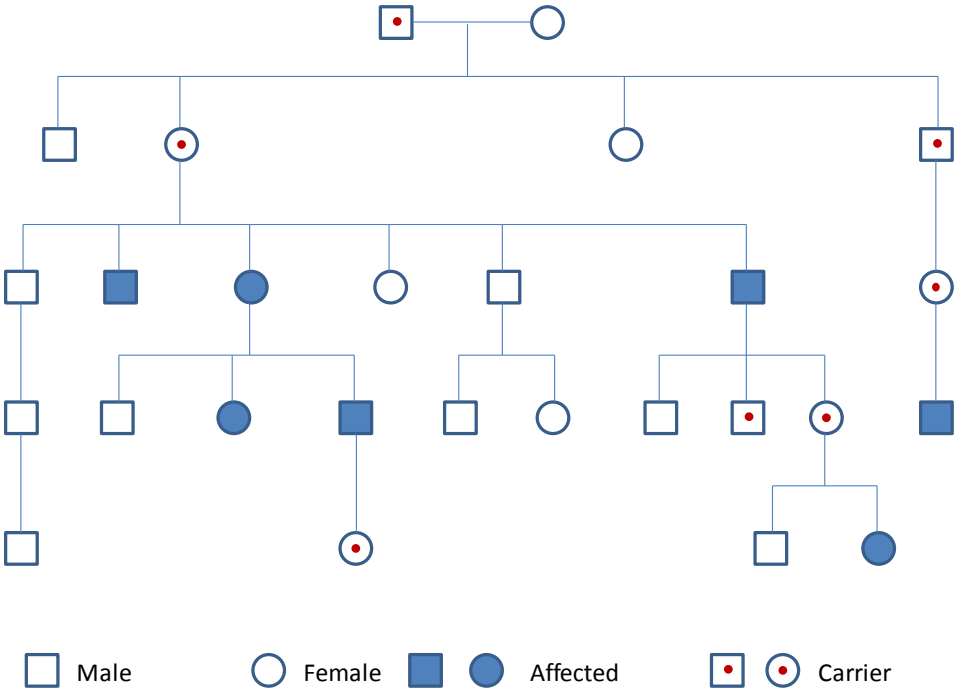


Figure 1.2: Hypothetical pedigree of inheritance of an imprinted disorder.

Figure modified from (Aitman et al., 2011)

Until today, especially imprinting in mice was studied and plenty of imprinted genes have been identified. A subset of these genes have even been validated in other mammalian and also in the human genome (Killian et al., 2000). Notably 80% of the imprinted genes annotated in the mouse cluster in specific genomic regions (Killian et al., 2000). Besides their affinity to appear in clustered groups, there is an enrichment of imprinted genes covering CpG islands that are located close to repeated genomic regions (Paulsen et al., 2000). Most of the imprinted genes show differences in maternal and paternal DNA methylation patterns. However, it should be distinguished between imprinted genes established from parental germ cells and maintained during cell differentiation (Stöger et al., 1993; Olek and Walter, 1997; Tremblay et al., 1997; Shemer et al., 1997) and imprinted genes, which initially exhibit equal parental DNA methylation patterns and evolve into tissue-specific imprinted genes (R Feil et al., 1994). Another property of the imprinting mechanism is observed during the cell division process. Kitsberg *et al.* observed a time-related property of imprinted genomic regions (Kitsberg et al., 1993) with a temporary shift in the replication of imprinted genomic regions during the cell cycle. The replication of the paternal copy usually occurs before the maternal one. Furthermore, it has been noted that

parental-specific meiotic recombination rates arise during the synthesis of imprinted genomic regions (Pàldi et al., 1995; Robinson and Lalande, 1995).

It is certainly of great interest to understand the impact of genomic imprinting on an organism's phenotype. A human disorder that shows parental-origin effects, due to genomic imprinting, is the Beckwith-Wiedemann syndrome (BWS) (Lubinsky et al., 1974). Lubinsky *et al.* reported: "affected offspring of either sex born only to female but not to male carriers" (Lubinsky et al., 1974). This disorder is additionally associated with an increased incidence of childhood tumors. However, BWS, a sporadic disease, is additionally associated with genomic alterations within the region 11p15 and is not only caused by imprinting (Michael R DeBaun et al., 2002). Chromosomal rearrangements, paternal uniparental disomy (maternal copy of chromosome 11 is replaced by the paternal one), or the presence of only one gene within the 11p15 region might also cause BWS (M R DeBaun and Tucker, 1998).

Neurological disorders, such as the Prader-Willi syndrome and the Angelman syndrome, have as well been identified as diseases regulated by imprinted genes (Buiting et al., 1995). Both disorders are caused by genetic and epigenetic defects within the same genomic domain on human chromosome 15. However, they differ in their parental-origin-specific manner. The Prader-Willi syndrome occurs in ~1 in 20,000 births and is characterized "by a failure to thrive during infancy, hyperphagia and obesity during early childhood, mental retardation, and behavioural problems" (Robertson, 2005). It involves a ~2 Mb imprinted genomic region that consist of a combination of maternally and paternally imprinted genes (Robertson, 2005). The Angelman syndrome occurs in ~1 in 15,000 births and is characterized "by mental retardation, speech impairment and behavioural abnormalities" (Robertson, 2005). The imprinting-defect is caused by the loss of maternally expression of the gene *UBE3A*, which is solely imprinted in the brain.

1.4 Epigenetics and the Environment

Although epigenetic modifications are predominantly stable, it is often discussed whether they might be influenced by environmental factors. Thereby, environmental influences might have a direct effect on an organism's gene expression level and its phenotype. In plants, persistent temperature changes can control epigenetic modifications (Chinnusamy and Zhu, 2009). Plants' vernalization might be regulated by epigenetic transitions in the

following manner. In temperate climates they are initiated to flower directly after having been exposed to the cold temperatures of winter (D.-H. Kim et al., 2009). In detail, Kim *et al.* show that vernalization of *Arabidopsis* is controlled by histone modifications of flowering suppressor genes (D.-H. Kim et al., 2009). Environmental-based epigenetic regulation in plants was also observed in *Linaria vulgaris* with a change of the fundamental symmetry of the blossom from bilateral to radial (Cubas et al., 1999). Cubas *et al.* identified *Lcyc*, a homologue of the *cycloidea* gene, as the regulating gene. Flowers, exhibiting a high level of DNA methylation within *Lcyc*, involve a change of symmetry. This DNA methylation regulated phenomenon is highly adaptable. A demethylation of *Lcyc* during somatic cell differentiation reverses the symmetry of the flower (Cubas et al., 1999).

Environmental-based epigenetic regulation can also be observed in mammalian organisms. There are different animal models, which show a correlation between environmental influences and changes in the epigenome (Rosenfeld, 2010). The agouti viable yellow allele (A^{vy}) in the mouse is an example of a metastable epiallele, an allele that can stably exist in more than one epigenetic state, resulting in different phenotypes (Rakyan et al., 2002). The epigenetic state of a metastable epiallele can switch and establishment is a probabilistic event. Once established, the state is mitotically inherited (Rakyan et al., 2002). The methylation level of the intracisternal A-particle retrotransposon of the A^{vy} locus is strongly associated with the coat colour of the mouse. A weakly methylated retrotransposon and therefore expressed *agouti* gene leads to a yellow coat colour, obesity and diabetes (Morgan et al., 1999). Folate (a B vitamin, which is abundant in green vegetables and fruits), and further compounds that affect one-carbon-transfer reactions, interfere with the DNA methylation level of A^{vy} of the developing offspring. The resulting coat color distribution of the offspring is shifted towards the brown pseudoagouti phenotype (Morgan et al., 1999). However, the methylation state of the mother remains unmodified (Waterland and Jirtle, 2003).

In humans, it is known that nutrition, emotional stress and toxic exposure might influence the phenotype by epigenetic changes (Gluckman et al., 2009) (see Table 3). Especially gestational effects can be observed in mammalians. An experiment gives proof that suboptimal nutrition during elementary gestation implicates increased incidence of type 2 diabetes in the offspring of rats (Sandovici et al., 2011). Ewes showed also adverse effects during gestation. The offspring of pregnant ewes, who were fed a restricted amount of

folate, methionine and vitamin B12, showed health problems that were caused by decreases in the DNA methylation levels of specific CpG islands (Sinclair et al., 2007). Environmental-based epigenetic regulations, based on histone modifications, have additionally been identified. The offspring of japanese macaques, fed high-fat diet during gestation, show a globally high level of acetylation of histone H3 (Aagaard-Tillery et al., 2008). In summary, fetal metabolic impairments due to nutritional restrictions are associated with epigenetic alterations, which affect the risk of chronic disorders throughout an organism's lifetime.

Table 3: Environmental-induced epigenetic alterations that affect health.

Table modified from (Robert Feil and Mario F Fraga, 2011)

Compound	*Species*	*Ontogenic stage*	*Epigenetic alteration*	*Tissues or cell types affected*	*Phenotypic alterations*	*References*
Tobacco smoke	Human	Adult life	DNA methylation and histone modifications	Lung, blood	Lung cancer?	(Pulling et al., 2004; Breitling et al., 2011; Hussain et al., 2009)
Particulate air pollution	Human, mouse	Adult life	DNA methylation	Blood, sperm	Unknown	(Baccarelli et al., 2009; Yauk et al., 2008)
Silica	Human	Adult life	DNA methylation	Blood	Silicosis	(Umemura et al., 2008)
Benzene	Human	Adult life	DNA methylation	Blood	Increased risk of AML	(Bollati et al., 2007)

Postnatal effects of environmentally-based epigenetic alterations can be observed as well. As already mentioned in the beginning of the introduction, Fraga *et al.* pointed out that monozygotic twins develop varying DNA methylation patterns during lifetime (Mario F Fraga et al., 2005). Moreover they showed that differences in DNA methylation correlate with increasing age. Some of these alterations might be explained by environmental factors. A further experiment approaching the same question was done by Wong *et al.* in 2010 (C. C. Y. Wong et al., 2010). They use DNA of 46 monozygotic and 45 dizygotic twin-pairs to generate locus specific DNA methylation levels of the dopamine receptor 4 gene (*DRD4*), the serotonin transporter gene (*SLC6A4/SERT*) and the X-linked monoamine oxidase A gene (*MAOA*) at both ages 5 and 10 years. The results of their study confirmed individual

differences in DNA methylation patterns, which are already established in early childhood, whereas they additionally might be altered during lifetime (C. C. Y. Wong et al., 2010). This highlights the dynamic property of epigenetic modifications induced by varying life conditions.

A direct association between altered global DNA methylation states and air pollution is given by an epidemiological study of Baccarelli *et al.* in 2009 (Baccarelli et al., 2009). Tissue-specific alterations in DNA methylation of several loci were associated with environmental influences, such as chronic exposure to sunlight, asbestos and tobacco smoke, consumption of alcohol and use of hair dye (Christensen et al., 2009; Langevin et al., 2011; Grönniger et al., 2010). Especially the interference of tobacco smoking on DNA methylation is of great interest, since it has been validated that smokers and former smokers show an enrichment of methylated promoter regions of tumor-suppressor genes in non-transformed lung cells (Pulling et al., 2004). Altogether, it has been shown that toxic environmental exposure might bias global and locus-specific DNA methylation patterns, whereas it should be differentiated between directly affected tissues and indirectly affected tissues exposed to chemical pollutants.

1.5 Transgenerational Epigenetic Inheritance

Transgenerational epigenetic effects require epigenetic alterations in the germ line, which are not erased by the reprogramming mechanism in the early embryo (see section 1.2). It is often discussed, whether disease risk, influenced by environmental-based epigenetic alterations, can be inherited (see section 1.4). Most of the environmental-based epigenetic effects that affect the offspring can be observed during gestation (Robert Feil and Mario F Fraga, 2011) and cannot be attributed to a transgenerational transmission mechanism. For example, when a female of the F_0 generation (initial parent generation in a multi-generation study) is exposed to toxic environmental influences, both the F_1 embryo and the F_2 generation germ line are also affected by the exposure (Jirtle and Skinner, 2007). Thus, transgenerational environmental-based epigenetic effects have to be found in the F_3 generation. Even studies analyzing postnatal or adult transgenerational effects have to assess the F_2 generation, since F_1 generation germ line is also directly affected by parental exposure (Ikeda et al., 2005; Blatt et al., 2003; Barber et al., 2006).

Several human and animal studies based on nutritional deficiency emphasize that an F_0 exposure might influence the phenotype of the F_2 generation (Pembrey et al., 2006; Csaba and Karabélyos, 1997; Ottinger et al., 2005; Anderson et al., 2006; Csaba and Inczefi-Gonda, 1998; Newbold et al., 2006; Popova, 1989; Zambrano et al., 2005). Specific chemical exposures to the F_0 generation show also effects in the F_2 generation (Dubrova, 2005; Csaba and Inczefi-Gonda, 1998; Popova, 1989; Turusov et al., 1990). A transgenerational effect that is even passed to the F_3 generation is caused by the endocrine disruptor vinclozolin, an antiandrogene (a blocker for steroid hormone that promotes male secondary sex characters), which causes spermatogenic defects, male infertility, breast cancer, kidney disease, prostate disease and immune abnormalities in up to four generations, but only when transmitted through the male germ line (Anway et al., 2006). Recent publications pointed out that these transgenerational effects were probably caused by epigenetic alterations (Jirtle and Skinner, 2007; Anway et al., 2005; Chang et al., 2006).

As already indicated, most of transgenerational epigenetic studies describe effects based on maternal exposure during gestation (Robert Feil and Mario F Fraga, 2011) (see section 1.4). In 2010, Carone *et al.* identified transgenerational effects in metabolic gene expression of mice based on paternal diet (Carone et al., 2010). They combined microarray and next-generation sequencing expression profiling analysis and MeDIP-seq (methylated DNA immunoprecipitation combined with next-generation sequencing) and RRBS (reduced representation bisulfite sequencing) to assess DNA methylation. On the transcriptional level, they showed that offspring of males, fed a low-protein diet, exhibit an upregulation of genes involved in lipid and cholesterol biosynthesis. Cholesterol ester levels were significantly downregulated compared to unexposed offspring. On the epigenetic level, they solely identified marginal alterations in liver methylomes of offspring depending on paternal diet. However, they indicated that a putative enhancer for a major lipid regulator, *Ppara*, is in low-protein offspring predominantly higher methylated compared to the offspring of males fed a control diet.

1.6 Thesis Structure

The remainder of this thesis is structured as follows:

Chapter 2 describes methodological considerations about NGS platforms for the determination of genome-wide DNA methylation patterns. Different biotechnological assays are discussed, whereas the main part focuses on bisulfite next-generation sequencing. Furthermore, primary analysis of raw sequencing data and their recommended quality controls are presented. Finally, secondary analyses including the alignment of bisulfite sequencing data are presented. All aspects considered, this chapter summarizes the biotechnological idea of next-generation sequencing assays, especially the one using bisulfite converted DNA and their bioinformatics challenges.

Chapter 3 outlines findings of the dissertation, which I published as first author (the manuscript of Appendix C was submitted in June 2012). An application note presents a bioinformatics approach for the primary analysis of bisulfite next-generation sequencing data. An overview about recommendations and pitfalls of methylome analysis, using short bisulfite sequencing data on different platforms, is presented in a review article. A biological study about the methylome of a human B-cell lymphoma makes use of the results of the application note and the review mentioned above.

Chapter 4 contains the discussion and conclusion of this thesis. Different bisulfite-based methods for genome-wide DNA methylation analysis are discussed and findings of this thesis are integrated into today's epigenetic research. Finally, future perspectives of computational epigenetics are presented.

Chapter 2 Methodological Considerations

For the Projects, incorporated in this thesis, next-generation sequencing (NGS) was carried out. The following section provides insight into different technologies and the respective challenges and limitations. The main part concerns bisulfite sequencing, a method for the determination of genome-wide DNA methylation levels at single-base resolution. Detailed information about the methods, used in each project, can be found in the supplement of the respective publication (see Chapter 3).

2.1 Next-Generation Sequencing

NGS enables the cost-efficient generation of large sequencing data sets for case-control and evolutionary studies based on whole genomes at single-base resolution. Today, NGS is especially used for variant detection by resequencing (personnel genomes), transcriptome analysis (RNA-seq), and the discovery of epigenetic variations (DNA methylation). Before NGS was established, research projects, including the generation of the human genome by the International Human Genome Consortium (IHGSC, 2004), were dependent on automated Sanger sequencing.

The biotechnological workflow of NGS basically involves three steps: template preparation, sequencing or alternatively imaging, and the bioinformatics analysis including sequence mapping or genome assembly. The latter is specifically highlighted in section 2.3. The fast technological development reveals new systems within short periods of time. Selected methods, relevant for the projects incorporated in this thesis, are presented in this chapter.

Two main assays are used for the preparation of templates, namely amplification-based methods and single molecule templates (the latter are discussed in section 4.5). For amplification-based template preparation, emulsion polymerase chain reaction (PCR) and solid-phase based methods are most often used (Dressman et al., 2003; Fedurco et al., 2006). Emulsion PCR uses beads equipped with a primer to bind amplified single-stranded DNA molecules. Therefore, millions of beads are fixed in a polyacrylamide gel for the upcoming sequencing step. A substantial advantage of emulsion PCR is the cell-free environment avoiding the arbitrary loss of sequences, which is a problem in bacterial cloning methods (Metzker, 2010). The alternative amplification-based approach for template preparation, namely solid-phase amplification, fixes clusters of immediately adjacent primers to a solid surface in order that added single-stranded DNA molecules result

in bridges between their corresponding primers. The subsequent amplification takes place along these bridges. It has to take into account that amplification by PCR might imply biases as for instance transcription errors or underrepresentation of specific sequences (Acinas et al., 2005).

The 454 technology by Roche was the first NGS platform established in 2005. It uses the emulsion PCR for the amplification step and the pyrosequencing technology (Ronaghi et al., 1996). Pyrosequencing by the 454 approach basically involves the complementation of single stranded DNA and the simultaneous sensing of the signal emitted from the respective nucleotide (see Figure 2.1). This technology avoids electrophoresis as the decoding can be proceeded during the sequence extension. In detail, each nucleotide is added by a polymerase chain reaction and a pyrophosphate for each nucleotide is transformed to an ATP by an ATP sulfurylase (Metzker, 2010). Afterwards, unincorporated nucleotides are washed away and the same process is started for the adjacent base. The average error rate of the 454 method is in the range of 10^{-3}-10^{-4} (Margulies et al., 2005; Quinlan et al., 2008), which is significantly higher than for Sanger sequencing. The error rate of the 454 approach increases towards the end of the underlying sequence, which is caused by a decrease of the productivity of the specifc enzyme resulting in weaker signals (Kircher & Kelso 2010).

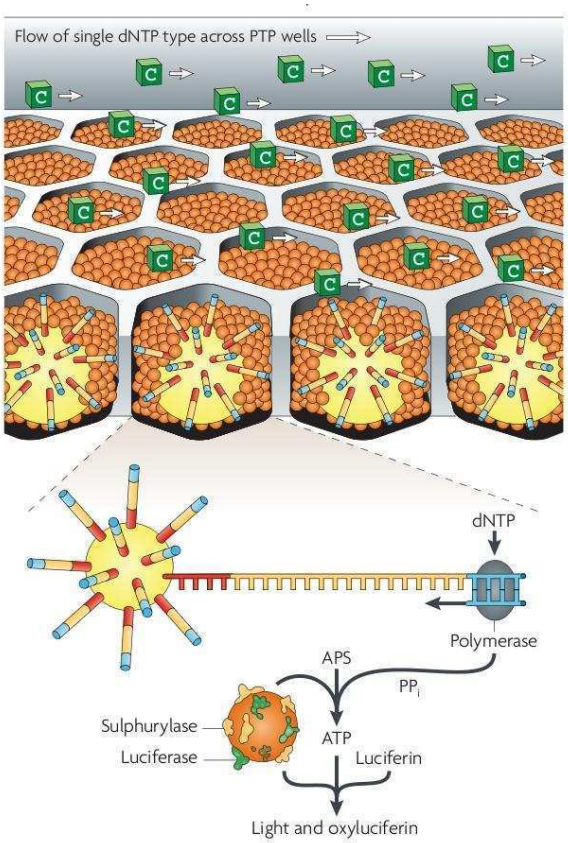


Figure 2.1: The idea of the 454 pyrosequencing approach.

Figure from (Metzker, 2010)

The HiSeq 2000 technology of Illumina uses the solid-phase amplification. Effectively, this sequencing by synthesis approach is similar to the idea of Sanger sequencing. A reversible dye terminator is used to control the incorporation of solely one nucleotide (Metzker, 2010). Free nucleotides are washed away and the respective integrated nucleotide is readout by four images using different filters and lasers to differentiate all genomic bases. Subsequently, dye terminators are removed and the procedure starts for the adjacent base. This approach exhibits a per base error rate of about 10^{-2}-10^{-3}, which is slightly higher compared to the 454 method (Kircher et al., 2009; Dohm et al., 2008). The per-base error rate also increases towards the end of the sequence, which is mainly due to phasing (not synchronized amplification of a population of DNA molecules). This phenomenon significantly increases the background noise (Kircher & Kelso 2010). More precisely, uni-directional phasing results in an incorrect reversible termination, which leads to an uncontrolled synthesis of further nucleotides (Kircher et al., 2009; Erlich et al., 2008).

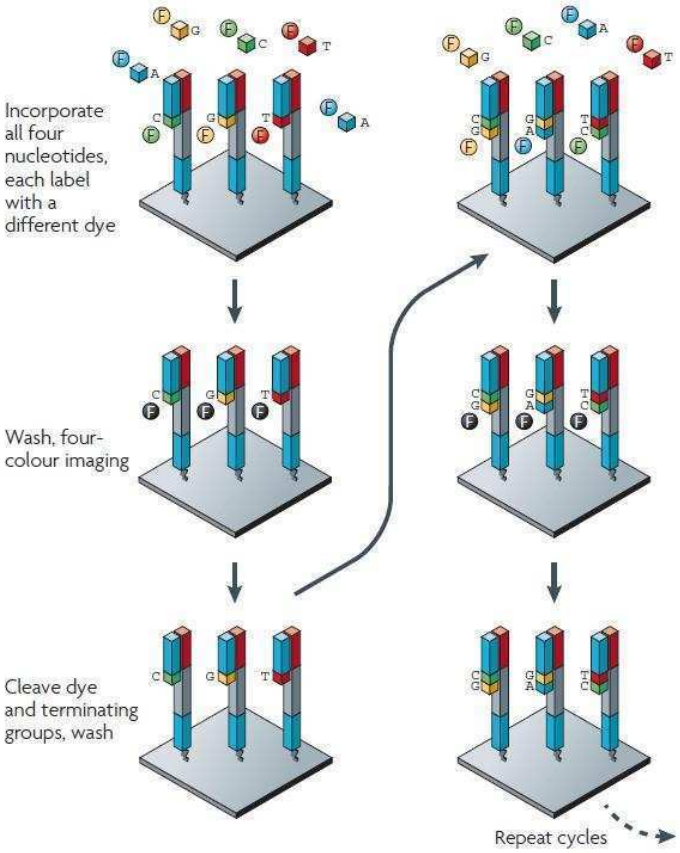


Figure 2.2: The concept of the HiSeq 2000 approach using a reversible dye terminator.

Figure from (Metzker, 2010)

2.1.1 Two-base Encoding Sequencing

The research projects, incorporated in this thesis, make use in particular of the SOLiD™ sequencing technology of Life Technologies (see Chapter 3). Therefore, this approach is comprehensively presented in the following section. SOLiD™ sequencing was developed by the Harvard Medical School and the Howard Hughes Medical Institute in 2005 (Shendure et al., 2005). It is the third NGS platform beyond the 454 platform and the Illumina Genome Analyzer (the predecessor of the HiSeq 2000 platform). The fundamental difference in SOLiD™ sequencing, compared to those mentioned above, is the fact that it uses a ligation reaction instead of a polymerase reaction (Shendure et al., 2005). Hereby, 8-mer probes modified with four different fluorescent labels are allocated for the ligation at single-stranded sequences hybridized with primers (LifeTechnologies, 2008). The two 3'-most nucleotides encoding the fluorophore are readout to determine the respective base. Subsequently, three bases including the dye are cleaved from the 5' end of the 8-mer. The

remaining 5-mer probe with a free 5' phosphate is used for the next ligation step. On average, 10 ligations are concatenated and the resulting sequence is then washed away to start the process for the next primer set. This upcoming primer set is solely shifted one base towards the 5' end of the underlying fragment (see Figure 2.3). Each primer exists in four modifications to ensure the presence of all four nucleotides in the first ligation step.

There are various sources of erroneous base call. Firstly, the emulsion PCR amplification step leads to a higher error rate than solid-phase amplification methods (Kircher & Kelso 2010). Secondly, an incomplete cleavage of the dye can result in a biased ligation process. The efficiency of a phosphate-based termination in this NGS approach minimizes the per base error rate of phasing. Altogether, the estimated per base error rate is in the range of 10^{-2}-10^{-4}, which depends on the availability of a reference genome to correct errors (Kircher & Kelso 2010) (see Figure 2.5).

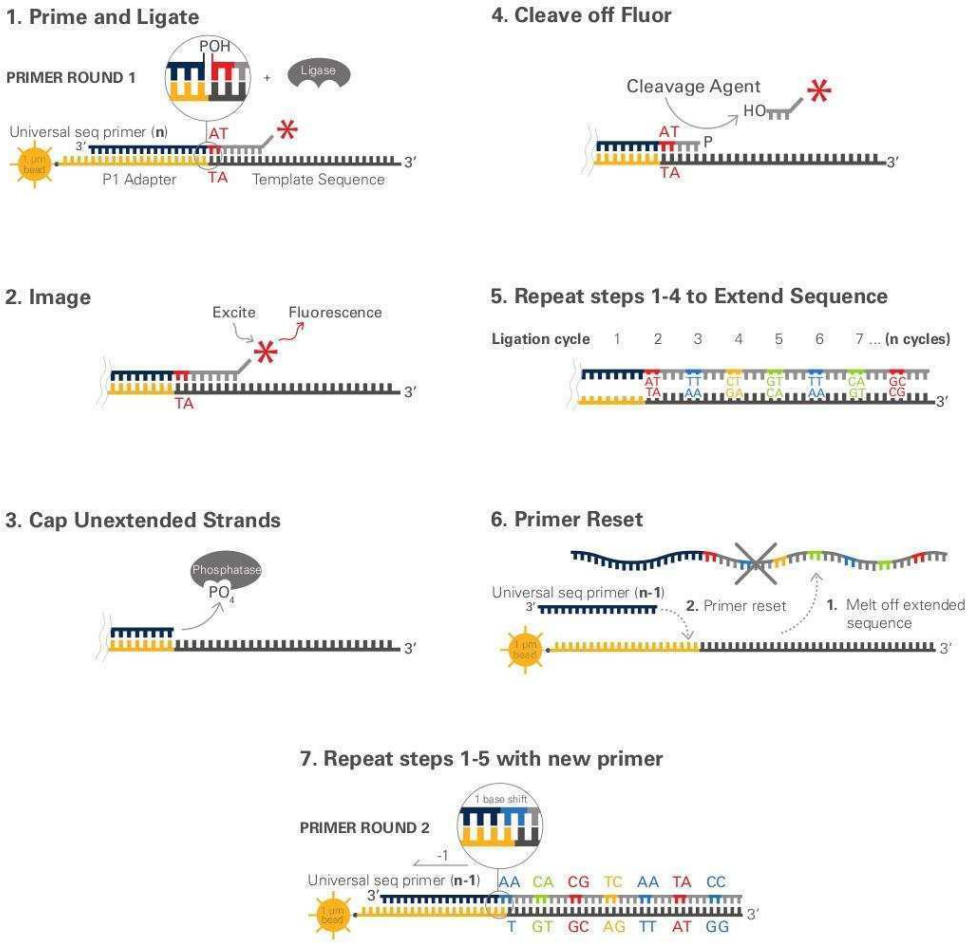


Figure 2.3: Sequencing by ligation using a two-base encoding technology by the SOLiD™ system.

Figure from (LifeTechnologies, 2008)

The sequence output of two-base encoding SOLiD™ sequencing is called color space. Sanger sequencing already encodes nucleotides by colors. However, SOLiD™ color space makes use of 4 colors, where each encodes 4 out of 16 transitions between all nucleotides (see Figure 2.4).

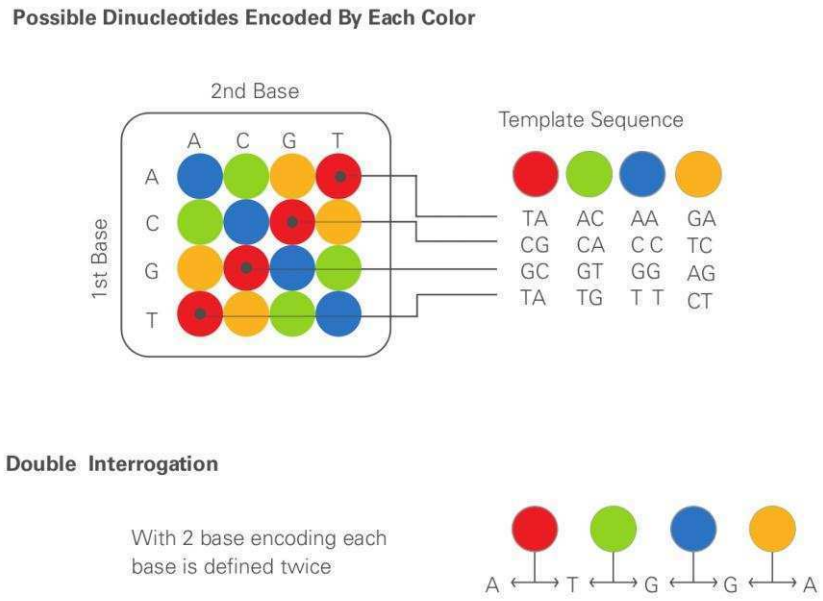


Figure 2.4: Scheme of SOLiD™ color space.

Figure from (LifeTechnologies, 2008)

Hence, two independent ligation steps determine each nucleotide, which results in a higher specificity. This fact enables the differentiation of sequencing errors and potential variants such as single nucleotide polymorphisms (SNPs). In detail, a single color change within a SOLiD™ sequence is typically traceable to a measurement error (LifeTechnologies, 2008). However, two adjacent color changes normally indicate a SNP. Furthermore, insertions and deletions can be detected in a similar manner (see Figure 2.5).

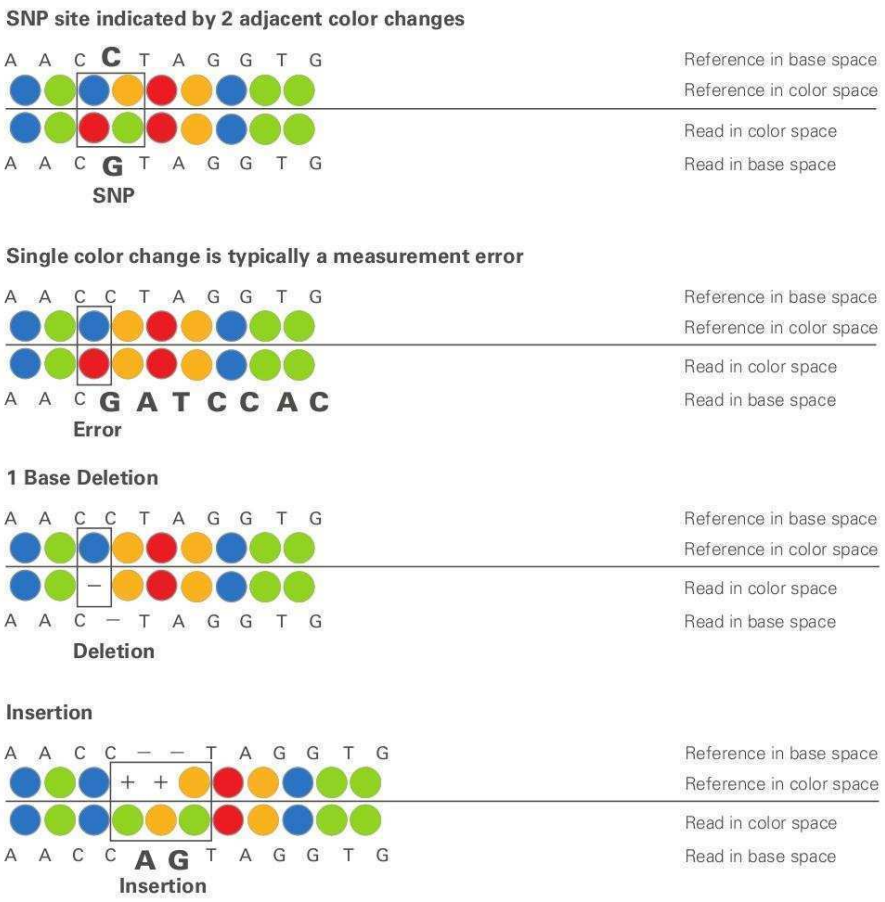


Figure 2.5: Properties of color space reads containing measurement errors and variants.

Figure from (LifeTechnologies, 2008)

2.2 Bisulfite Sequencing-based Methods to Profile DNA Methylation

2.2.1 Bisulfite Sequencing

Currently, there are two different approaches for the detection of DNA methylation based on NGS. Basically, they can be split into enrichment-based and bisulfite-based methods. Enrichment-based methods consist of methylated DNA immunoprecipitation sequencing (MeDIP-seq) and methylated DNA binding domain sequencing (MBD-seq) (Jacinto et al., 2008; Down et al., 2008; Serre et al., 2010). MeDIP-seq makes use of an anti-methylcytosine antibody to immunoprecipitate single-stranded DNA fragments. MBD-seq involves an enrichment of double-stranded DNA fragments via the MBD2 protein methyl-CpG binding domain. Recent publications have shown that enrichment-based and bisulfite-based methods generate comparable DNA methylation results (Harris et al., 2010; Bock et al., 2010).

Bisulfite sequencing (BS-seq) involves the bisulfite conversion of genomic DNA combined with NGS. The bisulfite treatment of DNA molecules enables a differentiation between methylated and unmethylated cytosines at single-base resolution. Thereby, unmethylated cytosines are converted to uracils, whereas methylated cytosines remain unmodified (Frommer et al., 1992). Uracils are read as thymines by DNA polymerase. Thus, the amplification of bisulfite-treated DNA by PCR yields products in which unmethylated cytosines appear as thymines. Consequently, differences in methylation states at single-base resolution can be inferred depending on the amount of cytosines and thymines assigned to a specific genomic position. BS-seq is nowadays the gold standard for genome-wide DNA methylation analysis because of its clear readout at each cytosine position. A limitation of BS-seq is the fact that it cannot distinguish between DNA hydroxymethylation (see section 1.2) and usual DNA methylation. In detail, bisulfite treatment converts unmethylated and non-hydroxymethylated cytosines to thymines and leaves methylated and hydroxymethylated sites unmodified (Krueger et al., 2012). Hence, the respective modification cannot be determined separately.

2.2.2 Reduced Representation Bisulfite Sequencing

BS-seq is an accurate method for the determination of genome-wide DNA methylation levels. However, it is still a cost-intensive approach especially for large genomes. Meissner *et al.* developed a genome-scale method, which also makes use of bisulfite converted DNA, providing insights into parts of the methylome (Meissner et al., 2005). Their reduced representation bisulfite sequencing (RRBS) approach enables the facilitation of case-control studies involving large sample sizes (Gu et al., 2010). The idea of RRBS is to digest genomic DNA with a methylation-insensitive restriction enzyme. Fragments of a specific length are selected to filter the most informative genomic subset. Then, a bisulfite conversion of end-repaired, A-tailed, and adapter ligated fragments is carried out to determine DNA methylation levels as described in section 2.2.1 (Gu et al., 2011). Hereby, DNA methylation patterns of parts of the genome can be obtained. It has been shown that these restricted fragments cover especially core promoters and CpG islands (Gu et al., 2010), which include essential regulatory parts of the genome. Altogether, RRBS composes only ~1% of the underlying whole genome (Meissner et al., 2008). It has been shown that DNA amounts of

10-300 ηg are sufficient to generate accurate DNA methylation levels with RRBS (Gu et al., 2011). Consequently, it is well-suited for many clinical samples, such as tumors, which can only provide a small amount of genomic input DNA material.

The following section describes selected steps of the preparation of a RRBS library using the digestion enzyme MspI (see Figure 2.6). Further bisulfite-based library preparation protocols are given in the respective supplemental information of projects incorporated in this thesis (see Chapter 3).

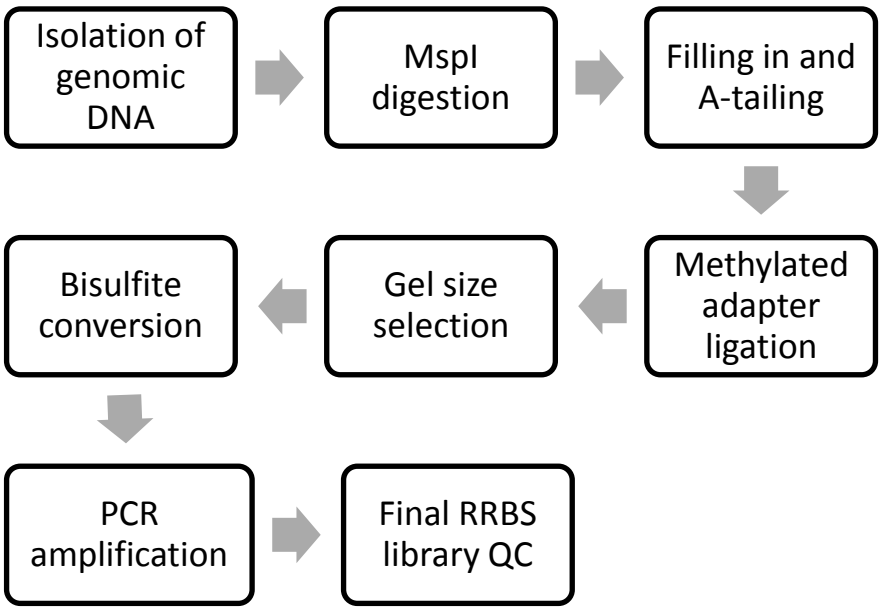


Figure 2.6 Workflow of a RRBS library preparation.

Figure modified from (Gu et al., 2011)

- **Isolation of genomic DNA:** It is mandatory to use highly purified genomic input DNA to generate a high-quality RRBS library (Gu et al., 2011). Contaminated DNA molecules might interact with restriction enzymes and interfere with the bisulfite conversion. Gu *et al.* additionally recommend the use of DNase-free RNase in the lysis buffer to avoid DNA degradation (Gu et al., 2011).
- **Digestion reaction:** Two different enzymes are commercially available by now: MspI (restriction motif: C↓CGG) and TaqI (restriction motif: T↓CGA) (Gu et al., 2011). Both, MspI and TaqI are insensitive for CpG DNA methylation, whereas MspI exclusively generates fragments containing CpG dinucleotides at both ends. A drawback of MspI is that a methylated cytosine at the first position of the restriction motif C↓CGG interferes

with the digestion reaction (Gu et al., 2011). However, this situation can rarely be observed, at least in human methylomes, since methylated non-CpG sites hardly ever occur (Pelizzola and Ecker, 2010) (see section 1.2.2).

- **Filling in and A-tailing:** 3'-terminal recessive ends containing an adenine are added, since they are required for the adapter ligation of the upcoming library preparation (Gu et al., 2011).
- **Methylated adapter ligation:** Both, single-end and paired-end sequencing can be carried out with RRBS libraries, whereas adapters have to consist of methylated cytosines to maintain their compatibility with the subsequent bisulfite conversion. Paired-end sequencing certainly has the advantage to increase the mapping efficiency by unique alignments. However, it can also bias DNA methylation levels, as overlapping pairs generate redundant DNA methylation information (Krueger et al., 2012).
- **Gel size selection:** Before fragments are bisulfite converted, they are size selected. *In-silico* analyses show that a size selection of 40-220 bp for fragments, containing the MspI restriction motif C↓CGG, covers most promoter sequences and CpG islands (Gu et al., 2011).
- **Bisulfite conversion, PCR amplification, and sequencing:** Digested and size-selected fragments are finally bisulfite converted. In the end, fragments are amplified by PCR for the sequencing process on a NGS platform. To date, RRBS is solely carried out on Illumina platforms.

The bioinformatics challenge of analyzing RRBS sequencing data is given by the fact that the alignment step requires an *in-silico* modified reference genome. Thus, the reference should consist of size-selected and enzyme specific digested genomic sequences. Apart from the alignment reference, RRBS depends on comparable bioinformatics primary and secondary analyses as they are applied to genome-wide BS-seq. Details about challenges and pitfalls of BS-seq data are given in section 2.3.1 and 2.3.2.

2.3 Analysis of Bisulfite Sequencing Data

2.3.1 Quality Control

Several quality control approaches for BS-seq data exist. Firstly, sequencing results can be validated by positive and negative controls incorporated into the sequencing library. Secondly, raw sequencing data can be controlled for quality by *in-silico* analyses to filter contaminations.

An effective quality control can be performed with lambda phage (*enterobacteria phage*) DNA, which is spiked-in during the library preparation. The lambda phage genome is originally completely unmethylated and is therefore used as a positive control of the bisulfite conversion. Hence, the overall DNA methylation level in the lambda phage genome assesses the quality of the bisulfite conversion. Recent publications identified 1% of methylated cytosines within the lambda phage genome (Lister et al., 2009, 2011; Hansen et al., 2011).

A critical part of genome-wide DNA methylation analyses is the determination of non-CpG sites. As already mentioned in section 1.2.1, ES cells involve the highest amount of methylated non-CpG sites in human cells. However, the principal part of these non-CpG sites involves DNA methylation levels of less than 0.4 (Lister et al., 2009). Hence, it is indispensable to distinguish between truly low methylated non-CpG sites and non-CpG sites, which involve a low DNA methylation due to an incomplete or defective bisulfite conversion. Considering that, Lister *et al.* developed the following approach using an estimation of the bisulfite conversion based on the lambda phage. They used the binomial distribution:

$$f(k;n,p) = \binom{n}{k} p^k (1-p)^{n-k}; \text{ for } 0 \leq k \leq n \text{ and } \binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$

Parameters are defined as follows: n is the sequence coverage at a specific genomic position; k is the number of sequenced cytosines at the corresponding genomic position; p is the fraction of cytosines sequenced in the lambda genome and the sum of all thymines and cytosines sequenced in the lambda genome i.e. the genome-wide fraction of DNA methylation in the lambda phage genome. The result of this binomial distribution is the probability that the DNA methylation level of a specific cytosine arises from an incomplete

bisulfite conversion. In summary, this method enables a quality control for the bisulfite conversion, which then can be used for the classification of all DNA methylation levels.

In the following section examples of bioinformatics quality controls are presented to validate bisulfite sequences obtained from NGS platforms. Detailed information about bioinformatics quality control of bisulfite sequencing data can be found in the publication incorporated as Appendix B. The following analyses are based on the publicly available tools *FastQC* (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and *Trim Galore* (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/).

Firstly, it is mandatory to visualize the distribution of quality scores at each position in bisulfite sequences with *FastQC*. Low base call qualities, which most often arise towards the end of sequences, can be eliminated with *Trim Galore*. Therefore partial sums from all positions of the bisulfite sequence to its end are computed. The bisulfite sequence is truncated at the position involving the minimal relative partial sum. Thus, low base call qualities, which frequently involve false positive base-calls, are reduced. It has been shown that this quality control results in more precise DNA methylation levels (see supplemental information of Appendix B).

Furthermore, it is recommended to control genome-wide frequencies of all four nucleotides at each position in bisulfite sequences. These frequencies are shifted for bisulfite sequences compared to usual genomic sequences due to the conversion of unmethylated cytosines to thymines. For the human genome, 67-82% of all CpG sites are methylated, whereas not more than 3% of all non-CpG sites are methylated. However, only 5% of all cytosines are located in CpG dinucleotides resulting in approximately 96% of unmethylated cytosines in the human genome. These unmethylated cytosines appear as thymines after the bisulfite conversion and the PCR (Pelizzola and Ecker, 2010). Thus, the percentage of cytosines is to the greatest extent very low, whilst the amount of thymines is cleary enriched. The analysis and adjustment of these frequencies can be performed with *FastQC* and *Trim Galore* (see supplemental information of Appendix B).

2.3.2 Mapping of Bisulfite Sequencing Data

The mapping of sequences generated by BS-seq poses the main challenge for this type of data. Mapping approaches for usual genomic sequencing combine the alignment to the Watson and its complementary Crick strand. However, bisulfite conversion of DNA sequences results in non-complementary strands. In detail, there are up to four distinct strands, which have to be included in the mapping reference (see Figure 2.7). This situation can be avoided by a specific library preparation, which predetermines the ligation of adapter sequences before the PCR amplification is carried out. In this case, only two distinct bisulfite treated strands are sequenced. This type of BS-seq library is called a directional BS-seq library, whilst a BS-seq library involving all four strands is called a non-directional BS-seq library (Krueger and Andrews, 2011). Further observations are based on BS-seq data obtained by directional libraries. In conclusion, the mapping of BS-seq data needs to be performed for the reference and additionally for its *in-silico* bisulfite converted modifications (see Appendix B).

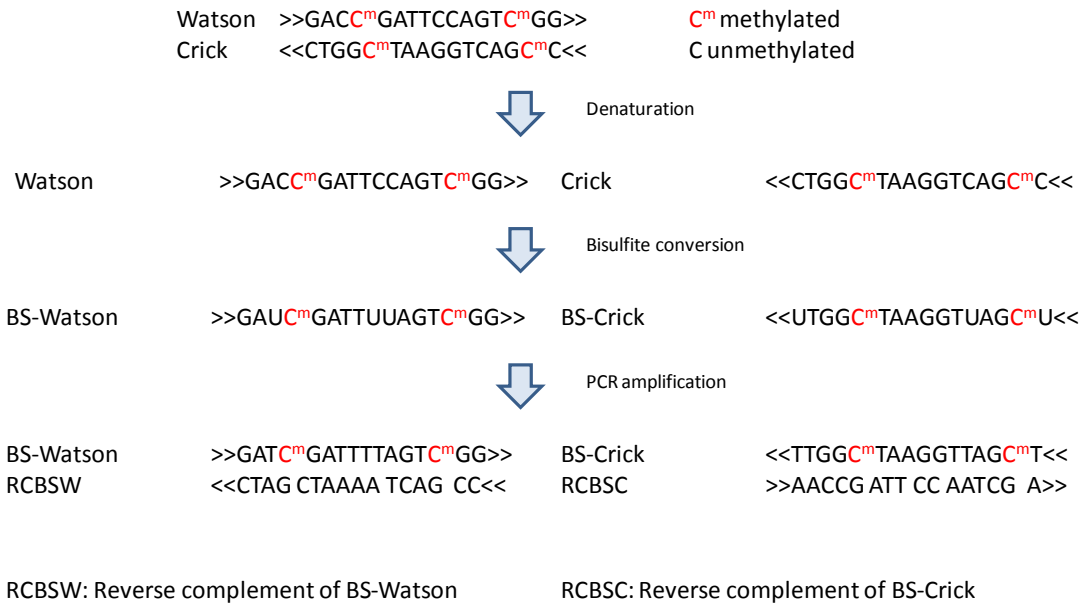


Figure 2.7: Impact of bisulfite conversion on double stranded DNA sequences.

A further limitation of bisulfite sequences is the fact that their complexity is significantly reduced. About 96% of all cytosines in human methylomes are unmethylated and by this all of these cytosines appear as thymines in bisulfite sequences. Consequently, bisulfite sequences predominantly consist of three nucleotides: adenines, guanines, and thymines. Therefore, resulting sequences are less differentiated and involve an increased frequency of ambiguous mapping results. BS-seq mapping approaches additionally have to take into account that thymines in bisulfite sequences might be assigned to referential cytosines and thymines, since they might be unmethylated cytosines or original genomic thymines (see Figure 2.8).

A detailed comparison of bioinformatics tools for the mapping of BS-seq data is described in Appendix B.

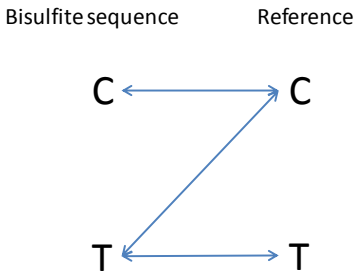


Figure 2.8: Mapping asymmetry of bisulfite sequences.

Chapter 3 Thesis Outline and Summary of Findings

This thesis comprises a bioinformatics tool for the analysis of two-base encoding BS-seq data, a review about pitfalls and challenges of BS-seq analyses on different NGS platforms, and a biological study about a methylome of a B-cell lymphoma.

3.1 Application Note (see Appendix A)

BS-seq is currently the gold standard for the analysis of DNA methylation at single-base resolution. Originally, it was used on the Genome Analyzer platform of Illumina (Lister et al., 2009). One of the main challenges within the framework of this thesis was the development of a bioinformatic tool for bisulfite sequencing analysis on the SOLiD™ platform of Life Technologies (see section 2.1.1). In the application note, I presented B-SOLANA, the first tool for the analysis of large SOLiD™ BS-seq data sets. It includes the alignment and determination of DNA methylation levels in CpG as well as non-CpG sequence contexts. B-SOLANA exhibited a high alignment efficiency compared to further approaches, which are available by now.

3.2 Review (see Appendix B)

BS-seq is a rapidly developing research field in the last few years. In this review article, bioinformatics aspects about BS-seq analyses were discussed. Therefore, challenges of BS-seq alignment as they apply to both base and color-space data were summarized. There are different contaminations within raw sequences, which might interfere with genome-wide DNA methylation levels. Potential sources of contaminations are for instance platform-specific sequencing errors and adapter sequences. This review article presented quality controls of BS-seq data and methods to minimize false positively detected DNA methylation levels. Finally, it gives a recommendation of the most appropriate way to analyze this type of data.

3.3 Study (see Appendix C)

This study analyzed the methylome of the DAUDI cell line, an archetypal endemic Burkitt's lymphoma. It combined genome-wide DNA methylation results obtained by BS-seq on the HiSeq 2000 and the SOLiD™ platform, whereas platform-independent data sets exhibited comparable results. DNA methylation levels of 91.1% of all referenced CpG sites and 90.2% of all referenced non-CpG sites of the DAUDI methylome were assessed. The genome-wide DNA methylation accounted for 68.99%, which is comparable to further human methylomes (Li et al., 2010; Lister et al., 2009). The study identified an enrichment of significantly methylated non-CpG sites within RefSeq genes, which was previously reported for the methylome of the ES cell line *H1* (Lister et al., 2009). Correlation analysis revealed that transcription levels were strongly associated with the amount of methylated CpG sites around the transcription start site (TSS), where present transcripts involved CpG sites with minimal DNA methylation levels immediately at the TSS. Interestingly, sharp transitions of DNA methylation levels at exon-intron boundaries of absent transcripts could be identified. It was previously shown that DAUDI involves an upregulation of the Epstein-Barr virus (EBV) (D. N. Kim et al., 2011). In our study, the EBV and the mitochondrial methylome and their transcriptomes were analyzed, which showed that methylation in CpG dinucleotides significantly varied between nuclear (68.99%), mitochondrial (6.43%) and EBV (80.18%) genomes. In conclusion, the analysis revealed that the mechanisms of DNA methylation associated with transcriptional regulation in endemic Burkitt's lymphomas go by far beyond the usually studied promoter methylation.

Chapter 4 Discussion and Conclusion

The first part of this chapter integrates the results of this thesis into the current state of knowledge concerning DNA methylation analyses. The development from array-based to NGS-based DNA methylation analyses is described and different NGS-based methods, involving computational approaches for the determination of interindividual DNA methylation differences, are compared. The second part of this chapter comprises conclusions drawn from this thesis and discusses future perspectives of computational epigenetic studies. The following observations are based on analyses of DNA methylation in human genomes.

4.1 From Array Technologies to Next-Generation Sequencing

In the past, most epigenetic studies examined DNA methylation patterns within promoter regions to identify regulatory effects related to gene expression. Advantages and disadvantages of array-based and NGS-based methods are presented in the following section.

In the following comparison, the *Infinium HumanMethylation BeadChip* of Illumina, an array method based on bisulfite converted DNA, is discussed. The current version assesses DNA methylation levels of about 485,000 CpG sites. It covers 99% of RefSeq genes, with an average of 17 CpG sites within the promoter, the 5' untranslated region, the first exon, the gene body, and the 3' untranslated region. Additionally, DNA methylation in 96% of all CpG islands, with additional coverage in island shores is assessed. DNA methylation levels are quantified by beta values. Beta values range from 0 (low DNA methylation level) to 1 (high DNA methylation level). This technology is a hybrid approach of two different chemical assays, the Infinium I and Infinium II assays. The previous version of the *HumanMethylation BeadChip* array, assessing only 27,000 CpG sites, used the Infinium I assay. However, one third of the DNA methylation levels measured by the current version are obtained by the Infinium II assay, whereas recent publications indicated that this chemical assay is less accurate and reproducible (Sandoval et al., 2011; Bibikova et al., 2011; Dedeurwaerder et al., 2011). It is mandatory to correct the Infinium II results to obtain comparable DNA methylation levels assessed by both chemical assays. Varying bioinformatics approaches are available for this task by now (Dedeurwaerder et al., 2011).

The choice of DNA methylation profiling technology certainly depends on which problem is studied. Whole genome DNA methylation analyses usually use sequencing-based methods.

However, gene-specific studies apply either array-based approaches or even candidate gene-specific methods, where the latter can be carried out with bisulfite Sanger sequencing. Genome-wide arrays, as described above, are a reliable and cost-efficient tool for the determination of quantitative DNA methylation levels at specific genomic loci. However, they should rather be used for large studies than for the measurement of different DNA methylation patterns between few samples. The relatively low density of array-annotated cytosines within a specific genomic domain implies a weak statistical power for the detection of DMRs. Even a small number of varying DNA methylation levels, potentially induced by technological biases, significantly influence test statistics (see section 4.2).

Independent comparisons between array-based and NGS-based methods indicated comparable results (Kreck et al., 2011; Bock et al., 2010) (see Appendix A). However, extreme DNA methylation levels are assessed differently (see supplemental information of Appendix A). For NGS-based methods, completely unmethylated and fully methylated sites correspond to DNA methylation levels of 0 and 1. The *HumanMethylation BeadChip* method merely assigns these extreme sites to DNA methylation levels close to 0 and 1. The following correlation analyses of a SOLiD™ BS-seq data set, consisting of cytosines covered by at least 10 bisulfite sequences, and DNA methylation levels assessed by the *HumanMethylation BeadChip* method target this phenomenon (the data belongs to the study of Appendix A (see Figure 4.1). Correlation analysis were carried out using residuals i.e. differences of DNA methylation levels ($(\text{Methylation level})_{\text{SOLiD™}}$ - $(\text{Methylation level})_{\textit{BeadChip}}$) assessed by both platforms. Both methods assess DNA methylation levels in a similar way, because most of the residuals are located next to 0. Nevertheless, the distribution of residuals exhibits two maxima next to 0 (see Figure 4.1). The left peak represents completely unmethylated sites, which are assigned to 0 by SOLiD™ BS-seq and to beta values close to 0 by the *HumanMethylation BeadChip*. The right peak depicts the reverse effect of fully methylated sites assessed by both methods. In conclusion, this technology-specific fact has to be taken into account for the determination of weakly differentially methylated sites of extreme DNA methylation levels.

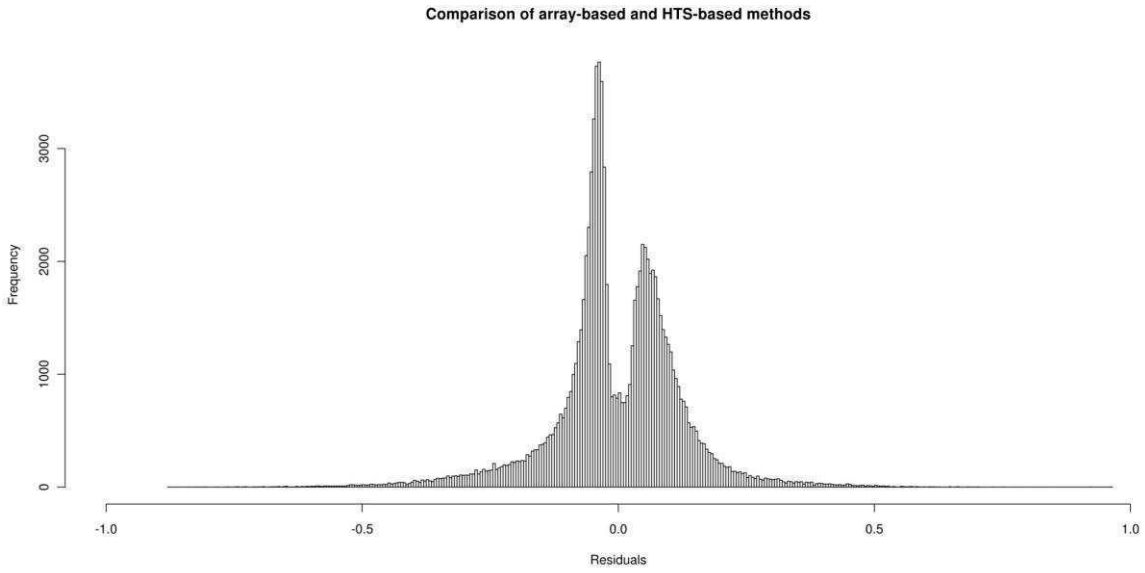


Figure 4.1: Residuals of DNA methylation levels assessed by SOLiD™ BS-seq and the *HumanMethylation BeadChip*.

Correlation analyses of DNA methylation vs. transcriptional levels are preferably carried out for regions upstream of genes and especially promoters. Array-based methods assess specific representative cytosines within these genomic regions. However, Appendix C pointed out that it is necessary to consider the majority of cytosines around the TSS to receive precise results of the correlation analyses. High transcriptional levels correlate with minimal DNA methylation at the TSS, which then rapidly increases towards downstream and upstream directions (see Appendix C). Thus, array-based methods only enable an insight into DNA methylation levels and are less conclusive concerning correlation analyses related to transcriptional levels.

In the human genome, 67-82% of all cytosines within CpG dinucleotides are methylated (Pelizzola and Ecker, 2010). Cytosines within CpG dinucleotides are common sites of polymorphisms due to the deamination of methylated cytosines to thymines during evolution (Venter et al., 2001). This fact might interfere with the extraction of DNA methylation levels by bisulfite-based methods, since unmethylated cytosines appear as thymines in bisulfite converted DNA. The array-based *Infinium HumanMethylation BeadChip* technology cannot distinguish, whether variations are a result of deamination or bisulfite conversion (Byun et al., 2009). For this purpose, non-bisulfite converted genomic DNA has to be sequenced to detect potential polymorphisms. In contrast, NGS-based methods are

able to differentiate these two types of variations in the following way: A polymorphism generated by deamination of a methylated cytosine to thymine now has an adenine nucleotide on the opposite strand. However, the conversion of cytosines to thymines, induced by bisulfite treatment, leaves the guanine on the complementary strand unmodified. As bisulfite sequences need to be independently mapped to the forward and reverse strand, polymorphisms and variations, caused by bisulfite conversion can be independently detected.

4.2 Epigenome-wide Association Studies

Although genome-wide association studies (GWASs) identified more than 1,449 genomic loci that were associated to 237 diseases and traits so far, only a small proportion of the underlying genetic architecture can be explained (Hindorff LA, MacArthur J (European Bioinformatics Institute), Wise A, Junkins HA, Hall PN, Klemm AK, and Manolio TA. A Catalog of Published Genome-Wide Association Studies. Available at: www.genome.gov/gwastudies. Accessed [June, 2012]). With the establishment of NGS technologies, resequencing of genomes and exomes (the entirety of exons) is expected to reveal further genetic alterations, which might help to explain the “missing” heritability (Manolio et al., 2009). However, increasing evidence points to epigenetic factors, especially to DNA methylation, which might influence the pathogenesis of complex diseases (Andrew P Feinberg and Irizarry, 2010; Petronis, 2010; Kulis and Manel Esteller, 2010). Most studies about DNA methylation so far have been carried out for either a few samples with high coverage, generated by NGS, or for many samples with low genome-wide coverage, generated by array methods. The recent biotechnological development increasingly enables large-scale DNA methylation analyses, so called epigenome-wide association studies (EWASs).

Chips used for GWASs allow genotyping of hundreds of thousands SNPs (Manolio et al., 2009). Statistical approaches subsequently test the genome-wide significance of individual SNPs. Comparable DNA methylation chips, as for instance the current *HumanMethylation BeadChip*, measuring DNA methylation of ~485,000 sites, are available by now. However, it is most often insufficient to compare DNA methylation levels of single sites. Potential variations in EWASs are: DMRs, variably methylated regions (VMRs), allele-specific DNA methylation (ASM), and haplotype-specific DNA methylation (HSM). VMRs are genomic

regions involving moderate alteration in DNA methylation, hence an attenuated type of DMRs. ASMs are genomic regions exhibiting variation in DNA methylation in a parent-of-origin specific manner or based on a nearby SNP. HSMs are DMRs defined by a combination of alleles within a genomic domain. These different types of epigenetic variation, involving single sites and even genomic regions, show the variability of distinguishing marks of DNA methylation. In summary, EWASs pose further challenges compared to GWASs.

DNA methylation is a highly dynamic epigenetic modification, which might change during lifetime and be altered by environmental influences or drug interventions (see section 1.4). Furthermore, it has been shown that genetic factors can considerably influence DNA methylation (Zhang et al., 2010; Kerkel et al., 2008; Hellman and Chess, 2010; Shoemaker et al., 2010). These observations have to be considered for the design and analysis of EWASs. A potential approach can be the integration of already existing GWASs results in upcoming EWASs to control the influence of genotypes on DNA methylation variation. In the end, a fundamental question of DNA methylation studies is: Is DNA methylation the cause or the consequence of a disease or even the consequence of a genotype causing the disease? So far, no study exists that specifies variation in DNA methylation as the cause of the disease (Rakyan et al., 2011)(S. Baylin and T. H. Bestor, 2002).

GWASs typically involve case-control studies consisting of unrelated individuals, who are clustered based on their phenotype. However, a potential study design of EWASs includes parent-offspring pairs, where combined genetic and DNA methylation analysis helps to identify truly associated variations in DNA methylation by filtering family-based genetic variations, which might alter DNA methylation (Rakyan et al., 2011). This can be achieved by either results of combined GWASs and EWASs or by methods, which simultaneously determine genetic and DNA methylation variations. A further possibility to minimize genetic confounders is to study monozygotic twins who are discordant for a disease. This study design of EWASs can be used to exclude any DNA methylation variations, which were caused by germline genetic variations (Kaminsky et al., 2009; Bell and Spector, 2011). Even this type of EWASs cannot decides whether alterations in DNA methylation are a cause or consequence of a specific disease. Besides confounding by genetic-epigenetic interaction, the high dynamic of DNA methylation, throughout an organism's lifetime, influences the outcome of EWASs. Longitudinal EWASs address this key aspect considering disease-free people (ideally from birth) over the course of many years. In conclusion, probably the best-

suited EWAS design, even though accompanied by great cost, is a longitudinal study of disease-discordant monozygotic twins, which integrates genetic-epigenetic interactions and environmental factors influencing DNA methylation.

Altogether, EWASs have the potential to explain parts of the genetic architecture of diseases and traits, even though the study design needs to consider several confounding factors to decide, whether DNA methylation is a cause or a consequence of a disease. In GWASs, tags SNPs based on high linkage disequilibrium are often utilized for comprehensive variation coverage. Such sites need to be explored for EWASs to facilitate cost-efficient studies. However, only whole methylomes of embryonic stem cells, fetal fibroblasts, peripheral blood mononuclear cells, colon cancer cells and B-cell lymphomas are available by now (Lister et al., 2009, 2011; Li et al., 2010; Hansen et al., 2011) (see Appendix C). Thus, it is necessary to analyze further methylomes to identify appropriate DNA methylation sites or even genomic domains, which can be used for comprehensive array-based EWASs.

4.3 Comparison of Sequencing-Based DNA Methylation Methods

Different NGS-based methods for the determination of genome-wide DNA methylation are available by now (Bock et al., 2010; Harris et al., 2010) (see section 2.2). The four most commonly used sequencing-based methods are BS-seq, RRBS, MeDIP-seq, and MBD-seq. BS-seq and RRBS enable an insight into DNA methylation at single-base resolution, whereas the enrichment-based methods, MeDIP-seq and MBD-seq, assess DNA methylation of genomic fragments ranging between 400-700 bp, which then are representative for all cytosines within the respective fragment (Down et al., 2008). All of these four methods exhibit comparable DNA methylation levels, but differ in their coverage, accuracy and cost (Bock et al., 2010; Harris et al., 2010). It has been shown that low sequencing coverage is most often sufficient to determine large differentially methylated domains or even global DNA methylation tendencies, but insufficient to detect loci-specific alterations. Depending on the problem, it is necessary to either sequence few samples more deeply or more samples less deeply.

In the following section advantages and disadvantages of different sequencing-based methods are specified. The ability of BS-seq to assess DNA methylation at single-base resolution certainly results in a high accuracy and enables the simultaneous readout of

further variants, such as SNPs (see Appendix C). However, a substantial increase of BS-seq coverage is accompanied by much higher costs compared to RRBS, MeDIP-seq, and MBD-seq. Enrichment-based methods benefitis from the fact that all four nucleotides are unmodified, which modestly increases the mapping efficiency compared to bisulfite-based methods, which predominantly involve adenines, guanines, and thymines in their bisulfite sequences. A further advantage of enrichment-based methods is the ability to determine DNA hydroxymethylation, the oxidation of methylated cytosines by the TET family (see section 1.2). Hydroymethylation cannot be detected by bisulfite-based methods (Krueger et al., 2012) (see section 2.2.1).

Although most studies analyzed CpG methylation so far, the interest in non-CpG methylation increases (Lister et al., 2009, 2011; Pelizzola and Ecker, 2010). This type of DNA methylation can be easily detected by bisulfite-based methods. However, enrichment-based methods generate several difficulties. It is a challenge to separate the level of CpG and non-CpG methylation because only entire genomic fragments are assessed. Furthermore, Lister *et al.* showed that methylated non-CpG sites are predominantly present in genomic regions with high CpG methylation. This fact additionally complicates the independent measurement of DNA methylation in different sequence contexts (Lister et al., 2009).

In the following section, I compare RRBS and BS-seq. The SOLiD™ BS-seq data set generated for the study of Appendix C and a RRBS data set obtained from the same cell line are analyzed. 79.9 gigabases (Gb) of SOLiD™ sequences and 5.8 Gb of HiSeq 2000 sequences were aligned to the human reference (hg19/NCBI 37). The log-scaled distribution of CpG coverage, assessed by both methods, is depicted in Figure 4.2. Obviously, the restriction of RRBS to genomic regions with high density of CpG dinucleotides results in considerably higher coverage. By absolute numbers, the mean coverage of RRBS accounts for 59.6× and 16.5× for SOLiD™ BS-seq.

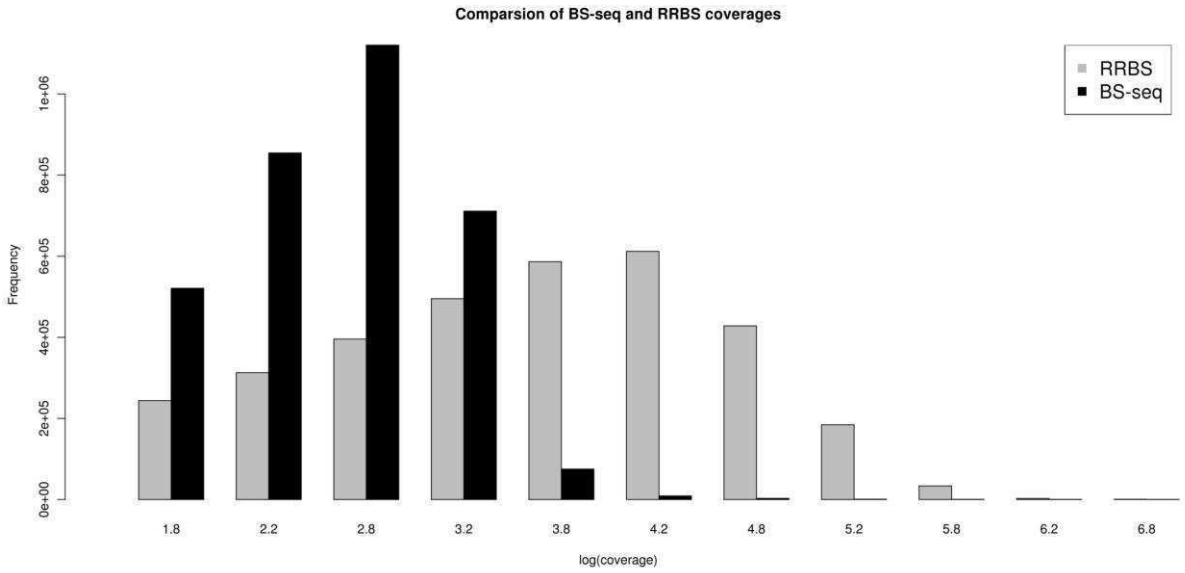


Figure 4.2: Distribution of coverage of CpG sites assessed by SOLiD™ BS-seq and RRBS.

It was previously shown that the accuracy of DNA methylation calls strongly depends on the coverage (Harris et al., 2010; Bock et al., 2010). This dependency was also analyzed for the RRBS and SOLiD™ BS-seq data sets discussed above. For the analysis, different groups of CpG sites, with increasing coverage, were generated. The Pearson correlation coefficient was used to compare DNA methylation levels of both methods (see Table 4). CpG sites, assessed by at least 5 bisulfite sequences, exhibit a Pearson correlation coefficient of r=0.902, which is in line with results of Bock *et al.* (Bock et al., 2010). Increasing coverage significantly results in higher correlations up to the threshold of 25 bisulfite sequences. Beyond that, coverage-specific variances are observed for single cytosines, but not for genomic domains. In the end, whole genome approaches, such as RRBS and BS-seq, should rather be used for the determination of DMRs than for DNA methylation alterations of single cytosines. For site-specific DNA methylation analysis, I would recommend genome-wide arrays, such as the *Infinium HumanMethylation BeadChip*.

Table 4: Pearson correlation of RRBS and SOLiD™ BS-seq based on increasing coverage.

Coverage	***≥5***	***≥7***	***≥9***	***≥11***	***≥13***	***≥15***	***≥17***
Pearson r	0.902	0.908	0.912	0.915	0.918	0.920	0.922
Coverage	***≥19***	***≥21***	***≥23***	***≥25***	***≥30***	***≥40***	***≥50***
Pearson r	0.924	0.925	0.927	0.928	0.930	0.935	0.938

Comprehensive analyses about platform-specific advantages and disadvantages are not published up to today. This issue is discussed in the remainder of this section**.** To date, bisulfite-based methods were carried out on the Genome Analyzer and HiSeq 2000 platforms (Illumina), the SOLiD™ platform (Life Technologies), and the 454 platform (Roche) (for details about NGS platforms see sections 2.1) (Lister et al., 2009; Kreck et al., 2011; Bormann Chung et al., 2010; Herrmann et al., 2011). Enrichment-based methods were only carried out on the Genome Analyzer and HiSeq 2000 platform (Illumina) (Bock et al., 2010).
We showed that both BS-seq data of the SOLiD™ and the HiSeq 2000 platform exhibit accurate and comparable CpG methylation measurements (Pearson correlation coefficient r=0.86) (see Appendix C). However, non-CpG methylation levels exhibit slightly different results. Non-CpG sites, assessed by SOLiD™ BS-seq, were slightly weaker methylated compared to DNA methylation levels of the HiSeq 2000 platform (non-CpG mean $\text{methylation}_{\text{SOLiD™}}$=0.16, non-CpG mean $\text{methylation}_{\text{HiSeq 2000}}$=0.47). This fact could also be observed for the same non-CpG sites assessed by RRBS, even though not to this extent (non-CpG mean $\text{methylation}_{\text{RRBS}}$=0.21). However, the latter observation may be related to the relatively low coverage of the HiSeq 2000 data set compared to the RRBS data set. In summary, both platforms are applicable to generate accurate DNA methylation measurements, whereas organisms with a high amount of methylated non-CpG sites should rather be analyzed by the HiSeq 2000 platform.

4.4 Interindividual DNA Methylation Differences

Epigenetic studies particularly aim to associate epigenetic alterations with specific phenotypes. In the case of DNA methylation, this is usually performed by the determination of DMRs. Sensitivity and specificity of approaches for the determination of DMRs notably vary between different technologies. Array-based methods only assess selected sites. The

low density of covered sites might significantly decrease the statistical power (see section 4.1). Thus DMRs obtained by array-based analysis have to be considered cautiously (see section 4.2). NGS-based methods provide an extensive insight into DNA methylation. However, there are fundamentally different computational approaches for the determination of DMRs in NGS data. Below, selected approaches are discussed.

Lister *et al.* at first published DMRs of methylomes of ES cells H1 and fetal fibroblasts IMR90 (Lister et al., 2009). They used a binomial distribution to determine DNA methylation sites in both CpG and non-CpG sequence contexts (see section 2.3.1). However, differentially methylated cytosines were solely determined for CpG sites, since hardly any significantly methylated non-CpG sites could be identified for IMR90. To exclude coverage biases, they compare only CpG sites of H1 and IMR90 that involve a ratio of coverage between 0.8 and 1.2. Subsequently, Lister *et al.* applied a two-tailed Fisher's Exact Test resulting in 6,023,738 CpG sites that were more highly methylated in H1 compared to IMR90, and 124,161 CpG sites that were more highly methylated in IMR90 compared to H1. Finally, they used a sliding window approach to select 1,000 bp regions containing at least 4 differential methylated sites. Adjacent differentially methylated regions were joined together.

A further publication by Lister *et al.* identified DMRs in iPSCs and ES cells (Lister et al., 2011). DMRs were detected based on more stringent test parameters than in their analysis of H1 and IMR90 (Lister et al., 2009). They generated smoothed DNA methylation levels in 100 bp windows, whereas regions comprising a set of 10 adjacent windows over a distance less than 1,100 bp were considered. A non-parametric Wilcoxon Test (or Kruskal-Wallis Test for more than two samples, $p<0.01$) for regions involving a 4-fold enrichment of DNA methylation level was applied.

Li *et al.* analyzed the methylome of human peripheral blood mononuclear cells (PBMCs) and compared it to the methylome of fetal fibroblasts IMR90 (Li et al., 2010). They used a sliding window approach combined with a Fisher Exact Test ($p<1e^{-20}$). Windows should at least contain 5 CpG sites with a 2-fold change in DNA methylation level. They additionally required that both tissues do not involve DNA methylation levels of less than 0.2. Adjacent differentially methylation regions were as well joined together.

Hansen *et al.* described methylomes of different cancer types, including colon, lung, breast, thyroid, and Wilms tumors (Hansen et al., 2011). They developed a bioinformatics approach

to smooth CpG methylation levels by their coverage and environmental CpG density. DMRs were determined using *t* statistics.

Although it is known that different tissues and cell types involve varying methylomes, the respective computational method significantly influences the amount of DMRs. The numbers of DMRs for the methylome analyses range from 1,175 (ES cells and 5 iPSCs) (Lister et al., 2011) to 240,856 (PBMCs and IMR90) (Li et al., 2010). In summary, it is a challenge to strike a balance between methods detecting too many false positive DMRs and those, which miss too many DMRs.

I developed a new sliding window approach for the determination of DMRs using smoothed DNA methylation levels considering environmental CpG density. Genomic windows containing at least 10 CpG sites within 1100 bp and 4-fold enrichment of mean DNA methylation levels are considered. CpG sites are then smoothed by a local polynomial regression fitting using the statistical software R (http://cran.r-project.org/). I use *t* statistics ($p<0.01$), since the Wilcoxon Rank-Sum Test is derived on the assumption that data consist of no ties. This assumption cannot be ensured a priori, as DNA methylation is usually bimodally distributed with a majority of clustered fully methylated sites (Pelizzola and Ecker, 2010).

For testing purposes of this DMR determination approach, methylome data of the DAUDI cell line (malignant B lymphocytes) and of PBMCs involving lymphocytes, monocytes, and macrophages were used (Li et al., 2010) (see Appendix C) (see Figure 4.3). 666 DMRs were identified, whereas the majority of DMRs involved higher DNA methylation in DAUDI compared to PBMCs (indicated as red boxes in the ideogram) and only 2 regions show a reverse effect (blue box in ideogram). DMRs were illustrated by a ideogram (the ideogram was generated with *ideographica* (Kin and Ono, 2007)) (see Figure 4.3). The heat map, incorporated in Figure 4.3, indicates densities of RefSeq genes within a specific genomic region, ranging from light gray (no RefSeq genes) to black (high density of RefSeq genes).

To validate the determined DMRs by independent technologies, HiSeq 2000 and *Infinium HumanMethylation BeadChip* data of Appendix C were used. The mean difference between DNA methylation levels within DMRs assessed by SOLiD™ BS-seq and HiSeq 2000 accounted for 0.02, emphasizing high concordance of both methods. The comparsion of SOLiD™ BS-seq and *HumanMethylation BeadChip* DNA methylation levels within DMRs required further criteria due to the notably lower density of sites covered in the array. Only DMRs, assessed

by at least 3 sites of the array, were considered. The mean difference between DNA methylation levels within DMRs assessed by SOLiD™ BS-seq and *HumanMethylation BeadChip* finally accounted for 0.12, which is marginally higher than the mean difference between the NGS methods. This larger difference is probably related to the sigmoidally distributed correlation between extreme DNA methylation levels of BS-seq and array data (see section 4.1) (see Appendix A).

Conclusions, based on these DMRs, should be drawn with care, since PBMCs are only partly suitable as a reference-methylome for DAUDI. At least it could be observed that the majority of DMRs were located within genomic regions involving a high density of RefSeq genes (see Figure 4.3).

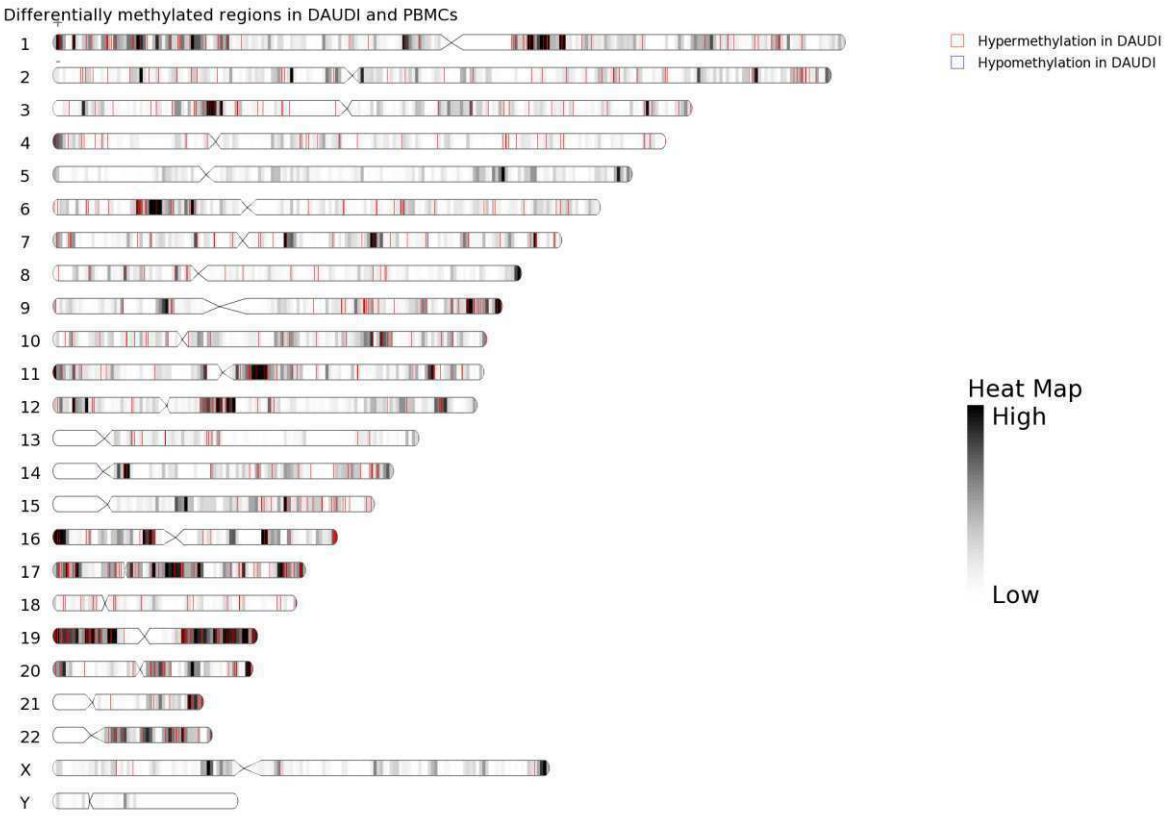


Figure 4.3: Ideogram of DMRs in DAUDI and PBMCs.

4.5 Conclusion and Future Perspectives

The generation of the first human methylomes at single base resolution by Lister *et al.* initiated a rapid development of NGS-based DNA methylation analysis over the past three years (Lister et al., 2009). After this publication and the establishment of associated biotechnological methods, the bottleneck of genome-wide DNA methylation analysis especially shifted from data production to data analysis. Findings, incorporated in this thesis, can be integrated into the recent development of BS-seq analysis in the following way. Before the publication of Lister *et al.*, BS-seq of human genomes was only carried out on NGS platforms using base space encoded sequences (see section 2.1) and DNA methylation levels could be only verified by array-based or loci-specific sequencing methods. A bioinformatics tool for the analysis of color-space (see section 2.1) BS-seq data henceforth represents an alternative for a platform-independent validation (see Appendix A). BS-seq analysis of both base and color-space data poses considerable challenges. These challenges and the most appropriate way to analyze this type of data are described in Appendix B. So far, whole methylomes are available for human embryonic stem cells, fetal fibroblasts, peripheral blood mononuclear cells, and colon cancer cells (Lister et al., 2009, 2011; Li et al., 2010; Hansen et al., 2011). The study, incorporated in Appendix C, analyzed a further human methylome derived from an endemic Burkitt's lymphoma cell line. BS-seq was carried out for base and color-space data, where a high correlation (r=0.86) could be observed. This study revealed new methylome characteristics for B-cell lymphomas and introduced new approaches for correlation analyses of DNA methylation and transcriptional levels. The DAUDI methylome might prove valuable as a reference methylome for future epigenetic studies.

Although whole methylome analysis can be carried out by BS-seq in a cost-efficient manner by now, there are still certain limitations. Firstly, current BS-seq methods analyze mixed populations of cell types. Consequently, BS-seq results only reflect the composition of methylomes of different cell types, which are difficult to interpret. Secondly, DNA molecules are amplified prior to the sequencing step (see section 2.1). This might lead to uneven distributions of bisulfite sequences, which interferes with the determination of DNA methylation levels. These limitations can be addressed by recently developed single-molecule real-time sequencing (SMRT) methods, which enable the direct detection of DNA

methylation without prior bisulfite conversion. In SMRT sequencing, DNA polymerases catalyze the incorporation of fluorescently labeled nucleotides into complementary DNA strands. Each nucleotide is attached to one of four different fluorescent dyes (Flusberg et al., 2010). During the incorporation of a nucleotide the fluorescent tag is detected by a zero-mode waveguide (an optical approach for studying single-molecule dynamics), and base calls are carried out based on the fluorescence of the underlying dye (Levene et al., 2003). The arrival times and durations of the resulting fluorescence pulses reveal information about polymerase kinetics, enabling the direct detection of epigenetic modifications, such as DNA methylation and hydroxymethylation (Flusberg et al., 2010; Song et al., 2012). These SMRT sequencing methods currently pass the end of the development phase and first test runs were already carried out. However, they still exhibit an increased per-base error rate and ambiguous results (Eid et al., 2009; Song et al., 2012). Once these methods are well-established, they will be very helpful and cost-effective to simultaneously readout genetic and different epigenetic variations.

DNA methylation was described to be associated with transcriptional silencing by quantitative studies comparing DNA methylation and transcriptional levels (Lister et al., 2009, 2011; Li et al., 2010; Hansen et al., 2011). However, the underlying regulatory network and its interaction partners are not yet clear. To address this issue, all DNA-binding proteins, affected or unaffected by DNA methylation, should be identified. Especially transcription factors, which are predominantly active in regions upstream of genes, are of great interest.

DNA methylation is the most studied epigenetic modification so far. Future studies need to analyze further epigenetic modifications and in particular histone modifications to a similar extent because possibly only the functional interaction of all epigenetic modifications clarifies their impact on transcriptional levels. This presumption can be exemplified by the following examples. On the one hand, Esteller stated that tumor-suppressor genes are inactivated by hypermethylation of CpG islands in their promoter regions, which directly influence the growth of tumors (Manel Esteller, 2008). This mechanism is reversible by so-called demethylating and methylating agents, which can awake and silence genes involving hypermethylation and hypomethylation in their promoter region (Manel Esteller, 2008). On the other hand, Lister *et al.* pointed out that DNA methylation might only be a consequence of closed chromatin structure, where DNA methylation in embryonic stem cells might be lost during differentiation, resulting in accumulation of repressive chromatin marks

(Pelizzola and Ecker, 2010; Lister et al., 2009). In summary, further functional studies on epigenetic modifications should be carried out to better understand the mechanism of DNA methylation and its regulatory effect on transcriptional levels.

In conclusion to my experience of the last three years, future computational epigenetic studies on DNA methylation should consider the following aspects. Firstly, it has to be explored to which extent DNA methylation is regulated by environmental factors. This phenomenon has to be studied regarding its stability or even inheritance, since it is also conceivable that a dynamic modification like DNA methylation is affected by environmental factors over a specific period of time and initial DNA methylation patterns are then re-established. In this regard, DNA hydroxymethylation, a possible pathway for de-methylation, might emerge as a helpful access. Secondly, DNA methylation needs to be analyzed in large-scale studies to assess its effect on different diseases. However, these studies involve critical challenges as for instance genetic-epigenetic interaction and the determination of appropriate features of variation in DNA methylation (single site alterations, DMRs, ASMs etc.). For this purpose, further whole methylomes of different cell types and tissues should be analyzed and categorized to specify appropriate genomic domains. Even though all these points minimize the amount of confounding factors for the analysis of DNA methylation, it will be very difficult to “control” such a dynamic modification and the following question will frequently come up: Is the observed alteration in DNA methylation the cause or the consequence of a specific disease? This question can possibly be addressed by the integration of all involved factors of the pathogenesis of a disease into for instance extended gene regulatory networks (GRNs), a system biological method to model DNA-encoded interaction partners. The dynamics of DNA methylation requires GRNs to distinguish different diseases involving the same genotype but not the same epigenotype. Such a dynamic GRNs approach should initially be explored for a phenotype, where genetics and epigenetics and their interaction is investigated as far as possible.

In summary, “dynamic times” are ahead for research in DNA methylation analysis!

References

Aagaard-Tillery,K.M. et al. (2008) Developmental origins of disease and determinants of chromatin structure: maternal diet modifies the primate fetal epigenome. *Journal of molecular endocrinology*, **41**, 91-102.

Acinas,S.G. et al. (2005) PCR-induced sequence artifacts and bias: insights from comparison of two 16S rRNA clone libraries constructed from the same sample. *Applied and environmental microbiology*, **71**, 8966-9.

Aitman,T.J. et al. (2011) The future of model organisms in human disease research. *Nature reviews. Genetics*, **12**, 575-82.

Akiyama,Y. et al. (2003) Cell-type-specific repression of the maspin gene is disrupted frequently by demethylation at the promoter region in gastric intestinal metaplasia and cancer cells. *The American journal of pathology*, **163**, 1911-9.

Anderson,C.M. et al. (2006) Placental insufficiency leads to developmental hypertension and mesenteric artery dysfunction in two generations of Sprague-Dawley rat offspring. *Biology of reproduction*, **74**, 538-44.

Anway,M.D. et al. (2006) Endocrine disruptor vinclozolin induced epigenetic transgenerational adult-onset disease. *Endocrinology*, **147**, 5515-23.

Anway,M.D. et al. (2005) Epigenetic transgenerational actions of endocrine disruptors and male fertility. *Science (New York, N.Y.)*, **308**, 1466-9.

Araujo,F.D. et al. (1998) Concurrent replication and methylation at mammalian origins of replication. *Molecular and cellular biology*, **18**, 3475-82.

Arber,W. and Linn,S. (1969) DNA modification and restriction. *Annual review of biochemistry*, **38**, 467-500.

Baccarelli,A. et al. (2009) Rapid DNA methylation changes after exposure to traffic particles. *American journal of respiratory and critical care medicine*, **179**, 572-8.

Barber,R.C. et al. (2006) Radiation-induced transgenerational alterations in genome stability and DNA damage. *Oncogene*, **25**, 7336-42.

Barlow,D.P. et al. (1991) The mouse insulin-like growth factor type-2 receptor is imprinted and closely linked to the Tme locus. *Nature*, **349**, 84-7.

Bartolomei,M.S. et al. (1991) Parental imprinting of the mouse H19 gene. *Nature*, **351**, 153-5.

References

Baylin,S. and Bestor,T.H. (2002) Altered methylation patterns in cancer cell genomes: cause or consequence? *Cancer cell*, **1**, 299-305.

Baylin,S B et al. (1998) Alterations in DNA methylation: a fundamental aspect of neoplasia. *Advances in cancer research*, **72**, 141-96.

Baylin,Stephen B and Ohm,J.E. (2006) Epigenetic gene silencing in cancer - a mechanism for early oncogenic pathway addiction? *Nature reviews. Cancer*, **6**, 107-16.

Bedford,M.T. and van Helden,P.D. (1987) Hypomethylation of DNA in pathological conditions of the human prostate. *Cancer research*, **47**, 5274-6.

Bell,J.T. and Spector,T.D. (2011) A twin approach to unraveling epigenetics. *Trends in genetics : TIG*, **27**, 116-25.

Bender,J. and Fink,G.R. (1995) Epigenetic control of an endogenous gene family is revealed by a novel blue fluorescent mutant of Arabidopsis. *Cell*, **83**, 725-34.

Bestor,T. et al. (1988) Cloning and sequencing of a cDNA encoding DNA methyltransferase of mouse cells. The carboxyl-terminal domain of the mammalian enzymes is related to bacterial restriction methyltransferases. *Journal of molecular biology*, **203**, 971-83.

Bibikova,M. et al. (2011) High density DNA methylation array with single CpG site resolution. *Genomics*, **98**, 288-95.

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

Bird,A.P. (1980) DNA methylation and the frequency of CpG in animal DNA. *Nucleic acids research*, **8**, 1499-504.

Bjerkvig,R. et al. (2005) Opinion: the origin of the cancer stem cell: current controversies and new insights. *Nature reviews. Cancer*, **5**, 899-904.

Blatt,J. et al. (2003) Ovarian carcinoma in an adolescent with transgenerational exposure to diethylstilbestrol. *Journal of pediatric hematology/oncology*, **25**, 635-6.

Bock,C. et al. (2010) Quantitative comparison of genome-wide DNA methylation mapping technologies. *Nature biotechnology*, **28**, 1106-14.

Bock,C. and Lengauer,T. (2008) Computational epigenetics. *Bioinformatics (Oxford, England)*, **24**, 1-10.

Bollati,V. et al. (2007) Changes in DNA methylation patterns in subjects exposed to low-dose benzene. *Cancer research*, **67**, 876-80.

Bormann Chung,C.A. et al. (2010) Whole methylome analysis by ultra-deep sequencing using two-base encoding. *PloS one*, **5**, e9320.

References

Breitling,L.P. et al. (2011) Tobacco-smoking-related differential DNA methylation: 27K discovery and replication. *American journal of human genetics*, **88**, 450-7.

Buiting,K. et al. (1995) Inherited microdeletions in the Angelman and Prader-Willi syndromes define an imprinting centre on human chromosome 15. *Nature genetics*, **9**, 395-400.

Byun,H.-M. et al. (2009) Epigenetic profiling of somatic tissues from human autopsy specimens identifies tissue- and individual-specific DNA methylation patterns. *Human molecular genetics*, **18**, 4808-17.

Bártová,E. et al. (2008) Histone modifications and nuclear architecture: a review. *The journal of histochemistry and cytochemistry : official journal of the Histochemistry Society*, **56**, 711-21.

Cadieux,B. et al. (2006) Genome-wide hypomethylation in human glioblastomas associated with specific copy number alteration, methylenetetrahydrofolate reductase allele status, and increased proliferation. *Cancer research*, **66**, 8469-76.

Carone,B.R. et al. (2010) Paternally induced transgenerational environmental reprogramming of metabolic gene expression in mammals. *Cell*, **143**, 1084-96.

Chang,H.-S. et al. (2006) Transgenerational epigenetic imprinting of the male germline by endocrine disruptor exposure during gonadal sex determination. *Endocrinology*, **147**, 5524-41.

Cheng,J.C. et al. (2004) Preferential response of cancer cells to zebularine. *Cancer cell*, **6**, 151-8.

Chinnusamy,V. and Zhu,J.-K. (2009) Epigenetic regulation of stress responses in plants. *Current opinion in plant biology*, **12**, 133-9.

Christensen,B.C. et al. (2009) Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CpG island context. *PLoS genetics*, **5**, e1000602.

Constantinides,P.G. et al. (1977) Functional striated muscle cells from non-myoblast precursors following 5-azacytidine treatment. *Nature*, **267**, 364-6.

Cooper,D.L. et al. (1993) Methyl-directed mismatch repair is bidirectional. *The Journal of biological chemistry*, **268**, 11823-9.

Cornacchia,E. et al. (1988) Hydralazine and procainamide inhibit T cell DNA methylation and induce autoreactivity. *Journal of immunology (Baltimore, Md. : 1950)*, **140**, 2197-200.

Costa,E. et al. (2002) REELIN and schizophrenia: a disease at the interface of the genome and the epigenome. *Molecular interventions*, **2**, 47-57.

Costello,J.F. et al. (2000) Aberrant CpG-island methylation has non-random and tumour-type-specific patterns. *Nature genetics*, **24**, 132-8.

Csaba,G. and Inczefi-Gonda,A. (1998) Transgenerational effect of a single neonatal benzpyrene treatment on the glucocorticoid receptor of the rat thymus. *Human & experimental toxicology*, **17**, 88-92.

Csaba,G. and Karabélyos,C. (1997) Transgenerational effect of a single neonatal benzpyrene treatment (imprinting) on the sexual behavior of adult female rats. *Human & experimental toxicology*, **16**, 553-6.

Cubas,P. et al. (1999) An epigenetic mutation responsible for natural variation in floral symmetry. *Nature*, **401**, 157-61.

DeBaun,M R and Tucker,M.A. (1998) Risk of cancer during the first four years of life in children from The Beckwith-Wiedemann Syndrome Registry. *The Journal of pediatrics*, **132**, 398-400.

DeBaun,Michael R et al. (2002) Epigenetic alterations of H19 and LIT1 distinguish patients with Beckwith-Wiedemann syndrome with cancer and birth defects. *American journal of human genetics*, **70**, 604-11.

DeChiara,T.M. et al. (1991) Parental imprinting of the mouse insulin-like growth factor II gene. *Cell*, **64**, 849-59.

Dedeurwaerder,S. et al. (2011) Evaluation of the Infinium Methylation 450K technology. *Epigenomics*, **3**, 771-84.

Dohm,J.C. et al. (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic acids research*, **36**, e105.

Down,T.A. et al. (2008) A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nature biotechnology*, **26**, 779-85.

Dressman,D. et al. (2003) Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 8817-22.

Dubrova,Y.E. (2005) Radiation-induced mutation at tandem repeat DNA Loci in the mouse germline: spectra and doubling doses. *Radiation research*, **163**, 200-7.

Duncan,B.K. and Miller,J.H. (1980) Mutagenic deamination of cytosine residues in DNA. *Nature*, **287**, 560-1.

Eid,J. et al. (2009) Real-time DNA sequencing from single polymerase molecules. *Science (New York, N.Y.)*, **323**, 133-8.

Erlich,Y. et al. (2008) Alta-Cyclic: a self-optimizing base caller for next-generation sequencing. *Nature methods*, **5**, 679-82.

Esteller,M, Corn,P.G., et al. (2001) A gene hypermethylation profile of human cancer. *Cancer research*, **61**, 3225-9.

Esteller,M, Fraga,M F, et al. (2001) DNA methylation patterns in hereditary human cancers mimic sporadic tumorigenesis. *Human molecular genetics*, **10**, 3001-7.

Esteller,M et al. (2000) Promoter hypermethylation and BRCA1 inactivation in sporadic breast and ovarian tumors. *Journal of the National Cancer Institute*, **92**, 564-9.

Esteller,Manel (2008) Epigenetics in cancer. *The New England journal of medicine*, **358**, 1148-59.

Fang,M.Z. et al. (2003) Tea polyphenol (-)-epigallocatechin-3-gallate inhibits DNA methyltransferase and reactivates methylation-silenced genes in cancer cell lines. *Cancer research*, **63**, 7563-70.

Fazzari,M.J. and Greally,J.M. (2004) Epigenomics: beyond CpG islands. *Nature reviews. Genetics*, **5**, 446-55.

Fedurco,M. et al. (2006) BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic acids research*, **34**, e22.

Feil,R et al. (1994) Developmental control of allelic methylation in the imprinted mouse Igf2 and H19 genes. *Development (Cambridge, England)*, **120**, 2933-43.

Feil,Robert and Fraga,Mario F (2011) Epigenetics and the environment: emerging patterns and implications. *Nature reviews. Genetics*, **13**, 97-109.

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

Feinberg,Andrew P and Irizarry,R.A. (2010) Evolution in health and medicine Sackler colloquium: Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. *Proceedings of the National Academy of Sciences of the United States of America*, **107 Suppl** , 1757-64.

Ferguson-Smith,A.C. (2011) Genomic imprinting: the emergence of an epigenetic paradigm. *Nature reviews. Genetics*, **12**, 565-75.

Ficz,G. et al. (2011) Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. *Nature*, **473**, 398-402.

Flintoft,L. (2010) Complex disease: Adding epigenetics to the mix. *Nature Reviews Genetics*, **11**, 94-95.

Flusberg,B.A. et al. (2010) Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nature methods*, **7**, 461-5.

Fraga,Mario F et al. (2005) Epigenetic differences arise during the lifetime of monozygotic twins. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 10604-9.

Frommer,M. et al. (1992) A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proceedings of the National Academy of Sciences of the United States of America*, **89**, 1827-31.

Gluckman,P.D. et al. (2009) Epigenetic mechanisms that underpin metabolic and cardiovascular diseases. *Nature reviews. Endocrinology*, **5**, 401-8.

Goll,M.G. et al. (2006) Methylation of tRNAAsp by the DNA methyltransferase homolog Dnmt2. *Science (New York, N.Y.)*, **311**, 395-8.

Goll,M.G. and Bestor,T.H. (2005) Eukaryotic cytosine methyltransferases. *Annual review of biochemistry*, **74**, 481-514.

Gonzalez-Zulueta,M. et al. (1995) Methylation of the 5' CpG island of the p16/CDKN2 tumor suppressor gene in normal and transformed human tissues correlates with gene silencing. *Cancer research*, **55**, 4531-5.

Graff,J.R. et al. (1995) E-cadherin expression is silenced by DNA hypermethylation in human breast and prostate carcinomas. *Cancer research*, **55**, 5195-9.

Greger,V. et al. (1989) Epigenetic changes may contribute to the formation and spontaneous regression of retinoblastoma. *Human genetics*, **83**, 155-8.

Grönniger,E. et al. (2010) Aging and chronic sun exposure cause distinct epigenetic changes in human skin. *PLoS genetics*, **6**, e1000971.

Gu,H. et al. (2010) Genome-scale DNA methylation mapping of clinical samples at single-nucleotide resolution. *Nature methods*, **7**, 133-6.

Gu,H. et al. (2011) Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. *Nature protocols*, **6**, 468-81.

Gupta,A. et al. (2003) Hypomethylation of the synuclein gamma gene CpG island promotes its aberrant expression in breast carcinoma and ovarian carcinoma. *Cancer research*, **63**, 664-73.

Hansen,K.D. et al. (2011) Increased methylation variation in epigenetic domains across cancer types. *Nature genetics*, **43**, 768-75.

Harris,R.A. et al. (2010) Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nature biotechnology*, **28**, 1097-105.

Hellman,A. and Chess,A. (2010) Extensive sequence-influenced DNA methylation polymorphism in the human genome. *Epigenetics & Chromatin*, **3**, 11.

Herman,J G et al. (1995) Inactivation of the CDKN2/p16/MTS1 gene is frequently associated with aberrant DNA methylation in all common human cancers. *Cancer research*, **55**, 4525-30.

Herman,J G et al. (1994) Silencing of the VHL tumor-suppressor gene by DNA methylation in renal carcinoma. *Proceedings of the National Academy of Sciences of the United States of America*, **91**, 9700-4.

Herman,James G and Baylin,Stephen B (2003) Gene silencing in cancer in association with promoter hypermethylation. *The New England journal of medicine*, **349**, 2042-54.

Herrmann,A. et al. (2011) Pipeline for large-scale microdroplet bisulfite PCR-based sequencing allows the tracking of hepitype evolution in tumors. *PloS one*, **6**, e21332.

Hindorff,L.A. et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 9362-7.

Hoffmann,M.J. and Schulz,W.A. (2005) Causes and consequences of DNA hypomethylation in human cancer. *Biochemistry and cell biology = Biochimie et biologie cellulaire*, **83**, 296-321.

Holliday,R. (1987) The inheritance of epigenetic defects. *Science (New York, N.Y.)*, **238**, 163-70.

Holliday,R. and Pugh,J.E. (1975) DNA modification mechanisms and gene activity during development. *Science (New York, N.Y.)*, **187**, 226-32.

Humpherys,D. et al. (2001) Epigenetic instability in ES cells and cloned mice. *Science (New York, N.Y.)*, **293**, 95-7.

Hussain,M. et al. (2009) Tobacco smoke induces polycomb-mediated repression of Dickkopf-1 in lung cancer cells. *Cancer research*, **69**, 3570-8.

IHGSC (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931-45.

Ikeda,M. et al. (2005) Repeated in utero and lactational 2,3,7,8-tetrachlorodibenzo-p-dioxin exposure affects male gonads in offspring, leading to sex ratio changes in F2 progeny. *Toxicology and applied pharmacology*, **206**, 351-5.

Jacinto,F.V. et al. (2008) Methyl-DNA immunoprecipitation (MeDIP): hunting down the DNA methylome. *BioTechniques*, **44**, 35, 37, 39 passim.

Jair,K.-W. et al. (2006) De novo CpG island methylation in human cancer cells. *Cancer research*, **66**, 682-92.

Jeltsch,A. (2002) Beyond Watson and Crick: DNA methylation and molecular enzymology of DNA methyltransferases. *Chembiochem : a European journal of chemical biology*, **3**, 274-93.

Jin,S.-G. et al. (2010) Examination of the specificity of DNA methylation profiling techniques towards 5-methylcytosine and 5-hydroxymethylcytosine. *Nucleic acids research*, **38**, e125.

Jirtle,R.L. and Skinner,M.K. (2007) Environmental epigenomics and disease susceptibility. *Nature reviews. Genetics*, **8**, 253-62.

Johannes,F. et al. (2009) Assessing the impact of transgenerational epigenetic variation on complex traits. *PLoS genetics*, **5**, e1000530.

Jones,P A and Taylor,S.M. (1980) Cellular differentiation, cytidine analogs and DNA methylation. *Cell*, **20**, 85-93.

Jones,Peter A and Baylin,Stephen B (2007) The epigenomics of cancer. *Cell*, **128**, 683-92.

Jones,Peter A and Baylin,Stephen B (2002) The fundamental role of epigenetic events in cancer. *Nature reviews. Genetics*, **3**, 415-28.

Kaminsky,Z.A. et al. (2009) DNA methylation profiles in monozygotic and dizygotic twins. *Nature genetics*, **41**, 240-5.

Kerkel,K. et al. (2008) Genomic surveys by methylation-sensitive SNP analysis identify sequence-dependent allele-specific DNA methylation. *Nature genetics*, **40**, 904-8.

Killian,J.K. et al. (2000) M6P/IGF2R imprinting evolution in mammals. *Molecular cell*, **5**, 707-16.

Kim,D.-H. et al. (2009) Vernalization: winter and the timing of flowering in plants. *Annual review of cell and developmental biology*, **25**, 277-99.

Kim,D.N. et al. (2011) The role of promoter methylation in Epstein-Barr virus (EBV) microRNA expression in EBV-infected B cell lines. *Experimental & molecular medicine*, **43**, 401-10.

Kin,T. and Ono,Y. (2007) Idiographica: a general-purpose web application to build idiograms on-demand for human, mouse and rat. *Bioinformatics (Oxford, England)*, **23**, 2945-6.

Kircher,M. et al. (2009) Improved base calling for the Illumina Genome Analyzer using machine learning strategies. *Genome biology*, **10**, R83.

Kircher,M. and Kelso,J. (2010a) High-throughput DNA sequencing--concepts and limitations. *BioEssays : news and reviews in molecular, cellular and developmental biology*, **32**, 524-36.

Kircher,M. and Kelso,J. (2010b) High-throughput DNA sequencing--concepts and limitations. *BioEssays : news and reviews in molecular, cellular and developmental biology*, **32**, 524-36.

Kisseljova,N.P. et al. (1998) De novo methylation of selective CpG dinucleotide clusters in transformed cells mediated by an activated N-ras. *International journal of oncology*, **12**, 203-9.

Kitsberg,D. et al. (1993) Allele-specific replication timing of imprinted gene regions. *Nature*, **364**, 459-63.

Knudson,A.G. (2001) Two genetic hits (more or less) to cancer. *Nature reviews. Cancer*, **1**, 157-62.

Koga,Y. et al. (2009) Genome-wide screen of promoter methylation identifies novel markers in melanoma. *Genome research*, **19**, 1462-70.

Kreck,B. et al. (2011) B-SOLANA: An approach for the analysis of two-base encoding bisulfite sequencing data. *Bioinformatics (Oxford, England)*, btr660-.

Krueger,F. et al. (2012) DNA methylome analysis using short bisulfite sequencing data. *Nature methods*, **9**, 145-51.

Krueger,F. and Andrews,S.R. (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics (Oxford, England)*, **27**, 1571-2.

Kucharski,R. et al. (2008) Nutritional control of reproductive status in honeybees via DNA methylation. *Science (New York, N.Y.)*, **319**, 1827-30.

Kulis,M. and Esteller,Manel (2010) DNA methylation and cancer. *Advances in genetics*, **70**, 27-56.

Langevin,S.M. et al. (2011) The influence of aging, environmental exposures and local sequence features on the variation of DNA methylation in blood. *Epigenetics : official journal of the DNA Methylation Society*, **6**, 908-19.

Levene,M.J. et al. (2003) Zero-mode waveguides for single-molecule analysis at high concentrations. *Science (New York, N.Y.)*, **299**, 682-6.

Li,Y. et al. (2010) The DNA methylome of human peripheral blood mononuclear cells. *PLoS biology*, **8**, e1000533.

References

LifeTechnologies (2008) Principles of Di-Base Sequencing and the Advantages of Color Space Analysis in the SOLiD System. *ABI. Tech. Rep.*, **139AP10-01**.

Lister,R. et al. (2011) Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature*, **471**, 68-73.

Lister,R. et al. (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, **462**, 315-22.

Lubinsky,M. et al. (1974) Letter: Autosomal-dominant sex-dependent transmission of the Wiedemann-Beckwith syndrome. *Lancet*, **1**, 932.

Maleszka,R. Epigenetic integration of environmental and genomic signals in honey bees: the critical interplay of nutritional, brain and reproductive networks. *Epigenetics : official journal of the DNA Methylation Society*, **3**, 188-92.

Mancini,D.N. et al. (1998) CpG methylation within the 5' regulatory region of the BRCA1 gene is tumor specific and includes a putative CREB binding site. *Oncogene*, **16**, 1161-9.

Manolio,T.A. et al. (2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747-53.

Margulies,M. et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376-80.

McGrath,J. and Solter,D. (1984) Completion of mouse embryogenesis requires both the maternal and paternal genomes. *Cell*, **37**, 179-83.

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

Meissner,A. et al. (2005) Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic acids research*, **33**, 5868-77.

Melki,J.R. et al. (1999) Concurrent DNA hypermethylation of multiple genes in acute myeloid leukemia. *Cancer research*, **59**, 3730-40.

Merlo,A. et al. (1995) 5' CpG island methylation is associated with transcriptional silencing of the tumour suppressor p16/CDKN2/MTS1 in human cancers. *Nature medicine*, **1**, 686-92.

Metzker,M.L. (2010) Sequencing technologies - the next generation. *Nature reviews. Genetics*, **11**, 31-46.

Morgan,H.D. et al. (1999) Epigenetic inheritance at the agouti locus in the mouse. *Nature genetics*, **23**, 314-8.

Nakamura,N. and Takenaga,K. (1998) Hypomethylation of the metastasis-associated S100A4 gene correlates with gene activation in human colon adenocarcinoma cell lines. *Clinical & experimental metastasis*, **16**, 471-9.

Newbold,R.R. et al. (2006) Adverse effects of the model environmental estrogen diethylstilbestrol are transmitted to subsequent generations. *Endocrinology*, **147**, S11-7.

Nicholls,R.D. and Knepper,J.L. (2001) Genome organization, function, and imprinting in Prader-Willi and Angelman syndromes. *Annual review of genomics and human genetics*, **2**, 153-75.

Okano,M. et al. (1998) Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases. *Nature genetics*, **19**, 219-20.

Olek,A. and Walter,J. (1997) The pre-implantation ontogeny of the H19 methylation imprint. *Nature genetics*, **17**, 275-6.

Oshimo,Y. et al. (2003) Promoter methylation of cyclin D2 gene in gastric carcinoma. *International journal of oncology*, **23**, 1663-70.

Ottinger,M.A. et al. (2005) Assessing the consequences of the pesticide methoxychlor: neuroendocrine and behavioral measures as indicators of biological impact of an estrogenic environmental chemical. *Brain research bulletin*, **65**, 199-209.

Paulsen,M. et al. (2000) Sequence conservation and variability of imprinting in the Beckwith-Wiedemann syndrome gene cluster in human and mouse. *Human molecular genetics*, **9**, 1829-41.

Pelizzola,M. and Ecker,J.R. (2010) The DNA methylome. *FEBS letters*, **585**, 1994-2000.

Pembrey,M.E. et al. (2006) Sex-specific, male-line transgenerational responses in humans. *European journal of human genetics : EJHG*, **14**, 159-66.

Petronis,A. (2010) Epigenetics as a unifying principle in the aetiology of complex traits and diseases. *Nature*, **465**, 721-7.

Popova,N.V. (1989) Transgenerational effect of orthoaminoasotoluol in mice. *Cancer letters*, **46**, 203-6.

Pulling,L.C. et al. (2004) Aberrant promoter hypermethylation of the death-associated protein kinase gene is early and frequent in murine lung tumors induced by cigarette smoke and tobacco carcinogens. *Cancer research*, **64**, 3844-8.

Pàldi,A. et al. (1995) Imprinted chromosomal regions of the human genome display sex-specific meiotic recombination frequencies. *Current biology : CB*, **5**, 1030-5.

Qiu,J. (2006) Epigenetics: unfinished symphony. *Nature*, **441**, 143-5.

Quinlan,A.R. et al. (2008) Pyrobayes: an improved base caller for SNP discovery in pyrosequences. *Nature methods*, **5**, 179-81.

Rakyan,V.K. et al. (2011) Epigenome-wide association studies for common human diseases. *Nature reviews. Genetics*, **12**, 529-41.

Rakyan,V.K. et al. (2002) Metastable epialleles in mammals. *Trends in genetics : TIG*, **18**, 348-51.

Reik,W and Walter,J. (2001) Genomic imprinting: parental influence on the genome. *Nature reviews. Genetics*, **2**, 21-32.

Reik,Wolf (2007) Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature*, **447**, 425-32.

Richardson,B. (2003) Impact of aging on DNA methylation. *Ageing research reviews*, **2**, 245-61.

Robertson,K.D. (2005) DNA methylation and human disease. *Nature reviews. Genetics*, **6**, 597-610.

Robinson,W.P. and Lalande,M. (1995) Sex-specific meiotic recombination in the Prader--Willi/Angelman syndrome imprinted region. *Human molecular genetics*, **4**, 801-6.

Rodenhiser,D. and Mann,M. (2006) Epigenetics and human disease: translating basic biology into clinical applications. *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne*, **174**, 341-8.

Ronaghi,M. et al. (1996) Real-time DNA sequencing using detection of pyrophosphate release. *Analytical biochemistry*, **242**, 84-9.

Rosenfeld,C.S. (2010) Animal models to study environmental epigenetics. *Biology of reproduction*, **82**, 473-88.

Sakai,T. et al. (1991) Allele-specific hypermethylation of the retinoblastoma tumor-suppressor gene. *American journal of human genetics*, **48**, 880-8.

Sandoval,J. et al. (2011) Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics : official journal of the DNA Methylation Society*, **6**, 692-702.

Sandovici,I. et al. (2011) Maternal diet and aging alter the epigenetic control of a promoter-enhancer interaction at the Hnf4a gene in rat pancreatic islets. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 5449-54.

Santi,D.V. et al. (1983) On the mechanism of inhibition of DNA-cytosine methyltransferases by cytosine analogs. *Cell*, **33**, 9-10.

Sasaki,H. et al. (1993) DNA methylation and genomic imprinting in mammals. *EXS*, **64**, 469-86.

Serre,D. et al. (2010) MBD-isolated Genome Sequencing provides a high-throughput and comprehensive survey of DNA methylation in the human genome. *Nucleic acids research*, **38**, 391-9.

Sharma,R.P. (2005) Schizophrenia, epigenetics and ligand-activated nuclear receptors: a framework for chromatin therapeutics. *Schizophrenia research*, **72**, 79-90.

Shemer,R. et al. (1997) Structure of the imprinted mouse Snrpn gene and establishment of its parental-specific methylation pattern. *Proceedings of the National Academy of Sciences of the United States of America*, **94**, 10267-72.

Shendure,J. et al. (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. *Science (New York, N.Y.)*, **309**, 1728-32.

Shoemaker,R. et al. (2010) Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. *Genome research*, **20**, 883-9.

Sinclair,K.D. et al. (2007) DNA methylation, insulin resistance, and blood pressure in offspring determined by maternal periconceptional B vitamin and methionine status. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 19351-6.

Singal,R. and Ginder,G.D. (1999) DNA methylation. *Blood*, **93**, 4059-70.

Skinner,M.K. (2011) Environmental epigenetic transgenerational inheritance and somatic epigenetic mitotic stability. *Epigenetics : official journal of the DNA Methylation Society*, **6**, 838-42.

De Smet,C. et al. (1999) DNA methylation is the primary silencing mechanism for a set of germ line- and tumor-specific genes with a CpG-rich promoter. *Molecular and cellular biology*, **19**, 7327-35.

Song,C.-X. et al. (2012) Sensitive and specific single-molecule sequencing of 5-hydroxymethylcytosine. *Nature methods*, **9**, 75-7.

Squires,J.E. et al. (2012) Widespread occurrence of 5-methylcytosine in human coding and non-coding RNA. *Nucleic acids research*.

Stein,R. et al. (1982) Clonal inheritance of the pattern of DNA methylation in mouse cells. *Proceedings of the National Academy of Sciences of the United States of America*, **79**, 61-5.

Stöger,R. et al. (1993) Maternal-specific methylation of the imprinted mouse Igf2r locus identifies the expressed locus as carrying the imprinting signal. *Cell*, **73**, 61-71.

Surani,M.A. et al. Development of reconstituted mouse eggs suggests imprinting of the genome during gametogenesis. *Nature*, **308**, 548-50.

Tahiliani,M. et al. (2009) Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science (New York, N.Y.)*, **324**, 930-5.

Ting,A.H. et al. (2006) Differential requirement for DNA methyltransferase 1 in maintaining human cancer cell gene promoter hypermethylation. *Cancer research*, **66**, 729-35.

Tremblay,K.D. et al. (1997) A 5' 2-kilobase-pair region of the imprinted mouse H19 gene exhibits exclusive paternal methylation throughout development. *Molecular and cellular biology*, **17**, 4322-9.

Turusov,V.S. et al. (1990) Increased multiplicity of lung adenomas in five generations of mice treated with benz(a)pyrene when pregnant. *Cancer letters*, **55**, 227-31.

Umemura,S. et al. (2008) Aberrant promoter hypermethylation in serum DNA from patients with silicosis. *Carcinogenesis*, **29**, 1845-9.

Venter,J.C. et al. (2001) The sequence of the human genome. *Science (New York, N.Y.)*, **291**, 1304-51.

Waddington,C.H. (1939) Preliminary Notes on the Development of the Wings in Normal and Mutant Strains of Drosophila. *Proceedings of the National Academy of Sciences of the United States of America*, **25**, 299-307.

Waterland,R.A. and Jirtle,R.L. (2003) Transposable elements: targets for early nutritional effects on epigenetic gene regulation. *Molecular and cellular biology*, **23**, 5293-300.

Weksberg,R. et al. (2003) Beckwith-Wiedemann syndrome demonstrates a role for epigenetic control of normal development. *Human molecular genetics*, **12 Spec No**, R61-8.

Wong,C.C.Y. et al. (2010) A longitudinal study of epigenetic variation in twins. *Epigenetics : official journal of the DNA Methylation Society*, **5**, 516-26.

Wong,N. et al. (2001) Hypomethylation of chromosome 1 heterochromatin DNA correlates with q-arm copy gain in human hepatocellular carcinoma. *The American journal of pathology*, **159**, 465-71.

Yauk,C. et al. (2008) Germ-line mutations, DNA damage, and global hypermethylation in mice exposed to particulate air pollution in an urban/industrial location. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 605-10.

Yoder,J.A. et al. (1997) Cytosine methylation and the ecology of intragenomic parasites. *Trends in genetics : TIG*, **13**, 335-40.

References

Zambrano,E. et al. (2005) Sex differences in transgenerational alterations of growth and metabolism in progeny (F2) of female offspring (F1) of rats fed a low protein diet during pregnancy and lactation. *The Journal of physiology*, **566**, 225-36.

Zhang,D. et al. (2010) Genetic control of individual differences in gene-specific methylation in human brain. *American journal of human genetics*, **86**, 411-9.

References

Appendices

Appendix A

"B-SOLANA: An approach for the analysis of two-base encoding bisulfite sequencing data"

Kreck B, Marnellos G, Richter J, Krueger F, Siebert R, Franke A

Bioinformatics (2011), 2012 Feb 1;28(3):428-9. Epub 2011 Dec 6

BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 3 2012, pages 428–429
doi:10.1093/bioinformatics/btr660

Sequence analysis

Advance Access publication December 6, 2011

B-SOLANA: an approach for the analysis of two-base encoding bisulfite sequencing data

Benjamin Kreck[1,*], George Marnellos[2], Julia Richter[3], Felix Krueger[4], Reiner Siebert[3] and Andre Franke[1]

[1]Institute of Clinical Molecular Biology, Christian-Albrechts-University, Kiel, Germany, [2]Life Technologies, Advanced Sequencing Applications, Carlsbad, CA 92008, USA, [3]Institute of Human Genetics, Christian-Albrechts-University, Kiel, Germany and [4]Bioinformatics Group, The Babraham Institute, CB22 3AT Cambridge, UK

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: Bisulfite sequencing, a combination of bisulfite treatment and high-throughput sequencing, has proved to be a valuable method for measuring DNA methylation at single base resolution. Here, we present B-SOLANA, an approach for the analysis of two-base encoding (colorspace) bisulfite sequencing data on the SOLiD platform of Life Technologies. It includes the alignment of bisulfite sequences and the determination of methylation levels in CpG as well as non-CpG sequence contexts. B-SOLANA enables a fast and accurate analysis of large raw sequence datasets.

Availability and implementation: The source code, released under the GNU GPLv3 licence, is freely available at http://code.google.com/p/bsolana/.

Contact: b.kreck@ikmb.uni-kiel.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on September 1, 2011; revised on November 17, 2011; accepted on November 24, 2011

1 INTRODUCTION

Methylation at position 5 of cytosines is a major epigenetic modification, which has an important impact on transcriptional and regulatory processes of DNA (Holliday, 1975). It is a stable modification of the genome which can be inherited from one generation to the next, even though it can also be dynamically changed by environmental influences. There are several methods based on high-throughput sequencing, such as methylated DNA immunoprecipitation sequencing (MeDIP-seq), methylated DNA capture by affinity purification (MethylCap-seq) and BS-Seq, which can provide good-quality genome-wide DNA methylation data (Bock, 2010).

Methods that currently provide genome-wide methylation patterns at single base resolution make use of bisulfite conversion and high-throughput sequencing. The treatment of DNA with sodium bisulfite has no effect on methylated cytosines, but it specifically converts unmethylated cytosines to uracils, which are converted to thymines during subsequent polymerase chain reaction amplification. As a result of bisulfite conversion, the Watson and Crick strands of bisulfite-treated DNA are no longer complementary to each other, they become essentially different genomes. This fact leads to an enlarged alignment reference space. The prevalence of T's that have replaced C's leads to reduced complexity in bisulfite sequences, which increases the bioinformatics challenge of BS-Seq analysis. Bioinformatics tools for BS-Seq have generally fallen into two categories: (i) methylation-aware alignment tools, which consider cytosines and thymines as potential matches to genomic cytosine positions and (ii) tools which convert any residual cytosines in bisulfite sequences and all cytosines of the reference genomes into thymines.

*To whom correspondence should be addressed.

2 COLORSPACE BISULFITE SEQUENCING

Due to the two-base encoding of SOLiD sequencing, *in silico* conversions of any residual bisulfite read cytosines into thymines, which can be carried out in basespace data to avoid bisulfite-mismatches during alignment, cannot be performed on bisulfite colorspace sequences, because sequencing errors would lead to the incorrect translation of colorspace to basespace (Supplementary Fig. 1). There are ways to align bisulfite colorspace sequences with methylation-aware alignment approaches, which convert bisulfite colorspace sequences to basespace and index all theoretically possible alignments by creating a hash table. Such an approach is implemented in SOCS-B, which is based on the iterative version of the Rabin–Karp algorithm (Ondov, 2010). Even though SOCS-B turns out to be an accurate tool for the analysis of colorspace BS-Seq datasets, it becomes very computationally intensive for complex genomes such as the human genome (~150 000 CPU hours for the analysis of 500 Million sequences). Therefore, it is not efficient for huge datasets like those produced in genome-wide methylation analyses with average coverage depths ≥10X and genome size ≥1000 MB.

Here, we present B-SOLANA, a tool which performs sequence alignment and methylation calling for colorspace bisulfite sequencing. It is based on the established short-read aligner Bowtie (Langmead, 2009) and SAMtools utilities for manipulating alignments (Li, 2009). B-SOLANA is divided into four individual steps: (i) indexing, (ii) mapping, (iii) determination of best alignment and (iv) methylation calling.

The idea of B-SOLANA is to use Bowtie to uniquely align bisulfite sequences to two different conversions of the reference genome and determine best alignments from the combined set of results. The analysis of whole methylomes of 23 eukaryotic organisms shows a variable percentage of methylation at CpG

Table 1. The 485 990 920 SOLiD BS-Seq reads (50 bp), taken from SRR204024 (Hansen, 2011), were analyzed with B-SOLANA and MethylCoder (one mismatch allowed) B-SOLANA exhibits a high correlation with the results of Hansen *et al*.

	Hansen *et al.*[a]	B-SOLANA	MethylCoder[b]
Uniquely mapped reads (%)	37.83	49.84	19.23
CpG positions: % C	69.84	72.83	67.07
CpG positions: % T	30.03	26.97	32.93
Non-CpG positions: % C	0.20	0.22	0.69
Non-CpG positions: % T	99.76	99.70	99.31

[a]Including post-processing quality control.
[b]Only cytosine and thymine base calls are included.

dinucleotides, whereas the percentage of methylated CHG and CHH is always lower (Pelizzola, 2010). The approach of B-SOLANA reduces the number of bisulfite-induced mismatches by considering the prevalence of methylated cytosines in their different sequence contexts.

In order to identify CpG and non-CpG methylation sites, B-SOLANA aligns bisulfite sequences to two *in-silico* conversions of the reference genome (Supplementary Fig. 2). In the first modified reference genome, all cytosines in a non-CpG context are converted to thymines (Conversion I). In the second, all cytosines, irrespective of their sequence context, are converted to thymines (Conversion II). After alignment to these converted genomes, B-SOLANA determines the best alignment for each bisulfite sequence in the following way: bisulfite sequences that are aligned to different genomic positions in Conversions I and II are assigned to the position with the lowest number of mismatches. Reads with the same number of mismatches at different positions are ignored. In its final step, B-SOLANA determines methylation levels. B-SOLANA is compatible with 50 bp directional single-end libraries and allows a simple adjustment for the upcoming read lengths.

B-SOLANA was designed to generate accurate results for methylomes with a low percentage of methylation in non-CpG sites (<5%). This includes most eukaryotic organisms, with mammalian genomes typically having methylation levels of <3% in CHG and <1% in CHH context (Pelizzola, 2010).

A detailed test of B-SOLANA was performed for genomic DNA extracted from a human lymphoma cell line. The library was prepared using a protocol for single-end SOLiD BS-Seq (Bormann Chung, 2010) and sequencing of the bisulfite-converted DNA was performed using SOLiD versions 3 Plus and 4. This generated 2.79 billion bisulfite reads of which ∼52% were mapped uniquely (genome build hg19/NCBI 37). The methylation results obtained by B-SOLANA were then compared to the Illumina Infinium HumanMethylation450 BeadChip (450 k) assay, an established methylation analysis method, as a quality control (Supplementary Fig. 3). We observed high concordance between the results of the two independent methods (99% of 450 k sites were also represented in the B-SOLANA results) and the methylation levels of cytosines, which were assayed by both methods displayed a very high correlation (Pearson correlation $r > 0.93$). As a proof of principle, we also generated different simulated datasets, reflecting varying CpG and non-CpG methylation levels. Encouragingly, the results generated by B-SOLANA closely match the expected methylation degrees (Supplementary Table 1).

A further approach for the analysis of colorspace BS-Seq was published with the tool MethylCoder (Pedersen, 2011). MethylCoder applies a conversion of any residual bisulfite read cytosines into thymines, which leads to erroneous alignments, as discussed above. Therefore, we compared B-SOLANA and MethylCoder (one mismatch allowed) by analyzing 485 990 920 SOLiD BS-Seq reads (50 bp), taken from SRR204024 (Hansen, 2011). We found a high concordance between methylation calls of Hansen *et al.*, analyzed by their yet unpublished and unavailable approach, and B-SOLANA. Moreover, B-SOLANA turns out to have a significantly higher mapping efficiency.

As a platform-independent benchmark, we demonstrate that the analysis of colorspace BS-Seq data of the fibroblast cell line IMR90 is comparable to methylome data published by Lister *et al.* (2009), who used a BS-Seq approach on the Illumina platform (Supplementary Information 1).

3 CONCLUSIONS

We present an efficient tool for the analysis of large colorspace BS-Seq data. B-SOLANA provides a fast and accurate all-in-one approach, including alignment and methylation calling. It is easy to use and generates an intuitive output, which can be used for genome-wide DNA methylation analysis.

ACKNOWLEDGEMENTS

We thank Ole Ammerpohl for helpful discussions (Institute of Human Genetics, Kiel), Gavin Meredith (Life Technologies, Foster City CA, USA) for providing SOLiD BS-Seq data of IMR90 cells and the sequencing facility at IKMB for helpful discussions and support.

Funding: Start-up grant from the Medizinausschuss Schleswig Holstein, the National Genome Research Network (NGFN) of Germany (BMBF-funded); DFG cluster of excellence 'Inflammation at Interfaces' (infrastructure support); BMBF in the framework of the ICGC MMML-Seq project (to R.S. and J.R.).

Conflict of Interest: none declared.

REFERENCES

Bock,C. *et al.* (2010) Quantitative comparison of genome-wide DNA methylation mapping technologies. *Nat. Biotechnol.*, **28**, 1106–1114.

Bormann Chung,C.A. *et al.* (2010) Whole methylome analysis by ultra-deep sequencing using two-base encoding. *PLoS One*, **5**, e9320.

Hansen,K. *et al.* (2011) Increased methylation variation in epigenetic domains across cancer types. *Nat. Genet.*, **43**, 768–775.

Holliday,R. *et al.* (1975) DNA modification mechanisms and gene activity during development. *Science*, **187**, 226–232.

Langmead,B. *et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

Lister,R. *et al.* (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, **462**, 315–322.

Li,H. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Ondov,B.D. *et al.* (2010) An alignment algorithm for bisulfite sequencing using the Applied Biosystems SOLiD System. *Bioinformatics*, **26**, 1901–1902.

Pedersen,B. *et al.* (2011) MethylCoder: Software Pipeline for Bisulfite-Treated Sequences. *Bioinformatics*, **27**, 2435–2436.

Pelizzola,M. *et al.* (2010) The DNA methylome. *FEBS Lett.*, **585**, 1994–2000.

Appendix A

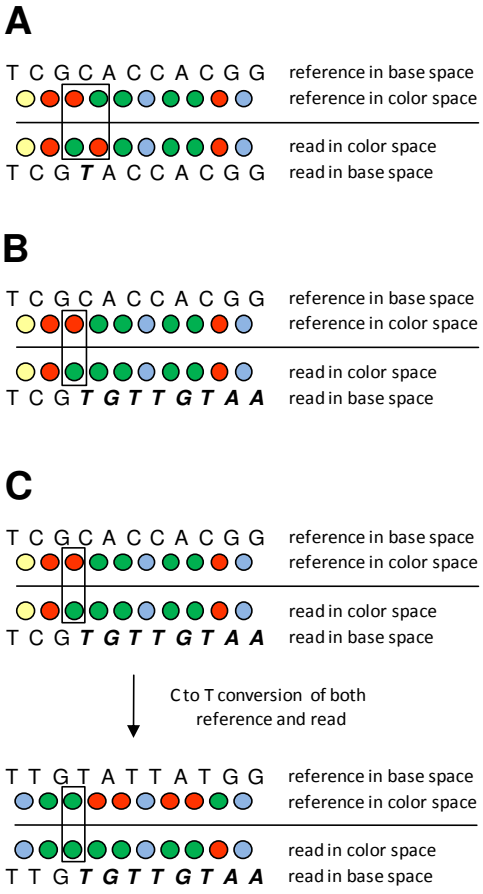
Supplemental Information

"B-SOLANA: An approach for the analysis of two-base encoding bisulfite sequencing data"

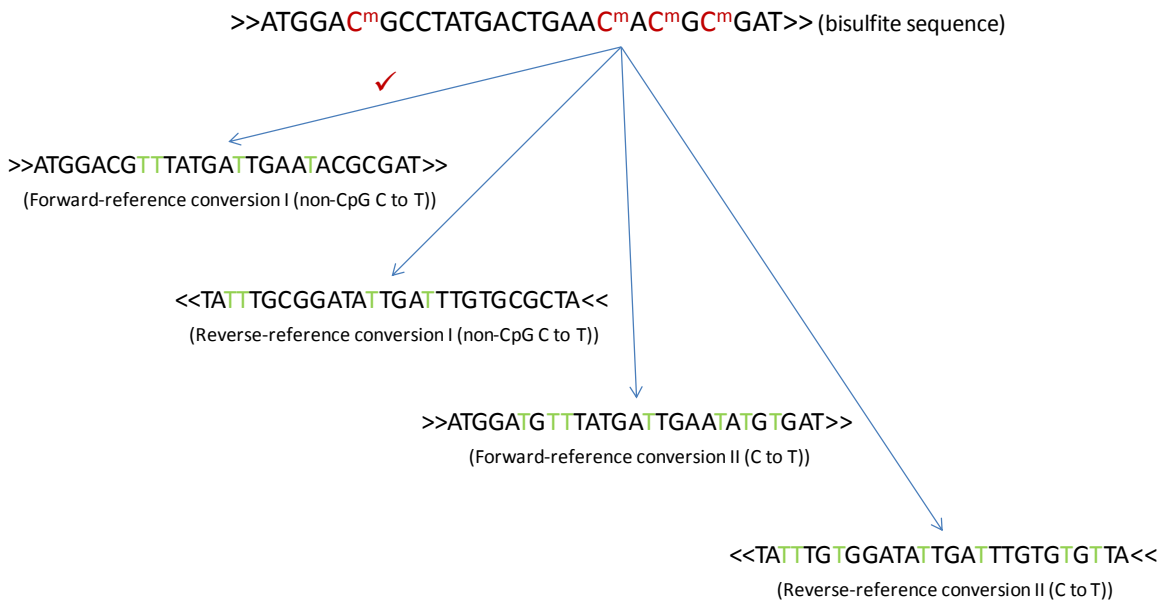
Kreck B, Marnellos G, Richter J, Krueger F, Siebert R, Franke A

Bioinformatics (2011), 2012 Feb 1;28(3):428-9. Epub 2011 Dec 6

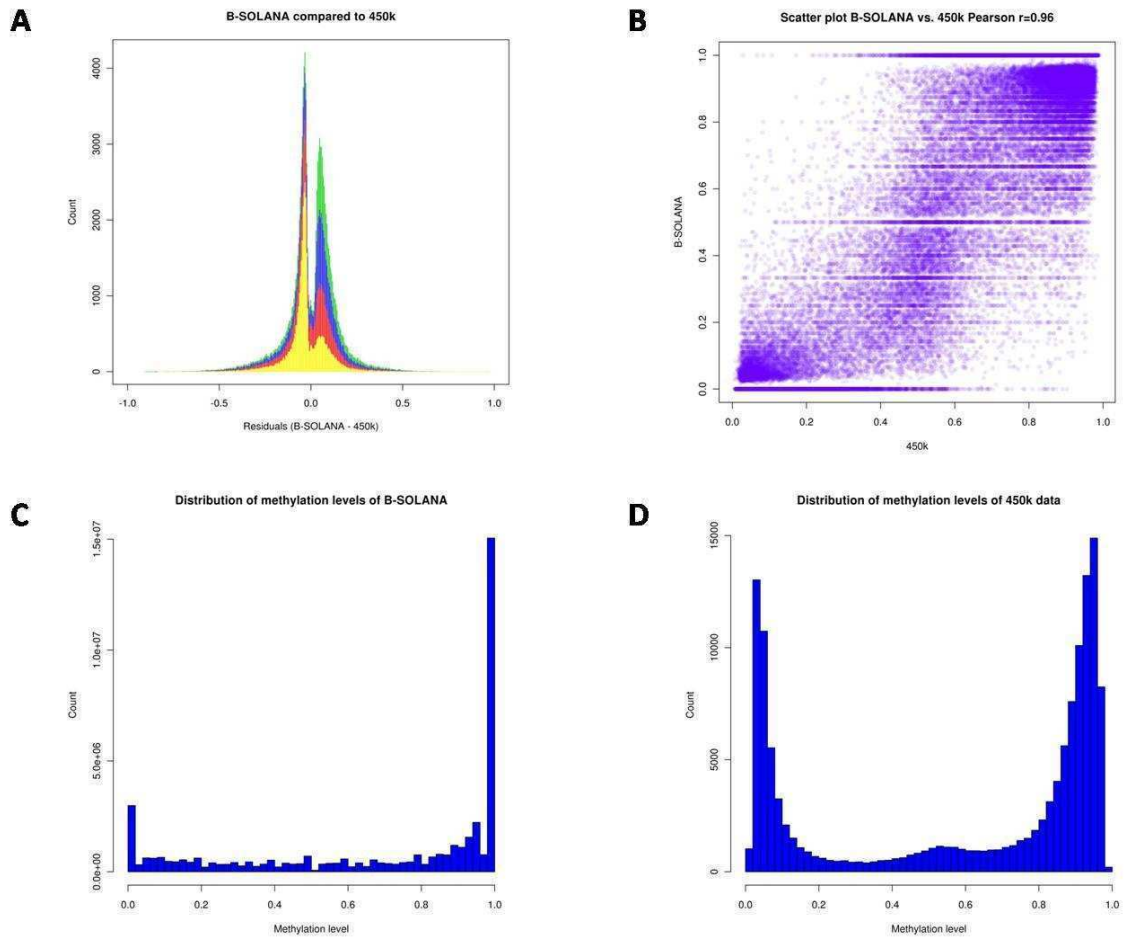
Supplementary Information



Supplementary Figure 1 (A) A SNP position in colorspace appears as two adjacent color transitions. **(B)** A measurement error in colorspace appears as a single color transition. **(C)** In silico conversion of C to T in bisulfite reads carrying a measurement error abrogates mapping to an equally converted reference sequence. This figure was adapted from Krueger F, Kreck B, et al., DNA Methylome Analysis Using Short Bisulfite Reads, in press.



Supplementary Figure 2 B-SOLANA performs four different mappings. Each bisulfite-treated sequence will be mapped to a modified reference genome in which all cytosines in non-CpG context are in silico converted to thymines and to a reference genome in which all cytosines, irrespective of their sequence context, are converted to thymines. B-SOLANA determines the best mapping in terms of lowest number of mismatches. Consequently, different methylation levels can be assessed.



Supplementary Figure 3 (A) Comparison of SOLiD BS-Seq to a single Illumina Infinium HumanMethylation450 BeadChip restricted to methylation calls of Typ I data. The distributions depict the residuals as the differences between SOLiD BS-Seq methylation levels and Infinium beta values, where values close to 0 mean that equal methylation levels were inferred by both methods. The green, blue, red and orange graphs represent the distribution of residuals of cytosines covered by at least 5, 10, 15 and 20 bisulfite sequences, respectively. Higher read coverage generally results in narrower residual profiles, i.e. better correlation. The bimodal distribution of the residuals has a technology specific explanation. Unmethylated and fully methylated sites assessed by the 450k assay are usually not detected with beta values of 0 and 1 but rather assigned to beta values next to ~0.05 and ~0.94 In contrast, in BS-Seq unmethylated and fully methylated sites correspond to methylation levels of 0 and 1. **(C,D)** This fact and the distribution of methylation levels **(C,D)** explain that the maxima in the distribution of residuals are located at ~-0.05 (0-0.05) and ~0.06 (1-0.94). Thus, the correlation of the 450k assay and BS-Seq is distributed sigmoidally. **(B)** Scatter plot of methylation sites, which were assessed by 450k assay and BS-Seq (coverage $\geq$ 5), shows a high correlation (Pearson correlation r=0.96). Bands at 0, 0.5, 1 (y-axis) correspond to the sigmoidal shift of the two assays at homozygous unmethylated, heterozygous methylated and homozygous methylated sites. **(C,D)** Different distribution profiles of methylation levels in the 450k assay and BS-Seq, explaining the sigmoidal form of **(A)**.

Simulated data	1% CH methylation	3% CH methylation	5% CH methylation
10% CG methylation	10.36 CG / 0.95 CH	9.79 CG / 2.62 CH	10.02 CG / 4.05 CH
20% CG methylation	20.35 CG / 0.96 CH	20.00 CG / 2.60 CH	19.82 CG / 4.07 CH
30% CG methylation	30.09 CG / 0.94 CH	29.83 CG / 2.59 CH	29.72 CG / 4.04 CH
40% CG methylation	40.70 CG / 0.96 CH	40.68 CG / 2.61 CH	40.47 CG / 4.03 CH
50% CG methylation	51.04 CG / 0.95 CH	50.77 CG / 2.66 CH	50.79 CG / 4.02 CH
60% CG methylation	60.79 CG / 0.94 CH	61.00 CG / 2.62 CH	61.27 CG / 4.03 CH
70% CG methylation	70.70 CG / 0.94 CH	70.98 CG / 2.61 CH	71.58 CG / 4.05 CH
80% CG methylation	80.55 CG / 0.95 CH	80.36 CG / 2.59 CH	81.17 CG / 4.09 CH
90% CG methylation	89.97 CG / 0.96 CH	89.92 CG / 2.63 CH	90.03 CG / 4.08 CH

Supplementary Table 1 The accuracy of B-SOLANA was tested using simulation data from Sherman (bisulfite-treated Read FastQ Simulator (http://www.bioinformatics.bbsrc.ac.uk/projects/sherman/)). One hundred thousand reads (genome build HG19/NCBI 37) with different rates of bisulfite conversion (10% ≤ CG methylation ≤ 90 %, 1% ≤ CH methylation ≤ 5 % - any combinations) were analyzed as indicated. B-SOLANA is able to accurately detect various levels of simulated methylation when methylation in non-CpG context is fairly low (<5%). The latter is the case for most eukaryotic genomes, with mammalian genomes typically having methylation levels of less than 3% in CHG and less than 1% in CHH context (Pelizzola, 2010).

Pelizzola, M., et al. (2010) The DNA methylome, FEBS Lett., **585**, 1994–2000

	Hansen et al.*	B-SOLANA	MethylCoder**
Uniquely mapped reads (in %) (adenoma I)	32.13	42.60	10.58
CpG positions: % C (adenoma I)	66.29	70.12	65.68
CpG positions: % T (adenoma I)	33.59	29.70	34.32
Non-CpG positions: % C (adenoma I)	0.24	0.25	1.31
Non-CpG positions: % T (adenoma I)	99.73	99.67	98.69
Uniquely mapped reads (in %) (cancer I)	37.48	49.66	18.47
CpG positions: % C (cancer I)	62.79	67.63	61.73
CpG positions: % T (cancer I)	37.08	32.17	38.27
Non-CpG positions: % C (cancer I)	0.23	0.24	0.69
Non-CpG positions: % T (cancer I)	99.74	99.68	99.17

* including post-processing quality control
** only cytosine and thymine base calls are included

Supplementary Table 2 Comparison of B-SOLANA and MethylCoder by analyzing SOLiD BS-Seq reads (50 bp), taken from (SRR204022 (adenoma I) and SRR204026 (cancer I)). The methylation calls of B-SOLANA show a high concordance with the results reported by Hansen et al., analyzed with their unpublished and unavailable approach. The mapping efficiencies of B-SOLANA and MethylCoder are substantially higher and dramatically lower, respectively, than reported by Hansen et al..

Supplementary Information 1 We include data of a SOLiD bisulfite sequencing run of IMR90, the same human diploid fibroblast strain which was also analyzed by Lister et al. (Lister 2009). Sequencing was performed using SOLiD version 3 which generated 515,699,253 bisulfite reads. The results were compared to those generated by Lister et al., who worked with BS-Seq data generated on the Illumina platform. For this purpose we used their publically available data at http://neomorph.salk.edu/human_methylome/ and used LiftOver (http://genome.ucsc.edu/cgi-bin/hgLiftOver) to convert the data from HG18/NCBI 36 to HG19/NCBI 37. We performed correlation analyses for two biological replicates of IMR90 generated by Lister et al. and for colorspace bisulfite sequencing data of IMR90 compared to the combined set of results of Lister et al.. Both correlation analyses use CG methylation calls which are covered by at least five sequences and were present in both samples (either the two replicates of the Lister et al. data or colorspace vs. Lister et al. data). Calculations of the correlations yield the following results:

	Pearson r
Replicates sequenced on the Illumina platform	0.80
Colorspace BS-Seq vs. BS-Seq on the Illumina platform	0.74

These similar results let us conclude that bisulfite sequencing data generated on the SOLiD platform and the Illumina platform are comparable, despite the fact that the samples were derived from separate biological preparations of the IMR90 cell line.

Lister, R., et al. (2009) Human DNA methylomes at base resolution show widespread epigenomic differences, Nature, **462**, 315-322.

Appendix B

"DNA methylome analysis using short bisulfite sequencing data."

Krueger F[*], **Kreck B**[*], Franke A, Andrews SR

Nature Methods (2012), Jan 30;9(2):145-51

[*]Authors that contributed equally to the manuscript.

DNA methylome analysis using short bisulfite sequencing data

Felix Krueger[1,3], Benjamin Kreck[2,3], Andre Franke[2] & Simon R Andrews[1]

Bisulfite conversion of genomic DNA combined with next-generation sequencing (BS-seq) is widely used to measure the methylation state of a whole genome, the methylome, at single-base resolution. However, analysis of BS-seq data still poses a considerable challenge. Here we summarize the challenges of BS-seq mapping as they apply to both base and color-space data. We also explore the effect of sequencing errors and contaminants on inferred methylation levels and recommend the most appropriate way to analyze this type of data.

DNA methylation involves the addition of a methyl group to the C5 carbon residue (5mC) of cytosines by DNA methyltransferases[1,2]. DNA methylation is an important epigenetic mechanism used by higher eukaryotes and is involved in several key physiological processes, including regulation of gene expression, X-chromosome inactivation, imprinting, and silencing of germline-specific genes and repetitive elements[3]. Patterns of methylation are stably maintained through somatic cell division and can be inherited across generations. These patterns are sometimes perturbed in important human diseases, such as imprinting disorders and cancer[3–5]. Understanding how methylation patterns are established and maintained is therefore of great importance.

The sequence context in which a cytosine occurs is a key factor in determining the regulation of its methylation. Cytosines that occur as part of a C-G dinucleotide (CpG) are often highly methylated (~60–80% in mammals[2]) and are regulated differently to cytosines in other contexts. CpG methylation usually occurs on both DNA strands[1] to maintain methylation at CpGs during DNA replication. In contrast, non-CpG methylation must be re-established *de novo* after each cell division. Although it is present at considerable levels during early development or in pluripotent cell types[6–8], most non-CpG cytosines are generally unmethylated in differentiated tissues (~0.3–3% in mammals[2]).

As methylated cytosines are susceptible to spontaneous conversion into thymines through chemical deamination, they tend to be generally underrepresented in the genome[9,10], and they are often grouped in dense patches termed CpG islands. These islands tend to be unmethylated in the germline and are consequently less vulnerable to spontaneous deamination[11]. CpG islands are frequently associated with promoters, and the regulation of promoter methylation has been shown to affect the expression of the corresponding transcripts. Traditionally, CpG islands were defined using sequence-composition analysis[12–14] which predicted that the mouse genome contained a substantially lower number of CpG islands than the human genome. However, a recent report demonstrated that the occurrence of functional CpG islands is in fact quite similar in the two organisms[15].

Measuring methylation

Several methods exist for measuring DNA methylation at specific genomic loci, and these have been reviewed recently[16,17]. They range from methylated DNA immunoprecipitation or methyl binding protein enrichment of methylated fragments[18–20] to digestion with methylation-sensitive restriction enzymes[21] and bisulfite modification of DNA[22]. Comparisons of these methods showed that they all can be used to produce accurate DNA methylation data[21,23,24].

During bisulfite sequencing the treatment of DNA with sodium bisulfite converts cytosines into uracils, whereas methylcytosines remain unmodified. Uracils are read as thymines by DNA polymerase, so

[1]Bioinformatics Group, The Babraham Institute, Cambridge, UK. [2]Institute of Clinical Molecular Biology, Christian Albrechts University, Kiel, Germany. [3]These authors contributed equally to this work. Correspondence should be addressed to A.F. (a.franke@mucosa.de) or S.R.A. (simon.andrews@babraham.ac.uk).

PUBLISHED ONLINE 30 JANUARY 2012; DOI:10.1038/NMETH.1828

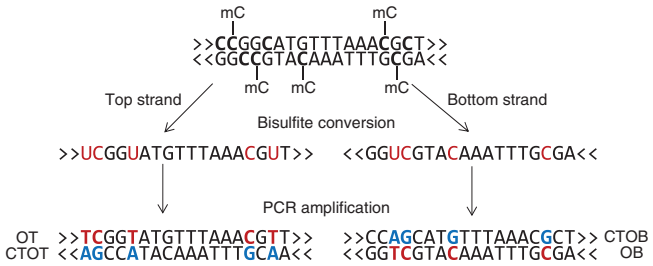


Figure 1 | Effect of bisulfite treatment of DNA. Bisulfite conversion of genomic DNA and subsequent PCR amplification gives rise to two PCR products and up to four potentially different DNA fragments for any given locus. (Hydroxy)methylated cytosine residues are resistant to bisulfite conversion and can be used as a readout of the DNA methylation state. mC, 5-methylcytosine; hmC, 5-hydroxymethylcytosine; OT, original top strand; CTOT, strand complementary to the original top strand; OB, original bottom strand; and CTOB, strand complementary to the original bottom strand.

amplifying bisulfite-treated DNA by PCR yields products in which unmethylated cytosines appear as thymines. By comparing the modified DNA with the original sequence, the methylation state of the original DNA can therefore be inferred. Bisulfite treatment of 5-hydroxymethylcytosine (5hmC) yields a similar intermediate to 5mC, meaning that BS-seq can be used to detect whether a position is (hydroxy-) methylated but not to determine the exact type of modification[21,25] (**Fig. 1**). This limitation does not apply to antibody-based techniques, which can be used to specifically enrich 5hmC[26–28].

Capillary electrophoresis–based bisulfite sequencing was considered the gold standard for methylation analysis because of its clear readout and single-base resolution[22], but it could only be applied to relatively small regions. New sequencing technologies mean that BS-seq is now a viable option for the sequencing of entire mammalian methylomes[6–8,29–32] (**Supplementary Table 1**).

For researchers primarily interested in CpG island methylation, the cost of bisulfite sequencing can be reduced by enriching CpG-dense regions by digesting genomic DNA with a methylation-insensitive restriction enzyme containing a C-G as part of its recognition site and selecting short fragments[6,30,33]. Even though the selected fragments are used to interrogate only a few percent of the genome, these data are informative for the majority of CpG islands. This approach, termed reduced representation BS-seq (RRBS), has been extensively described and compared to other techniques[23,33–35], and several genome-wide methylation maps based on RRBS have been reported[6,30].

In this Review we provide an overview of the computational analysis of bisulfite sequencing data. We highlight points to consider when designing a BS-seq experiment and point out pitfalls that can occur during the initial analysis. We also discuss different alignment strategies and their implementation by current bioinformatic tools. In particular, we present the main differences between the analysis of base space (Illumina) and color space (SOLiD, Applied Biosystems) BS-seq data.

Challenges of BS-seq data mapping

As the methylation state of bisulfite-treated DNA must be inferred by comparison to an unmodified reference sequence, a correct alignment is of critical importance. This is challenging because the aligned sequences do not exactly match the reference, and the complexity of the libraries is reduced. Also, as cytosine methylation is not symmetrical, the two strands of DNA in the reference genome must be considered separately. A single site can have a different methylation state in different cells. Thus, when sequencing cell mixtures or tissue fractions, the percentage of methylation at each site needs to be determined[36].

When performing an alignment one must discriminate between different types of bisulfite-treated DNA libraries (for a schematic drawing, see ref. 16). In the first, termed directional libraries, adapters are attached to the DNA fragments such that only the original top or bottom strands will be sequenced[7,30]. Alternatively, all four DNA strands that arise through bisulfite treatment and subsequent PCR amplification can be sequenced with the same frequency in nondirectional libraries[32,37,38]. BS-seq mapping may therefore require up to four different strand alignments to be analyzed for each sequence. Because of the complexity of BS-seq alignments, standard sequence alignment software cannot be used. However, several different tools for BS-seq analysis have been developed.

Base-space BS-seq data alignments

Methylation-'aware' alignment tools consider both cytosine and thymine as potential matches to a genomic cytosine. This strategy provides the highest possible mapping efficiency (high sensitivity) because it makes optimal use of the information present in the reads. However, a drawback of this technique is that methylated sequences will be aligned with greater efficiency because they carry more information than their unmethylated counterparts, leading this type of aligner to overestimate methylation levels.

Alternatively, in unbiased approaches usually any residual cytosines in the BS-seq read and all cytosines in the reference genome are converted into thymines before the alignment is performed[7,30]. This means that the read sequence to be aligned is unaffected by its methylation state. It also means that there will be an exact match between the converted read and converted genome sequence so that standard sequence alignment tools can be used to perform the mapping[39,40]. This approach, however, comes at the cost of slightly reduced mapping efficiencies (**Fig. 2a**).

BS-seq in color space

In contrast to the intuitive base-space sequence generated by Illumina sequencers, SOLiD sequencing (Applied Biosystems) encodes its reads in color space such that each color resembles the transition from one base to the next[41]. Single-nucleotide polymorphisms (SNPs) can be called with high confidence because they will result in two adjacent color changes, whereas technical errors are indicated by a single color change (**Supplementary Fig. 1a,b**). Owing to the way color-space encoding works, residual cytosines are correctly converted into thymines in the bisulfite reads *in silico* before the mapping only if the reads are completely error-free. A single measurement error in the read would lead to incorrect conversions throughout the rest of the read (**Supplementary Fig. 1c**). As a consequence, the *in silico* cytosine to thymine conversion, which guarantees unbiased alignments, should not be performed on color-space datasets.

Current tools to align color space BS-seq data to a reference genome either use methylation-aware alignments (SOCS-B[42]), which can be computationally intensive for complex genomes,

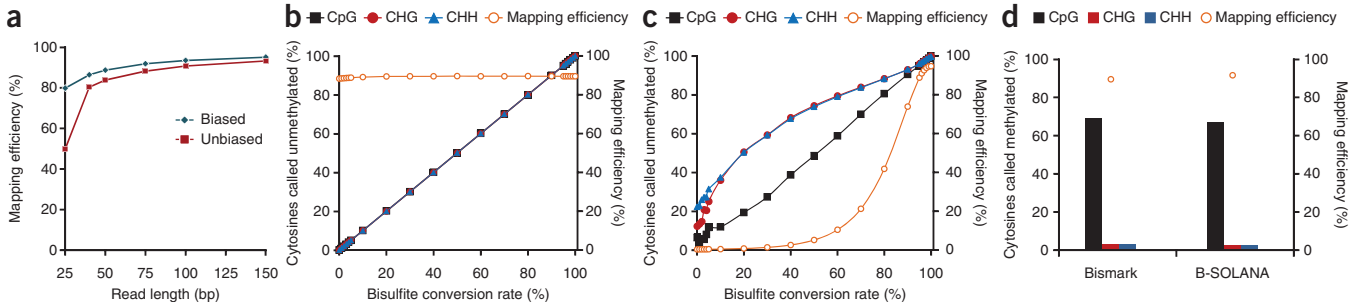


Figure 2 | Performance and accuracy of unbiased base-space and color-space BS-seq alignment tools. (**a**) A total of 10^6 random mouse genomic sequences of different lengths were aligned to the mouse genome (NCBIM37) with Bowtie as an example of methylation-aware mapping (biased) or with Bismark as an example of unbiased mapping (unbiased). Non-unique alignments were discarded. (**b**,**c**) A total of 10^6 random mouse base-space (Bismark; **b**) or human color-space (B-SOLANA; **c**) reads (75 base pairs) were simulated with different rates of bisulfite conversion (context is indicated) and aligned to the mouse (NCBIM37) or human (NCBI37) genomes. Bismark accurately detected various simulated methylation levels at a constant mapping efficiency. Alignment of color-space reads with B-SOLANA was efficient, and methylation calls were accurate only when methylation in non-CpG context was fairly low (ideally less than 5%). H (in CHG and CHH) stands for C, T or A. (**d**) Reads as in **b**,**c** were simulated with typical mammalian methylation levels (CpG context, 70%; CHG and CHH context, 3%) using Sherman (http://www.bioinformatics.bbsrc.ac.uk/projects/sherman/).

or align reads to *in silico*–converted versions of the reference genome, with bisulfite-induced mismatches are treated as normal mismatches. As different levels of methylation can result in increased numbers of mismatches to the reference genome, this approach is, however, not free of bias. One possibility to reduce mapping bias is to apply different *in silico* conversions to the reference genome and determine best alignments from the combined set of results of different mapping runs. This approach, however, requires prior knowledge of the methylation characteristics of the organism to be analyzed.

BS-seq data alignment tools

Several tools have been developed for the analysis of BS-seq datasets[17]. These not only differ considerably regarding their alignment speed, flexibility and ease of use but also in the information they report. Many older BS-seq data aligners only reported a bisulfite read mapping output, and the user had to extract methylation information from the alignments. More recent tools provide a comprehensive methylation output, which enables the end user to explore the biological effects of methylation more quickly[39,40,43]. Most recent tools, such as Bismark[40], BS-Seeker[39] or B-SOLANA[44], use existing short-read aligners (Bowtie[45] for the mentioned tools) and handle the requirements unique to BS-seq data analysis internally.

An example for a color-space alignment tool is B-SOLANA, with which reads are initially aligned to a reference genome in which all cytosines in non-CpG context had been *in silico*–converted into thymines and are then aligned to a second reference genome in which cytosines in all sequence contexts are converted into thymines. Unlike bisulfite alignments in base space (**Fig. 2b**), this method is not suited to accurate detection of arbitrary methylation levels in unknown samples because a high degree of methylation in the non-CpG context would produce too many mismatches in the alignment step (**Fig. 2c**). This would lead to a dramatic decrease in mapping efficiency and an apparent bias toward hypomethylation in the non-CpG context. However, for the majority of eukaryotic genomes with less than 5% methylation in non-CpG context[2], alignments can be generated efficiently and accurately (**Fig. 2d**).

In the rest of this Review we use Bismark to illustrate different aspects of BS-seq analysis. Bismark can accurately detect the simulated methylation state of cytosines in any sequence context while the mapping efficiency is completely unaffected (**Fig. 2b**). We summarize details of different software packages for BS-seq data analysis in **Table 1**.

Once a dataset consisting of best alignments has been determined based on predefined alignment criteria, the methylation state of positions involving cytosines in the reference sequence can be inferred. Then these methylation calls can serve to determine the ratios of methylated versus unmethylated cytosines at every position assayed. Later, analyses of the methylation data could include looking at minimum read depths, determining methylation states of individual cytosines or genomic features, or estimating cytosine-conversion errors or false discovery rates. The biological analysis of methylome data is manifold and beyond the scope of this review.

Factors affecting the accuracy of methylation calls

Two key factors are crucial when determining the methylation state of a read from a BS-seq experiment. First, the sequence of the read must be correct and derive entirely from a bisulfite-converted sequence in the original genome. Second, the read must be correctly mapped to the corresponding position of the targeted genome. Failure to meet either of these criteria will result in the generation of incorrect methylation calls and, in extreme cases, the noise from these miscalls can adversely affect the conclusions drawn from the whole experiment. If a base is misaligned or miscalled, then on average it will display a methylation rate of 50% because both cytosine and thymine are equally likely to be misplaced against a genomic cytosine. If the true methylation level is close to 0% or 100%, then a relatively small number of errors can disproportionally shift the predicted overall level of methylation.

Base-call qualities

In real data, the quality of base calls tends to fall as the length of the reads increases (**Supplementary Fig. 2a**). As base-call errors are random, the frequency for each base will tend toward 25% each at positions with high error rates. Another source of contamination that can lead to a change in base composition is the presence of (methylated) adaptor sequences, which we discuss below. Such deviations of the average nucleotide distribution toward later cycles of a library can usually be spotted in a base-composition analysis (**Supplementary Fig. 2b**). A tradeoff can be

Table 1 | Software packages for BS-seq analysis and their performance parameters

	Bismark	BRAT	BSMAP	BS-Seeker	MethylCoder	RMAP-BS	SOCS-B	B-SOLANA
Matching tool	Bowtie[44]	Reference hashing and wildcard matching	SOAP[48]	Bowtie[45]	Bowtie/GSNAP[44,49]	Wildcard/position-weight-matrix matching	Robin-Karp hashing	Bowtie[45]
Reference	40	43	50	39	51	52	42	44
Version	0.5.0	1.2.2	0.2.1	N/A	0.14.1	2.05	2.1	0.1.0
Language	Perl	C++	C++	Python	Python	C++	Perl	Python
Library type[a]	D and ND	D and ND[b]	D and ND	D and ND[c]	D and ND	D and ND[b]	D and ND	D
Sequencing technology	Base space	Base space	Base space	Base space	Base space	Base space	Color space	Color space
Sequencing mode	Single-end and paired-end	Single-end and paired-end	Single-end and paired-end	Single-end	Single-end and paired-end	Single-end	Single-end	Single-end
Best alignment criteria	Lowest number of non-BS[d] mismatches	Lowest number of non-BS mismatches	Lowest number of mismatches	Lowest number of mismatches	Lowest number of non-BS mismatches	Lowest number of mismatches	Lowest number of non-BS mismatches	Lowest number of mismatches
Output	Mapping output including methylation calls and extra tools	Mapping output and extra tools for methylation calls	Mapping output	Mapping output including methylation calls	Mapping output and methylation call output	Mapping output	Mapping output including methylation calls	Mapping output including methylation calls
Advantages	Unbiased mapping; performance[e]	Unbiased mapping; performance	–	Unbiased mapping; performance	Unbiased mapping; performance	–	Ignores bisulfite-induced color-space mismatches	Performance
Drawbacks	–	Inflexible parameters	Performance[e]; biased mapping	–	–	Unbiased mapping only in C-G context; biased mapping in non-CG context	Performance[e]	–

[a]D, directional library; ND, nondirectional library. [b]Requires two separate runs. [c]Requires presence of a tag sequence. [d]Non-BS, not bisulfite induced. [e]Performance here signifies run time on a reasonable time scale (that is, a few hours, as compared to several days or even weeks for the same technique with a human dataset).
–, not applicable.

made, in which longer reads increase coverage but also increase the number of incorrect methylation calls. Although it would be possible to weight a bisulfite methylation call based on the quality of the original base call, this is not currently done by any of the commonly used analysis tools and would only be of benefit for miscalled bases rather than misaligned reads.

To quantify the effect of miscalled bases, we simulated a 75–base-pair read dataset containing no methylation and added random miscalls at rates between 0.01% and 10% following an exponential decay model over the length of the sequence (**Supplementary Fig. 2c**). As the error rate increased, so did the apparent methylation level.

A way to counteract methylation miscalls or mismapping events as a consequence of base call errors in the reads, is to select strict alignment parameters. To demonstrate this, we simulated bisulfite reads with sequences carrying varying numbers of false base calls and aligned this dataset to the mouse reference genome using increasingly stringent cutoffs. Increasing the mapping stringency prevented sequences with several mismatches from aligning, thus reducing the number of erroneously inferred methylation states (**Supplementary Fig. 2d**) but at the cost of reduced mapping efficiency. A better way of decreasing methylation call errors from such poor quality data is to trim off low-quality base calls before read alignments are carried out. Such adaptive quality trimming can be performed with several publically available tools, such as cutadapt (http://code.google.com/p/cutadapt/), the FASTX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html), PRINSEQ[46] (http://prinseq.sourceforge.net/), SolexaQA[47] (http://solexaqa.sourceforge.net/), Trimmomatic (http://www.usadellab.org/cms/index.php?page=trimmomatic) and others.

Sequencing into the adaptor

In many libraries, a proportion of reads will run through the insert and begin to sequence the adaptor on the 3′ end. Including adaptor sequence in a read will dramatically decrease the mapping efficiency of the read and will add a subset of random methylation calls.

We simulated the addition of varying lengths of Illumina adaptor sequence onto a BS-seq library containing no base call errors and measured the effect on both mapping efficiency and methylation calls (**Supplementary Fig. 3a**). The mapping efficiency decreased steadily with increasing adaptor contamination, but methylation errors were tightly linked to the sequence of the adaptor. Each addition of a cytosine in the adaptor caused a dramatic spike in the observed level of methylation (data not shown). Nondirectional libraries are even more susceptible to adaptor contamination as the introduction of guanine or adenine into reads aligning to the complementary bisulfite strands can introduce additional errors. Appropriate steps to identify and remove adaptor contamination, such as *k*-mer analysis (**Supplementary Fig. 3b**) with tools such as FastQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/) and adaptor-trimming software (for example, cutadapt, the FASTX toolkit, Trimmomatic, FAR (http://sourceforge.net/projects/theflexibleadap/) and others), should therefore always be taken before read alignments are carried out.

When we introduced both base call quality degradation and adaptor contamination into simulated BS-seq data reads, we observed a greatly reduced mapping efficiency compared to perfect genomic sequences (**Supplementary Fig. 3c**). When we universally trimmed the same sequences to shorter read lengths,

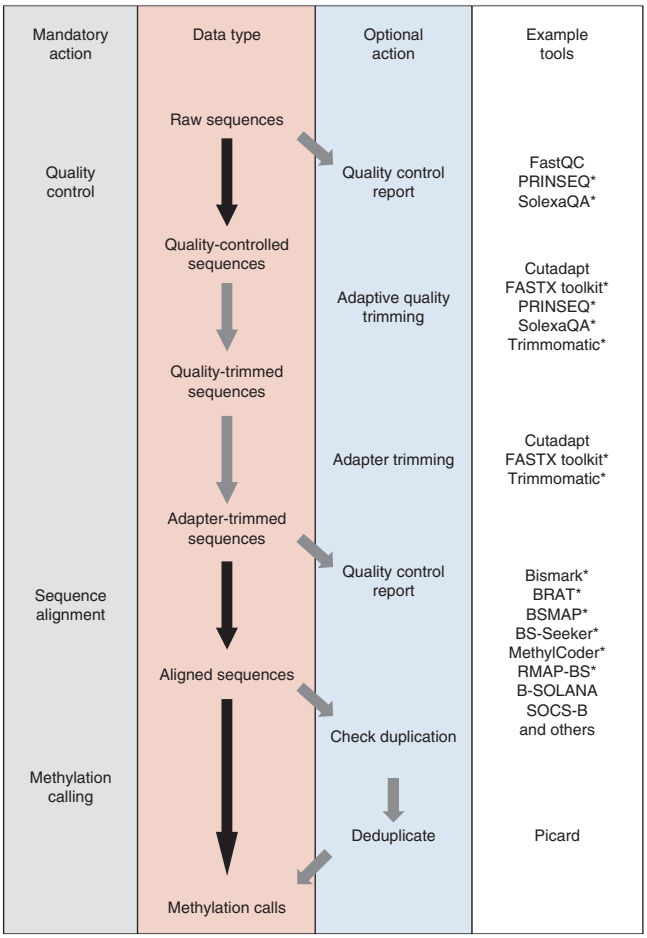


Figure 3 | Recommended workflow for the primary analysis of BS-seq data. Black arrows depict required steps, gray arrows indicate optional steps. *, only works with base-space data.

the mapping efficiency increased, reaching a maximum between 50 and 75 base pairs. This demonstrates that increasing the read length of bisulfite sequences does not necessarily translate into a linear increase in methylation information gained from an experiment. Similarly, paired-end reads do not automatically yield twice the amount of methylation data compared to single-end experiments because they result in a considerable amount of redundant methylation calls where both reads overlap.

Bisulfite conversion rate

In a BS-seq experiment we implicitly assume that all unmethylated cytosines are converted into thymines. However, this conversion may not run to completion. Incomplete conversion of unmethylated cytosines is indistinguishable from methylation and can thus introduce false positive methylation calls. In contrast, prolonged bisulfite treatment causes the sample to degrade in a way which enriches the small amount of remaining material for methylated reads.

Some studies have tried to avoid non-conversion errors by removing reads that exceeded an arbitrary threshold of methylation in a non-CpG context[7,30,37]; however, this procedure assumes very low methylation in a non-CpG context and hence introduces a potential bias against methylated reads. One option for estimating the bisulfite conversion rate is to use spike-in controls of nonnative DNA with a known methylation state. However, it should be kept in mind that such controls might not necessarily have the same conversion properties as the DNA sample to be analyzed.

End repair

It is crucial that the DNA methylation state of each fragment is not artificially modified before treatment with bisulfite because any amplification by a polymerase will erase any methylation marks that were present. In RRBS experiments, for instance, each fragment is generated by the digestion of a genome with a restriction endonuclease. The most commonly used enzyme for this type of library is MspI, which, upon cleavage, leaves a 5′ C-G overhang on the ends of each fragment[33]. To allow the addition of sequencing adapters, the overhangs are end-repaired using either methylated or unmethylated cytosines. These filled-in bases will align perfectly against the reference genome but will not maintain the original methylation state, and care must therefore be taken to exclude these bases from methylation calling. This problem only affects the 3′ end of reads when the read length is longer than the fragment to be sequenced (RRBS is probably more affected as fragments are usually size-selected to be 40 to 220 base pairs[24,33]). In such cases, reads should be screened for the occurrence of cytosine residues or a second MspI site just before reading into a potential adaptor contamination toward the 3′ end and trimmed back until the modified bases have been removed. In addition, paired-end or nondirectional RRBS libraries may also contain reads originating from filled-in MspI sites at the beginning of the reads, which consequently need to be excluded from downstream analysis[38].

Single-nucleotide variants

Any SNPs that are a cytosine in the reference genome but a thymine in the experimental sample would appear as consistent calls of unmethylated cytosines. Such errors would be impossible to detect from the quality of the reads because the base calls would be good and only the isolated nature of the effect seen might suggest that it is a technical rather than a biological effect. Both BS-seq alignments and methylation calls assume that the genomic reference sequence that reads are compared to is correct. Thus, if no SNP information is available, one has to expect a certain extent of systematic errors. These effects could be minimized by integrating available genomic-variation data, for example, from SNP databases into the reference sequence before bisulfite alignments are carried out or by using nucleotide information of the opposing genomic strand.

Conclusions

The primary analysis of BS-seq data should always start with a thorough assessment of the raw sequence data. Reads with low base call qualities or adaptor contamination should be identified and trimmed rigorously, even if this entails the risk of losing a few base pairs of real data, because the gain of confidence in correct alignments and methylation calls outweighs this minor data loss. Considering that mapping efficiencies of BS-seq and standard genomic reads converge quickly for read lengths greater than 40 base pairs (**Fig. 1a**), single-end reads of 50–75 base pairs seem to offer a reasonable compromise, providing good mapping capacity without running into problems associated with longer read lengths.

For base-space data, it is critical not to tolerate a high level of non-bisulfite mismatches (mismatches not induced by bisulfite treatment—that is, all mismatches other than (unmethylated) C-to-T mismatches) during the alignments because this allows reads to align to incorrect positions in the genome, resulting in false methylation calls. This can become especially relevant for reads originating from highly repetitive sites or from regions that are not yet part of the genome assembly. Stringent alignments are equally important for color-space data, but because color-space mapping approaches work differently, the strategy may have to be adapted to the individual needs of specific tools.

Many of the aspects of BS-seq discussed here, taken on their own, do not seem to have particularly drastic effects. Their combination, however, could easily lead to several million false methylation calls, which might have profound effects on the biological conclusions drawn from an experiment. Additional attention should be paid to reducing the number of artifacts that can only be spotted after the alignments have been performed. Duplicate reads or regions displaying abnormally high read coverage should be excluded from the analysis because they can comprise a sizeable proportion of the experiment and thereby introduce considerable bias.

Given one has a choice before starting an experiment, it is currently most convenient to opt for a platform generating base-space data because it can measure methylation over a wide dynamic range of methylation levels in any cytosine context with equal efficiencies, and most available tools are tailored to this kind of data. For small genomes or genomes fulfilling certain criteria regarding their methylation state, there are now also good tools available to handle color-space reads.

The best practices recommended here (**Fig. 3**) apply to data generated on current sequencing platforms. It will be interesting to see whether forthcoming single-molecule sequencing technologies will be able to live up to their promise to revolutionize the way in which methylation is measured.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

This work was funded by the Biotechnology and Biological Sciences Research Council, UK. A.F. and B.K. received infrastructure support from the Deutsche Forschungsgemeinschaft Excellence Cluster 'Inflammation at Interfaces'.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturemethods/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Law, J.A. & Jacobsen, S.E. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat. Rev. Genet.* **11**, 204–220 (2010).
2. Pelizzola, M. & Ecker, J.R. The DNA methylome. *FEBS Lett.* **585**, 1994–2000 (2010).
3. Robertson, K.D. DNA methylation and human disease. *Nat. Rev. Genet.* **6**, 597–610 (2005).
4. Doi, A. *et al.* Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts. *Nat. Genet.* **41**, 1350–1353 (2009).
5. Esteller, M. Epigenetics in cancer. *N. Engl. J. Med.* **358**, 1148–1159 (2008).
6. Bock, C. *et al.* Reference Maps of human ES and iPS cell variation enable high-throughput characterization of pluripotent cell lines. *Cell* **144**, 439–452 (2011).
7. Lister, R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315–322 (2009).
This was the first human methylome analyzed at single-base resolution using whole-genome bisulfite next-generation sequencing.
8. Lister, R. *et al.* Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature* **471**, 68–73 (2011).
9. Bird, A.P. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res.* **8**, 1499–1504 (1980).
10. Coulondre, C., Miller, J.H., Farabaugh, P.J. & Gilbert, W. Molecular basis of base substitution hotspots in *Escherichia coli*. *Nature* **274**, 775–780 (1978).
11. Weber, M. *et al.* Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat. Genet.* **39**, 457–466 (2007).
12. Lander, E.S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
13. Suzuki, M.M. & Bird, A. DNA methylation landscapes: provocative insights from epigenomics. *Nat. Rev. Genet.* **9**, 465–476 (2008).
14. Waterston, R.H. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
15. Illingworth, R.S. *et al.* Orphan CpG islands identify numerous conserved promoters in the mammalian genome. *PLoS Genet.* **6**, e1001134 (2010).
16. Lister, R. & Ecker, J.R. Finding the fifth base: genome-wide sequencing of cytosine methylation. *Genome Res.* **19**, 959–966 (2009).
17. Laird, P.W. Principles and challenges of genomewide DNA methylation analysis. *Nat. Rev. Genet.* **11**, 191–203 (2010).
18. Down, T.A. *et al.* A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nat. Biotechnol.* **26**, 779–785 (2008).
19. Jacinto, F.V., Ballestar, E. & Esteller, M. Methyl-DNA immunoprecipitation (MeDIP): hunting down the DNA methylome. *Biotechniques* **44**, 35–39 (2008).
20. Serre, D., Lee, B.H. & Ting, A.H. MBD-isolated Genome sequencing provides a high-throughput and comprehensive survey of DNA methylation in the human genome. *Nucleic Acids Res.* **38**, 391–399 (2010).
21. Li, N. *et al.* Whole genome DNA methylation analysis based on high throughput sequencing technology. *Methods* **52**, 203–212 (2010).
22. Frommer, M. *et al.* A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc. Natl. Acad. Sci. USA* **89**, 1827–1831 (1992).
23. Bock, C. *et al.* Quantitative comparison of genome-wide DNA methylation mapping technologies. *Nat. Biotechnol.* **28**, 1106–1114 (2010).
24. Harris, R.A. *et al.* Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nat. Biotechnol.* **28**, 1097–1105 (2010).
A detailed comparison of different sequencing-based technologies to analyze DNA methylation genome-wide.
25. Huang, Y. *et al.* The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing. *PLoS ONE* **5**, e8888 (2010).
26. Ficz, G. *et al.* Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. *Nature* **473**, 398–402 (2011).
27. Pastor, W.A. *et al.* Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. *Nature* **473**, 394–397 (2011).
28. Song, C.X. *et al.* Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. *Nat. Biotechnol.* **29**, 68–72 (2011).
29. Li, Y. *et al.* The DNA methylome of human peripheral blood mononuclear cells. *PLoS Biol.* **8**, e1000533 (2010).
30. Meissner, A. *et al.* Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454**, 766–770 (2008).
This study reported the first genome-wide DNA methylation in mouse cells generated by RRBS.
31. Feng, S. *et al.* Conservation and divergence of methylation patterning in plants and animals. *Proc. Natl. Acad. Sci. USA* **107**, 8689–8694 (2010).
32. Popp, C. *et al.* Genome-wide erasure of DNA methylation in mouse primordial germ cells is affected by AID deficiency. *Nature* **463**, 1101–1105 (2010).
33. Gu, H. *et al.* Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. *Nat. Protoc.* **6**, 468–481 (2011).
34. Gu, H. *et al.* Genome-scale DNA methylation mapping of clinical samples at single-nucleotide resolution. *Nat. Methods* **7**, 133–136 (2010).

35. Smith, Z.D., Gu, H., Bock, C., Gnirke, A. & Meissner, A. High-throughput bisulfite sequencing in mammalian genomes. *Methods* **48**, 226–232 (2009).
36. Song, F. *et al.* Association of tissue-specific differentially methylated regions (TDMs) with differential gene expression. *Proc. Natl. Acad. Sci. USA* **102**, 3336–3341 (2005).
37. Cokus, S.J. *et al.* Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature* **452**, 215–219 (2008).
This study reported a methylome of *Arabidopsis thaliana* at single-base resolution generated via a nondirectional bisulfite sequencing library.
38. Smallwood, S.A. *et al.* Dynamic CpG island methylation landscape in oocytes and preimplantation embryos. *Nat Genet.* **43**, 811–814 (2011).
39. Chen, P.Y., Cokus, S.J. & Pellegrini, M.B.S. Seeker: precise mapping for bisulfite sequencing. *BMC Bioinformatics* **11**, 203 (2010).
40. Krueger, F. & Andrews, S.R. Bismark: A flexible aligner and methylation caller for Bisulfite-seq applications. *Bioinformatics* **27**, 1571–1572 (2011).
41. McKernan, K.J. *et al.* Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res.* **19**, 1527–1541 (2009).
42. Ondov, B.D. *et al.* An alignment algorithm for bisulfite sequencing using the Applied Biosystems SOLiD System. *Bioinformatics* **26**, 1901–1902 (2010).
43. Harris, E.Y., Ponts, N., Levchuk, A., Roch, K.L. & Lonardi, S. BRAT: bisulfite-treated reads analysis tool. *Bioinformatics* **26**, 572–573 (2010).
44. Kreck, B. *et al.* B-SOLANA: An approach for the analysis of two-base encoding bisulfite sequencing data. *Bioinformatics* published online, doi:10.1093/bioinformatics/btr660 (6 December 2011).
45. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
46. Schmieder, R. & Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**, 863–864 (2011).
47. Cox, M.P., Peterson, D.A., Biggs, P.J. & Solexa, Q.A. At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics* **11**, 485 (2010).
48. Li, R., Li, Y., Kristiansen, K. & Wang, J. SOAP: short oligonucleotide alignment program. *Bioinformatics* **24**, 713–714 (2008).
49. Wu, T.D. & Nacu, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**, 873–881 (2010).
50. Xi, Y. & Li, W. BSMAP: whole genome bisulfite sequence MAPping program. *BMC Bioinformatics* **10**, 232 (2009).
51. Pedersen, B., Hsieh, T.F., Ibarra, C. & Fischer, R.L. MethylCoder: software pipeline for bisulfite-treated sequences. *Bioinformatics* **27**, 2435–2436 (2011).
52. Smith, A.D. *et al.* Updates to the RMAP short-read mapping software. *Bioinformatics* **25**, 2841–2842 (2009).

Appendix B

Supplemental Information

"DNA methylome analysis using short bisulfite sequencing data."

Krueger F[*], **Kreck B**[*], Franke A, Andrews SR

Nature Methods (2012), Jan 30;9(2):145-51

[*]Authors that contributed equally to the manuscript.

Nature Methods

DNA methylome analysis using short bisulfite read data

Felix Krueger, Benjamin Kreck, Andre Franke & Simon R Andrews

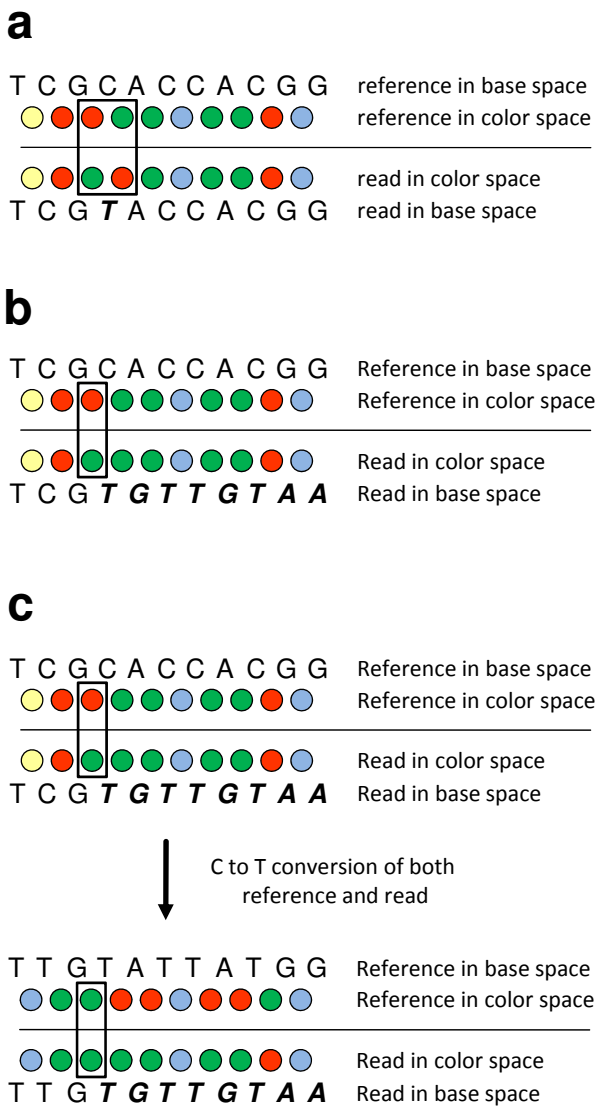
Supplementary Figure 1: Effects of inter-converting bisulfite treated color space and base space sequences.

Supplementary Figure 2: Influence of low base call qualities and errors on BS-seq alignments.

Supplementary Figure 3: Influence of adapter contamination on BS-seq alignments.

Supplementary Table 1: Whole genome shotgun bisulfite datasets to date with a mean coverage of >2-fold.

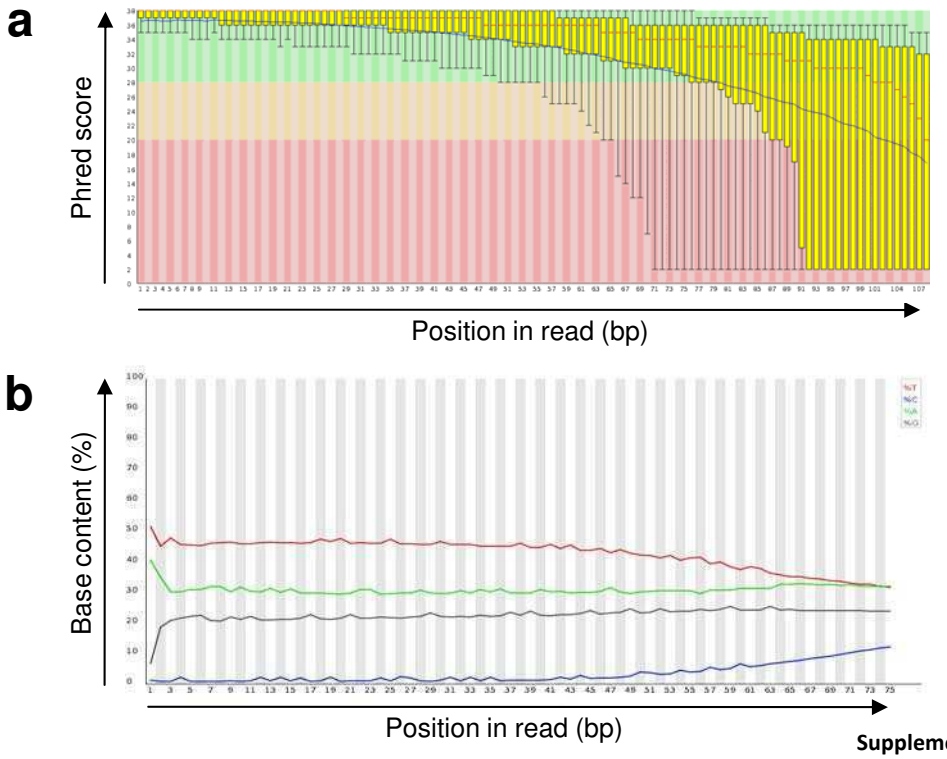
Supplementary Figure 1



Supplementary Figure 1: Effects of inter-converting bisulfite treated color space and base space sequences. (**a**) A SNP position in color space appears as two adjacent color transitions. (**b**) A measurement error in color space normally appears as a single color transition. Translating a read into base space would result in a wrong sequence, however, mapping in color space is still possible. (**c**) *In silico* conversion of C to T in bisulfite reads carrying a measurement error abrogates mapping to an equally converted reference sequence in both color and base space.

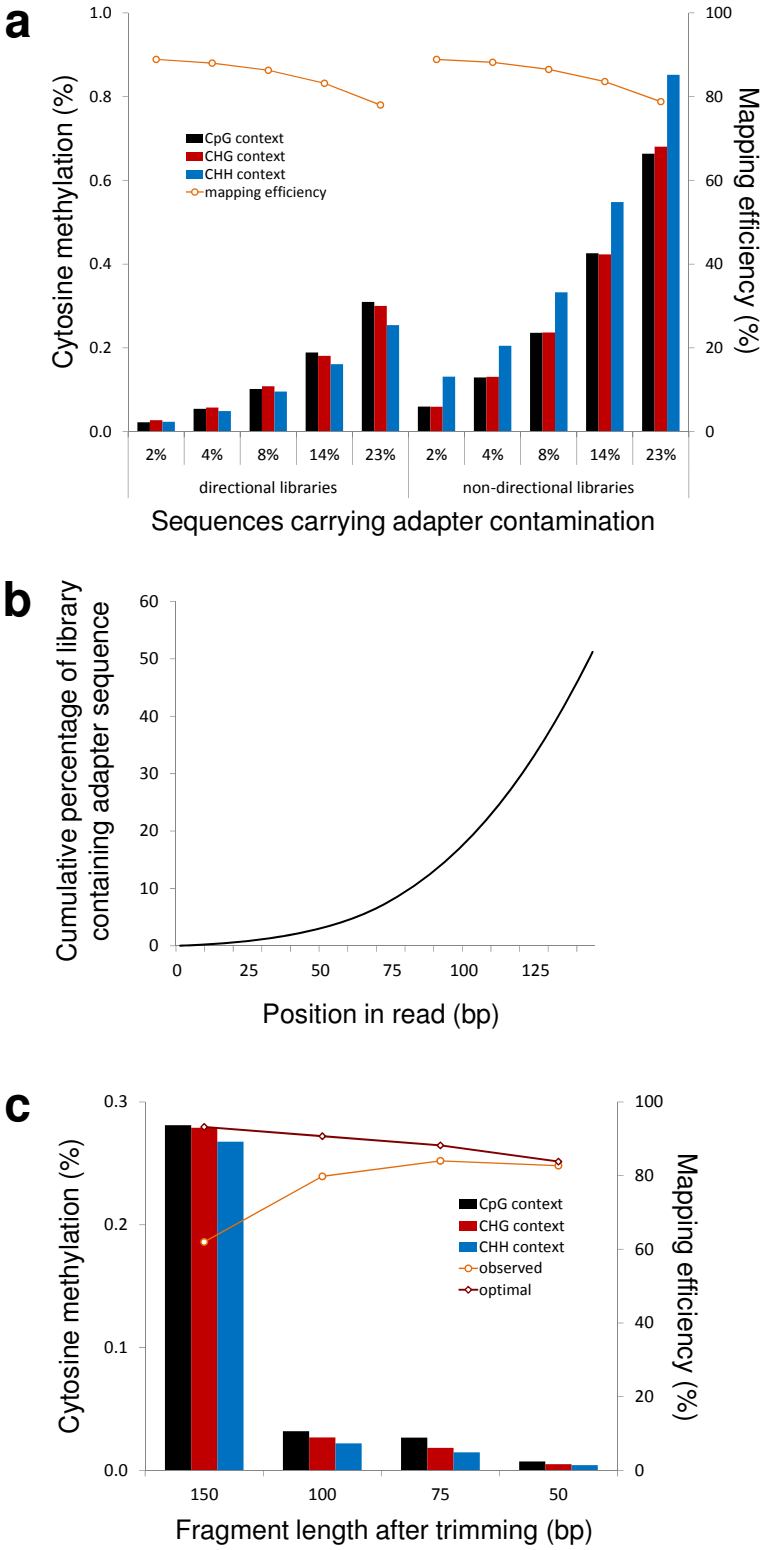
Nature Methods: doi:10.1038/nmeth.1828

Supplementary Figure 2



Supplementary Figure 2: Influence of low base call qualities and errors on BS-Seq alignments. (**a**) Decreasing base call qualities of a real 108 bp BS-Seq dataset. (**b**) Deviation of the average base composition in a real 75 bp dataset over the length of a read as a consequence of increased error rates and reading into adapter sequence on the 3' end. (**c**) 75 bp reads were simulated with varying mean per-base error rates as indicated (10^6 reads for each condition with a bisulfite conversion rate of 100%; mapping with Bismark was performed tolerating many mismatches (default options with -e 500)). (**d**) A simulated dataset containing 75 bp sequences with various amounts of miscalls was mapped to the mouse genome (NCBIM37). Increasingly strict mapping parameters reduce the number of erroneous methylation calls at the expense of mapping efficiency (the dataset contained 10^6 reads for each of the following error rates: 0.01%, 0.1%, 0.2%, 0.5%, 1%, 2% and 5%; bisulfite conversion rate was 100%). n: number of tolerated mismatches across the entire sequence length; e: mismatch ceiling of Bowtie, i.e. potentially many (low-quality) mismatches are tolerated.

Supplementary Figure 3



Supplementary Figure 3: Influence of adapter contamination on BS-Seq alignments. (**a**) Various amounts of adapter sequence (variable but with increasing length) were introduced into a 75 bp BS-Seq data set (10^6 reads, bisulfite conversion rate 100%). (**b**) Adapter sequences of variable length were introduced into 150 bp reads assuming normally distributed fragment sizes with a mean of 160 bp. Adapter contamination was detected using a custom 5-mer analysis script for the adapter starting sequence AGATC. (**c**) Effect of read length on the mapping efficiency in a BS-Seq setting. A realistic 150 bp data set was simulated with a mean per-bp error rate of 0.5% and assuming some adapter contamination of variable length (160 bp mean fragment length) (10^6 reads, bisulfite conversion rate 100%). Sequences were subsequently universally trimmed to shorter read lengths and mapped with Bismark using the same mapping parameters (-n 2, -l 50, -e 150). observed: mapping efficiencies for the realistic dataset; optimal: unbiased alignment efficiencies as in **Fig. 1a**.

Supplementary Table 1

Sample	Organism	Average coverage	Platform	Reference
ADS	*Homo sapiens*	11.5 x	Illumina	[1]
ADS-adipose	*Homo sapiens*	12.3 x	Illumina	[1]
ADS-iPSC	*Homo sapiens*	13.1 x	Illumina	[1]
FF	*Homo sapiens*	8.4 x	Illumina	[1]
FF-iPSC 6.9	*Homo sapiens*	4.8 x	Illumina	[1]
FF-iPSC 19.7	*Homo sapiens*	4.7 x	Illumina	[1]
FF-iPSC 19.11	*Homo sapiens*	4.1 x	Illumina	[1]
FF iPSC 19.11-BMP4	*Homo sapiens*	8.5 x	Illumina	[1]
IMR90	*Homo sapiens*	14.7 x	Illumina	[2]
IMR90-iPSC	*Homo sapiens*	4.5 x	Illumina	[1]
H1	*Homo sapiens*	14.0 x	Illumina	[2]
H1-BMP4	*Homo sapiens*	16.5 x	Illumina	[1]
H9	*Homo sapiens*	4.6 x	Illumina	[1]
fibroblasts (newborn)	*Homo sapiens*	9.0 x	Illumina	[3]
fibroblasts (hESC-derived)	*Homo sapiens*	9.0 x	Illumina	[3]
WA09 hESC	*Homo sapiens*	9.0 x	Illumina	[3]
Peripheral Blood MC	*Homo sapiens*	12.3 x	Illumina	[4]
HSF1 hESC	*Homo sapiens*	2.6 x	Illumina	[5]
Arabidopsis	*Arabidopsis thaliana*	20 x	Illumina	[6]
Arabidopsis	*Arabidopsis thaliana*	8.0 x	Illumina	[7]
Honey Bee	*Apis mellifera*	20 x	Illumina	[8]
Silkworm	*Bombyx mori*	7.4 x	Illumina	[9]
Escherichia coli	*Escherichia coli*	> 270 x	SOLiD	[10]

Supplementary Table 1: Overview of current whole genome shotgun bisulfite datasets with a mean coverage of >2-fold.

1. R. Lister, M. Pelizzola, Y. S. Kida et al., *Nature* **471** (7336), 68 (2011).
2. R. Lister, M. Pelizzola, R. H. Dowen et al., *Nature* **462** (7271), 315 (2009).
3. L. Laurent, E. Wong, G. Li et al., *Genome Res* **20** (3), 320 (2010).
4. Y. Li, J. Zhu, G. Tian et al., *PLoS Biol* **8** (11), e1000533 (2010).
5. R. K. Chodavarapu, S. Feng, Y. V. Bernatavichute et al., *Nature* **466** (7304), 388 (2010).
6. S. J. Cokus, S. Feng, X. Zhang et al., *Nature* **452** (7184), 215 (2008).
7. R. Lister, R. C. O'Malley, J. Tonti-Filippini et al., *Cell* **133** (3), 523 (2008).
8. F. Lyko, S. Foret, R. Kucharski et al., *PLoS Biol* **8** (11), e1000506 (2010).
9. H. Xiang, J. Zhu, Q. Chen et al., *Nat Biotechnol* **28** (5), 516 (2010).
10. C. A. Bormann Chung, V. L. Boyd, K. J. McKernan et al., *PLoS One* **5** (2), e9320 (2010).

Appendix C

"Analysis of a Base-Pair Resolution DNA Methylome from an Endemic Burkitt Lymphoma"

Kreck B, Richter J, Ammerpohl O, Barann M, Esser D, Pedersen BS, Vater I, Murga Penas EM, Chung CA, Seisenberger S, Boyd VL, Smallwood S, Drexler HG, MacLeod RAF, Hummel M, Krueger F, Häsler R, Schreiber S, Rosenstiel P, Franke A, Siebert R

Submitted to **Leukemia** (August 2012)

Analysis of a Base-Pair-Resolution DNA Methylome from an Endemic Burkitt Lymphoma

Running title: Bisulfite-sequenced Methylome of Burkitt Lymphoma

Benjamin Kreck[1], Julia Richter[2], Ole Ammerpohl[2], Matthias Barann[1], Daniela Esser[1], Britt-Sabina Petersen[1], Inga Vater[2], Eva Maria Murga Penas[2], Christina A. Bormann Chung[3], Stefanie Seisenberger[4], Victoria Lee Boyd[3], Sebastien Smallwood[4], Hans G. Drexler[5], Roderick A.F. MacLeod[5], Michael Hummel[6], Felix Krueger[4], Robert Häsler[1], Stefan Schreiber[1,7], Philip Rosenstiel[1], Andre Franke[1*], Reiner Siebert[2*]

[1]Institute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel (Kiel, Germany); [2]Institute of Human Genetics, Christian-Albrechts-University of Kiel & University Hospital Schleswig-Holstein, Campus Kiel (Kiel, Germany); [3]Life Technologies (Foster City, CA, United States of America); [4]The Babraham Institute (Cambridge, United Kingdom); [5]Department of Human and Animal Cell Cultures, German Collection of Microorganisms and Cell Cultures, Braunschweig, Germany; [6]Institute of Pathology, Charité – University Medicine Berlin, Germany; [7]Department of General Internal Medicine, University Hospital Schleswig-Holstein, Campus Kiel (Kiel, Germany)

*AF and RS share senior authorship

Corresponding authors:

Andre Franke, PhD, Institute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel, Schittenhelmstr. 12, D-24105 Kiel, mail: a.franke@mucosa.de, Tel.: +49 431 597 4138, Fax: +49 431 597 2196; and

Reiner Siebert, MD, Institute of Human Genetics, Christian-Albrechts-University of Kiel & University Hospital Schleswig-Holstein, Campus Kiel, Schwanenweg 24, D-24105 Kiel (Germany), mail: rsiebert@medgen.uni-kiel.de, Tel.: +49 431 597 4701, Fax: +49 431 597 1841

Sources of any support for the work are declared in the Acknowledgments.

Abstract

Recent epigenomic studies suggest Burkitt lymphoma (BL), in contrast to its few chromosomal alterations, to carry a high number of DNA methylation changes. Therefore, we here generated and analyzed the DNA methylome of an archetypal endemic BL cell line (DAUDI) at base-pair resolution using different platforms. The extent of cytosine methylation in CpG dinucleotides significantly varied between nuclear (68.99%), mitochondrial (6.43%) and EBV (80.18%) genomes. Despite being rare on the genome-wide level, non-CpG methylation was significantly enriched within genes. Gene expression was strongly associated with lacking CpG methylation immediately at transcription start sites. Expressed and non-expressed genes were associated with distinct patterns of gene body methylation with remarkably sharp transitions at exon-intron borders. At lower resolution, we could confirm presence of this pattern also in primary sporadic BL. Our findings show that the mechanisms of DNA methylation in lymphomas, which are associated with transcriptional regulation, extend by far the usually studied promoter methylation.

Keywords: Burkitt Lymphoma, Epigenetics, DNA Methylation, Bisulfite Sequencing

Introduction

The Burkitt translocation t(8;14), first identified in the 1970s in biopsies and cell lines from BL,[1,2] and its variants juxtapose the *MYC* oncogene to one of the immunoglobulin (*IG*) loci.[3] Nowadays, it is assumed that (nearly) all BL carry an *IG-MYC* translocation, rendering this somatic mutation a diagnostic marker for all three subtypes of BL (endemic, sporadic and immunodeficiency-related BL).

In contrast to many other lymphomas, BL show a quite simple karyotype, i.e. with few if any chromosomal changes in addition to the *IG-MYC* translocation.[4] Though there is evidence for some few recurrent secondary genetic changes the number of epigenetic alterations in BL as compared to normal B-cell subsets seems to outnumber the genetic changes by far. Indeed, we and others have identified several hundred genes showing *de novo* DNA methylation in aggressive B-cell lymphoma, including BL as compared to normal B-cell subsets.[5-7] Nevertheless, the mentioned DNA methylation studies focused on a maximum of probably 10% (by HELP assays) of the CpGs of the genome, and were biased towards promoter regions and CpG islands and did not systematically analyze non-CpG methylation.[5-8] Therefore, we here aimed at generating a complete DNA-methylome of a BL, allowing for unbiased analyses of all cytosines in the genome. To this end, we chose the archetypal DAUDI cell line, established from an endemic BL that was derived from a 16-year-old African male patient in 1967.[9,10] We selected this cell line as it has been pivotal for the identification of t(8;14), still carries a simple karyotype despite being many years in culture and because it shows the prototypic features of eBL. Moreover, considering the strong association of eBL with Epstein-Barr virus (EBV) infection, the EBV-positive

DAUDI cell line offers the opportunity for a direct comparison of its lymphoma and EBV methylomes.

Methods

Genomic characterization of DAUDI cells

DAUDI cells (ACC-078) and DNA were provided by the "Deutsche Sammlung von Mikroorganismen und Zellkulturen" (DSMZ). Chromosomal R-banding analysis and Genome-Wide Human SNP Array 6.0 (Affymetrix, Santa Clara, CA, USA) analysis were performed according to standard methods. Whole exome capture and sequencing were carried out using Illumina's TruSeq Exome Enrichment Kit (Illumina, San Diego, CA, USA). A subset of mutations identified by exome sequencing was verified by Sanger Sequencing (see **Supplementary Data**).

DNA methylation profiling using Bisulfite-Sequencing (BS-seq)

We performed genome-wide BS-seq on the SOLiD™ (Life Technologies, Carlsbad, CA) and the HiSeq 2000 platform.

For the prior, two bisulfite-converted SOLiD™ fragment libraries were constructed as described previously.[11,12] Briefly, 15 µg of genomic DNA were sheared to approximately 125 bp using a Covaris S2 system (Life Technologies, Carlsbad, CA, USA). After end-repair of the DNA fragments, methyl-P1 and P2 adaptors were ligated (for details on methyl-P1 and P2 see Ranade *et al.*[12]). The DNA was then size selected on an agarose gel and nick-translated with a modified dNTP Mix containing methyl-dCTPs instead of regular dCTPs in order to protect the adaptor sequences

during bisulfite conversion. Bisulfite conversion was carried out in solution as described previously[12] and recovered DNA fragments were PCR amplified using 8 cycles. The bisulfite converted fragment library was clonally amplified on SOLiD P1 beads using emulsion PCR. Templated (P2 positive) beads were then enriched and deposited on a slide for sequencing.

For the HiSeq 2000 analyses, genomic DNA was sonicated using the Diagenode Bioruptor (Diagenode, Denville, NJ) to a final size distribution ranging from 100 bp to 800 bp. Libraries were prepared from the sonicated DNA using the NEBNext Sample Prep Master Mix Set 1 (New England Biolabs, Ipswich, MA) according to the manufacturers' instructions. Illumina's Early Access Methylation Adapter Oligo was used for the ligation. The adapter-ligated DNA was treated with sodium bisulfite using the Imprint DNA Modification Kit (Sigma, St. Louis, MO) according to the manufacturers' instructions. The bisulfite-treated product was amplified with 16 cycles using a uracil stalling-free polymerase (Agilent, Santa Clara, CA, USA) followed by size selection on a gel (200 bp – 250 bp) and purification with the Qiagen Gel Extraction Kit (Qiagen, Hilden, Germany).

Sequencing using SOLiD™ v4.0 chemistry according to manufacturer's instructions yielding 50 bp reads, which were analyzed with B-SOLANA.[13,14]

For the HiSeq 2000 analyses, V3 chemistry was used according to manufacturer's instructions yielding 100 bp reads, which were analyzed with Bismark.[14,15]

DNA methylation profiling using universal BeadArrays

DNA from DAUDI and four proto-typic sporadic BL was bisulfite-converted using the Zymo EZ DNA Methylation Kit (Zymo Research, Orange, CA, USA) and hybridized to the HumanMethylation450 DNA Analysis BeadChip (Illumina). Hybridization signals were analyzed using GenomeStudio software (ver. 2011.1, Methylation Analysis Module ver. 1.9.0; Illumina Inc).

Gene expression analyses

We combined expression data from the U133A gene chip (Affymetrix, Santa Clara, CA) and RNA-seq using the SOLiD™ platform (Life Technologies, Carlsbad, CA). The RNA-seq library was prepared by a modified Whole Transcriptome Analysis Kit (WTAK) (Life Technologies, Carlsbad, CA) and analyzed as previously described,[16] with the modification that we used TopHat[17] to carry out alignments.

See the provided **Supplemental Material and Methods** section for details.

Results and Discussion

To obtain a base-pair resolution DNA methylome of a prototypic eBL we subjected the widely-used t(8;14)- and EBV-positive DAUDI cell line to full bisulfite-sequencing, using two different platforms. By karyotyping, SNP-array analysis and exome sequencing we confirmed that the cells under study show the typical features of BL, including the t(8;14) plus a few secondary chromosomal changes and mutations in genes like *ID3* and *B2M* (**Supplemental Table 1**).

We aligned 79.9 Gb of bisulfite sequences of the SOLiD™ and 7.8 Gb of the HiSeq 2000 platform and compared the results to the DNA methylation levels determined

by HumanMethylation450 BeadChip analysis (**Supplemental Figure S1**). We observed high correlation of the SOLiD™ data with both the sequence-based HiSeq 2000 (Pearson r=0.86; **Supplemental Figure S2**) and the array-based (Pearson r=0.96, **Supplemental Figure S3**) methylation levels. This led us to focus our further analyses on the most extensive dataset derived from SOLiD™ BS-seq.

In total, 91.1% of all CpG sites and 90.2% of all non-CpG sites of the genome were covered by at least five SOLiD™ reads (**Supplemental Figure S4**). On the genome-wide level 68.99% cytosines in CpG dinucleotides were methylated which is in line with previous pyrosequencing-based determinations using LUMA (**Figure 1**). In contrast, the 450K BeadArray shows a mean methylation level of 59.24%, which is mostly due to the selection bias of the array-loci, which are predominantly located within regions upstream of genes. We observed a mean CpG methylation level of 70.88% in high-complex regions, whereas mean CpG methylation in repeats accounted for 65.79% in LINEs and 78.84% in SINEs (**Supplemental Figure S5**). Bisulfite sequencing shows the DNA methylation patterns on the forward and reverse strand to be comparably established (Pearson r=0.90).

Considering the recent description of non-CpG methylation in ESC, and the fact that the *MYC* oncogene deregulated in BL is also one of the four factors used to induce a stem cell-like phenotype in differentiated cells,[18] we analyzed the level of non-CpG methylation. The genome-wide fraction of methylated cytosines in a non-CpG context does not exceed the respective threshold of 0.003 given by the unmethylated lambda control DNA that was tested in parallel.[8] Moreover, we confirmed absence of non-CpG methylation at hallmark sites described in ESC[8] by BS pyrosequencing (**Supplemental Figure S6**). Despite this overall low frequency of non-

CpG methylation, we could identify a remarkable 6.7-fold enrichment of non-CpG methylcytosines within genes ($p<2.2\times10^{-16}$; **Figure 2**). Such non-CpG methylation might be linked to transcriptional activity (**Supplemental Figure S7**).

We next determined the sequence-based methylation status of 969 genes recently shown by us to exhibit *de novo* promoter hypermethylation in mature aggressive B-cell lymphoma (including BL) as compared to normal B-cells.[5] We could confirm that in DAUDI cells 91.21% of these genes have a DNA methylation level ≥60% in their promoter region and lack transcription. As compared to all other RefSeq genes, the mean CpG methylation level within promoter regions of the 969 genes was significantly higher (84% vs. 41%; **Supplemental Tables S2**).

Gene expression analyses confirmed that DAUDI cells show the typical signature of molecular Burkitt Lymphoma (mBL).[19] Correlating methylation and expression patterns in our data revealed that significant presence of transcripts is associated with absence of DNA methylation particular at and closely around the transcription start site (TSS). In contrast, DNA methylation exactly at the TSS correlates with lack of transcription (**Figure 3**). Whereas the group of non-expressed genes showed an overall high mean DNA methylation level across the whole gene with highest methylation levels in exons, genomic regions comprising expressed genes were characterized by particular high methylation levels in the first intron. Moreover, the patterns of both expressed and non-expressed genes were characterized by sharp transitions of methylation levels at exon-intron borders (**Figure 3**).

In order to exclude that this pattern of DNA methylation is limited to the DAUDI cell line, we analyzed 450K BeadArray data from four primary prototypic sporadic BL (sBL) Though DAUDI contains significantly more methylated CpGs than the primary

sBL (59.24% vs. 47.50%) the targets of DNA methylation are coincident suggesting a significant epigenomic (and regulation thereof) analogy (**Supplemental Figure S8**). We observed a similar association of transcriptional activity and DNA methylation in sBL like in DAUDI cells, both around the TSS and in the gene body (**Supplemental Figure S9**). These findings indicate that correlation analyses between DNA methylation and expression in BL strongly depend on the localization of the CpGs under study.

Finally, we studied the DNA methylation of the mitochondrial and EBV genomes of DAUDI cells.[20-22] We estimated 80 EBV and 370 mitochondrial copies per DAUDI cell based on coverage analyses, which is in accordance with previous studies.[23] Whereas mitochondrial DNA is mostly unmethylated (mean methylation 6.43%; **Figure 1**), CpG methylation in the human and EBV genome is comparably distributed, though the EBV genome exhibits hardly any fully methylated sites (**Figure 1**). Overall, EBV shows a high level of DNA methylation (mean methylation 80.18%), as it was previously shown for BL cell lines.[24] Nevertheless, DNA methylation within the EBV genome correlates with expression only at high transcript levels (FPKM>=15) (**Supplemental Figure S10**).

In summary, we have characterized the nuclear DNA methylome of an endemic BL along with its mitochondrial and EBV methylome. We unravel significant differences between the different sub-methylomes and moreover show that gene transcription is associated with complex patterns of methylation, extending beyond simple promoter and CpG methylation. As the DAUDI cell line has been used over decades in many laboratories in the world, the obtained methylome data might serve as a "reference epigenome" for future studies.

Note: Supplementary information available

Acknowledgement

This study was financed through a grant from the “Medizinausschuss Schleswig-Holstein”. The project received support through the BMBF ICGC MMML-Seq project (FKZ 01KU1002A and 01KU1002E) and infrastructure through the DFG Cluster of Excellence “Inflammation at Interfaces”. Bisulfite pyrosequencing was supported by KinderKrebsInitiative Buchholz/Holm-Seppensen and gene expression data were obtained in the framework of the Deutsche Krebshilfe-Network “Molecular Mechanisms of Malignant Lymphoma (MMML)”.

Authorship Contributions and Disclosure of Conflict of Interest

Contribution: B.K., J.R., O.A., P.R., A.F., and R.S. designed the research, interpreted data and wrote the paper; P.R., A.F., O.A., S.Sc. and R.S. provided infrastructure and obtained grant support; I.V., E.M.M.P, H.G.D. and R.A.F.M. performed cell line characterization; M.H. performed array-based gene expression profiling; B.K., M.B., and F.K. analyzed BS-seq data; B.K., J.R., and O.A. analyzed pyrosequencing data and array-based DNA methylation data; B.K., D.E., and R.H. analyzed RNA-seq data; B.-S.P. analyzed exome sequencing data; C.B.C. and V.L.B. prepared the SOLiD BS-seq library; S.Se. and S.Sm. prepared the Hi-Seq 2000 BS-seq libraries; and all authors critically reviewed the paper.

Conflict-of-interest disclosure: O.A. and R.S. received research support by Illumina on BeadChip Analyses. C.B.C is and V.L.B. used to be employed by Life Technologies.

References

1. Manolov G, Manolova Y. Marker band in one chromosome 14 from Burkitt lymphomas. *Nature*. 1972; **237:** 33 - 34.

2. Zech L, Haglund U, Nilsson K, Klein G. Characteristic chromosomal abnormalities in biopsies and lymphoid-cell lines from patients with Burkitt and non-Burkitt lymphomas. *Int J Cancer* 1976; **17:** 47 - 56.

3. Salaverria I, Siebert R. The gray zone between Burkitt's lymphoma and diffuse large B-cell lymphoma from a genetics perspective. *J Clin Oncol* 2011; **29:** 1835 - 1843.

4. Boerma EG, Siebert R, Kluin PM, Baudis M. Translocations involving 8q24 in Burkitt lymphoma and other malignant lymphomas: a historical review of cytogenetics in the light of todays knowledge. *Leukemia* 2009; **23:** 225 - 234.

5. Martín-Subero JI, Kreuz M, Bibikova M, Bentink S, Ammerpohl O, Wickham-Garcia E, *et al*. New insights into the biology and origin of mature aggressive B-cell lymphomas by combined epigenomic, genomic, and transcriptional profiling. *Blood* 2009; **113:** 2488 - 2497.

6. Shaknovich R, Geng H, Johnson NA, Tsikitas L, Cerchietti L, Greally JM, *et al*. DNA methylation signatures define molecular subtypes of diffuse large B-cell lymphoma. *Blood* 2010; **116:** e81 - 89. Epub 2010 Jul 7.

7. Shaknovich R, Cerchietti L, Tsikitas L, Kormaksson M, De S, Figueroa ME, *et al.* DNA methyltransferase 1 and DNA methylation patterning contribute to germinal center B-cell differentiation. *Blood* 2011; **118:** 3559 - 3569.

8. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 2009; **462:** 315 - 322.

9. Klein E, Klein G, Nadkarni JS, Nadkarni JJ, Wigzell H, Clifford P. Surface IgM-kappa Specificity on a Burkitt Lymphoma Cell In Vivo and in Derived Culture Lines. *Cancer Res* 1968; **28:** 1300 - 1310.

10. Nadkarni JS, Nadkarni JJ, Clifford P, Manolov G, Fenyö EM, Klein E. Characteristics of new cell lines derived from Burkitt lymphomas. *Cancer* 1969; **23:** 64 - 79.

11. Bormann Chung CA, Boyd VL, McKernan KJ, Fu Y, Monighetti C, Peckham HE. Whole methylome analysis by ultra-deep sequencing using two-base encoding. *PloS ONE* 2010; **5:** e9320.

12. Ranade SS, Bormann Chung C, Zon G, Boyd VL. Preparation of genome-wide DNA fragment libraries using bisulfite in polyacrylamide gel electrophoresis slices with formamide denaturation and quality control for massively parallel sequencing by oligonucleotide ligation and detection. *Anal Biochem* 2009; **390:** 126 - 135.

13. Kreck B, Marnellos G, Richter J, Krueger F, Siebert R, Franke A. B-SOLANA: an approach for the analysis of two-base encoding bisulfite sequencing data. *Bioinformatics* 2012; **28:** 428 - 429.

14. Krueger F, Kreck B, Franke A, Andrews SR. DNA methylome analysis using short bisulfite sequencing data. *Nat Methods*. 2012; **9**: 145 - 151.

15. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* 2011; **27:** 1571 - 1572.

16. Klostermeier UC, Barann M, Wittig M, Häsler R, Franke A, Gavrilova O, et al. A tissue-specific landscape of sense/antisense transcription in the mouse intestine. *BMC Genomics* 2011; **12:** 305.

17. Trapnell C, Pachter L, Salzberg SL. Tophat: discovering splice junctions with RNA-Seq. *Bioinformatics* 2009; **25:** 1105 - 1111.

18. Takahashi K, Yamanaka S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* 2006; **126:** 663 - 676.

19. Hummel M, Bentink S, Berger H, Klapper W, Wessendorf S, Barth TF, *et al.* A biologic definition of Burkitt's lymphoma from transcriptional and genomic profiling. *N Engl J Med* 2006; **354:** 2419 – 2430.

20. Lin Z, Xu G, Deng N, Taylor C, Zhu D, Flemington EK. Quantitative and qualitative RNA-Seq-based evaluation of Epstein-Barr virus transcription in type I latency Burkitt's lymphoma cells. *J Virol* 2010; **84:** 13053 - 13058.

21. Flower K, Thomas D, Heather J, Ramasubramanyan S, Jones S, Sinclair AJ. Epigenetic control of viral life-cycle by a DNA-methylation dependent transcription factor. *PloS ONE* 2011; **6:** e25922.

22. Dresang LR, Teuton JR, Feng H, Jacobs JM, Camp DG 2nd, Purvine SO, *et al*. Coupled transcriptome and proteome analysis of human lymphotropic tumor viruses: insights on the detection and discovery of viral genes. *BMC Genomics* 2011; **12:** 625.

23. Leenman EE, Panzer-Grümayer RE, Fischer S, Leitch HA, Horsman DE, Lion T, *et al.* Rapid determination of Epstein-Barr virus latent or lytic infection in

single human cells using in situ hybridization. *Mod Pathol* 2004; **17:** 1564 - 1572.

24. Kim DN, Song YJ, Lee SK. The role of promoter methylation in Epstein-Barr virus (EBV) microRNA expression in EBV-infected B cell lines. *Exp Mol Med* 2011; **43:** 401 - 410.

Figure Legends

Figure 1. Distribution of CpG DNA-methylation in DAUDI cells

The graphs show genome-wide distributions of CpG methylation of the human nuclear, EBV and mitochondrial genomes. The y-axis indicates DNA methylation levels assessed by SOLiD™ BS-seq. Green: Human nuclear CpG methylation, Red: EBV CpG methylation, Blue: Mitochondrial CpG methylation.

Figure 2. Distribution of non-CpG DNA-methylation in DAUDI cells

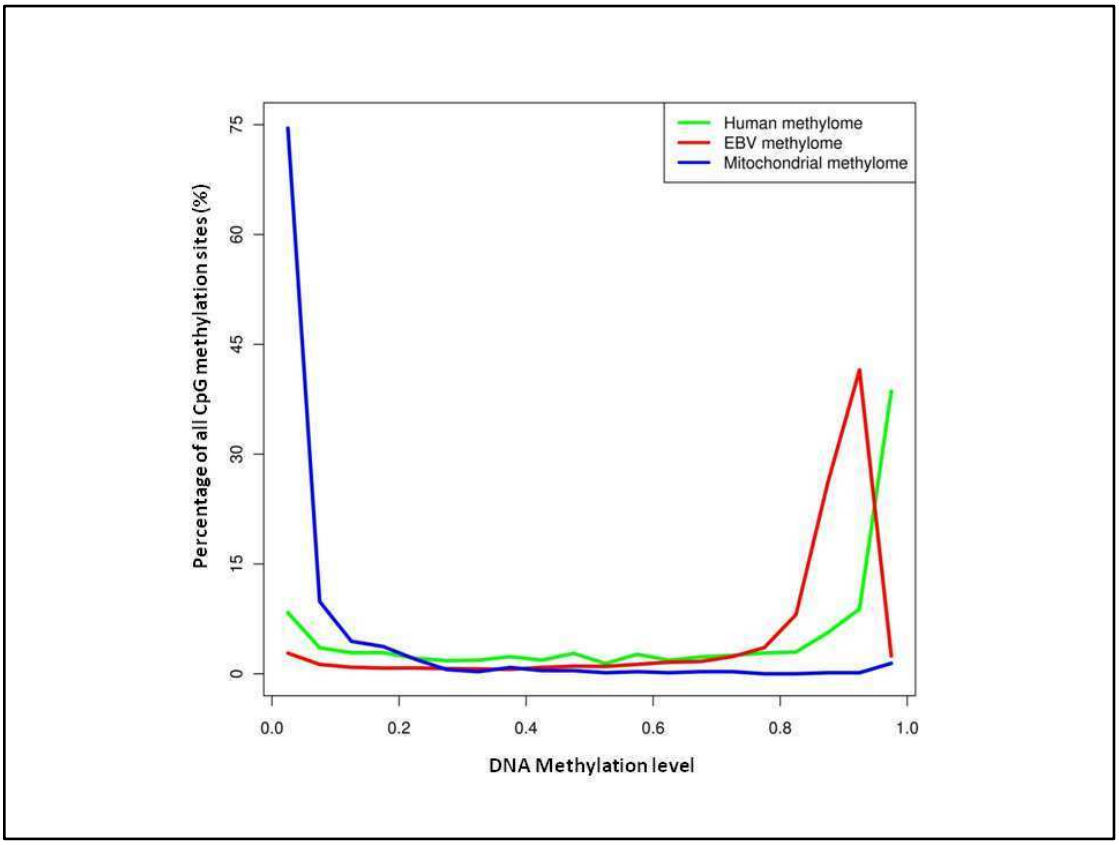
Significantly methylated nonCpG sites of DAUDI within RefSeq genes (comprising 424,969,306 non-CpGs) are 6.7-fold enriched compared to those outside of RefSeq genes (comprising 689,750,660 non-CpGs). Red: Fraction of significantly methylated non-CpGs of DAUDI within RefSeq genes, Blue: Fraction of significantly methylated non-CpGs of DAUDI outside of RefSeq genes.

Figure 3. Correlation of DNA methylation levels and transcriptional states

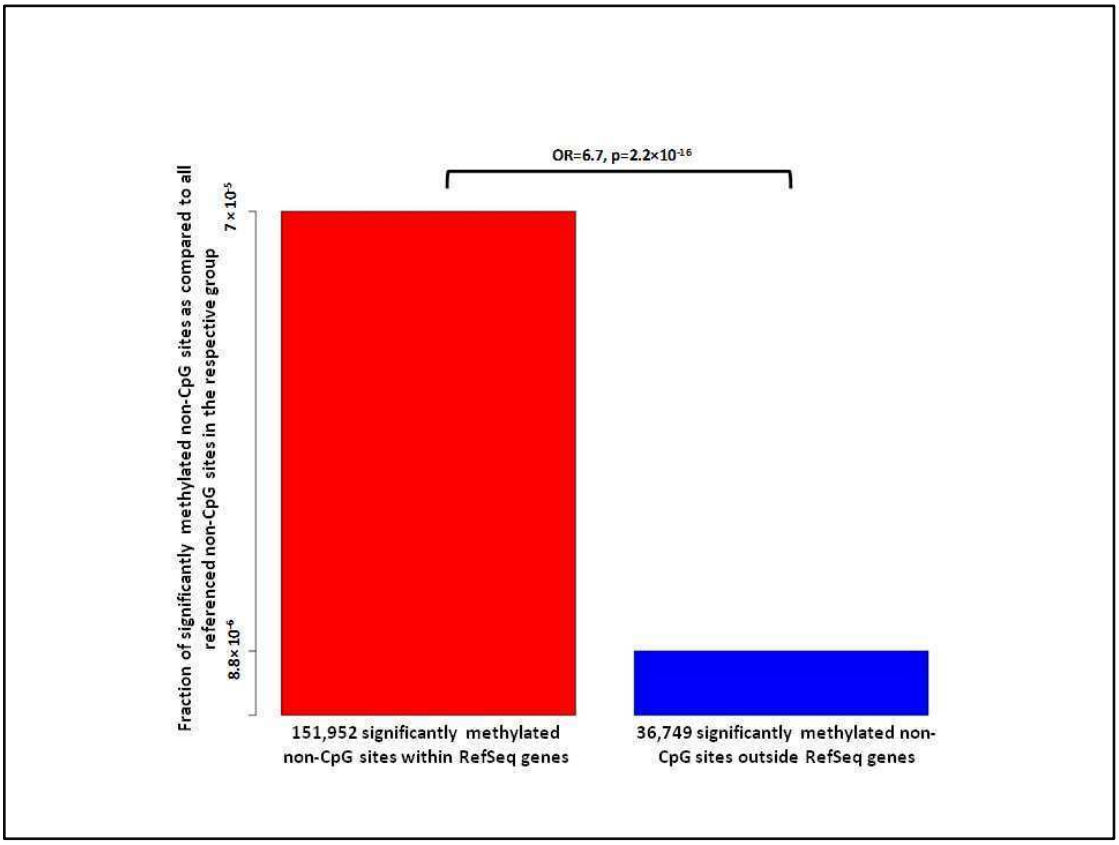
CpG methylation levels were averaged for annotated RefSeq gene regions and transcripts are clustered by their expression level in present (n=7662 transcripts) and absent (n=5429 transcripts) calls. A strong dependency of the location of CpGs related to their distance to the TSS and the transcript expression level can be observed. Green: Average methylation pattern for present transcripts, Red: Average methylation pattern for absent transcripts.

Figures

Figure 1

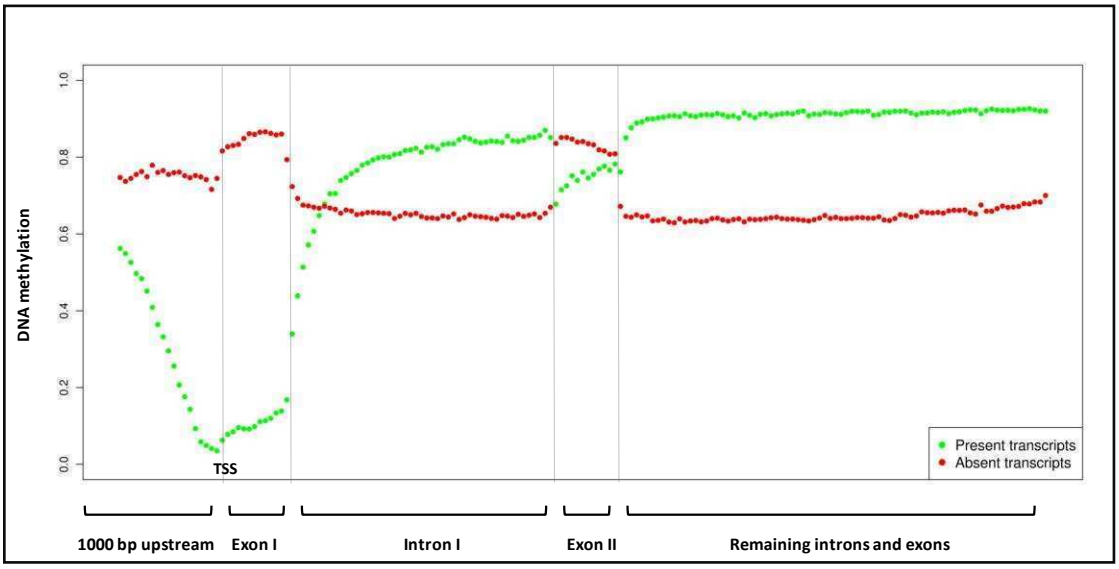


342 **Figure 2**



343

Figure 3


DNA methylation
0.0
0.2
0.4
0.6
0.8
1.0
TSS
Present transcripts
Absent transcripts
1000 bp upstream
Exon I
Intron I
Exon II
Remaining introns and exons

Appendix C

Supplemental Information

"Analysis of a Base-Pair Resolution DNA Methylome from an Endemic Burkitt Lymphoma"

Kreck B, Richter J, Ammerpohl O, Barann M, Esser D, Pedersen BS, Vater I, Murga Penas EM, Chung CA, Seisenberger S, Boyd VL, Smallwood S, Drexler HG, MacLeod RAF, Hummel M, Krueger F, Häsler R, Schreiber S, Rosenstiel P, Franke A, Siebert R

Submitted to **Leukemia** (August 2012)

SUPPLEMENTARY INFORMATION

Analysis of a Base-Pair-Resolution DNA Methylome from an Endemic Burkitt Lymphoma

Running title: Bisulfite-sequenced Methylome of Burkitt Lymphoma

Benjamin Kreck[1], Julia Richter[2], Ole Ammerpohl[2], Matthias Barann[1], Daniela Esser[1], Britt-Sabina Petersen[1], Inga Vater[2], Eva Maria Murga Penas[2], Christina A. Bormann Chung[3], Stefanie Seisenberger[4], Victoria Lee Boyd[3], Sebastien Smallwood[4], Hans G. Drexler[5], Roderick A.F. MacLeod[5], Michael Hummel[6], Felix Krueger[4], Robert Häsler[1], Stefan Schreiber[1,7], Philip Rosenstiel[1], Andre Franke[1*], Reiner Siebert[2*]

[1]Institute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel (Kiel, Germany); [2]Institute of Human Genetics, Christian-Albrechts-University of Kiel & University Hospital Schleswig-Holstein, Campus Kiel (Kiel, Germany); [3]Life Technologies (Foster City, CA, United States of America); [4]The Babraham Institute (Cambridge, United Kingdom); [5]Department of Human and Animal Cell Cultures, German Collection of Microorganisms and Cell Cultures, Braunschweig, Germany; [6]Institute of Pathology, Charité – University Medicine Berlin, Germany; [7]Department of General Internal Medicine, University Hospital Schleswig-Holstein, Campus Kiel (Kiel, Germany)

*AF and RS share senior authorship

Supplementary Material and Methods

Characterization of the DAUDI cell line

The cell line DAUDI has been established in 1967 from an endemic BL (eBL) presenting in the left orbita of a 16-year-old African boy.[1] The cell line is positive for EBV (HHV-4) but lacks expression of immediate-early protein BZLF-1 and lately expressed capsid protein. By PCR, the cell line is negative for HBV, HCV, HHV-8, HIV, HTLV-I/II and SMRV. The immunophenotype has been determined as CD3 -, CD10 +, CD19 +, CD20 +, CD37 +, CD38 +, cyCD79a +, CD80 +, CD138 -, HLA-DR +, sm/cyIgM +, sm/cyIgG -, sm/cykappa +, sm/cylambda – (data available from DSMZ at www.dsmz.de).

Cytogenetic analysis using chromosomal R-banding revealed a karyotype 46,XY,t(8;14)(q24;q32). These findings are in line with those from Multicolor Fluorescence In Situ Hybridization (M-FISH) of the DAUDI cell line and showing that it has remained karyotypically stable along decades of continuous cultivation (E.M.M.P *et al.*, manuscript in preparation).

The Genome-Wide Human SNP Array 6.0 was performed according to manufacturer's protocol (Affymetrix, Santa Clara, CA) using the Fluidics Station 450 and the GeneChip Scanner 3000 (Affymetrix, Santa Clara, CA). The Birdseed v2 algorithm was used to genotype tumour samples. Copy number analysis, Loss of heterozygosity (LOH) analysis and segmentation was calculated using Genotyping Console software version 3.0 (Affymetrix, Santa Clara, CA). Segments with significant imbalances were considered as copy number aberration only if they consisted of at least 20 sequential probes, comprised a minimal size of 100 kb, and mapped outside

known copy number polymorphisms. Data analysis revealed two gains and four chromosomal losses: arr 4q13.3(75,205,069-75,765,823)x1,5p11.1(46,361,933-49,591,883)x1,7q31.32q31.33(121,093,529-123,683,639)x1, 8q24.21q24.3(128,682,912-146,268,947)x3,14q32.31q32.33(101,614,613-105,400,262)x4,15q12q21.1(25,151,945-45,290,444)x1 (NCBI36/hg 18)

Whole Exome Sequencing

We performed whole exome capture using Illumina's TruSeq Exome Enrichment Kit and sequencing of 2x100 bp paired-end reads was performed on one quarter lane of an Illumina HiSeq2000. Reads were mapped against the human reference genome build hg18 using BWA,[2] followed by the removal of PCR duplicates with Picard (http://picard.sourceforge.net). Variant calling was performed with SAMtools mpileup and GATK,[3,4] for SNP annotation and filtering we applied our own in-house tool *snp*Acts. InDels were annotated using ANNOVAR.[5]

A total of 62,578 on target SNPs were filtered against 8 exomes of healthy individuals, allowing for a maximum frequency of 1% in the 1000 genomes project and keeping only non-synonymous and splice-site SNPs that were not present in dbSNP130 resulting in 2,313 SNP after filtering.

Sanger Sequencing

Selected SNVs detected by exome sequencing were verified by Sanger Sequencing on an ABI Sequencer 3100 (Applied Biosystems).

Primer sequences used for validation of potentially protein changing mutations:

Gene	Primer name	Sequence 5´-3´
B2M	B2M_FP	TCCCTCTCTCTAACCTGGCAC
	B2M_RP	ACTTGGAGAAGGGAAGTCACG
TET2	TET2_FP	TGCATGCAAAATACAGGTTTC
	TET2_RP	CAGCTTGCAGGTGGATTCTC
ID3	ID3_FP	TCCAGGCAGGCTCTATAAGTG
	ID3_RP	CCGAGTGAGTGGCAATTTTT
KIT	KIT_FP	CACAGACCCAGAAGTGACCA
	KIT_RP	TACCTGGCCTCACTTTCAGG

Gene expression analyses

U133A raw data was analyzed using the *panp* package of the R statistical software (Peter Warren, panp R package version 1.20.1.) and the Affymetrix Microarray Analysis Suite version 5.0 (MAS).[6]

The paired-end RNA-seq library was prepared using the RiboMinus™ Eukaryote Isolation Kit and the RiboMinus™ Concentration Kit (Life Technologies, Carlsbad, CA). Subsequently, the SOLiD™ Whole Transcriptome Analysis Kit (WTAK) was performed and sequencing was carried out at 50 bp in the forward and 35 bp in the reverse direction using SOLiD™ v4.0 chemistry according to manufacturer's instructions.

A combined set of results was generated in the following manner. Present transcripts are the intersection of U133A-determined present calls and RNA-seq calls consisting of a FPKM value >0.01. However, transcripts including an absent assignment by the U133A assay and a FPKM value ≤0.01 were considered as an absent call. In total, we could assess 7662 present calls and 5429 absent calls.

Expression analyses of EBV are restricted to RNA-seq data, due to missing EBV-annotations on the U133A-chip. We raise the FPKM threshold for transcripts within the EBV genome from 0.01 to 1 estimated by the number of ≈100 copies (compared to ≈80 copies assessed by coverage analyses) of virus particles per cell.[7]

Identification and analyses of significant non-CpG methylation sites

Potential methylcytosines in a non-CpG context were detected as described by Lister *et al.*.[8] Hereby, the binomial distribution was used to exclude false positive non-CpG methylation sites arisen by incomplete bisulfite conversion. Furthermore, we corrected for potential non-CpG sites including a mutation (N to G) in their adjacent base (+1). Although the genome-wide amount of methylated cytosines in non-CpG context does not exceed the expected threshold, given by an estimation based on spiked-in lambda phage, we were able to identify a local enrichment of non-CpG methylcytosines within gene regions (Figure 1B). In detail, we observed 9533 RefSeq genes containing significant non-CpG sites. The ratio of these present and absent transcripts is slightly enriched compared to the genome-wide level (OR=1.11). However there is no significant difference between absent and present transcripts in terms of the ratio of significant and existing non-CpG sites within RefSeq genes (Table S2).

HumanMethylation450 DNA Analysis BeadArray

DNA of DAUDI cells and of four primary sporadic BL was subjected to 450K BeadArray-analysis. Primary samples were proto-typic *IG-MYC* positive BL derived from children. Bisulfite conversion was performed using the Zymo EZ DNA

Methylation Kit (Zymo Research, Orange, CA, USA) according to the manufacturer's instruction. Subsequent analysis steps were performed according to the manufacturer's protocol measuring DNA methylation at >485,000 CpG sites selected from more than 21,000 genes in parallel. Hybridization signals were analyzed using GenomeStudio software (default settings; GenomeStudio ver. 2011.1, Methylation Analysis Module ver. 1.9.0; Illumina Inc) and internal controls for normalization.

Bisulfite Pyrosequencing

Bisulfite pyrosequencing of two regions identified by Lister et al., to carry non-CpG methylation in differentiated and stem cells were pyrosequenced.[8] Bisulfite pyrosequencing was carried out as described by Lamprecht *et al.*.[9] Briefly, genomic DNA was bisulfite converted using the EpiTect Bisulfite Conversion Kit (Qiagen). In a subsequent PCR amplification locus-specific primers were used with one primer biotinylated at the 5´ end (sequencing primer sequences are shown below). Amplification was verified by agarose gel electrophoresis. Using the VacuumPrep Tool (Biotage, Uppsala, Schweden) single strands were prepared followed by a denaturation step at 85°C for two minutes and final sequencing primer hybridization. Pyrosequencing was performed using the Pyrosequencer ID and the DNA methylation analysis software Pyro Q-CpG 1.0.9 (Biotage), which was also used to evaluate the ratio T:C (mC:C) at the CpG sites analyzed. All assays were optimized and validated using commercially available completely methylated DNA (Millipore) and pooled DNA isolated from peripheral blood of 10 healthy male and female controls, respectively.

Primer sequences used for bisulfite pyrosequencing:

Primer name	5´-3´sequence	5´modification
methyl_chr1_FP	AAATTTGGTTTTTTTATATGG	
methyl_chr1_RP	CTAAAACCTCTTAAACTTTTATCA	Biotin
methyl_chr1_seq	GGTTTTTTTATATGGTTA	
methyl_chr10_FP	GATGGGTGATTTTTTAGA	
methyl_chr10_RP	ACATTTCCTACAATTTCAA	Biotin
methyl_chr10_seqa	TGGGTGATTTTTTAGAGTT	
methyl_chr10_seqb	GATTTGTGGAAGATAGA	

Luminometric Methylation Assay (LUMA)

To analyze global genomic DNA methylation, LUMA was performed as previously described.[10]

Data availability: Methylome data are available at sftp://134.245.63.215/export/home/daudi (login: daudi; password: daudismethylome2012)

Supplementary Results

Supplemental Table 1: SNP filtering of Exome sequencing data (a) and potentially protein changing mutations detected by Exome sequencing and validated by Sanger Sequencing (b).

Supplementary Table 1a

SNPs on target	62,578
	↓
SNPs not in healthy controls*	25,752
	↓
SNPs involving a frequency in 1000 genomes pilot of max. 1%	18,822
	↓
SNPs not in dbSNP130	10,650
	↓
non-synonymous SNPs**	**2,313**

*Healthy controls are taken from published datasets[11,12], whereas we solely used the HapMap probes published by Ng *et al.*.[12] Additionally, we included exome data generated in-house. **Due to lack of germline control from the patients from which the DAUDI cell line has been established it is not possible to differentiate somatic (lymphoma-associated) mutations from germline variants.

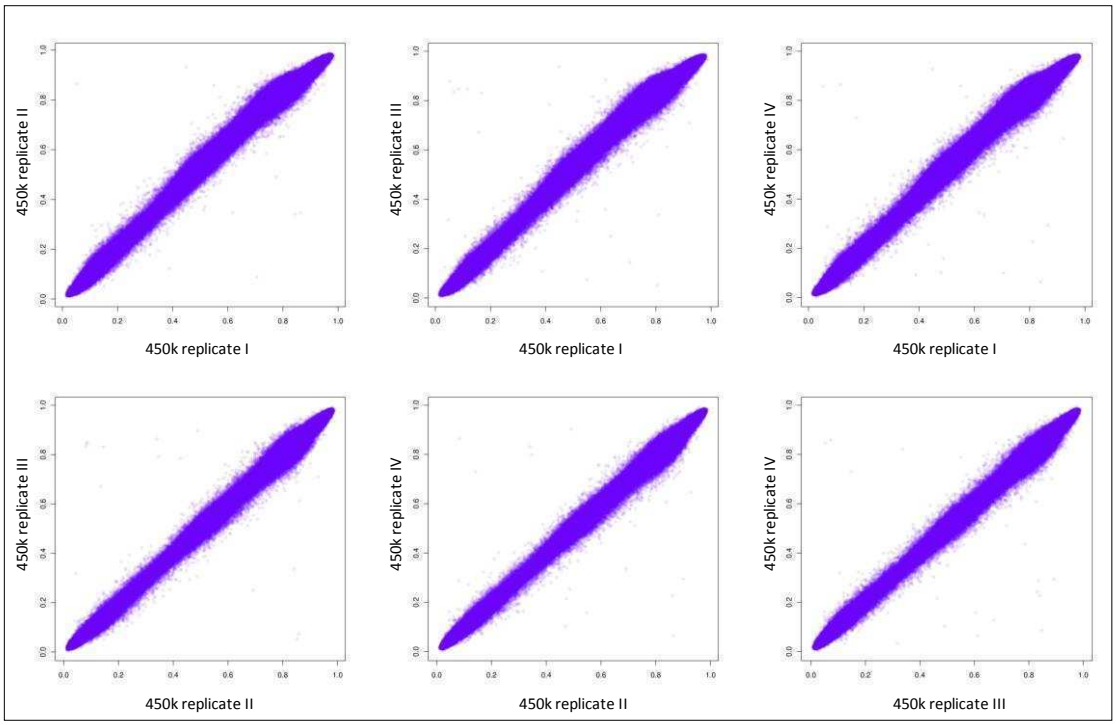
Supplementary Table 1b

Gene	chr	position	ref/ alt allele	Aa consequence	note
B2M	15	42791039	G/C	p.Met1Ile	Confirmed

TET2	4	106377217	C/T	p.Ser911Leu	Confirmed
ID3	1	23758264	G/A	p.Leu52Val	Confirmed
	1	23758345	G/C	p.Gln81*	Confirmed
KIT	4	55259372	C/T	p.Ala67Ser	Confirmed

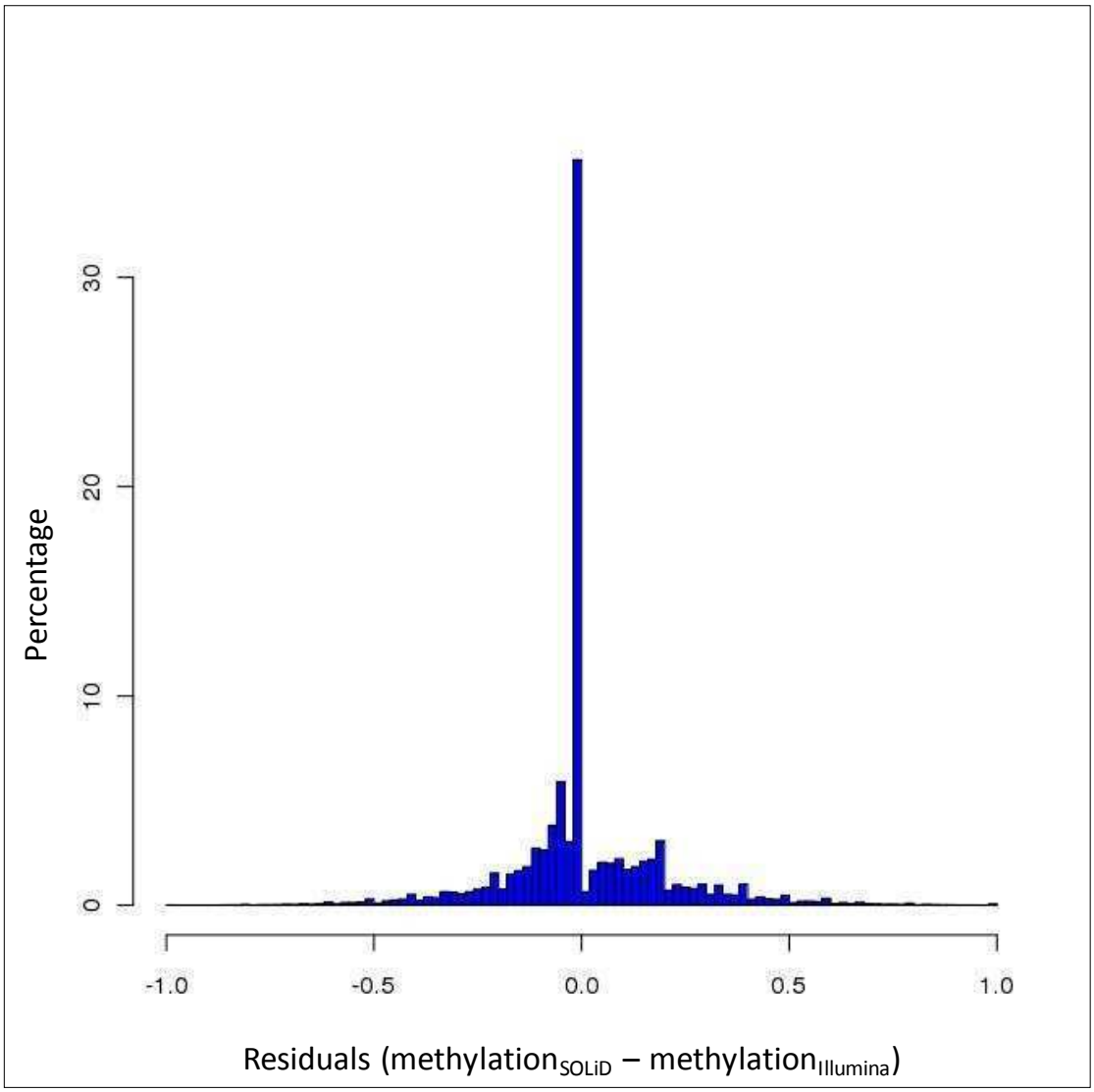
161

162



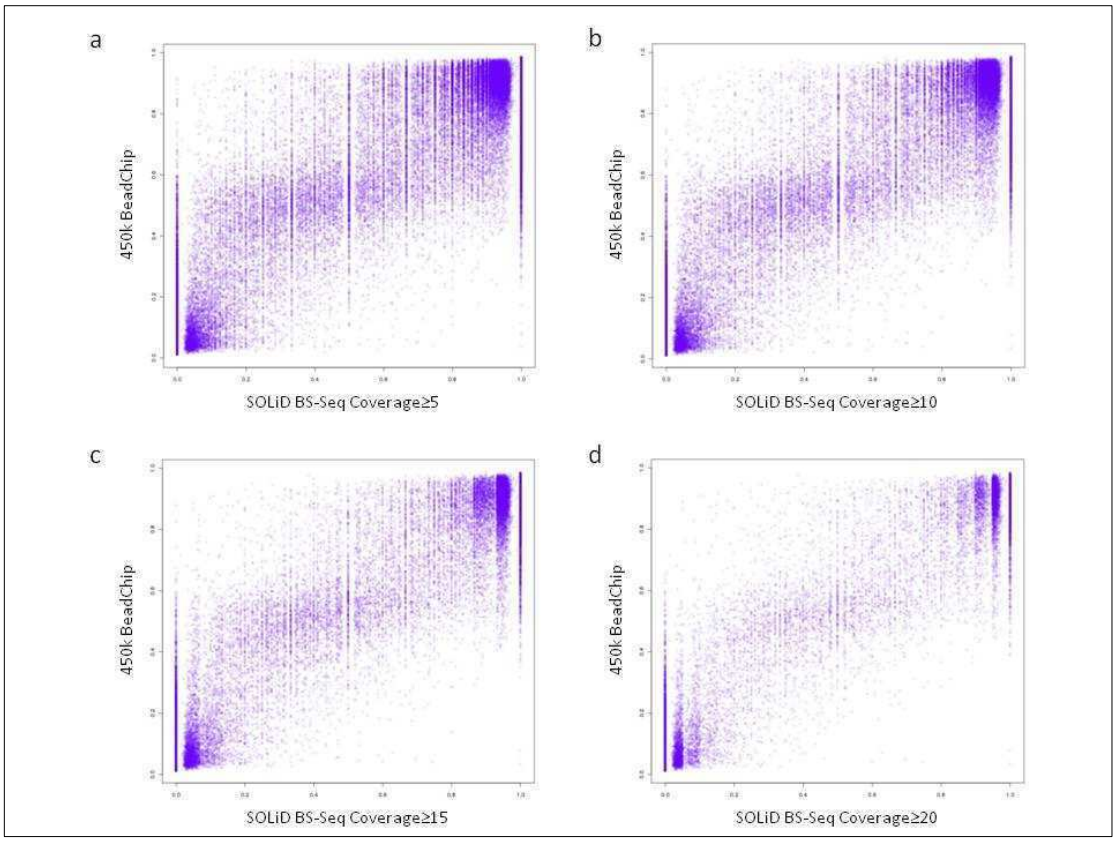
Supplemental Figure S1: Comparison of four HumanMethylation450 BeadArray replicates

Scatter plots depict the comparison of all four replicates of 450K runs of DAUDI cells. All of them show high correlations among each other r≥0.99.

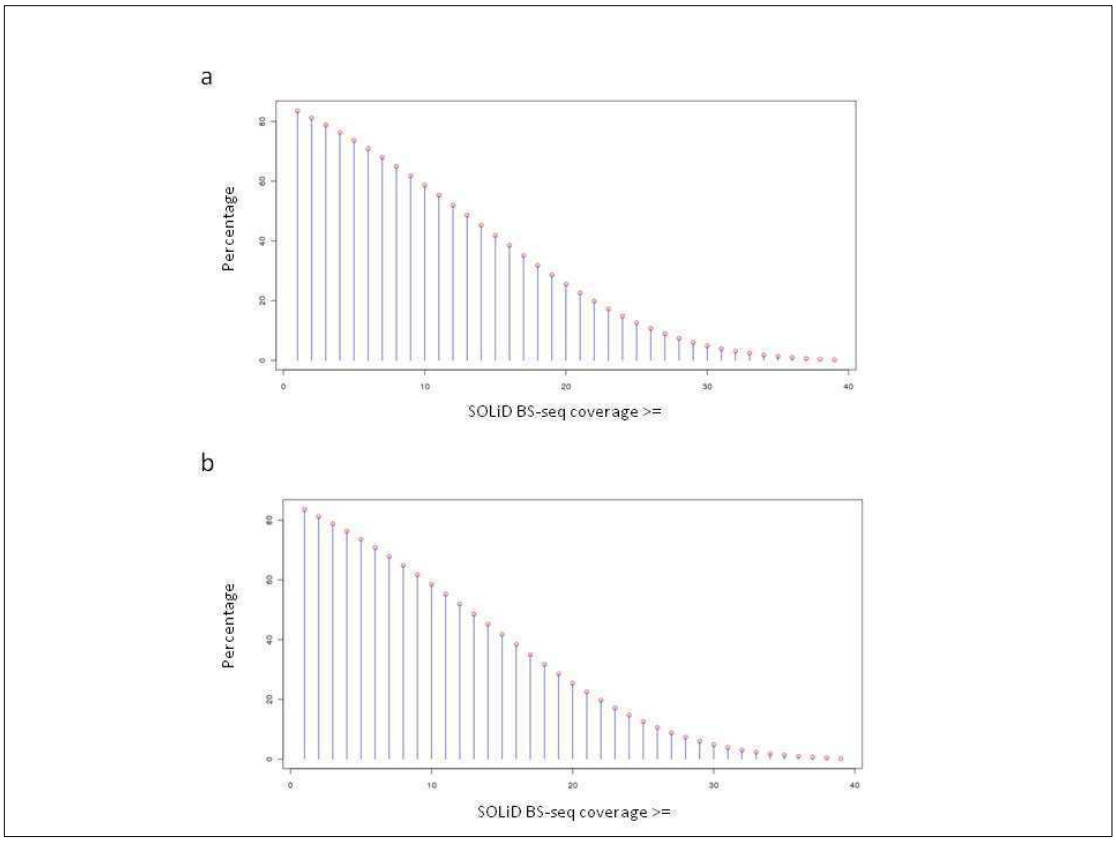


Supplemental Figure S2: Comparison of SOLiD™ and Illumina BS-seq

The histogram depicts the differences (SOLiD™ BS-seq – Illumina BS-seq) in methylation levels ((methylated reads)/(unmethylated reads + methylated reads)) for CpGs with coverage of at least 5 reads. Values close to 0.0 indicate that equal methylation levels were inferred by both methods. Both approaches show comparable genome-wide DNA methylation levels r=0.86.

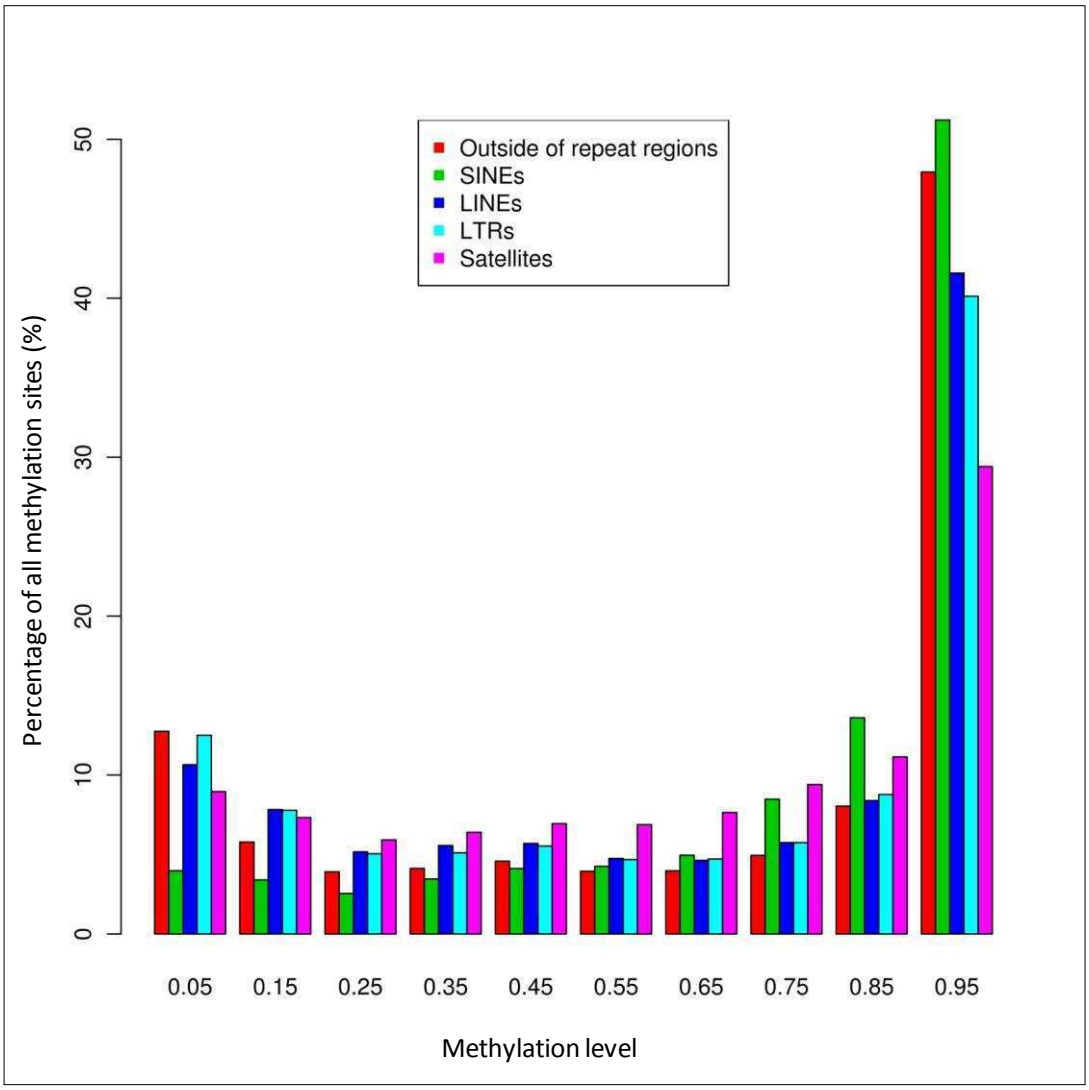


Supplemental Figure S3: Comparison of SOLiD™ BS-seq and HumanMethylation450 Beadchip data. The scatter plots depict the comparison of SOLiD™ BS-seq and 450k methylation levels. Higher read coverage generally results in better correlation. (a) includes SOLiD™ BS-Seq data with coverage≥5, (b) includes SOLiD™ BS-Seq data with coverage≥10, (c) includes SOLiD™ BS-Seq data with coverage≥15 and (d) includes SOLiD™ BS-Seq data with coverage≥20. All of them show high correlation r≥0.94.



Supplemental Figure S4: Coverage for SOLiD™ BS-seq

Percentage of uniquely aligned SOLiD™ BS-Seq reads to the human reference (hg19/NCBI 37) for the forward (a) and the reverse (b) strand. Both strand coverages are equally distributed.



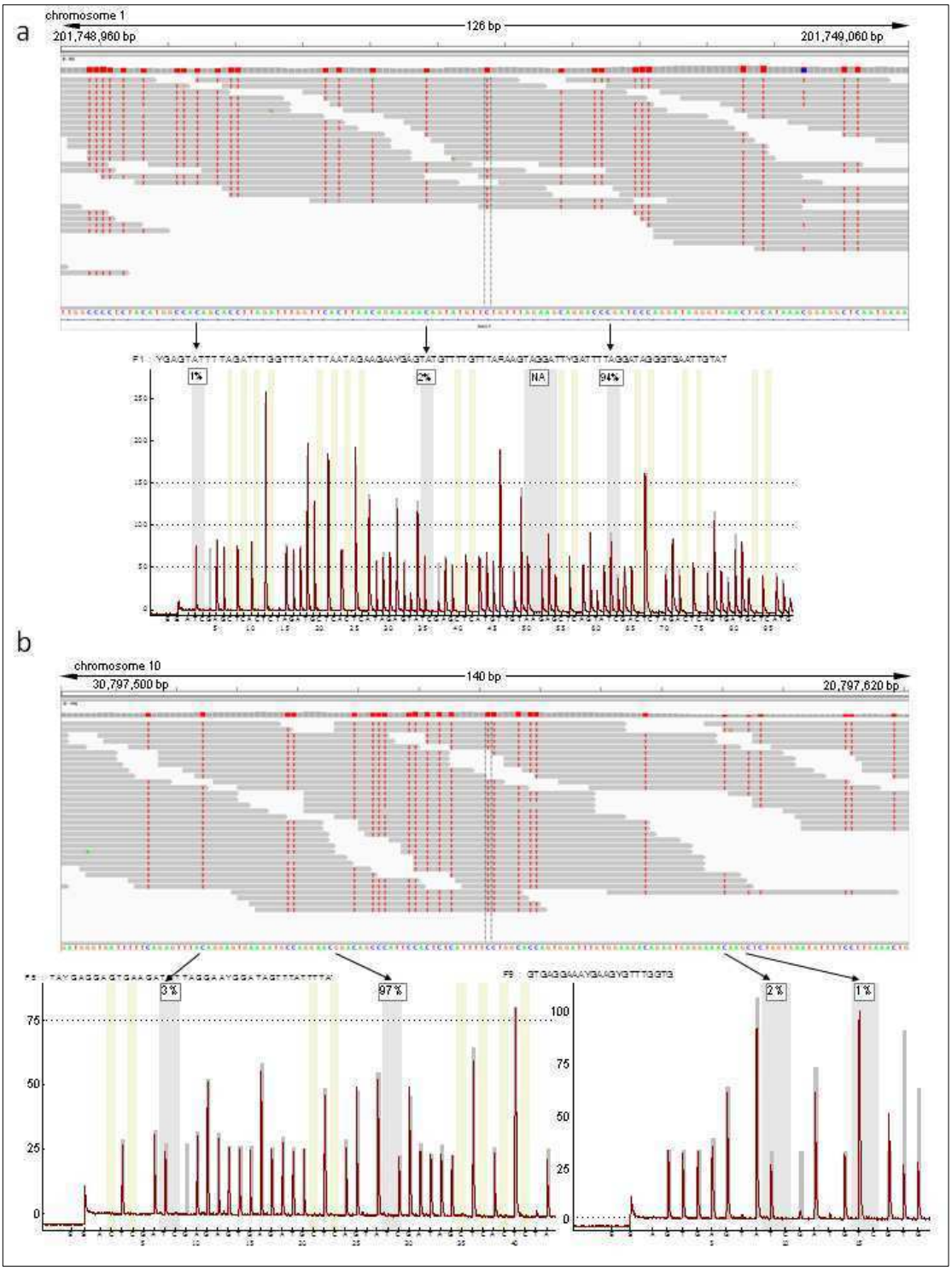
Supplemental Figure S5: Genome-wide comparison of DNA methylation for different annotated repeat regions. In total, 64.82% of all CpG sites within annotated repeat regions and 84.05% of all CpG sites outside of repeat regions were covered. Bar plots show distributions of DNA methylation levels within non-repeat regions, SINEs, LINEs, LTRs and Satellites. The y-axis indicates DNA methylation levels ((methylated reads)/(unmethylated reads + methylated reads)) assessed by SOLiD™ BS-seq.

MeanMethylation(Outside of repeat regions)=0.68

MeanMethylation(SINEs)=0.79, MeanMethylation(LINEs)=0.66

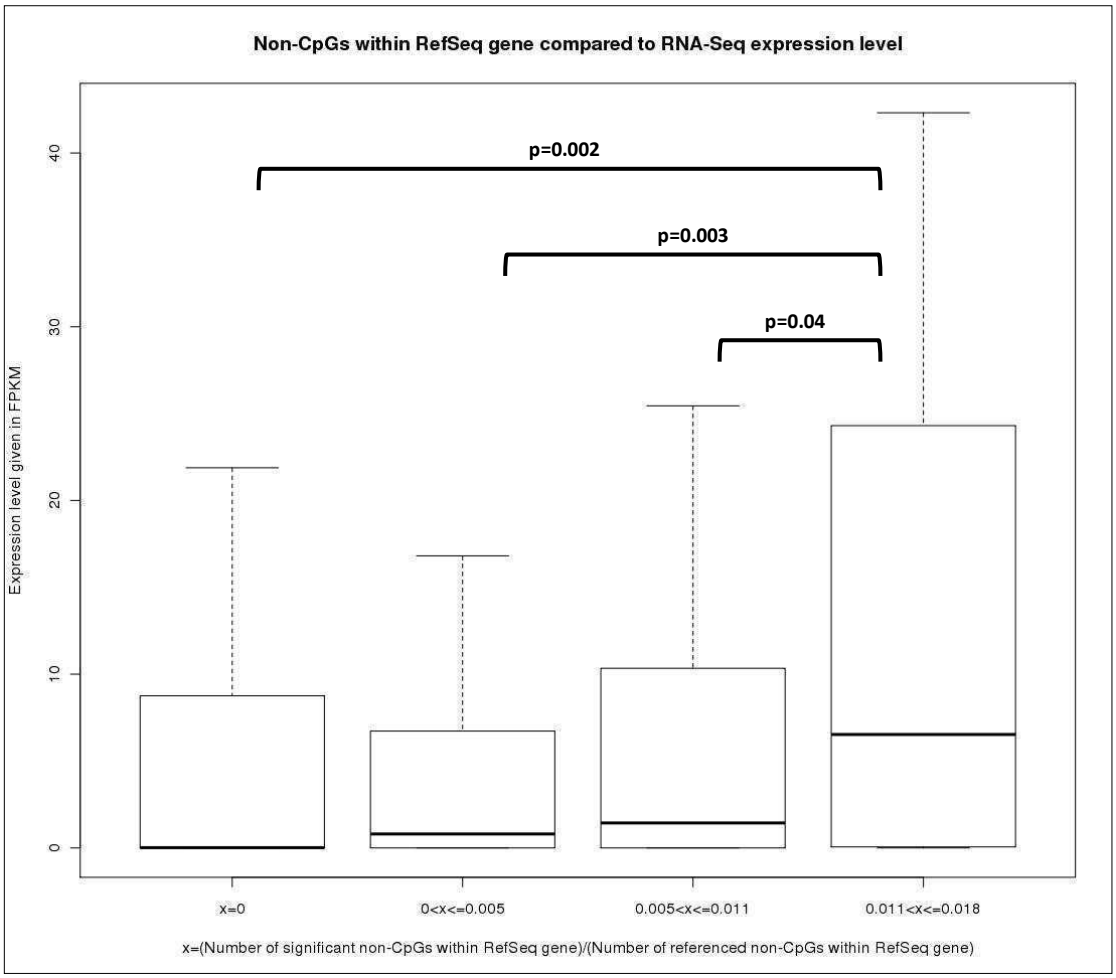
MeanMethylation(LTRs)=0.64, MeanMethylation(Satellites)=0.63

DNA methylation patterns on the forward and reverse strand were comparably established (Pearson r=0.90).



Supplemental Figure S6: Comparison of hallmark non-CpG sites showing significant non-CpG methylation in ESC cells[3] analyzed by BS pyrosequencing and methylome sequencing. No significant non-CpG methylation could be identified for hallmark sites on chromosome 1 (a) and chromosome 10 (b). For each region above the IGV plot of the bisulfite methylome data and below the BS pyrosequencing data are

shown. The results of pyrosequencing are indicated as percentiles above each potentially methylated position (colored in grey). In each region one CpG site was analyzed which showed near to complete methylation.



Supplemental Figure S7: Correlation analysis of significantly methylated non-CpG sites and transcriptional levels within RefSeq genes in DAUDI cells: RefSeq genes were clustered ($x=0$, $0<x\leq0.005$, $0.005<x\leq0.011$, $0.011<x\leq0.018$) by their fraction of significantly methylated non-CpG sites related to the number of referenced non-CpGs within the respective RefSeq gene. The y-axis depicts transcriptional levels measured by FPKM values. A modest positive correlation between significantly methylated non-CpG sites and respective transcriptional levels can be observed. Similar results could be observed for transcriptional levels assessed by the Affymterix U133A chip ($p_{(x=0,\ 0.011\leq x\leq0.018)}=0.005$, $p_{(0<x\leq0.005,\ 0.011\leq x\leq0.018)}=0.004$, $p_{(0.005<x\leq0.011,\ 0.011\leq x\leq0.018)}=0.15$).

Supplemental Tables S2: Hypermethylation in mature aggressive B-cell lymphoma

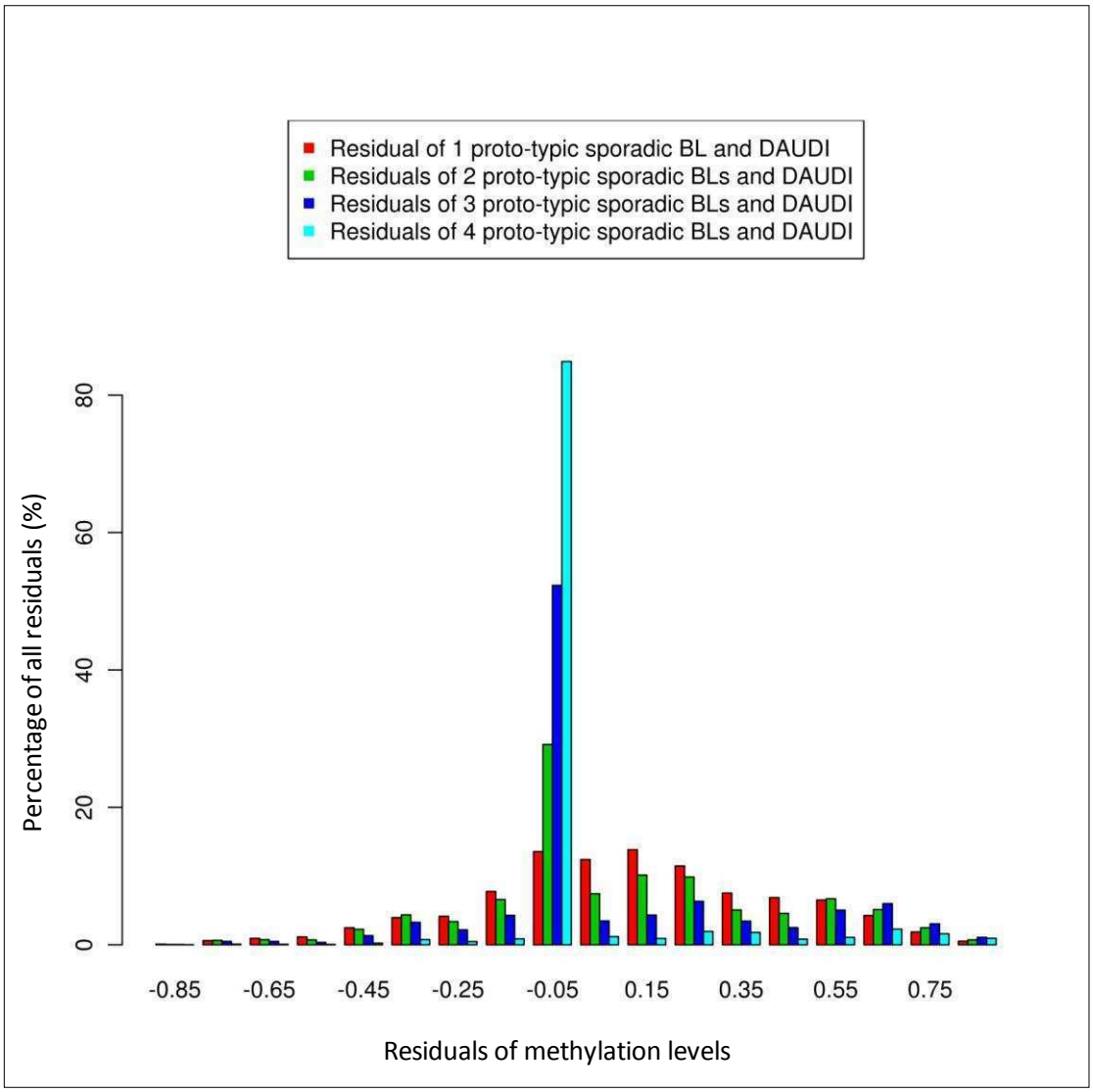
Comparison of genome-wide DNA methylation within RefSeq genes (Table S2a) and hypermethylated genes associated with mature aggressive B-cell lymphoma including BL (Table S2b).

Supplemental Table S2a

RefSeq genes	**1000 bp upstream**	**First exon**	**First intron**	**Second exon**	**Second exon to last exon**
Mean methylation (watson strand)	0.40	0.40	0.70	0.80	0.78
Mean methylation (crick strand)	0.41	0.39	0.69	0.81	0.78

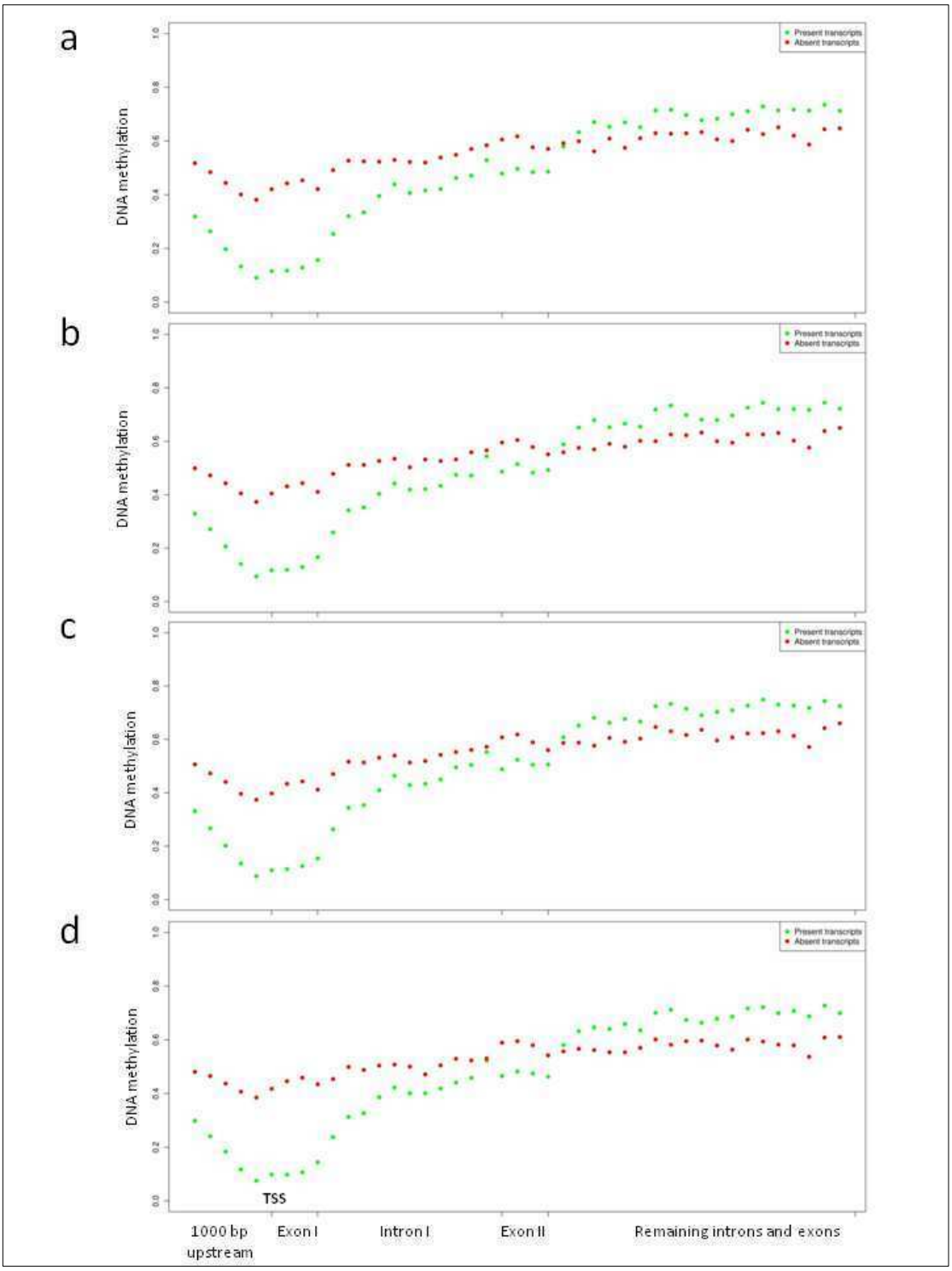
Supplemental Table S2b

Genes de novo methylated in mature aggressive B-cell lymphoma[13]	**1000 bp upstream**	**First exon**	**First intron**	**Second exon**	**Second exon to last exon**
Mean methylation (watson strand)	0.84	0.90	0.63	0.81	0.60
Mean methylation (crick strand)	0.84	0.90	0.66	0.84	0.63



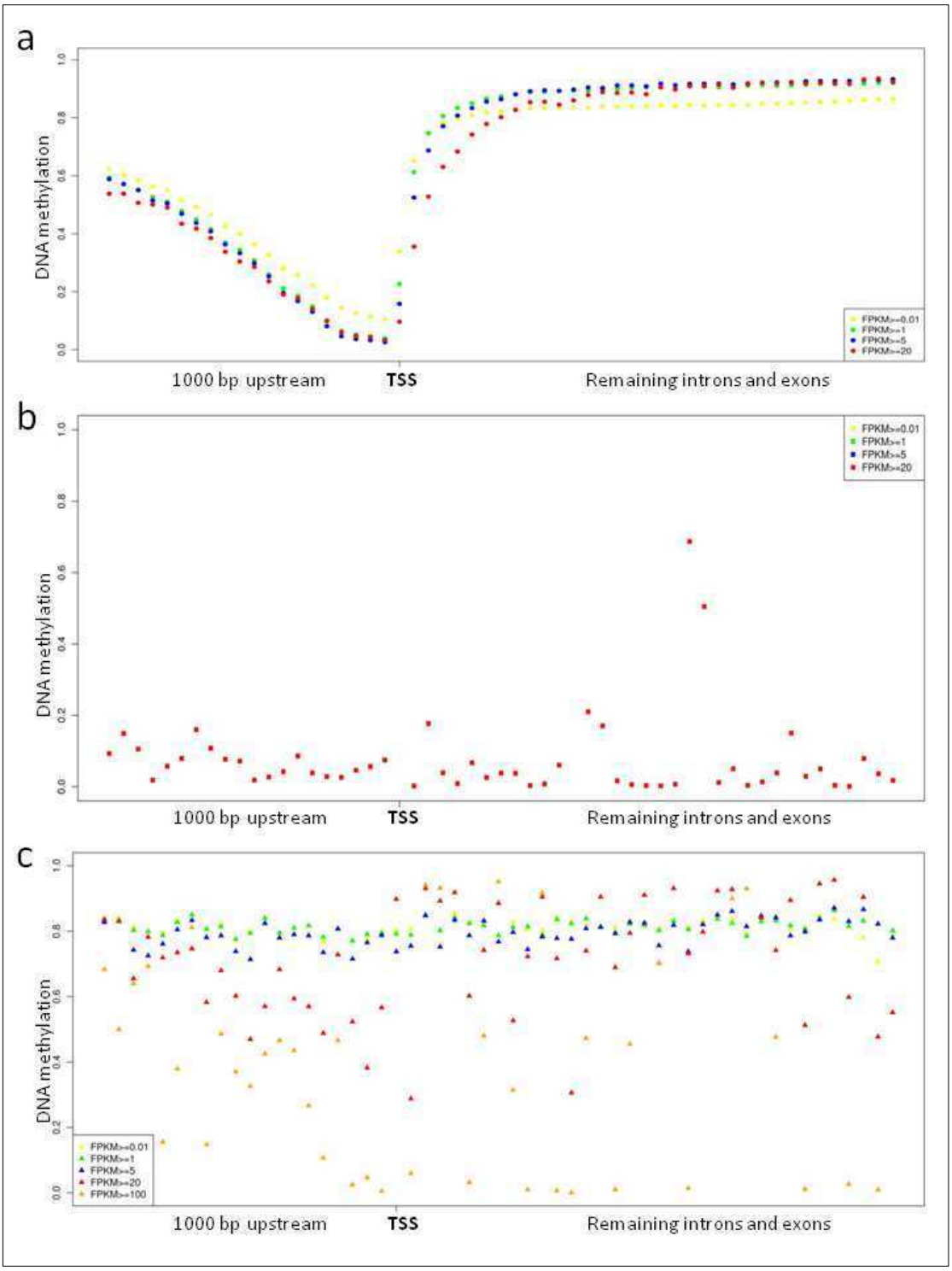
Supplemental Figure S8: Comparison of DAUDI and four primary proto-typic sporadic BLs based on 450K data. Barplots depict residuals (MethylationLevel(DAUDI)-MethylationLevel(primary proto-typic sporadic BLs)) of 450k data based on DAUDI and four primary proto-typic sporadic BLs. Residuals are clustered by their frequency i.e. red barplots involve CpG sites having the same residual (rounded to the next decimal place) for DAUDI and one BL, green barplots involve CpG sites having the same residual (rounded to the next decimal place) for DAUDI and two BLs, darkblue barplots involve CpG sites having the same residual (rounded to the next decimal place) for DAUDI and three BLs, lightblue barplots

249 involve CpG sites having the same residual (rounded to the next decimal place) for
250 DAUDI and four BLs. A high similarity of DAUDI and at least one primary proto-typic
251 sporadic BL can be observed.
252



Supplemental Figure S9: Analysis of the DNA methylation status in four (a-d) prototypic sporadic Burkitt lymphomas using Illumina 450K Methylation arrays with regard to presence and absence of transcription of the respective genes in DAUDI cells. The analysis shows that the pattern of methylation in the gene groups defined by expression in DAUDI cells is quite conserved also in primary sporadic BL with

259 expressed genes in DAUDI being characterized by absence of DNA methylation
260 around the TSS and increased methylation in the gene body.
261



Supplemental Figure S10: Methylation-expression correlation analyses

CpG Methylation levels were averaged for annotated RefSeq gene regions and transcripts are clustered by their expression level in present and absent calls. A strong dependency of the location of CpGs related to their distance to the TSS and

267 the transcript expression level can be observed for human data (a). Regarding the
268 mitochondria (b), we observed an overall low DNA methylation degree independent
269 from their expression levels. However, EBV (c) DNA methylation does not correlate
270 within transcripts until an expression level of FPKM≥15.

271

REFERENCES

1. Nadkarni JS, Nadkarni JJ, Clifford P, Manolov G, Fenyö EM, Klein E. Characteristics of new cell lines derived from Burkitt lymphomas. *Cancer* 1969; **23:** 64 - 79.
2. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009; **25:** 1754 - 1760.
3. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* 2009; **25:** 2078 - 2079.
4. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*. 2011; **43:** 491 - 498.
5. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 2010; 38: e164. Epub 2010 Jul 3.
6. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 2004; **5:** R80. Epub 2004 Sep 15.
7. Leenman EE, Panzer-Grümayer RE, Fischer S, Leitch HA, Horsman DE, Lion T, *et al.* Rapid determination of Epstein-Barr virus latent or lytic infection in single human cells using in situ hybridization. *Mod Pathol*. 2004; **17:** 1564 - 1572.

8. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 2009; **462:** 315 - 322.
9. Lamprecht B, Walter K, Kreher S, Kumar R, Hummel M, Lenze D, *et al.* Derepression of an endogenous long terminal repeat activates the CSF1R proto-oncogene in human lymphoma. *Nat Med* 2010; **16:** 571 - 579, 1p following 579.
10. Karimi M, Johansson S ET. Using LUMA: a Luminometric-Based Assay for Global DNA-Methylation. *Epigenetics* 2006; **1:** 45 - 48.
11. Li Y, Vinckenbosch N, Tian G, Huerta-Sanchez E, Jiang T, Jiang H, *et al.* Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nat Genet* 2010; **42:** 969 - 972.
12. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*. 2009; **461:** 272 - 276.
13. Martín-Subero JI, Kreuz M, Bibikova M, Bentink S, Ammerpohl O, Wickham-Garcia E, *et al*. New insights into the biology and origin of mature aggressive B-cell lymphomas by combined epigenomic, genomic, and transcriptional profiling. *Blood* 2009; **113:** 2488 - 2497.

Summary

The scientific contribution of this thesis consists of three articles that have been published in *Bioinformatics* (Oxford Journals) and *Nature Methods* and the third article being under review at *Leukemia* (Nature Publishing Group), respectively. The implications of these articles for the field of computational epigenetics and future perspectives of this research area are discussed.

The main challenge within the framework of this thesis was the development of a bioinformatics tool for bisulfite sequencing analysis. The article in *Bioinformatics* presents the bioinformatics tool B-SOLANA for the analysis of DNA methylation data generated by two-base encoding bisulfite sequencing on the SOLiD™ platform of Life Technologies. Additionally, benchmark analyses revealed that B-SOLANA exhibits a significantly higher sensitivity and specificity compared to other software approaches which were developed at the same time. The review article in *Nature Methods* summarizes challenges of bisulfite sequencing analysis as they appear on different high-throughput sequencing platforms. Especially primary analyses including the quality control and mapping of raw sequences are discussed. Furthermore, the article debates the effect of sequencing errors and contaminations on inferred DNA methylation levels and recommends the most appropriate way to analyze this type of data. This review is a helpful reference for the analysis of DNA methylation by high-throughput sequencing, a currently rapidly developing research area.

The third article, which has been submitted to *Leukemia*, comprises the analysis of a DNA methylome of the DAUDI cell line at single base resolution. On the genetic level, this endemic Burkitt Lymphoma cell line is characterized by the presence of the hallmark *IG-MYC* translocation. Recent publications about this cell line suggested a high number of DNA methylation changes. However, until now only array-based studies were published, which have concentrated their focus on loci-specific DNA methylation patterns. We showed that the mechanisms of DNA methylation associated with transcriptional regulation in lymphomas go by far beyond the usually studied promoter methylation. Furthermore, we characterized the DNA methylome of the mitochondria and the Epstein-Barr virus, whereas upregulation of the latter has already been identified in DAUDI before. As the DAUDI cell line is used over decades in many laboratories throughout the world, the obtained methylome data prove valuable as a “reference epigenome” for future studies.

Zusammenfassung

Die Grundlage dieser Dissertation bilden drei Publikationen, die in den Fachjournalen *Bioinformatics* (Oxford Journals) und *Nature Methods* erschienen sind, während der dritte Artikel zur Zeit bei *Leukemia* (Nature Publishing Group) begutachtet wird. Die Ergebnisse der beiden Veröffentlichungen und des eingereichten Manuskriptes werden in die aktuelle epigenetische Forschung integriert. Abschließend wird ein Ausblick auf mögliche zukünftige epigenetische Forschungsschwerpunkte gegeben.

Die Problemstellung dieser Arbeit war die Entwicklung eines bioinformatischen Programmes zur Analyse von Bisulfit-Sequenzierungsdaten der SOLiD™ Hochdurchsatztechnologie von Life Technologies. Dazu wurde das Programm B-SOLANA entwickelt, welches in *Bioinformatics* publiziert wurde. Benchmark-Analysen zu B-SOLANA und weiteren Programmen belegen die hohe Sensitivität und Spezifizität unseres Ansatzes. Der Artikel in *Nature Methods* fasst die Herausforderungen von Bisulfit-Sequenzierungsanalysen mittels unterschiedlicher Technologien zusammen. Insbesondere werden Primäranalysen zur Qualitätskontrolle und dem Mapping von Rohsequenzen diskutiert. Hierbei können Sequenzierungsfehler und Kontaminationen zu fehlerhaften DNA Methylierungsergebnissen führen. Wir erläutern Ansätze zur Detektion und Qualitätskontrolle dieser negativen Einflussfaktoren.

Der bei *Leukemia* eingereichte Artikel beinhaltet die Analyse des basengenauen DAUDI Methyloms, einer endemischen Burkitt-Lymphom Zelllinie. Genetische Eigenschaften, wie die *IG-MYC* Translokation, konnten bereits für DAUDI identifiziert werden. Aktuelle Forschungsergebnisse weisen auf epigenetische Eigenschaften dieser Zelllinie hin. Allerdings wurden bisher nur Array-Studien durchgeführt, die lediglich Lokus-spezifische DNA-Methylierungsmuster untersucht haben. Korrelationsanalysen zeigen, dass Gentranskription komplexen Methylierungsmustern innerhalb der Promotorregion unterliegt, welche mit array-basierten Studien nicht identifiziert werden können. Außerdem haben wir das Methylom der Mitochondrien und des Epstein-Barr Viruses, welches in DAUDI im besonderen Ausmaß vorliegt, analysiert. Das in diesem Artikel publizierte Methylom könnte als „Referenz-Epigenom“ für eine Zelllinie dienen, die in den letzten Jahrzehnten in vielen Laboren untersucht wurde.