**Dissertation**

**submitted to the**

**Combined Faculties for the Natural Sciences and for Mathematics**

**of the Ruperto-Carola University of Heidelberg, Germany**

**for the degree of**

**Doctor of Natural Sciences**

**Presented by**

**M.Sc. Lei Gu**

**born in: Ningbo, Zhejiang, China**

**Oral-examination: 2013-12-19**

# Integrative Analysis of Prostate Cancer Methylome and Smoking-induced Transgenerational Epigenomic Reprogramming

**Referees:**

**Prof. Dr. Roland Eils**

**Prof. Dr. Christoph Plass**

# Contributions

Chapter 2 was based on the publication Qi Wang*, **Lei Gu***, *et al,* 2013. Tagmentation-based whole genome bisulfite sequencing. ***Nat Protoc*** **8(10): 2022-2032 (*contributed equally).**

Chapter 3 was based on the publication Weischenfeldt, J., *et al* 2013. Integrative genomic analyses reveal an androgen-driven somatic alteration landscape in early-onset prostate cancer. ***Cancer Cell*** **23(2): 159-170.**

Chapter 4 was based on the manuscript **Lei Gu***, Christopher C. Oakes*, Ronald Simon*, *et al* 2013. BAZ2A links epigenetic remodeling and recurrence in prostate cancer. **(submitted) (*contributed equally).**

Chapter 5 was based on the manuscript **Lei Gu***, Tobias Bauer*, Saskia Trump*, Mario Bauer*, Qi Wang*, *et al* 2013. Environmentally induced epigenetic reprogramming in mothers and their newborn children. **(in preparation) (*contributed equally).**

These chapters contain suggestions and contributions from co-authors.

# Declarations

Declarations according to § 8 (3) b) and c) of the doctoral degree regulations:

a) I hereby declare that I have written the submitted dissertation myself and in this process have used no other sources or materials than those expressly indicated,

b) I hereby declare that I have not applied to be examined at any other institution, nor have I used the dissertation in this or any other form at any other institution as an examination paper, nor submitted it to any other faculty as a dissertation.

# Summary

Epigenetic factors such as DNA methylation, histone modification and noncoding RNAs are highly associated with early developmental processes, later environmental adaption and diseases development such as cancer. With the availability of current high throughput assays (microarray and next generation sequencing), one can already produce comprehensive picture of the epigenetic profile, especially the DNA methylome, in normal and tumor/diseased cells. However, managing and analyzing such vast datasets is challenging. In addition, interpretation of the observations from (epi)genetic information is also a limiting factor due to the lack of understanding epigenetic mechanisms and the interactions between genetic and epigenetic factors under environmental selection.

Thus, during my PhD studies, two pipelines were developed to process genome-wide methylation data generated by Methyl-CpG-immunoprecipitation sequencing (MCIP-seq) for the ICGC early onset prostate project and whole genome bisulfite sequencing (WGBS) for the environment induced transgenerational epigenetic remodeling project. The WGBS pipeline was adjusted later for a modified WGBS protocol, tagementaion-based WGBS, which allows to investigate the whole methylome (around 27 million CpGs) at single base resolution by using only 10-20 ng of input DNA compared to 3-5 ug required for traditional WGBS.

Developing these computational tools, provided an opportunity to look closely at methylation changes in prostate cancers. With an integrative meta-analysis of public prostate (epi)genomic data and a large cohort of 7682 prostate cancer specimens, *BAZ2A* was found to be overexpressed in a large subset of prostate tumors that are characterized by early post-operative PSA recurrence and high tumor grades. In multivariate analyses, *BAZ2A* was found

to be an independent factor predicting recurrence. Furthermore, high levels of *BAZ2A* were tightly associated with a distinct molecular subtype demarked by aberrant genome-wide DNA methylation and elevated numbers of genetic alterations suggesting a CpG island-methylator phenotype (CIMP) to selectively occur in *BAZ2A*-upregulated tumors. In summary, this study showed the clinical impact of *BAZ2A* as a key epigenetic regulator linking aberrant DNA methylation and outcome in prostate cancer.

In addition, epigenetic changes is not only important for the diseased individuals including cancer, but also for the healthy individuals to adapt the external environmental stimulus such as smoking. In order to investigate the interaction between the methylome and environmental factor in a human prospective mother-child study at single base resolution, tobacco smoke-induced changes to epigenetic programming during the prenatal period was studied by WGBS and targeted methylation analysis. In mothers and children a distinct, genome-wide epigenetic response is induced. While mothers showed a genome-wide hypomethylation profile, children revealed tobacco-smoke induced hyper- and hypomethylation. By focusing on chromatin regulators, differential DNA methylation with functionally deregulated histone modifiers was linked, which together induce epigenetic reprogramming upon exposure to tobacco smoking. Together with the observed deregulation of a number of disease related pathways, the identified aberrant DNA methylation was suggested as a possible molecular mechanism linking between prenatal exposure and disease outcomes later in life.

In summary, comprehensive epigenomic analyses were performed on both diseased and healthy individuals in order to shed a light on how epigenetic factors influence the tumor development and interact with external environmental stimulus.

# Zusammenfassung

Epigenetische Faktoren wie DNA-Methylierung, Histonmodifikation und nicht-kodierende RNS sind stark assoziiert mit Prozessen der Frühentwicklung, der späteren Anpassung an Umwelteinflüsse oder der Krankheitsentwicklung wie z.B. Krebs. Moderner Hochdurchsatzmethoden (Microarrays und Tiefensequenzierung) ermöglichen eine ganzheitlicheres Bild epigenetischer Profile, insbesondere des DNA-Methyloms, in normalen und erkrankten oder Tumorzellen. Die Analyse solch riesiger Datensätze stellt allerdings eine besondere Herausforderung dar. Die Interpretation (epi)genetischer Information ist aufgrund mangelndem Verständnis epigenetischer Mechanismen und den Interaktionen zwischen genetischen und epigenetischen Faktoren im Bezug auf Umwelteinflüsse ebenfalls ein limitierender Faktor.

Daher wurden während meiner Promotionsarbeit zwei Pipelines entwickelt: zum einen für Sequenzierungen aus Methyl-CpG-Immunopräzipitationen (MCIP-seq) des ICGC Prostataprojekts und zum zweiten für genomweite Bisulfitsequenzierungen (WGBS) zur Analyse umweltbeeinflusster, generationsübergreifender epigenetischer Remodellierung. Die WGBS Pipeline wurde im Verlauf an ein modifiziertes, auf *tagmentation* basierendes WGBS-Protokoll angepasst, das die Untersuchung des Gesamtmethyloms (ca. 27 Mio CpGs) auf Einzelnukleotidebene ermöglicht mit nur 10-20 ng DNS-Materialbedarf im Vergleich zu 3-5 µg der herkömmlichen Methoden.

Die Entwicklung computergestützter Methoden bot die Gelegenheit zur detaillierten Untersuchung von Methylierungsveränderungen in Prostatatumoren. Mittels einer integrativen Metaanalyse publizierter (epi)genetischer Prostatadaten und einer großen Kohorte von 7682

Prostatatumorproben wurde ermittelt, dass das Gen *BAZ2A* überexprimiert wird in einem großen Anteil der Prostatatumore, für frühes Wiederauftreten postoperativen PSAs und ein hoher Tumorgrad charakteristisch ist. *BAZ2A* erwies sich in multivariater Analyse als unabhängiger prädiktiver Faktor für das Wiederauftreten. Des Weiteren ist ein hohes *BAZ2A*-Niveau eng assoziiert mit einem ausgeprägten molekularen Subtypen, der sich abgrenzt durch aberannte genomweite DNA-Methylierung und erhöhte Anzahl genetischer Veränderungen, was darauf hindeutet dass ein sog. *CpG island methylator* Phänotyp (CIMP) selektiv in *BAZ2A*-hochregulierten Tumoren auftritt. Zusammenfassend zeigt diese Studie die klinische Bedeutung von *BAZ2A* als Schlüsselfaktor epigenetischer Regulation, der die aberrante DNS-Methylierung mit dem klinischen Verlauf von Prostatatumoren verbindet.

Epigenetische Veränderungen sind nicht nur wichtig für Personen mit Erkrankungen wie Krebs sondern auch für gesunde Individuen bei der Anpassung an externe Umwelteinflüsse wie beispielsweise das Rauchen. Der Einfluss von Umweltfaktoren (Tabakrauch) auf das Methylom auf der Einzelnukleotidebene wurde in einer langfristigen ausgelegten Mutter-Kind-Studie in der pränatalen Phase mittels WGBS untersucht. Sowohl bei Müttern als auch den Kindern wird ein individuelles Methylierungsmuster durch das Rauchen induziert. Während bei den Müttern ein genomweites Hypomethylierungsprofil sichtbar wurde, zeigten sich bei den Kindern sowohl Hyper- als auch Hypomethylierung. Durch Fokus auf Chromatinregulatoren konnte eine Verbindung zwischen differentieller DNA-Methylierung und funktionell deregulierten Histonmodifizierern hergestellt werden, durch die eine epigenetische Reprogrammierung in Folge des Rauchens induziert wird. Die Deregulation einer Reihe von krankheitsrelevanten Signalübertragungswegen zusammen mit den beobachteten aberranten DNA-Methylierung deutet hin auf einen möglichen molekularen Wirkungsmechanismus zwischen pränataler

Exposition und einer Krankheitsentwicklung im späteren Leben.

In der Zusammenfassung wurden epigenomische Analysen durchgeführt sowohl auf erkrankten wie gesunden Personen, die zur Aufklärung beitragen, wie epigenetische Faktoren die Tumorgenese beeinflussen oder auf Umwelteinflüsse reagieren.

# 摘要

表观遗传因子，比如 DNA 甲基化，组蛋白修饰以及非编码 RNA 等，已经被揭示与早期的发育过程，对环境的适应过程以及各种疾病（包括癌症）的发生都有密切关联。 随着各种高通量技术的发展与成熟（从微阵列到第二代测序技术）我们现在已经可以得到病变细胞和正常细胞的一幅比较完整的表观遗传图谱，尤其是 DNA 甲基化图谱。但是，对于这类高通量数据的有效管理和分析还存在很大的挑战。不仅如此，由于我们对表观遗传调控的机理和它在各种环境选择下与遗传规律之间的互作的了解甚少，导致我们在面对这些数据时能做出的解释也变得很有限。

所以，在我的博士期间，我建立了两套软件流程用来分析在国际癌症基因组项目中产生的全基因组 DNA 甲基化数据，以及在另一个关于吸烟诱导的可传代表观遗传重编程的课题中产生的数据。后来我又改良了全基因组重亚硫酸盐测序的方案，传统的技术则需要 3-5 微克的 DNA，而改良的版本只需要 10-20 纳克的 DNA 就能得到全基因组的甲基化数据。

在开发相关的技术和软件的过程中，使得我有机会研究前列腺癌的全基因组的甲基化的变化过程。根据对已经存在的公共数据的综合分析（包括）以及对我们在汉堡的合作单位提供的 7682 个样本的检测，我发现 BAZ2A 这个基因在一大批前列腺癌症样本中过度表达， 而这批前列腺癌样本基本都具有高复发和高的肿瘤等级的特征。在多变量分析中，我证明了 BAZ2A 可以作为一个独立的诊断预测标记用来预测前列腺癌的复发。而且，这批具有 BAZ2A 过度表达的前列腺癌样本还具有非常高的全基因组范围的拷贝数变异。这些样本不仅在遗传机制层面和其他 BAZ2A 正常表达的癌症样本有显著差异，而且在表观遗传层面，也就是 DNA 甲基化水平上也呈现出 2 种不同的模式。所以我们认为这个基因可以把前列腺癌更进一步分为 2 个亚型，一个是 BAZ2A 过度表达，所以导致后期的复发和扩散，而另一个亚型是 BAZ2A 正常表达，所以比较癌症比较不易复发和扩散。

另外，我的另一个课题是关于研究母亲在怀孕期间吸烟，是否对新生儿

的基因组有影响，并且能否体现在 DNA 的甲基化水平上。我们分别对吸烟母亲和不吸烟母亲已经她们的新生儿产生了一批精度能达到单个核苷酸水平的甲基化测序数据。这是第一次在人类样本中获得此类数据。综合组蛋白编码分析和转录组分析，我们发现吸烟对母亲和新生儿产生了完全不同的影响，具体表现在完全不同的全基因组 DNA 甲基化变化上。并且我们还观察到组蛋白活性也有显著的差异。所以我们提出一个假说，吸烟能通过改变基因组的 DNA 甲基化来影响人的后期生活的各种疾病的易感性。

# Acknowledgements

This is one of the most difficult part in this thesis because there are many people who helped me and are still helping me at this time point. It's impossible for me to start, work on and finish the PhD project without all of you. Every word here is not enough to express my deep gratitude.

First, I would like to thank Prof. Christoph Plass and Prof. Benedikt Brors who offered me the joined position so that I can learn from both groups. I particularly thank Prof. Christoph Plass to guide me to step into the exciting field of epigenetics and trained me in many skills including how to generate hypotheses from public dataset, how to search for supports and resources to test and validate hypotheses, and finally how to design and write a scientific paper. This systematic training will definitely help me throughout my later scientific career.

Second, I would like to thank Prof. Roland Eils to give me the opportunity to explore the fascinating human transgenerational epigenetic reprogramming study which has never ever been touched by any other group before. I enjoy a lot working with you and learned a lot from this project. I also very much appreciate all your supports to the BAZ2A project. This project will never ever come to this point without you. And I really enjoy the eilslabs gatherings and eilslabs retreats where I enjoy not only the terrific scientific communications but also the great fun from each social activity. I will never forget the Scottish group dancing, dragon boat race and the football match between Germany and Netherlands in the last European Cup and so on and so on...

Third, my special thanks go to Dr. Dieter Weichenhan who is one of my best friends during my PhD study. You always try your best to support me not only in the scientific field but also in the daily life. I am sure we will never forget the wonderful time in Kyoto where we sit on a river levee, drink Japanese

beers, share opinions from politics to football, from religion to science, talk everything we can imagine. And we also have a very nice cooperation which in turn comes up with a paper published in Nature Protocols.

Fourth, I would like to take this chance to thank all members in C010 and B080. Dr. Khelifa Arab, Dr. Christopher Oakes, Dr. Dieter Weichenhan， Dr. Rainer Claus, Dr. David Scherf, Hanna Jacobsson, Christopher Schmidt, David Brocks, Oliver Mücke, it's so nice to play football with your guys every week. Thanks Dr. Christopher Oakes for many useful discussions for the research itself as well as the current severe scientific environment. And of couse thanks to recruit me into the attractive poker night. Thanks Dr. Qi Wang and Dr. Zuguang Gu, it's nice to work with you and learn from you, especially the R programming skills from Dr. Zuguang Gu. Thanks Dr. Aoife Ward, Dr. Christopher Oakes, Dr. Dieter Weichenhan for the proofreading of my thesis. Thanks Dr. Tobias Bauer and Christoph Weigel for the help on German summary.

Fifth, I would like to thank Prof. Li Jin who brought me into science in the very beginning and Prof. Andreas Dress who recommended me to come to Germany. It's hard to imagine how I can start my scientific study at that time if without your supports.

Finally, I cannot thank my family enough. 我想特别感谢我的父母，我的妻子，我的外公外婆以及舅舅，感谢你们一直对我的支持，让我做自己喜欢的事情，一直对我有信心。自从大学开始离开宁波，后来又到上海中科院，到现在的德国，已经有差不多 10 年很少陪伴在你们身边，过了这个春节，又要去哈佛医学院继续我的科研工作，肯定还是聚少离多，千言万语难以表达我对你们的感激。 我为能拥有这样的家人而感到幸福，希望你们每一天都能健康，快乐！

# Contents

# List of Figures

# List of Tables

# Chapter 1: Introduction

## 1.1 Epigenetics

The term, epigenetics, is derived from the word epigenesis. The early embryo is undifferentiated. As development proceeds, increasing levels of complexity emerge giving rise to the larval stage or to the adult organism. In 1942, Conrad Waddington introduced the term epigenetics, which was defined as "the branch of biology which studies the causal interactions between genes and their products, which bring the phenotype into being"[1].

The consensus definition of epigenetics nowadays is non-sequence dependent inheritance. The most important features of epigenetics are thought to be heritable and reversible[2,3]. This phenomenon was first described in plants and has been expanded to yeast, Drosophila, mouse and, possibly, humans[4-7].

Epigenetic mechanisms mainly include DNA methylation, histone modifications, chromatin remodeling and regulation of non-coding RNAs. Epigenetic processes are heavily involved in diverse biological functions, such as genomic imprinting, X-chromosome inactivation, stem cell differentiation, tissue/organ regeneration and aging. Aberrations of epigenetic processes are found in many diseases, including cancer, cognitive dysfunction, and cardiovascular, reproductive, autoimmune, and neurobehavioral disorders.

DNA methylation is one of the most important and best characterized epigenetic processes which involves the addition of a methyl group to a cytosine. In mammals and other vertebrates, nearly all DNA methylation occurs at the cytosine in the context of CpG dinucleotides[8]. The DNA methyltransferase (DNMT) gene family plays a critical role in mediating DNA methylation[9]. Methylation of DNA is catalyzed by three members of the DNA

methyltransferase family including DNMT1, DNMT2 and DNMT3[10]. The maintenance methyltransferase, DNMT1, adds methyl groups to hemi-methylated DNA during DNA replication[11,12]. DNMT2 has been reported to catalyze RNA methylation[13,14]. The DNMT3 subfamily has three members: DNMT3A, DNMT3B and DNMT3L. DNMT3A and DNMT3B are responsible for the methylation pattern establishment without a template during embryonic development[15,16]. DNMT3L is thought to enhance the activity of DNMT3A and DNMT3B[17,18]. Genomic regions with at least 50% CG content and a ratio of observed CpGs to expected CpGs larger than 0.6 are known as CpG islands[19] which comprise of normally unmethylated CpGs are located in around 60% of human gene promoters and correlate with transcriptional regulation[12,20-24]. A small proportion of CpG islands are methylated during developmental processes involved in genomic imprinting and X chromosome inactivation[25]. De novo methylation is active in germ cells or early embryo stages[26]. A large fraction of highly methylated CpGs are found in repetitive sequences which is needed to maintain genomic stability by preventing the activation of mobile elements[27,28]

Histone modifications contribute another important epigenetic alteration. Chromatin is the complex of histones and DNA that forms the scaffold for nuclear processes including transcription, replication and DNA repair[29]. Nucleosomes are the basic units of chromatin, consisting of a segment of DNA (147bp) wrapped around an octamer of histone protein cores (H2A, H2B, H3 and H4). The amino-terminal of the histone proteins has a flexible tail which is conserved among species and is subject to different post-transcriptional modifications. There are at least eight kinds of modifications: acetylation, methylation, phosphorylation, ubiquitination, sumoylation, ADP ribosylation, deimination and proline isomerization[30-32]. All of these modifications form a set of combinations known as the "histone code" which act as markers that can be

read by other proteins to control the expression, replication, DNA repair, alternative splicing and chromosome condensation, which leads to distinct cellular outcomes[33-37]. The histone code may be heritable. There are two forms of chromatin. One is heterochromatin which is a condensed form and is characterized by a low level of acetylation and high levels of H3K9, H3K27 and H4K20 methylation which generally codes for transcriptional repression[38]. The other form is called euchromatin which has a looser structure and is often characterized by overall high levels of histone acetylation and trimethylated H3K4, H3K36 and H3K79 and thus provides the environment for active transcriptional processes[39-41]. Mounting evidences have suggested that histone modifications and histone-modifying complexes play critical roles in cellular processes and human cancer development. Furthermore, the dynamic regulation of histone modifications may have the potential to be molecular targets for human cancer treatment.

Non-coding RNAs (ncRNAs) are RNAs transcribed normally, but are not translated into proteins. Long non-coding RNAs (typically > 200 nt) have been implicated in variety of biological functions[42]. During the last few years, more and more epigenetic control systems have been found to be mediated by long non-coding RNAs[43-45]. X-chromosome inactivation[46-52] and genomic imprinting[53-56] are two classical systems mediated by long non-coding RNAs which have been known for many years. However, the details of how these long non-coding RNAs are generated and regulated are still largely unknown. In summary, all these non-coding RNAs form a network to not only spatially but also temporally regulate transcriptional activity.

## 1.2 Cancer Epigenetics

### 1.2.1 Overview of cancer epigenetics

Cancer has been defined as a complex disease with both genetic and epigenetic components. Many genetic driver mutations have been found by sequencing efforts in large patient cohorts for many cancer types. In recent years there has been a growing interest in the rapidly advancing field of cancer epigenetics and the interplay between genome and epigenome[57].

Historically, there are three main models which address the origin of cancer defined in the early 1970s[58]. One model considered cancer as a disease of abnormal differentiation[59]. The second model suggested that cancers are caused by viruses, such as avian sarcoma virus[60,61]. The third model pointed out that cancer is a result of an accumulation of mutations[62]. Actually, the abnormal differentiantion is probably coupled by the two others. Thus, it might explain better when all three models are integrated into a single framework. Later, Kundson's two-hit model was proposed. One classical example for Kundson's hypothesis is the *Rb-1* locus in retinoblastoma[63]. Numerous oncogenes and tumor suppressor genes were then identified in the following years[64]. However, mutations do not account for all alterations found in cancers. Later it was found that non-mutational (epigenetic) activation and inactivation of oncogenes and tumor suppressor genes were frequently observed in cancers[65-67]. Thus, epigenetic mechanisms are proposed to be highly responsible for a significant portion of the alterations in cancer initiation, development and metastasis[68-70].

In general, focal promoter hypermethylation and global hypomethylation are two patterns that play an important role in many cancer types. Changes to methylation do not only occur in CpG islands, but also in the peripherial

regions called CpG shores which are shown to possess a high degree of tissue-specific variation in DNA methylation[71]. Loss of imprinting (LOI), represented by biallelic expression or silencing of the imprinted allele, is another type of methylation change occurring in almost all tumor types[72-74] and is currently considered as the most common early event in cancer[75]. In histone modifications, loss of monoacetylation and trimethylation of H4 appear early and accumulated during the tumor development[76]. This pattern has been observed in many other cancers[77,78] and has been considered as a common cancer hallmark like global hypomethylation and CpG island promoter hypermethylation.

### 1.2.2 Prostate cancer epigenetics

One of the major cancers of older men is prostate cancer. Epigenetic alterations have been documented in most of human cancer development and progression. In prostate cancer, genes silenced by promoter hypermethylation are involved in DNA repair, apoptosis, cell cycle control, steroid hormone response and metastasis[79]. One of the best characterized genes is *GSTP1* which is consistently hypermethylated in the promoter region in the early stage of prostate tumorigenesis[80-84]. Similarly, the negative regulator of the Ras signaling pathway, *RASSFIA*, is also commonly downregulated by promoter hypermethylation[79,85,86]. Methylation levels of *APC* are highly related to biochemical recurrence in some prostate cancer studies[87-89]. In addition, polycomb target genes are preferentially hypermethylated in prostate cancer[90]. Global hypomethylation affecting repetitive elements has been observed in various cancer types[91,92]. In prostate cancer, hypomethylaton is likewise associated with progression rather than initiation, thus it is usually associated with advanced tumor stages[93].

Besides, long non-coding RNAs are also contributed to the development of prostate cancer. For example, PTENP1 is a pseudo gene of the tumor suppressor gene PTEN and upregulate PTEN expression by binding to microRNAs that downregulate PTEN transcription[94]. Additionally, a recent study reported that two long non-coding RNAs, PRNCR1 and PCGEM1, enhance the androgen receptor associated transcriptional programs to promote the growth of prostate cancer[95].

### 1.2.3 CpG island methylator phenotype

The CpG island methylator phenotype (CIMP) was first identified in colorectal cancer[96]. With the help of high throughput technology, it now refers a phenomenon that an exceptionally high frequency of CpG island hypermethylation occurs in a subset of tumors which suggests a potential epigenetic defect in this tumor subgroup. Later, this term was repeatedly used over the last several years in other tumor types including glioma[97], breast[98-100], renal[101] and gastric cancers[102-104]. But for others, such as ovarian cancer[105], no CIMP was identified. It was shown that CIMP is usually highly associated with clinical and pathological outcomes and thus is useful for the classification of prognosis in varies tumor types. Several studies suggested a third group of CIMP in colorectal cancer, namely, CIMP-high and CIMP-low. Although CIMP-low colorectal tumors have repeatedly been associated to KRAS mutations, this subgroup has many common clinical and pathologic features with non-CIMP colorectal tumors. Thus, no significant evidence could demenstrate that this is a distinct phenotype so far.

One significant feature of CIMP is that it is tightly linked to somatic mutations, such as mutations of the *BRAF* oncogene[106] in colorectal cancer, mutations of the *IDH1* gene in glioblastoma[107] and mutations of the *TET* gene

in leukemia[108]. So far, the best characterized CIMP is the *IDH1* defined G-CIMP. Mutatant *IDH1* can catalyzes the reduction of α-ketoglutarate to 2-hydroxyglutarate (2-HG) which is a potential oncometabolite[109-112]. Then, 2-HG can inhibit the TET family that convert 5mC to 5-hydroxyl-methylcytosine (5hmC) via direct competition with α-ketoglutarate which leads to an accumulation of 5mC and therefore influences the transcription of many genes. However, despite a clear rationale for the association of *IDH1* mutation with G-CIMP, the molecular mechanism of CIMP is still not fully understood for almost all tumor types with CIMP identified and will remain an active area of investigation. With the help from the varies kinds of genome-wide analysis, the causal relationship between somatic mutations in chromatin remodeling genes and altered genome-wide DNA methylation profiles is a promising clue on the cause of CIMP[113-116].

In order to better define CIMP, a quantitative method should be used for the methylation frequency and extent measurement. In addition, genes with high methylation level in normal tissues have to be excluded to define the phenotype. This may be problematic in tumors such as breast and prostate tumors, in which a considerable fraction of the tissue is from the relevant normal cells. Third, a large sample size is needed to check whether CIMP is really existing and what are the best markers to define it. Fourth, appropriate statistical methods should be developed for the analysis of data from microarray or NGS. One possible approach could be to perform a k-means censuses clustering combined with unsupervised clustering to identify a minimal set of markers and then confirm it in a separate group of tumors.

## 1.3 Environmental Epigenetics

Biological science is undergoing a paradigm shift away from the fixed

genetic determinism of the 20th century and toward an understanding that environmental factors can alter gene expression and activity in a heritable manner. Genetic factors interact with the environment to contribute to disease risk. In gene-environment interactions, the genetic polymorphisms that modify the effects of environmental exposures are transmitted transgenerationally according to Mendelian genetics. A second interplay are the mutations induced by environmental exposures. It has been reported that genotoxic agents could cause mutations to increase disease the risk[117] and these environmentally-induced DNA mutations can have a transgenerational effect (the consequence of genetic alterations in one generation can be inheritanted into the next generation) when occurring in the germline[118,119].

Similar to genetic polymorphisms, epigenetic aberrations could also make individuals more vulnerable to environmental insults. Animal studies have provided us with some examples, suggesting that epigenetic marks established during life can be passed onto the next generations[120-125]. This phenomena has been challenging to prove in humans and few debatable examples exist to suggest the inheritance of epigenetic states. For example, the DNA mismatch repair genes *MLH1* and *MSH2* were initially found to be deactivated by promoter hypermethylation in several generations with familial colorectal cancer[126-128]. However, underlying genetic mechanisms for these effects have been uncovered.

In the last few years, many studies have investigated the correlation between environmental exposures and epigenetic changes including DNA methylation and histone modifications[129], and identified several toxicants which can directly modify epigenetic marks. One example of an epigenetic toxicant is bisphenol A (BPA) which was frequently used in manufacturing of polycarbonate plastics. Exposure to BPA is reported to be associated with higher body weight, increased breast and prostate cancer development and

altered reproductive function. In mouse models, it has been shown that maternal BPA exposure shifted the coat color of spotted yellow agouti ($A^{vy}$) mouse offspring toward to complete yellow by hypomethylation of an retrotransposable intra-cisternal A particle (IAP) sequence upstream of the Agouti gene[130,131].

Exposure to air pollution, such as particulate matter (PM), was associated with increased rate of cardiorespiratory disease and lung cancer risk[132-136]. It has also been shown that the inducible Nitric Oxide Synthase (iNOS) gene was upregulated due to the promoter hypomethylation in samples with exposure to PM with aerodynamic diameter < 10 um ($PM_{10}$)[137]. The upregulation of iNOS can contribute to inflammation and oxidative stress generation, which are primary mechanisms linking inhalation of air pollutants to their acute health effects[138-140]. Other exposures, such as to persistent organic pollutants (POPs), have been associated with hematopoietic malignancies mediated by the methylation changes in repetitive elements[141]. Another well studied environmental exposure is tobacco smoke. Several lines of evidence indicated that fetal exposure to maternal smoking during pregnancy is not only associated with hypomethylation in repetitive sequences including Sat2[142], Alu and LINE1[143], but also associated with hypermethylation of specific genes, such as *AXK*, *PTPRO*[144] and *IGF2*[145].

In summary, accumulating evidence suggests that epigenetic processes could potentially mediate effects of environmental exposures to influence disease susceptibility. We now need to better understand the basic epigenetic mechanisms that operate and maintain proper epigenetic states in order to identify the most relevant periods and biomarkers of exposure. It is clear that statistical and bioinformatic approaches will be required to enable the efficient comprehension of these analyses, especially as we expand to the genome-wide scale.

## 1.4 High throughput assays for methylome analysis

With the availability of current high throughput technologies (microarray and next generation sequencing), one can already produce a comprehensive picture of the epigenetic profile, especially the methylome, in normal and tumor/diseased cells[146]. Numerous epigenomic projects, such as the Human Epigenome Project and the NIH Epigenomic Roadmap Initiative, have been launched to uncover epigenetic mechanisms and to integrate epigenetic factors into regulatory networks[147-150]. For methylome data generation and analysis, there are several main approaches including microarray, chromatin immunoprecipitation (ChIP) and next generation sequencing (NGS) summarized in the following sub sections.

### 1.4.1 Infinium HumanMethylation450 BeadChip

The array based Illumina Infinium HumanMethylation450 BeadChip is a comprehensive platform for human methylome analysis. The CpGs on the chip are selected by experts in field and cover CpG islands and shores, non-CpG methylated sites identified in human stem cells, differentially methylated sites identified in tumor versus normal and across several tissue types and microRNA promoter regions. The low price and input DNA make it a powerful tool in epigenetics research, especially in epigenome-wide association studies (EWAS). For the data analysis, there are many R packages available for processing, normalization and downstream analysis[151-153].

### 1.4.2 Methyl-CpG-immunoprecipitation followed by sequencing

Methylated CpG enrichment approaches such as MeDIP[154], MethylCap[155] and MCIP[156] followed by next generation sequencing or microarray analysis are widely used methods for methylation profiling. Those methods provide enrichment values for different methylation states of genomic regions by counting read numbers or assessing relative fluorescence ratios of regional sequences. One drwaback is the readout from such approaches do not give quantitative values of CpG methylation levels. This can only be determined for regions of interest by additional follow up analysis like bisulfite sequencing of cloned sequences, pyrosequencing or mass spectrometric analysis. Generally, MCIP allows rapid enrichment of methylated CpGs in DNA. The affinity is increased with the density of methylated CpGs and lowered with higher salt concentrations in the buffer. After the enrichment, NGS can be performed to get unbiased genome-wide qualitative methylation profile.

### 1.4.3 Whole genome bisulfite sequencing

Comprehensive understanding of the role of genome-wide DNA methylation patterns, requires quantitative determination of the methylation states of all CpGs in a genome. Thus, we have to sequence the bisulfite converted genomic DNA to obtain the complete insight into the DNA methylome. The conventional WGBS protocol was described by Lister et al. in 2009[157]. Generally, bisulfite treatment will convert all cytosines to uracil apart from 5-methyl-cytosines which helps us to distinguish methylated cytosines from unmethylated ones. The pool of DNA is then subjected to NGS and followed by bioinformatic analysis. Although the price is still high for this

technique, the advantage is that we can evaluate the methylation level of all potential cytosines including both CpG and CpH contexts in our genome, and allele specific differences in epigenetic patterns can be also detected. More recently, a tagmentation based whole genome bisulfite sequencing protocol was developed as a less time and input DNA consuming alternative approach to the conventional generation of next generation libraries [158]. This protocol was later modified and used to investigate the whole methylome (around 27 million CpGs) at single base resolution by using only 10-20 ng of input DNA, equivalent to around 1700-5100 cells, compared to 3-5 ug required for traditional WGBS.

It is generally difficult to align bisuflite treated DNA sequences back to the genome, since the complexity of bisuflite treated reads is effectively reduced to 3 bases which means that Cs in the read may also align to T positions in the genome. So far, there are two main algorithms designed for the mapping of bisulfite treated DNA sequences. One is 3-nt alignment which convert C to T and G to A in the reference genome[159]. In bisulfite sequencing, only T in reads could be mapped to C in the references, not the other way around. It seems that simply treat C and T equally introduces false mappings which need to be filtered in post-alignment processing. Actually, the post processing could not fully eliminate the mapping biases since some alignment information, such as the multiple hits information, is only available in alignment stage, but not fully recorded in the alignment output. So the 3-nt alignment algorithm has the advantage of speed but at the price of accuracy. However, if a read aligns to a wrong position (e.g. a read containing Cs aligns to a genomic position containing Ts) it might be indeed a mis-alignment. Thus, no methylation call would be made for these positions since they are no Cs in the genome. In essence, these reads should normally have no influence on the later estimation of methylation levels. The best way to avoid mis-alignments and

increase accuracy is to use high quality data (appropriately trimmed) and use stringent mapping parameters. Another one is called wildcard algorithm which uses a native algorithm to do the C->T alignment[160]. It has its own bias as well. If positions which are a C in the read but a T in the genome would receive a penalty when the wildcard algorithm is used, we would probably not see such mis-alignments. Hence it may bias the entire mapping output in favour of methylated reads over unmethylated reads. Here is an example to show that the wildcard algorithm may give rise to biases:

Scenario 1:

ATTGATCTGATTA (read sequence) (C methylated)

ATTGATCTGATTA (genome position 1)

ATTGATTTGATTA (genome position 2)

The wildcard algorithm would align the read sequence uniquely to genome position 1, but genome position 2 would not be a valid alignment (mapping asymmetry).

Scenario 2 (same sequence but with a T in the middle (C unmethylated)):

ATTGATTTGATTA (read sequence)

ATTGATCTGATTA (genome position 1)

ATTGATTTGATTA (genome position 2)

In this case, the read could either be derived from genome position 1 if the C was not methylated (and thus converted), or it could be derived from genome position 2. Thus, this read would be booted since it cannot be mapped unambiguously. By doing so, the wildcard algorithm would favour mapping of methylated reads so that potentially bias the methylation results depending on the methylation state of the read.

In order to evaluate the accuracy, coverage, speed and the sensitivity/specificity for DMR calling for these two algorithms, a systematic benchmarking should be performed on a real WGBS dataset and a simulated

data.

After mapping, DMR calling should be performed to detect the changes of methylation pattern between tumor and normal or other cases. Normally, DMR detection can be performed using a sliding window approach followed by Fisher's exact test when the coverage is relatively high. However, it is more common to have a low coverage WGBS data due to the high cost of sequencing. Thus, a smoothing function was applied to improve the accuracy of DMR calling accounting for biological variability when replicates are available even with low coverage[161]. Although It is true that in some parts of the genome, methylation is less smooth, so it is not all CpGs we expect smoothing to be extremely close to single CpG estimates, but it does not matter if we are interested in regional differences.

Recent studies also pointed out that methylation change could be defined by the binding of transcription factor[162]. Thus, it's now possible to detect the potential active regulatory regions from high resolution methylation datasets. For this purpose, a computational method called MethylSeekR was developed to precisely detect the footprints from methylomes[163]. With this tool, partially methylated domains (PMDs) can be identified using a two-state Hidden Markov Model (HMM) with Gaussian emissions. In addition, this tool can reliably detect unmethylated regions (UMR) and lowly methylated regions (LMR) which are usually associated to proximal and distal regulatory regions across varies cell types and tissues.

The identification of single-nucleotide polymorphisms (SNPs) from bisulfite sequencing data is challenging and important for accurate quantification of methylation levels due to the fact that 65% of all SNPs in dbSNP occur in CpG context[164]. In order to solve this problem, a probabilistic SNP caller, Bis-SNP, was developed for the SNP detection for bisulfite sequencing data. It uses Bayesian inference to evaluate a model of strand

specific base calls and base call quality scores, along with prior information on population SNP frequencies, experiment specific bisulfite conversion efficiency, and site specific DNA methylation estimates[165]. It has been shown that the accuracy for the DNA methylation calling and heterozygous SNPs identification from bisulfite sequencing data is significantly improved by using this tool.

# Chapter 2: Computational evaluation of T-WGBS

**Note:**

Dieter Weichenhan, Wei Wang and Marion Bähr performed the experiments. Bernhard Radlwimmer, Wei Wang, Jay Shendure, Volker Hovestadt and Andrew Adey contributed data. The DKFZ Genomics and Proteomics Core Facility provided technical support for the sequencing.

## 2.1 Aim of the study

T-WGBS technique is able to generate the genome-wide DNA methylation patterns at single CpG resolution using only 10-20 ng of input DNA, compared to 3-5 µg required for traditional WGBS. Since T-WGBS uses a hyperactive Tn5 transposase to fragment the DNA and to append sequencing adapters, it is highly important to investigate its reliability and reproducibility. In addition, it should be systematic evaluated that whether T-WGBS induces sequence dependent biases into the final methylation estimate.

## 2.2 Methods and materials

In order to keep the comparison bias as low as possible, DNA isolated from a human glioblastoma multiforme tumor biopsy was subjected to T-WGBS and conventional WGBS. Two independent tagmentations with 30ng input DNA each were carried out. Each tagmentation was used to build two libraries. The conventional WGBS was performed as described previously with 5 ug input DNA for a single library. The four T-WGBS libraries were sequenced one lane per library, while the conventional WGBS library was loaded onto three lanes.

A recently published mapping pipeline[166] with modifications was used to

adapt for the T-WGBS data. Briefly, the human reference genome (37d5) was transformed in silico for both the top strand (C to T) and bottom strand (G to A). Before alignment, adaptor sequences were trimmed using SeqPrep (https://github.com/jstjohn/SeqPrep). The first read in each read pair was then C-to-T converted and the 2nd read in the pair was G-to-A converted. The converted reads were aligned to a combined reference of the transformed top and bottom strands using BWA[167] using default parameters with disabling the quality threshold for read trimming (-q) of 20 and the Smith-Waterman for the unmapped mate (-s). After alignment, reads were converted back to the original states, and reads mapped to the antisense strand of the respective reference were removed. Duplicate reads were further removed, and the complexity was then determined by Picard (http://picard.sourceforge.net/). Reads with alignment scores less than 1 were filtered before subsequent analysis. Total genome coverage was calculated using the total number of bases aligned from uniquely mapped reads over the total number of mappable bases in the genome. At each cytosine position, reads that maintained the cytosine status were considered methylated, and the reads which were detected as thymine were considered unmethylated. Only bases with Phred-scales quality score $\geqslant$ 20 were considered. In addition, the 5 bp at the two ends of the reads were excluded from methylation calling according to M-bias plot quality control. For T-WGBS libraries, the first 9 bp of the second read and the last 9 bp before the adaptor of the first read were excluded before the methylation calling step.

## 2.3 Results

The four T-WGBS libraries were almost identical to each other and performed similarly well compared to the conventional WGBS library with

respect to the percentage of mapped reads, the overall methylation level assessment and the conversion frequency as shown below (**Table 1**).

The relative higher duplication level in T-WGBS is probably due to the higher PCR cycle number used in T-WGBS (ten or eleven cycles) than in conventional WGBS (eight cycles). Compared to the CpG coverage (13.7X) from the conventional WGBS, T-WGBS provided only slightly lower CpG coverage (12.1X), when reads from three lanes for each were merged. The bisulfite conversion frequency of the T-WGBS libraries was marginally lower than that of the conventional WGBS library, 99.5% vs. 99.9%, and, in line with this, the average CpG methylation level in T-WGBS was slightly higher, 77.2% vs 75.8%; both differences likely reflect a better bisulfite treatment performance in the conventional WGBS rather than a better overall performance of the conventional method. High similarity in the performance between the two protocols was further supported by the high correlation of the methylation levels (Pearson correlation 0.95; **Figure 1**).

To further quantify the consistency between the two WGBS protocols, a concordance metric was defined as the percentage of CpG sites (at least 30X coverage) with less than 20% difference in methylation level. The concordance between the two protocols was 97.3% (**Figure 1**). Such reliability was further supported from two human blood samples (**Figure 2**).

As determined in the same manner, the concordance between two independent T-WGBS experiments was 97.8 % (r = 0.92; **Figure 3**), indicating high robustness and reliability of the T-WGBS protocol.

T-WGBS and conventional WGBS also display similar sequencing coverage at CpG sites as a function of CpG density (**Figure 4**). Comparative analysis of sequencing coverage versus density of cytosines in CpG, CHG and CHH context (H can be A, C or T) or versus local GC content revealed similar patterns from T-WGBS and conventional WGBS.

| Protocol | Library | Total Read Pairs | Mapped | Uniquely Mapped out of All Mapped | Duplication Frequency | Coverage per Strand at CG | Coverage per Strand at CH[3] | Average Methylation Level at CG | Average Methylation Level at CH[3] | Conversion Frequency[4] |
|---|---|---|---|---|---|---|---|---|---|---|
| Tagmentation-based | T-WGBS1_lib1 + T-WGBS1_lib2 +T-WGBS2_lib1[1] | 608,802,912 | 97.1% | 92.7% | 10.8% | 12.1 | 12.4 | 77.2% | 0.45% | 99.50% |
| | T-WGBS1_lib1[2] | 199,181,042 | 97.2% | 92.6% | 11.2% | 4.0 | 4.0 | 77.2% | 0.50% | 99.45% |
| | T-WGBS1_lib2[2] | 201,237,175 | 97.1% | 93.0% | 10.7% | 4.0 | 4.2 | 77.1% | 0.41% | 99.55% |
| | T-WGBS2_lib1[2] | 208,384,695 | 97.1% | 92.4% | 10.4% | 4.0 | 4.1 | 77.2% | 0.45% | 99.50% |
| | T-WGBS2_lib2[2] | 185,492,592 | 97.4% | 92.9% | 8.8% | 3.8 | 3.9 | 77.1% | 0.38% | 99.58% |
| Conventional | WGBS_lib1[1] | 606,295,337 | 96.5% | 92.5% | 4.3% | 13.7 | 15.6 | 75.8% | 0.18% | 99.91% |
| | WGBS_lib1[2] | 205,933,339 | 96.3% | 92.5% | 1.8% | 4.7 | 5.3 | 75.8% | 0.18% | 99.91% |
| | WGBS_lib1[2] | 213,939,066 | 96.8% | 92.6% | 1.2% | 5.0 | 5.6 | 75.6% | 0.19% | 99.91% |
| | WGBS_lib1[2] | 186,422,932 | 96.2% | 92.4% | 1.1% | 4.2 | 4.8 | 75.9% | 0.18% | 99.91% |

[1]Total read pairs and coverage refer to sum of 3 HiSeq 2000 lanes.

[2]Total read pairs and coverage refer to a single HiSeq 2000 lane.

[3]H can be A or C or T.

[4]Conversion frequency determined with spiked phage λ DNA for the conventional library and with the 9 bp filled-in unmethylated gaps for the T-WGBS libraries.

**Table 1 Sequencing statistics between WGBS and T-WGBS.** Reads numbers, duplication levels, coverages, methylation levels and conversion rates are compared between libraries from WGBS and T-WGBS protocols.

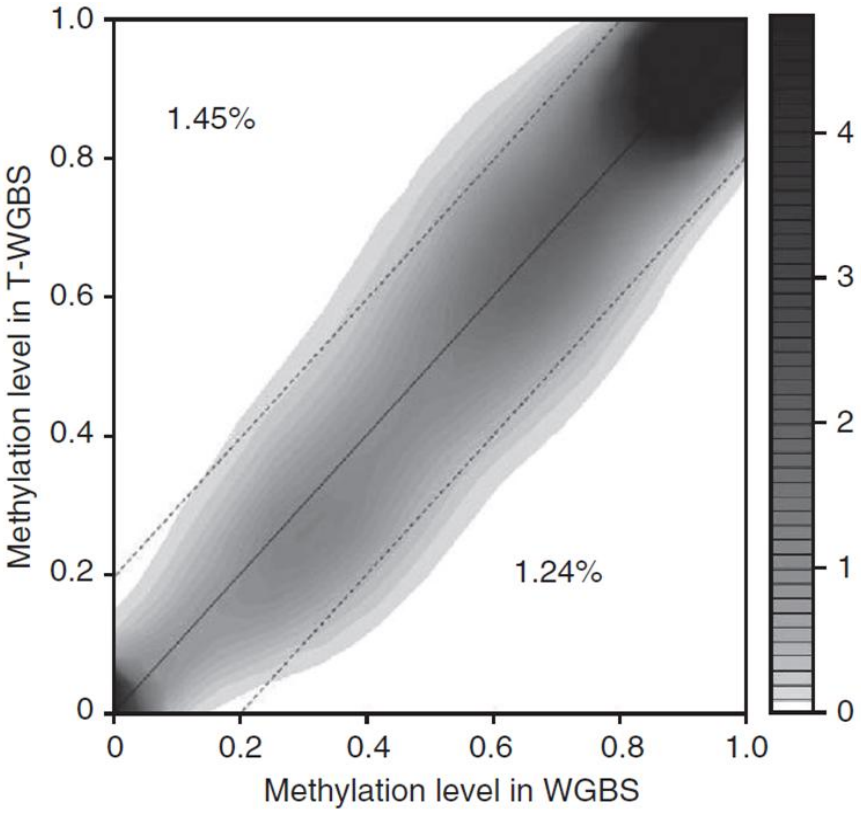



**Figure 1 Methylation level between WGBS and T-WGBS libraries.** High consistency with Pearson correlation of r = 0.95 between the methylation levels of corresponding single CpGs covered at least 30-fold in T-WGBS and conventional WGBS.

For methylome characterization, genomic features like promoters, CpG islands, exons, introns and intergenic regions are of particular interest. The proportions of CpGs covered at least 10-fold in these features were all above 90% and nearly identical between T-WGBS and conventional WGBS (**Figure 5**;

**Table 2**).

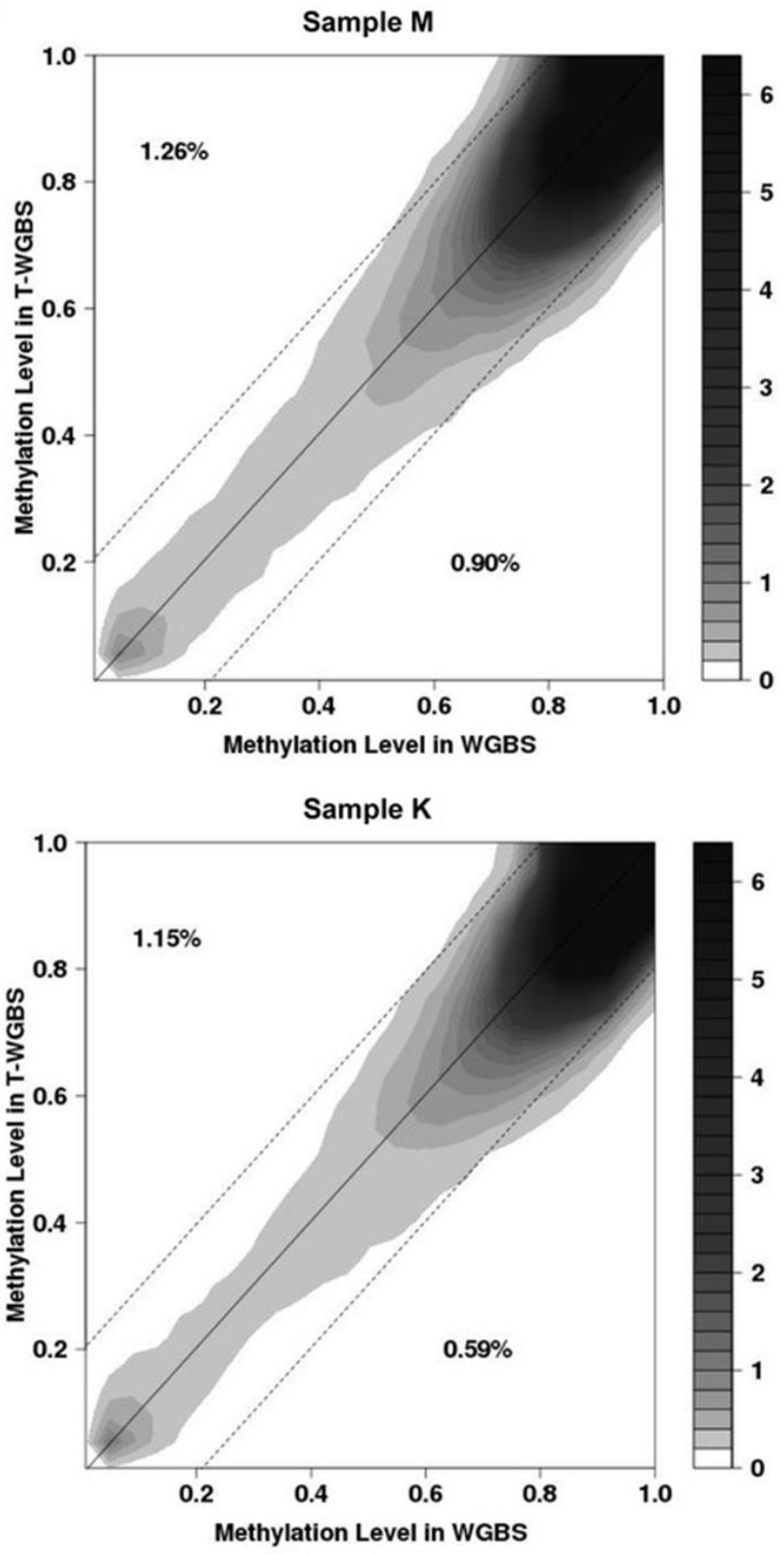



**Figure 2 High consistency between T-WGBS and WGBS methylation data from two human blood samples.** For each sample, two lanes of T-WGBS data and three lanes of

conventional WGBS were compared. Methylation levels were calculated based on scanning windows of 5 CpGs with at least 30-fold coverage and displayed in density plots.

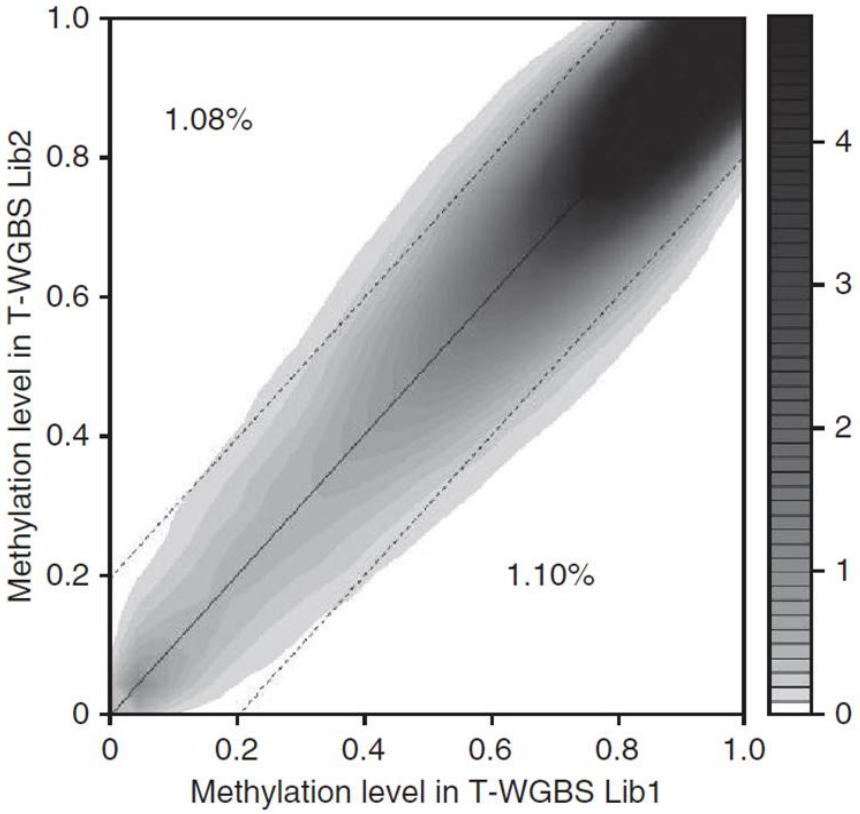



**Figure 3 High reproducibility of T-WGBS.** High reproducibility of T-WGBS indicated by strong agreement of the methylation levels (r = 0.92) in windows of 5 CpGs (read numbers too low for single CpG analysis) in libraries from 2 independent tagmentations analyzed on a single HiSeq2000 lane each.

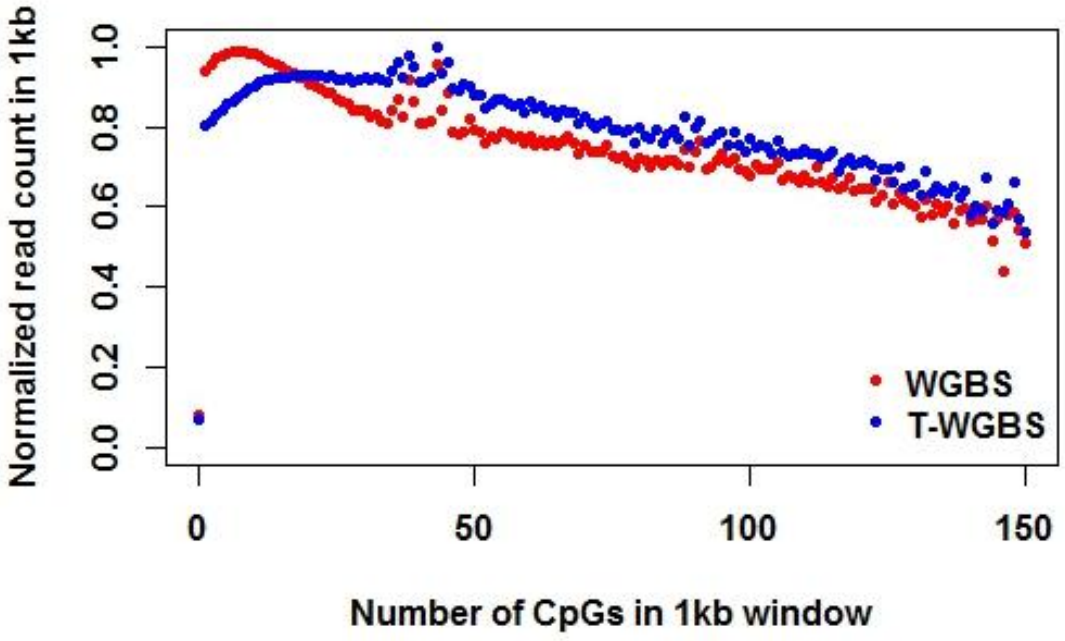



**Figure 4 Coverage vs. CpG density plot for both WGBS and T-WGBS.** Nearly identical sequencing coverage of CpGs as a function of CpG density between T-WGBS and conventional WGBS.

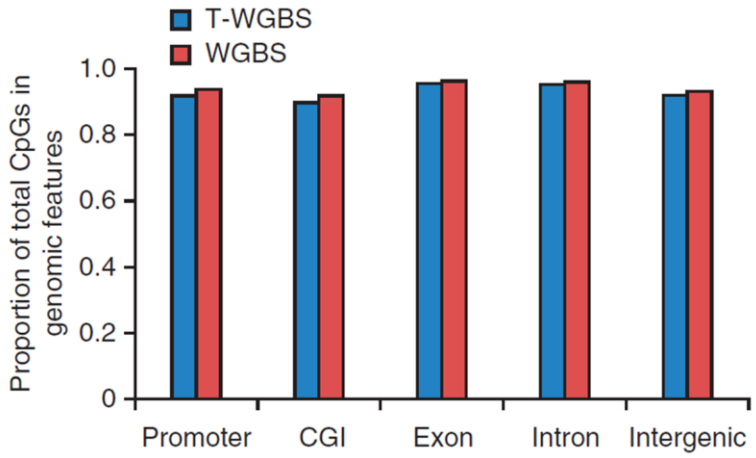



**Figure 5 Genomic coverage between WGBS and T-WGBS.** 90% or higher and almost identical proportions of CpGs covered at least 10-fold in 5 genomic features. T-WGBS and conventional WGBS reads from 3 lanes each were compared.

| | Coverage of CpGs | 1x | 5x | 10x | 15x | 20x |
|---|---|---|---|---|---|---|
| Total (T-WGBS) | 28,217,448 | 27,512,847 | 27,136,922 | 26,364,166 | 24,246,290 | 19,798,878 |
| Percentage | 100 | 97.5 | 96.2 | 93.4 | 85.9 | 70.2 |
| Total (conv. WGBS) | 28,217,448 | 27,454,762 | 27,161,682 | 26,577,078 | 25,180,920 | 22,228,929 |
| Percentage | 100 | 97.3 | 96.3 | 94.2 | 89.2 | 78.8 |

**Table 2 Comparison of CpG coverage between WGBS and T-WGBS.** Number of CpGs covered by at least 1x, 5x, 10x, 15x and 20x is nearly the same between WGBS and T-WGBS.

## 2.4 Discussion

The power of T-WGBS is to generate complete methylomes from ultra low amounts of input DNA which substantially improves the practicality of the whole methylome sequencing and removes a key advantage of less encompassing methods such as RRBS[168,169]. This method particularly allows the comprehensive interrogation of methylation in many contexts where DNA quantity is a bottleneck, e.g., developing anatomical structures, microdissected tissues, or pathologies such as cancer, where the methylation profile is of interest but tissue quantity limits high-resolution WGBS.

Although the first bases of the T-WGBS reads show a base composition bias, there is a high consistency in base composition of sequencing reads between T-WGBS and conventional WGBS (**Figure 6**). This bias may only become problematic if it has a considerable impact on genomic coverage; such an impact, however, is not observed in this study.

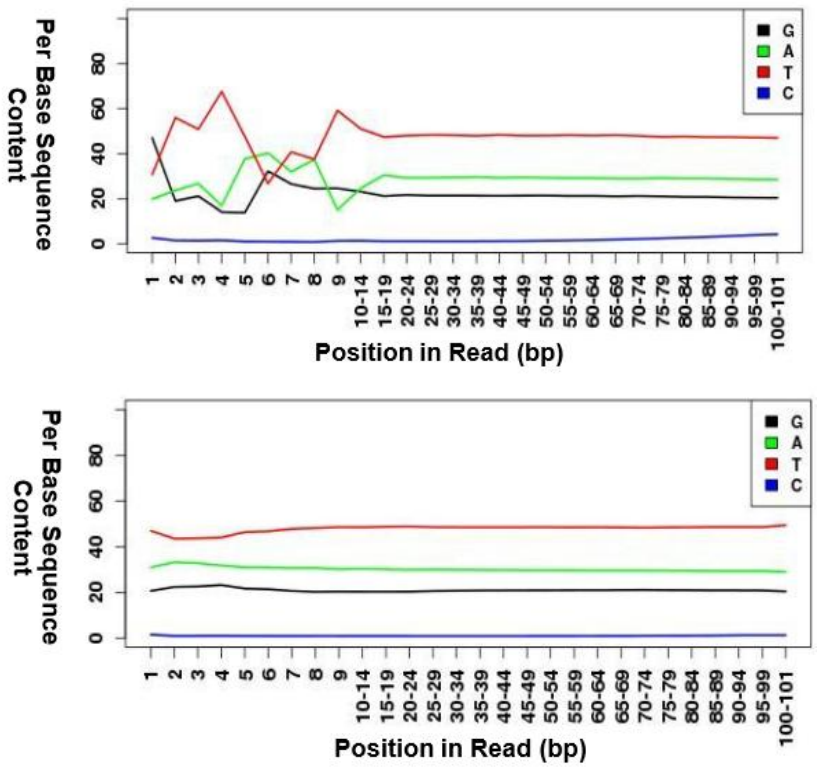



**Figure 6 Base composition of sequencing reads between T-WGBS (upper) and conventional WGBS (down)**

In order to further investigate the consistency, robustness and reproducibility of the T-WGBS method, a comparison between labs would be useful.

In summary, the methylome data from T-WGBS is highly reliable and reproducible. By comparing the coverages of different genomic features and levels of DNA methylation between T-WGBS and conventional WGBS, no significant sequence dependent bias is observed.

# Chapter 3: Identification of Genome-wide Methylation Alterations in Early Onset Prostate Cancer

**Note:**

Dieter Weichenhan provided sequencing library preparation. The DKFZ Genomics and Proteomics Core Facility provided technical support for the MCIP-seq. German ICGC early onset prostate consortium provided 11 prostate tumor samples and 1 normal sample.

## 3.1 Aim of the study

This aim of this project is to establish a pipeline that can be used for sequencing-based epigenomic data analysis for any other complex diseases and quantitative phenotypes and to profile the methylome of early onset prostate cancer by using MCIP-seq.

## 3.2 Methods and materials

The computational pipeline has been established by using the 11 tumor and 1 unmatched normal data. The whole pipeline starts with the alignment of raw reads and ends with the DMR calling and further downstream analysis including the model for validation analysis (the accuracy is around 87% for the early onset prostate cancer data) (**Figure 7**).

Briefly, reads are mapped to the human genome reference sequence (Build 37) using the alignment software BWA[167]. Two types of quality control are performed: (1) duplication reads and reads with a MAQ score of <20 are removed; (2) samples with a saturation coefficient of <0.95 are re-sequenced

in order to make sure that reads covered all regions that can be captured by MCIp[170]. To detect regions of differential methylation between tumor and normal, three criteria (i.e., q value, coverage and fold change) are applied both when using locus-specific analyses (focused approach) and unbiased analyses (genome-wide approach)[171].

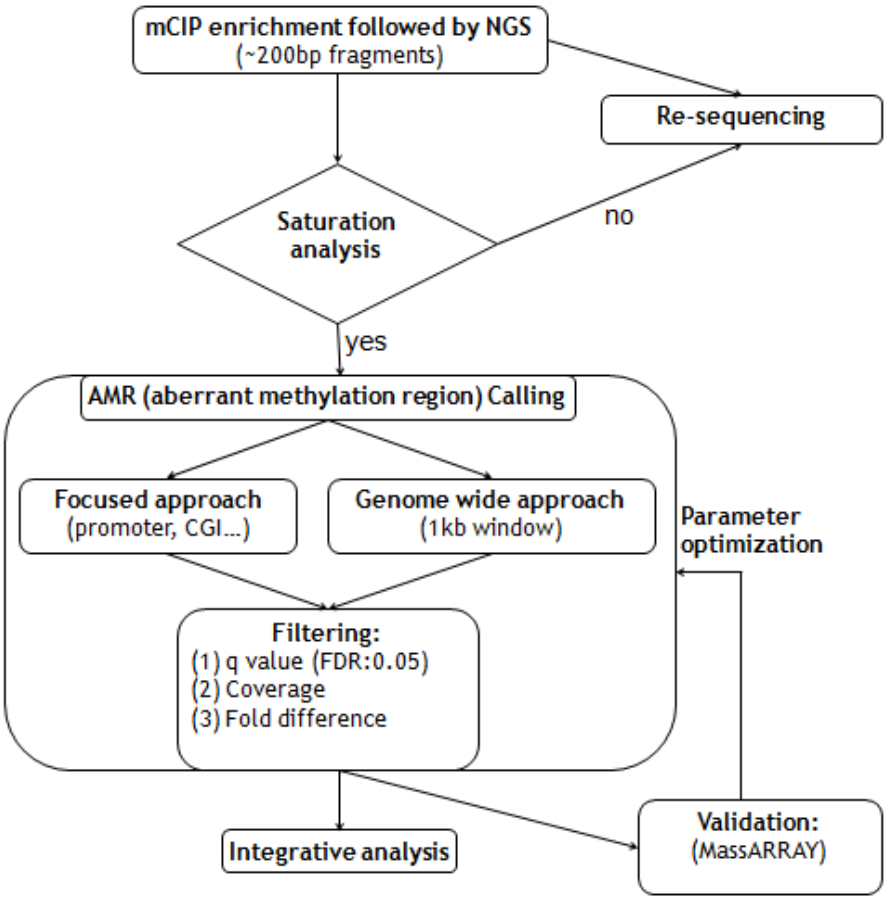



**Figure 7 Computational pipeline for DMR detection.** The first quality control step is the saturation analysis to check whether the number of reads is enough to capture all regions from MCIP enrichment. Two approaches are performed to detect DMRs genome-wide or in certain genomic features. The detected DMRs validated by MassARRAY are then used for downstream analysis.

## 3.3 Results

MCIP-seq was used to identify altered DNA methylation in tumor samples versus one normal epithelial control. Early onset prostate tumors were characterized by an average of 46095 DMRs (range 23585-61489). Focusing on promoter sequences identified an average of 10125 DMRs (range 6740-12744). A total of 1,319 DMRs were common to all 11 tumor samples. Of these DMRs, 1,245 were hypermethylated and only 74 were hypomethylated, indicating a clear preponderance of genomic hypermethylation in the non-repetitive tumor sequences. The distribution of common DMRs on Chromosomal-wise indicated the difference between observed and expected proportion of common (in all 11 tumor samples) hypermethylated (red: observed; yellow: expected) and hypomethylated (blue: observed; green: expected) regions. P value is calculated by Chi-square test with Monte Carlo simulation (**Figure 8**).

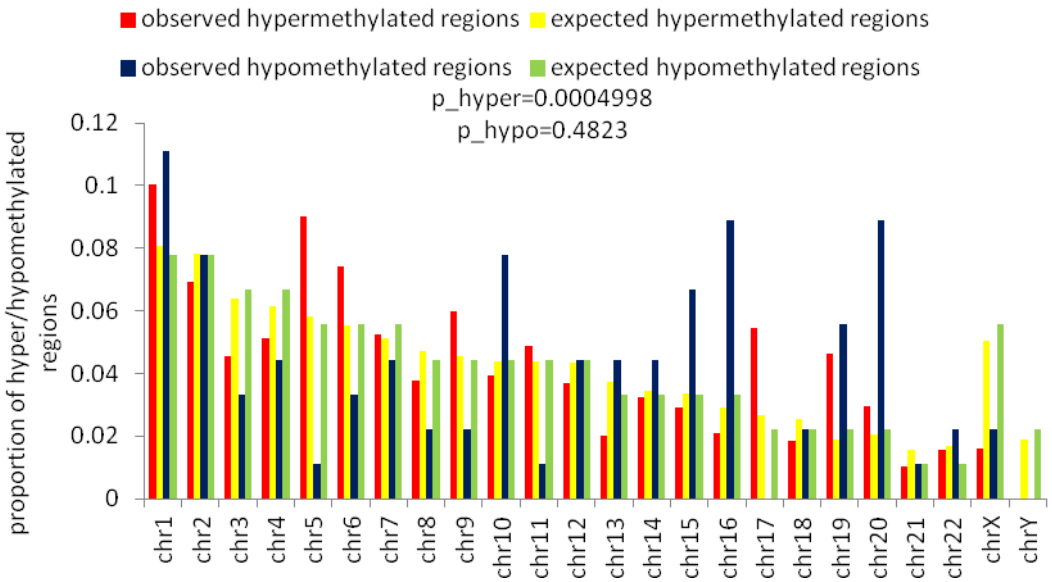



**Figure 8 Chromosomal-wise common DMR distribution in early onset prostate tumors.**

The majority of hypermethylated and hypomethylated regions locate in intergenic and intronic sequences. The proportion of hypermethylated as compared to that of hypomethylated regions is much higher in CGIs, CGI vicinal sequences, promoters and DNaseI-hypersensitive areas, whereas these proportions are more alike or even reversed for repetitive sequences (**Figure 9**).

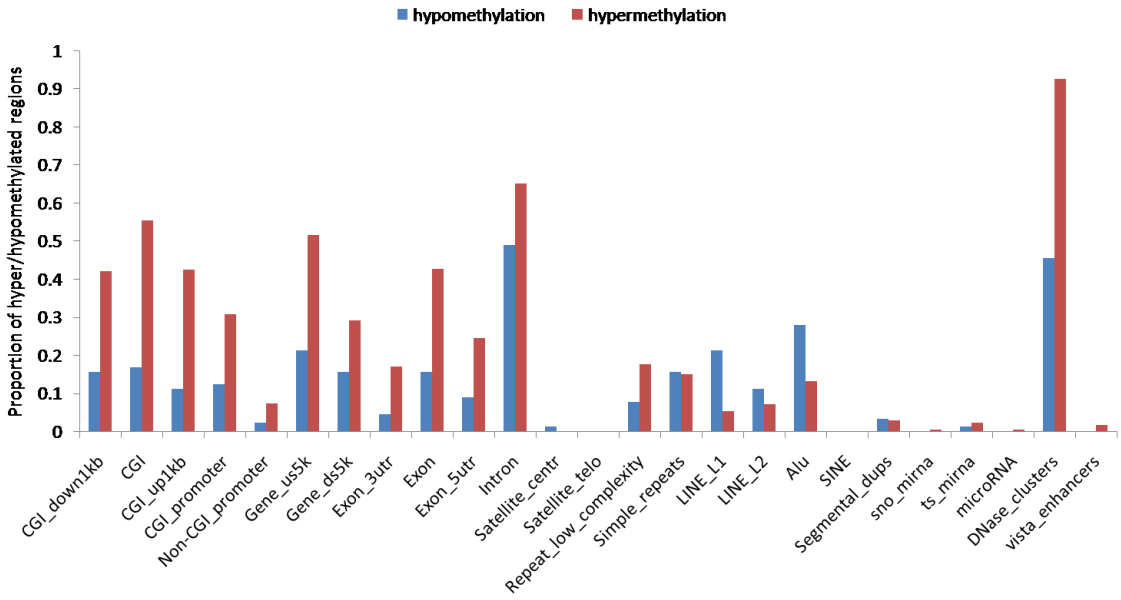



**Figure 9 Proportion of common hypermethylated (red) and hypomethylated (blue) regions among 25 different genomic features**

It was further demonstrated that the observed occurrence of differentially methylated promoters among the 11 tumor samples deviates significantly (p-value < 0.0001) from a random distribution (**Figure 10**).

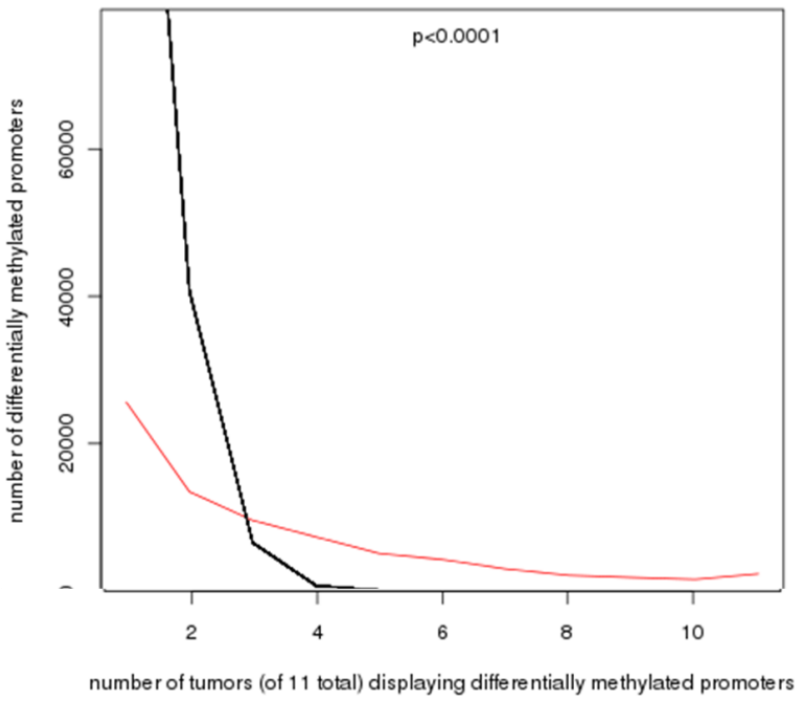



**Figure 10 Non-random distribution of differentially methylated promoters throughout the prostate cancer genome.** X-axis displays the number of tumor samples, Y-axis indicates the number of differentially methylated promoters. Black and red curves show the expected and the observed distribution, respectively. The empirical P value is calculated based on 10,000 permutations.

Among the hypermethylated and hypomethylated promoters, 92% and 85%, respectively, are high CpG promoters (HCPs) with a CpG ratio >0.75. None of the hypermethylated and hypomethylated promoters belongs to the low CpG promoter (LCP) type with a CpG ratio <0.48 (**Figure 11**).

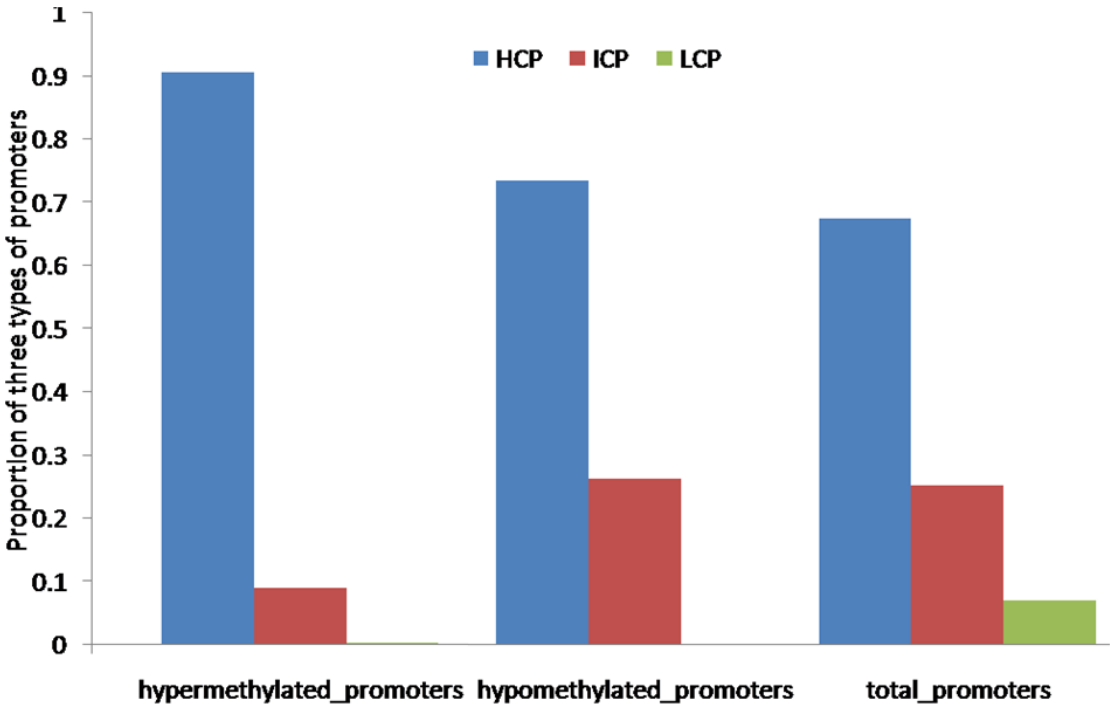



**Figure 11 Frequency of differentially methylated promoters depending on GC content and CpG ratio**

To validate the methylome data, we chose 15 regions and validated the methylation status by MassARRAY in the twelve tumors and one normal epithelium. There was a concordance of 87% (155/178) strengthening the quality of the data set. A tumor-suppressor gene, APC, was token as an example to show the methylation pattern in its promoter region (**Figure 12**).

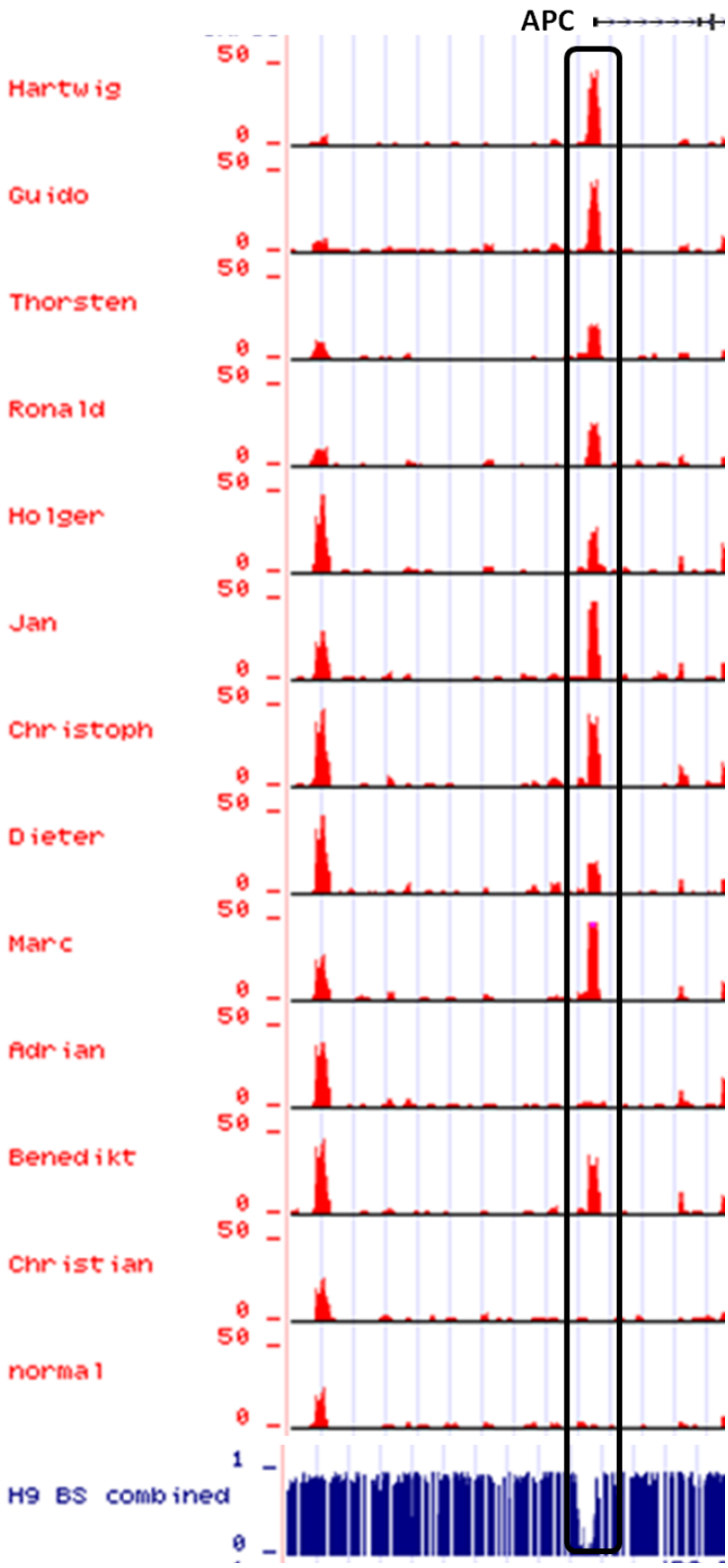



**Figure 12 Promoter hypermethylation in APC.**

## 3.4 Discussion

Similar to previous reported[172-176], a large number of DMRs were detected indicating the dramtic epigenetic reprogramming in cancers. The genomic distribution of DMRs and its non-random distribution pattern suggested that DMRs may have the functional potentials in the early onset prostate cancer development. Due to the normal contamination and MCIP intrinsic limitation, it is hard to detect hypomethylation in this study and it is difficult to evaluate the validation model due to the lack of normal samples. Thus, it would be interesting to use WGBS to further validate the DMR detected by MCIP-seq and increase the normal samples with high purity to refine the DMRs.

# Chapter 4: BAZ2A links epigenetic remodeling and recurrence in prostate cancer

**Note:**

Christopher Oakes, Constance Baer and Melanie Weiss performed experimental work. Ruprecht Kunert, Guido Sauter, Katharina Grupp and Ronald Simon provided clinical samples or data. Anna Postępska-Igielska, Nina Schmitt, Christopher Schmidt, Daniela Wuttig, David Brocks and Olga Bogatyrova for assistance with experiments and data. The DKFZ Genomics and Proteomics Core Facility provided technical support for illumina 450k array data production.

## 4.1 Aim of the study

Epigenetic regulatory genes have emerged to be vital in cancer due to new insights from genomic and expression studies[177]. MicroRNA-based modulation of numerous onco- and tumor-suppressor genes is now recognized as a key aspect of the establishment and maintenance of the tumor phenotype[178]. Thus, an integrative analysis was performed in order to identify novel prostate cancer-relevant genes and their potential impact on prostate cancer development and treatment.

## 4.2 Methods and materials

### 4.2.1 Bioinformatic identification of mir:target pairs

Six samples with RNA-seq and microRNA-seq data were downloaded from the European Genome-phenome archive database (hosted at the EBI) with accession number EGAS00001000258. In order to filter the low abundant microRNAs and genes, microRNAs with coverage of at least 1000 reads and

genes with at least 12 rpkm were kept for the following analysis. Then, 1.5 was set as the cut-off of fold change between tumor and normal. Five prediction tools (TargetScan, miRNAorg, PITA, PicTar and miRDB) were used for the microRNA target prediction and targets predicted by all 5 tools were extracted. MiR:target pairs not showing an inverse pattern of expression were filtered. Finally, genes that were found to involve other alterations (mutations, CNAs, SVs, LOH and promoter DMRs) were removed to enhance the likelihood that the target gene dysregulation was influenced by the microRNA. The general workflow is shown below (**Figure 13**). The large validation data was downloaded from the GEO database (GSE29079) and a t-test was used to calculate the significance of differential expression of *BAZ2A* and miR-133a (data used for validation of microRNA expression is not available in public databases).

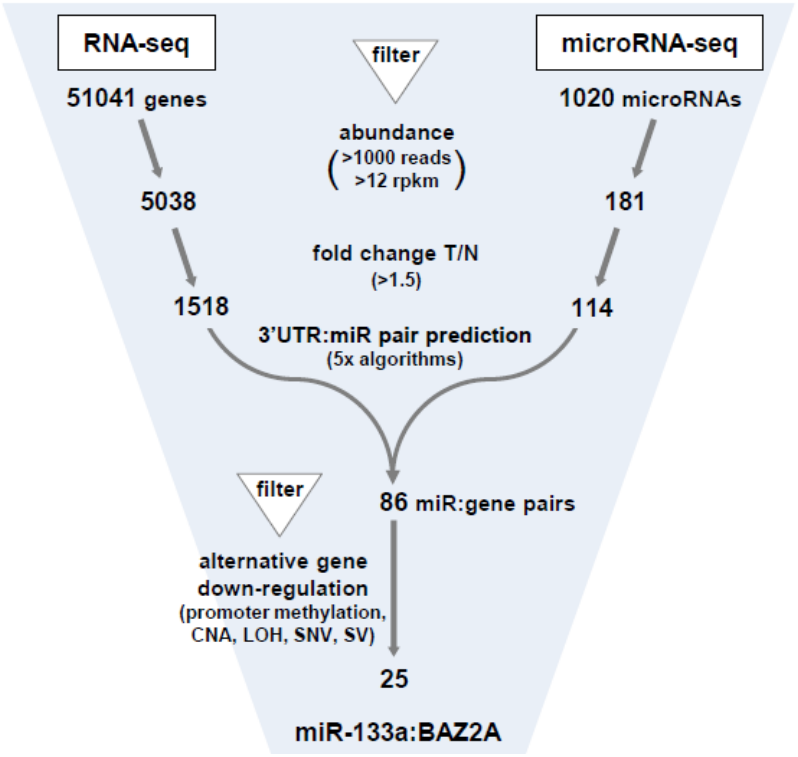



**Figure 13 Computational pipeline for the detection of mir:target pairs.** Analysis pipeline

comparing integrated RNA-seq and microRNA-seq data to identify high confidence, reciprocal expression of miRNA:gene pairs. Of the 25 pairs which fit all criteria, overexpression of the epigenetic regulator, BAZ2A, and downregulation of the tumor suppressor, miR-133a, was found.

### 4.2.2 Evaluation of microRNA targeting

Luciferase assays were performed in HEK293T cells grown in Dulbecco's Modified Eagle's Medium. 5 nM of miRNA mimics (Qiagen) or the non-targeting control (AllStar Negative Control, Qiagen) were transfected using DharmaFECT1 (Thermo Fisher Scientific) into cells grown in 384 well plates. BAZ2A 3'UTR fragments of 300-1609 bp were cloned into the pMIR-Report vector (Ambion) 3' of the firefly luciferase gene. After 24 hrs, 0.3 ng of each pMIR-Report-BAZ2A 3'UTR construct was mixed with 10 ng of the TK-Renilla plasmid (Promega) and were transfected using TransIT-LT1 transfection reagent (Mirus Bio) with 6 replicates per construct. The read-out was preformed 48 h after reporter transfection as previously described[179]. Firefly luciferase activity was normalized to Renilla luciferase activity and the average of technical replicates was calculated. Each experiment was performed in triplicate. The effect overexpression of miRNAs on the endogenous expression level of BAZ2A was performed using miRNA mimics (Qiagen) transfected using INTERFERin transfection reagent (Polyplus transfection). DU145 and BPH1cells were grown in RPMI. Cells were grown 72 hours following transfection and RNA was isolated using RNeasy columns (Qiagen). *BAZ2A* expression was measured using the Universal probe library system (Roche) in a LightCycler 480 real-time PCR machine (Roche). Expression was measured by three independent primer-probes relative to the average of GAPDH, ACTB and HPRT. Each experiment was performed in triplicate for each cell line.

### 4.2.3 Methylation data analysis

The minfi package was used to extract the raw methylation intensity data and perform the Subset-quantile within array normalization (SWAN). Probes with detection P-value <0.01 were excluded from the further analysis. The 5000 most variable probes were selected for the k-means consensus clustering by ConsensusClusterPlus package with Spearman distance and average linkage over 1000 resampling iterations with random restart. The optimal number of clusters was determined by the Consensus Cumulative Distribution Function (CDF). Hierarchical clustering was then performed to visualize the methylation patterns within 35 samples. The CNV profile was detected using 450k data as previous described[180].

### 4.2.4 Tissue microarray

Radical prostatectomy specimens were obtained from 11,152 patients undergoing surgery undergoing surgery between 1992 and 2011 at the Department of Urology and the Martini Clinics the Martini Clinics at the University Medical Center Hamburg-Eppendorf. Follow-up data were Follow-up data were available for a total of 9,628 patients with a median follow-up of 36.8 months follow-up of 36.8 months (range: 1 to 228 months;

**Table 3**).

Prostate specific antigen values were measured following surgery and recurrence was defined as a postoperative PSA of 0.2 ng/ml and increasing at first of appearance. All prostate specimens were analyzed according to a standard procedure, including a complete embedding of the entire prostate for histological analysis[181]. The TMA manufacturing process was described earlier in detail[182]. All hematoxylin and eosin-stained histological sections from all prostatectomy specimens were reviewed for the purpose of this study and the tumors were marked on the slides. One 0.6 mm tissue core was punched from a preselected area of each tumor and transferred in a tissue microarray. The

punch site was selected to contain the highest possible fraction of tumor cells. The tissues were distributed among 24 TMA blocks, each containing 144 to 522 tumor samples. Presence or absence of cancer tissue was validated by immunohistochemical AMACR and 34BE12 analysis on adjacent TMA sections. For internal controls, each TMA block also contained various control tissues, including normal prostate tissue.

### 4.2.5 Immunohistochemistry

Freshly cut TMA sections were immunostained in a single day and as one experiment. Primary antibody specific for *BAZ2A* (polyclonal; rabbit, Abnova cat.# PAB21919; at 1/150 dilution) was applied, slides were deparaffinized and exposed to heat-induced antigen retrieval for 5 minutes in an autoclave at 121°C in pH 7.8 Tris-EDTA buffer. Bound antibody was then visualized using the EnVision Kit (Dako). All stainings were analyzed by a single, experienced individual (K.G.). *BAZ2A* expression was predominantly localized in the nucleus with lower expression-levels in the cytoplasm of the cells. Nuclear *BAZ2A* staining was evaluated according to the following scoring system: The staining intensity (0, 1+, 2+, and 3+) and the fraction of positive tumor cells were recorded for each tissue spot. A final IHC score was built from these parameters as previously described[183-185]. Negative scores had complete absence of staining, weak scores had staining intensity of 1+ in ≤70% of tumor cells or staining intensity of 2+ in ≤30% of tumor cells; moderate scores had staining intensity of 1+ in >70% of tumor cells, staining intensity of 2+ in >30% but in ≤70% of tumor cells or staining intensity of 3+ in ≤30% of tumor cells; strong scores had staining intensity of 2+ in >70% of tumor cells or staining intensity of 3+ in >30% of tumor cells. As cytoplasmatic *BAZ2A* staining was rare and typically associated with high nuclear staining levels, it was thus not considered for analysis.

| | No. of patients | |
|---|---|---|
| | **Study cohort on TMA, n=11,152 *** | **Biochemical relapse rate in category, n=1,824 **** |
| **Follow-up (mo)** | | |
| Mean | 53 | - |
| Median | 37 | - |
| **Age (y)** | | |
| <50 | 318 (3%) | 49 (18%) |
| 50-60 | 2,768 (25%) | 460 (19%) |
| 60-70 | 6,548 (59%) | 1,081 (19%) |
| >70 | 1,439 (13%) | 232 (19%) |
| **Pretreatment PSA (ng/ml)** | | |
| <4 | 1,407 (13%) | 142 (11%) |
| 4-10 | 6,735 (61%) | 827 (14%) |
| 10-20 | 2,159 (20%) | 521 (28%) |
| >20 | 720 (7%) | 309 (49%) |
| **pT category (AJCC 2002)** | | |
| pT2 | 7,370 (66%) | 570 (9%) |
| pT3a | 2,409 (22%) | 587 (28%) |
| pT3b | 1,262 (11%) | 618 (55%) |
| pT4 | 63 (1%) | 49 (80%) |
| **Gleason grade** | | |
| ≤3+3 | 2,859 (26%) | 193 (8%) |
| 3+4 | 6,183 (56%) | 849 (16%) |
| 4+3 | 1,565 (14%) | 573 (42%) |
| ≥4+4 | 482 (4%) | 208 (50%) |
| **pN category** | | |
| pN0 | 6,117 (92%) | 1,126 (21%) |
| pN+ | 561 (8%) | 291 (59%) |
| **Surgical margin** | | |
| negative | 8,984 (82%) | 1,146 (15%) |
| positive | 1,970 (18%) | 642 (37%) |

* / ** numbers do not always add up to 11,152/1,824 in categories because of cases with missing data.

Abbreviation: AJCC, American Joint Committee on Cancer. *** p value not significant (ns) >0.05

**Table 3 Composition of the prostate prognosis tissue microarray containing 11,152 prostate cancer specimens.** The number and fraction of samples in each category, as well

as the number and fraction of samples with biochemical relaps within the different categories, are shown.

### 4.2.6 Multivariate analysis

Four multivariate analyses were performed evaluating the clinical relevance of *BAZ2A* expression in different scenarios. Scenario 1 was utilizing all post-operatively available parameters including pT, pN, margin status, pre-operative PSA value and Gleason grade obtained on the resected prostate. Scenario 2 was utilizing all postoperatively available parameters with the exception of nodal status. The rational for this approach was that lymphadenectomy is not a routine procedure in the surgical therapy of prostate cancer and that excluding pN in multivariate analysis increases case numbers. The next two scenarios tried to better model the pre-operative situation. Scenario 3 included the *BAZ2A* expression, pre-operative PSA, clinical stage (cT) and the Gleason grade obtained on the prostatectomy specimen. Because the post-operative Gleason grade varies from the pre-operative Gleason grade, another multivariate analysis was added as scenario 4. In this scenario, the pre-operative Gleason grade obtained on the original biopsy was combined with pre-operative PSA, clinical stage and *BAZ2A* expression. All four scenarios suggest a tendency towards *BAZ2A* representing an independent predictor of prognosis.

## 4.3 Results

By comparing the expression of all expressed genes with all expressed microRNAs, and combining somatic genomic variant (CNAs, LOH, SNVs) and DNA methylation data to filter for alternative (non-microRNA-mediated)

mechanisms of gene dysregulation, a list of significant microRNA:target gene pairs was generated. Among these pairs, miR-133a and *BAZ2A* were found to be downregulated and overexpressed, respectively. MiR-133a, a tumor suppressor in several cancer types[186,187] has been recently reported to have tumor-suppressive properties in prostate cancer[188,189]. Its predicted target, *BAZ2A*, is a key component of the nucleolar remodeling complex and is known to interact with DNMTs[190] and HDACs[191] to establish epigenetic silencing of rDNA. Corresponding downregulation of miR-133a and overexpression of *BAZ2A* were confirmed in a larger second dataset[192] (**Figure 14**) and the direct miR-133a:*BAZ2A* interaction was also able to be validated in vitro. Mir-133a was found to selectively suppress the expression of a luciferase-*BAZ2A* construct via an interaction with a single, highly conserved site within the 3'UTR (**Figure 15**). Furthermore, overexpression of miR-133a significantly reduced *BAZ2A* levels in the normal prostate cell line BPH1 and the prostate cancer cell line DU145 (**Figure 16**).

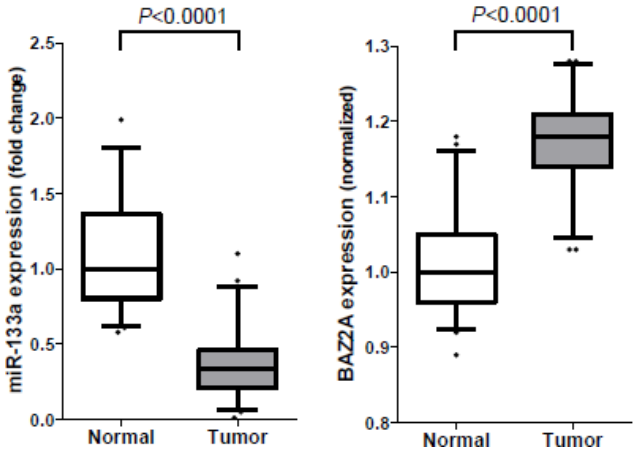



**Figure 14 miR-133a and *BAZ2A* expression in tumor and normal.** Validation of miR-133a downregulation and *BAZ2A* overexpression in secondary, larger datasets.

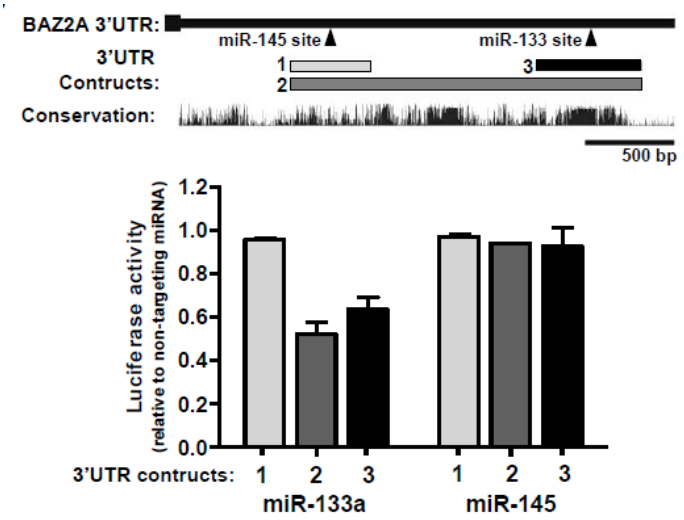



**Figure 15 Validating mir:target interaction *in vitro*.** Luciferase assay evaluating the direct interaction of miRNAs with the *BAZ2A* 3'UTR. MiR-133a specifically interacts with a distal conserved the site in the 3'UTR, versus miR-145, another miRNA predicted to target the 3'UTR.

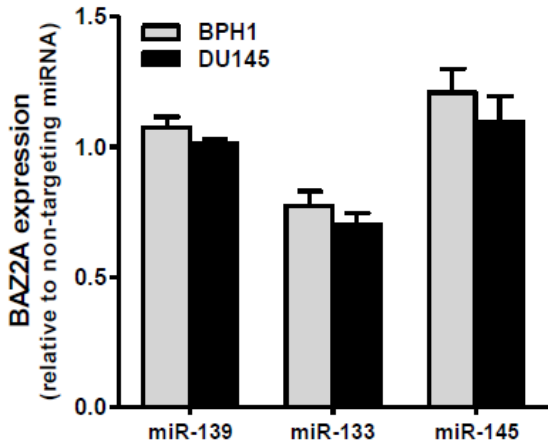



**Figure 16 miR-133a and *BAZ2A* expression in cancer cell lines.** Overexpression of miR-133a results in downregulation of *BAZ2A* in BPH1 and DU145 prostate cell lines versus other miRNAs, miR-139 and miR-145.

To investigate the role of *BAZ2A* protein expression,

immunohistochemistry was performed on a pilot tissue-microarray (TMA) of 384 clinical prostate tumor samples. Indeed, *BAZ2A* immunostaining was variable within prostate cancers with strong staining in 59 (20.7%), moderate in 55 (19.4%), and weak in 72 (25.4%) cancers while normal prostate epithelium did not show relevant staining. As *BAZ2A* is known to establish epigenetic silencing via the recruitment of DNMTs, whether upregulation of *BAZ2A* is associated with altered global DNA methylation was then investigated. From the pilot TMA, 22 and 13 prostate tumors with high and low *BAZ2A* levels were selected, respectively. Samples were also selected to have high (>70%) tumor content. DNA methylation analysis was performed using Illumina 450k Infinium arrays. Genome-wide analysis revealed that 32,707 CpGs were significantly altered (>±20%; q-value<0.05) versus 6 normal prostate samples, with 24,497 and 8,210 CpGs being hyper- and hypomethylated, respectively. Unsupervised clustering of the 3,000 most variable of these CpGs identified two distinct DNA methylation subtypes. A statistical evaluation testing the optimal number of methylation subtypes confirmed the existence of two subtypes (**Figure 17**; **Figure 18**).

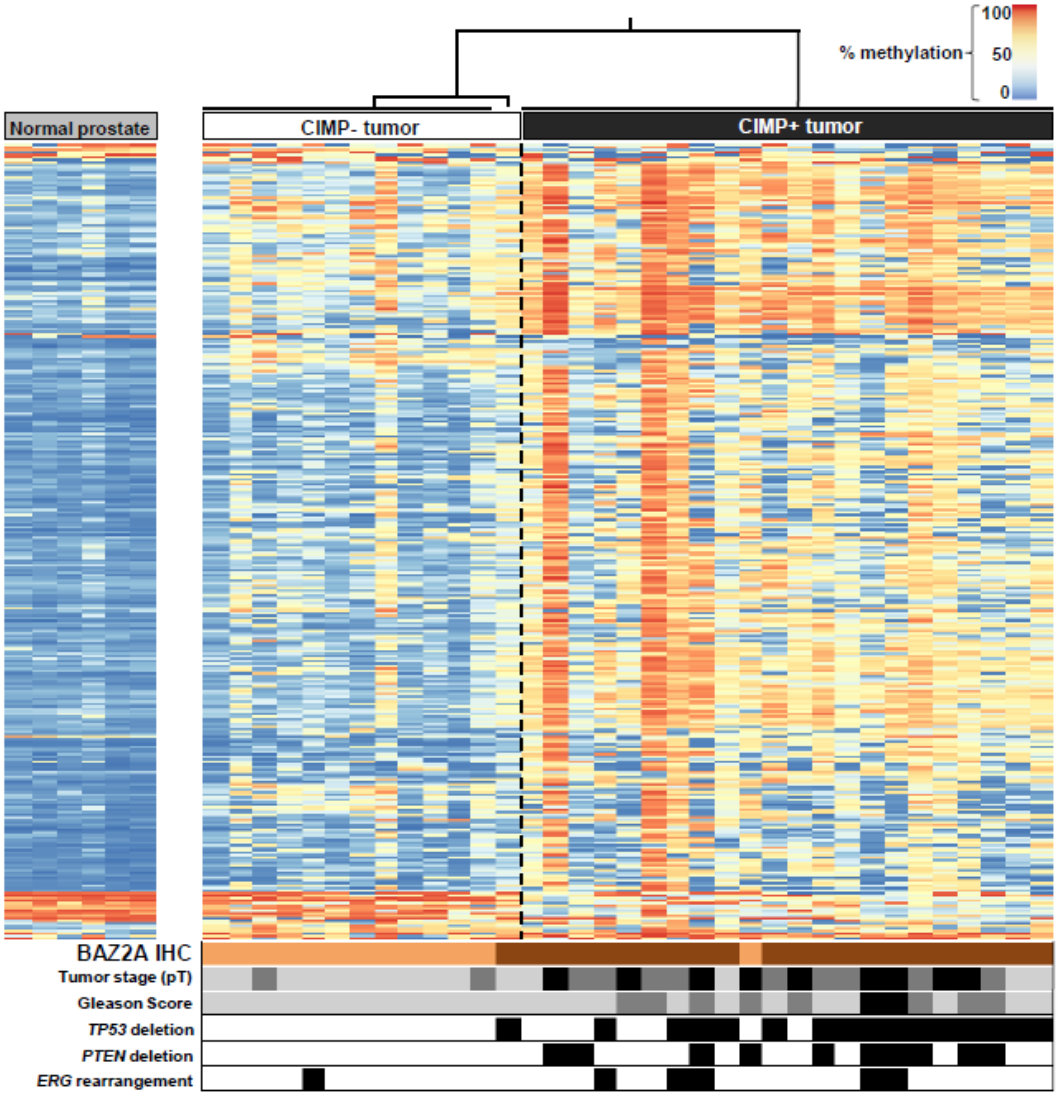



**Figure 17 Unsupervised clustering of tumor samples based on methylation level.** DNA methylation heatmap of the most variable 3000 CpGs. Hierarchical clustering of tumors identifies two DNA methylation subtypes displaying relatively high and low levels of methylation (termed CIMP+ and CIMP-, respectively). Tumors displaying high and low levels of *BAZ2A* from immunohistochemical (IHC) evaluation are illustrated by dark and light brown color, respectively. Increasing tumor (pT) stage (pT2, pT3a and pT3b) and Gleason score (3+4, 4+3, ≥4+4) are illustrated by increasingly light grey, dark grey and black indicators, respectively.

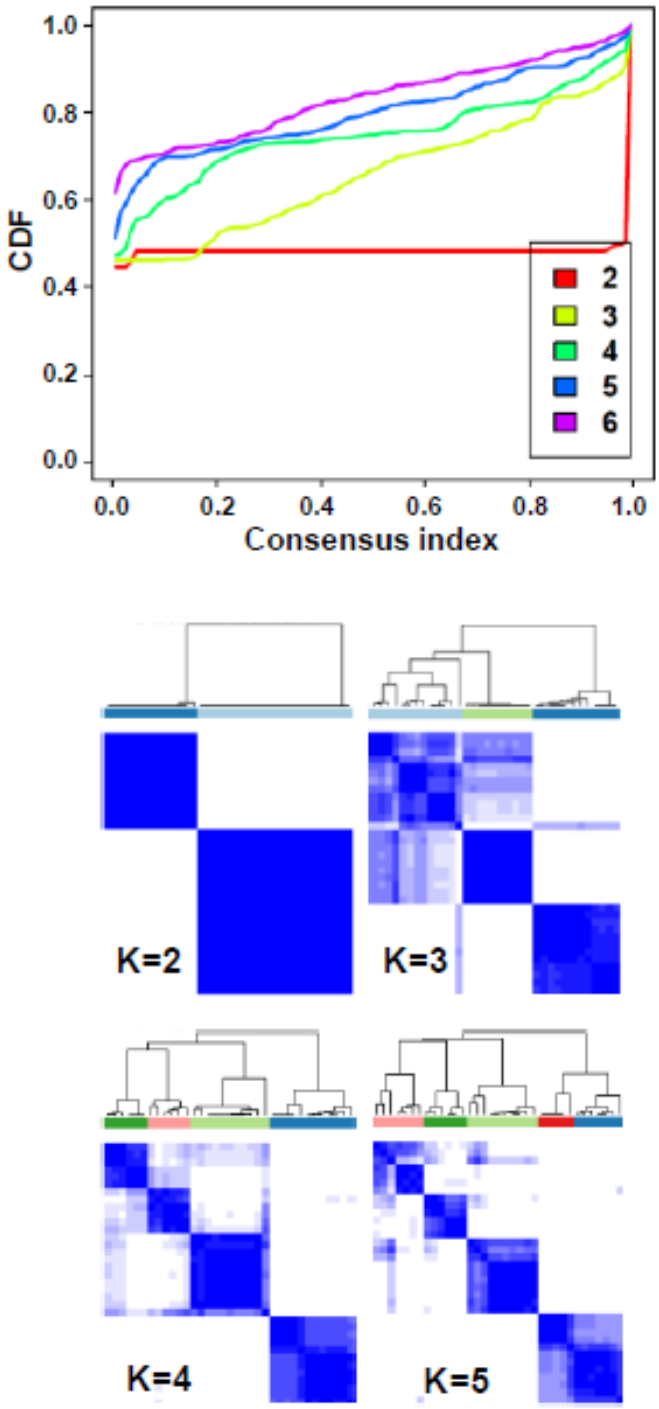



**Figure 18 Consensus clustering of tumor samples.** The optimal number of methylation subtypes was determined by the Consensus Cumulative Distribution Function (CDF).

One of the subtypes is characterized by a higher degree of hypermethylation within CGIs and sites associated with polycomb repression,

while simultaneously demonstrating hypomethylation of repetitive elements (such as LINEs) (**Figure 19**).

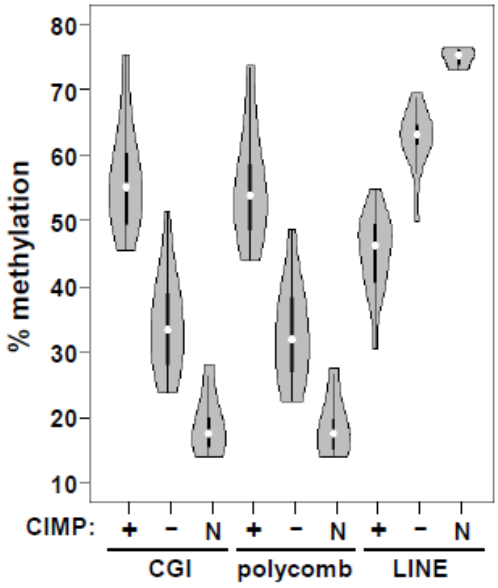



**Figure 19 Bean plot of methylation level in CGI, polycomb and LINE for two subgroups in prostate tumors and normals.** DNA methylation levels of CpG island, polycomb-associated and LINE regions in CIMP+ and CIMP- subtypes as well as in normal tissue (CGI, CpG island; LINE, long interspersed element).

Thus, based on descriptions of similar findings in other cancer types[193-195], this subtype is termed as a CpG island hypermethylator phenotype (CIMP+) and, conversely, CIMP- for tumor samples from the other subtype. Strikingly, in the CIMP+ subtype, 21/22 samples have high *BAZ2A* levels, and for CIMP-, 12/13 have low *BAZ2A* levels, linking *BAZ2A* to a high degree of abnormal methylation in prostate tumors. Relative to the CIMP- subtype, 6,155 CpGs were hypermethylated and 1,679 CpGs were hypomethylated (>± 20%) in the CIMP+/*BAZ2A*-high subtype. Hypermethylated CpGs were enriched in CpG islands (CGIs) as well as CGI shore regions (**Figure 20**). Along with enrichment of CpGs in transcription factor binding and DNase-hypersensitive

sites, enrichment at CGIs indicates that hypermethylation targets functionally-relevant regions of the genome in prostate cancer. Although less frequent, hypomethylation is also significantly enriched at non-CGI-associated promoters and enhancers. Thus, elevated *BAZ2A* levels are associated with widespread epigenetic remodeling, including functional regions, such as promoter and enhancer regions, and polycomb-associated domains.

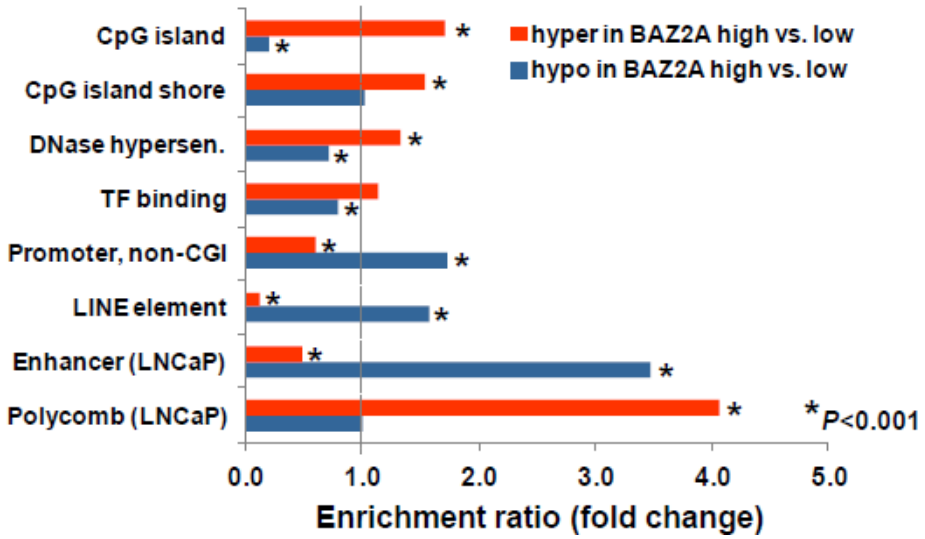



**Figure 20 Enrichment plot for two subgroups in prostate tumors.** Enrichment and/or depletion of genomic features in *BAZ2A*-high versus *BAZ2A*-low tumors. Annotation of enhancer and polycomb features are derived from ChIP-seq profiles from the prostate cancer cell line, LNCaP (TF, transcription factor).

*BAZ2A*-associated alterations to DNA methylation were found to occur at numerous genes that are associated with prostate cancer as well as other malignancies. As expected from other studies that have identified omnipresent *GSTP1* hypermethylation in prostate cancer[196,197], the *GSTP1* promoter CGI was found hypermethylated in the majority (>90%) of prostate tumors (**Figure 21**). Importantly, hypermethylation was found to occur at similar levels in

CIMP+/*BAZ2A*-high and CIMP-/*BAZ2A*-low prostate tumors, confirming that differential methylation between subtypes does not result from differential sample tumor content. Similarly, the tumor-suppressor gene, *APC*, along with several other genes, was found to be hypermethylated at equal frequency in *BAZ2A*-high and low subtypes.

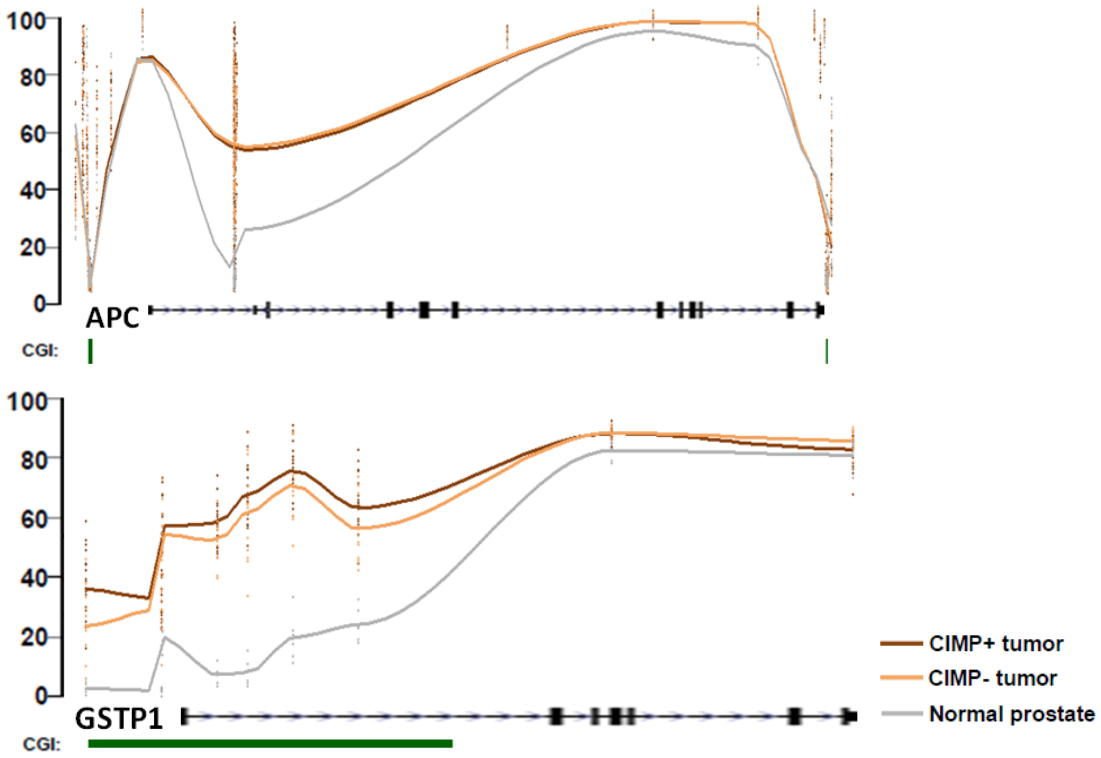



**Figure 21 DNA methylation profiles of *GSTP1* and *APC*.** *GSTP1* and *APC* are consistently hypermethylated in all prostate tumor samples relative to normals, demonstrating that tumor content does not appreciably differ between tumor samples.

Specific to *BAZ2A*-high tumors, hypermethylation of several tumor-suppressor genes with known roles in prostate cancer were observed, such as *PAX6*[198], *WT1*[199], *GATA3*[200] and *SFRP2*[201] (**Figure 22; Figure 23**). In addition, hypermethylation of microRNAs 9-1, 9-3, 34b/c and 124-2, all previously identified to inhibit androgen receptor expression[202-204], were found

to predominate in *BAZ2A*-high tumors (**Figure 24**). Together, these findings demonstrate that in addition to epigenetic changes that broadly occur in prostate tumors, additional epigenetic remodeling occurs in *BAZ2A*-high expressing tumors.

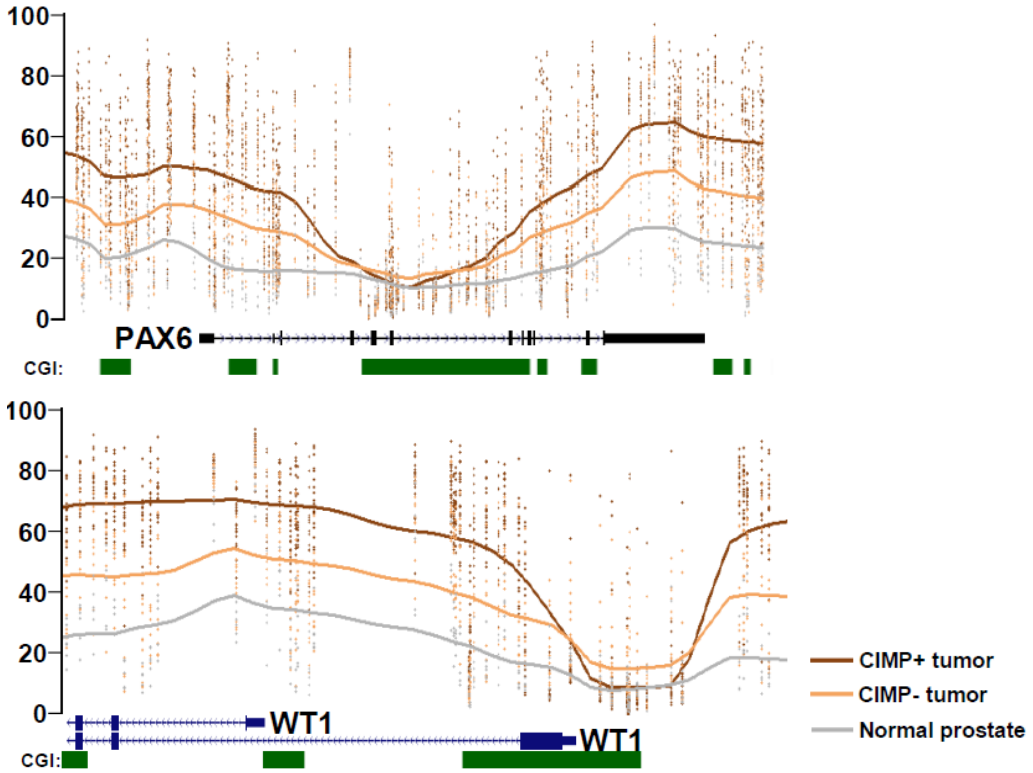
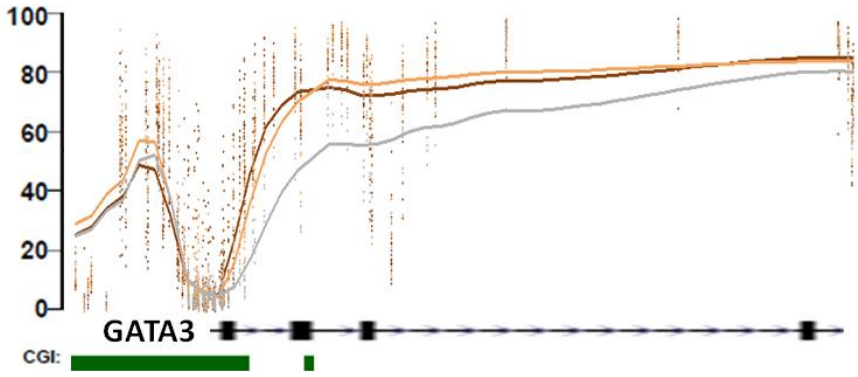



**Figure 22 DNA methylation profiles of *PAX6* and *WT1*.** DNA methylation profiles of the promoter regions of the tumor-suppressor genes *PAX6* and *WT1* in CIMP+ and CIMP- tumor subtypes versus normal prostate tissue.



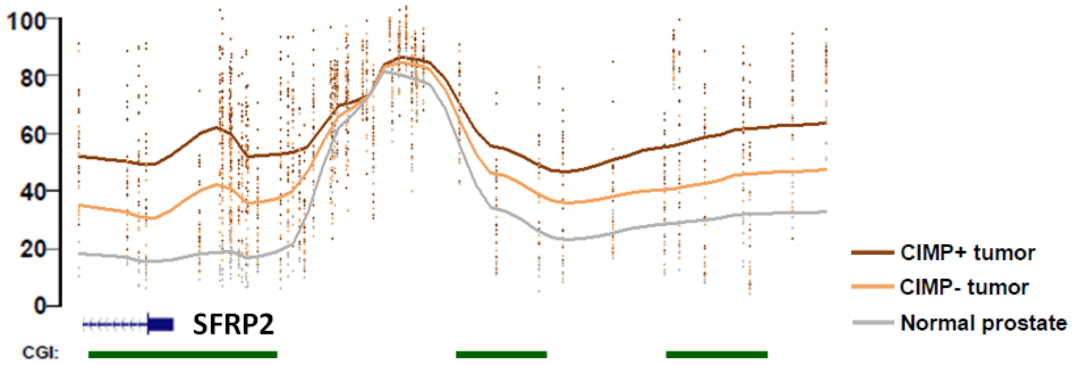



**Figure 23 DNA methylation profiles of *GATA3* and *SFRP2*.** Hypermethylation of known tumor suppressors *GATA3* and *SFRP2* occurs primarily in the CIMP+/*BAZ2A*-high subtype.

Recent data has demonstrated that altering *BAZ2A* expression levels modifies the telomeric and centromeric chromatin configurations leading to genomic instability[205]. Thus, whether CIMP+/*BAZ2A*-upregulated and CIMP-/*BAZ2A*-normal subtypes were also associated with variable amounts of genomic alterations was investigated. For this purpose, copy number alterations (CNAs) were inferred from the Illumina 450k Infinium array data. A dramatic increase in the number of CNAs in the *BAZ2A*-high subtype was observed, while few CNAs were present in *BAZ2A*-low tumors (P<0.001, **Figure 25**). BAZ2A-upregulated tumors also are enriched for *ERG* fusions, as well as *PTEN* and *TP53* deletions (P<0.01, **Figure 17**). These findings are also found across all samples analyzed by TMA (**Figure 26**). Together, these findings show that in addition to epigenetic alterations, *BAZ2A* levels are also associated with genetic instability in prostate cancer.

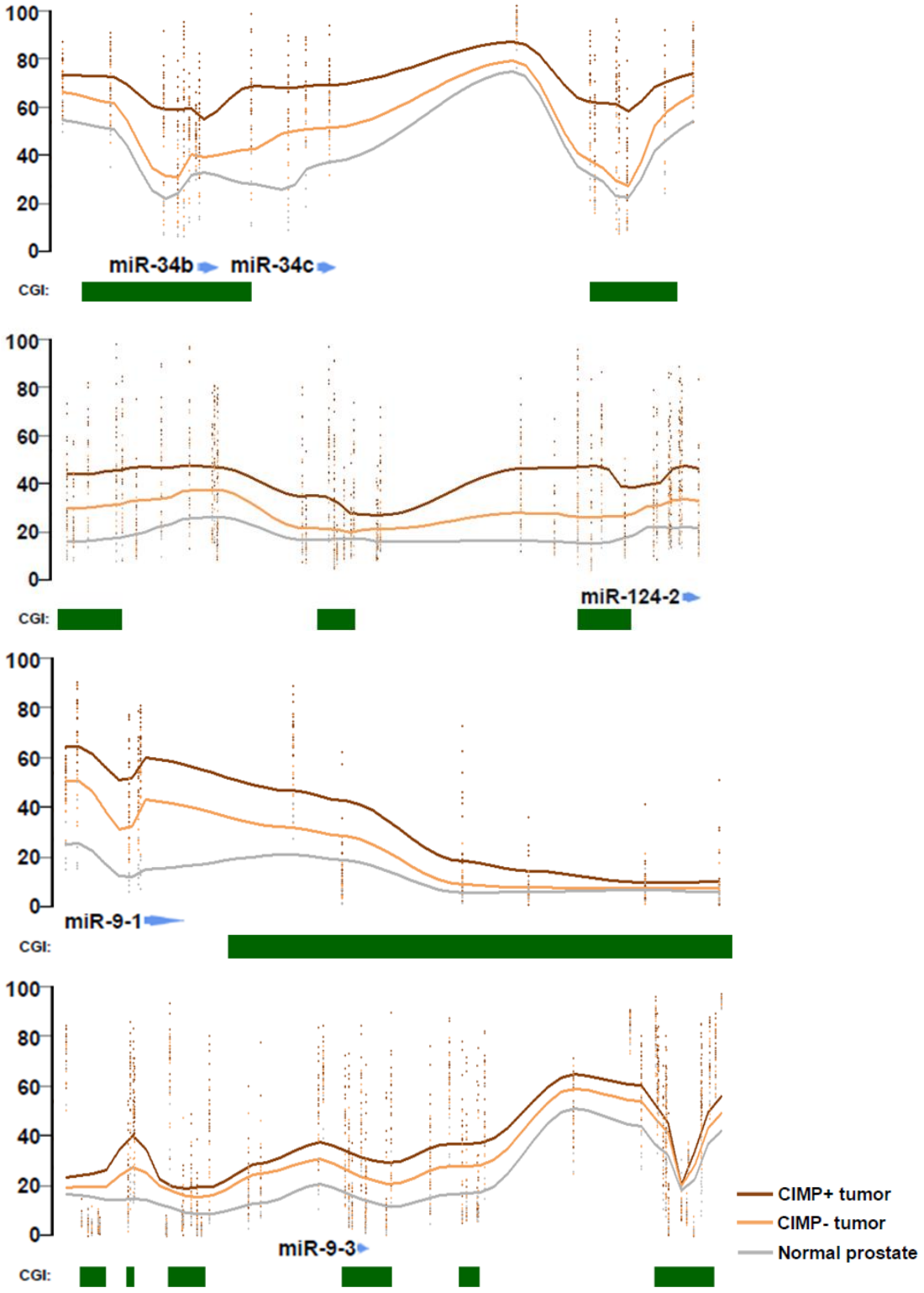



**Figure 24 DNA methylation profiles of prostate tumor associated microRNAs.** Hypermethylation of microRNAs 9-1, 9-3, 124-2, 34b and 34c, known to regulate androgen receptor expression, occurs primarily in the CIMP+/*BAZ2A*-high subtype.

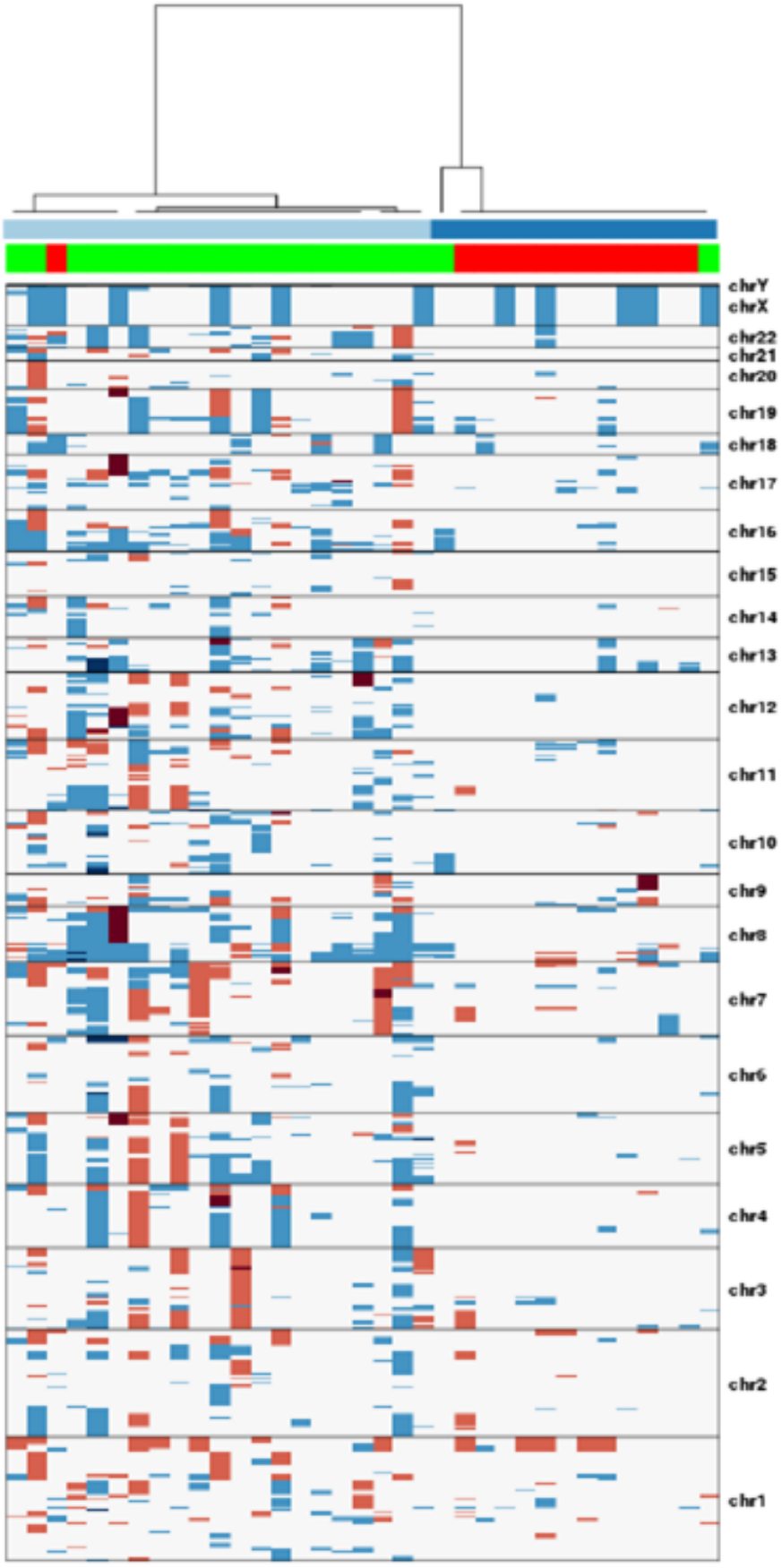



**Figure 25 CNV profiles in two subgroups of prostate tumors.** Samples are clustered

according to the copy number alteration profile. (deletion = blue; amplification= red)

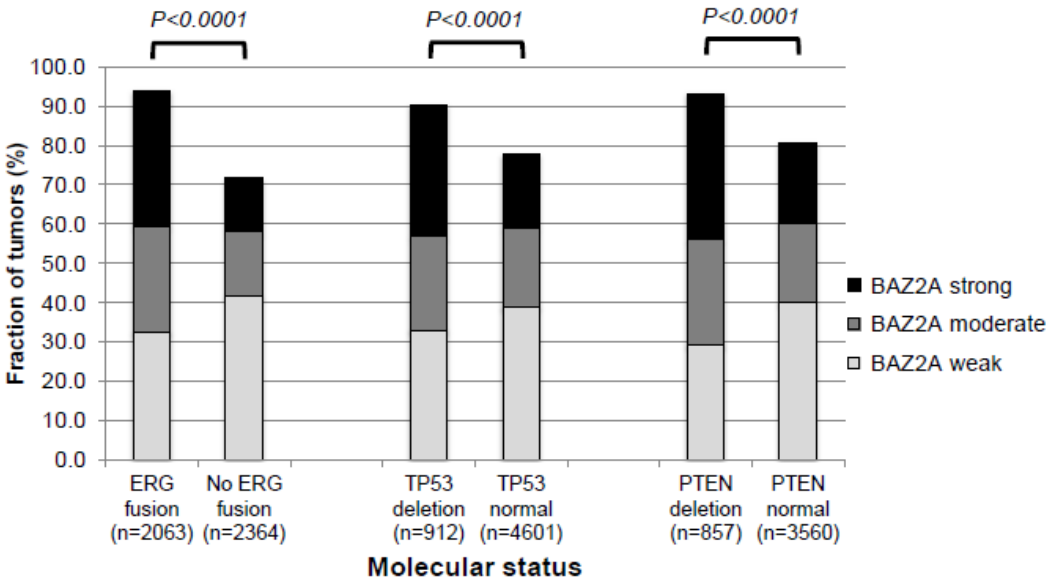



**Figure 26 *BAZ2A* expression correlates with *ERG* fusion, *TP53* and *PTEN* deletion.** The proportion of tumors that have strong, moderate or weak levels of *BAZ2A* staining from TMA analysis separated by either *ERG* fusion status, *TP53* deletion or *PTEN* deletion.

To further investigate the potential clinical impact of *BAZ2A* expression, a tissue microarray tissue microarray containing samples from >10,000 prostate cancers was investigated by means of investigated by means of immunohistochemistry. This analysis resulted in 7,682 informative 7,682 informative samples for which clinical follow up data were available. Patient characteristics Patient characteristics and clinical data are displayed in

**Table 3**. *BAZ2A* immunostaining was again categorized as negative (26.1%), weak (36.7%), moderate (18.5%) and strong (19.0%) (**Figure 27**; **Figure 28**).


**Figure 27 Tissue microarray analysis of *BAZ2A* level.** Representative examples of TMA histological sections showing negative, weak, moderate and strong *BAZ2A* staining classifications.

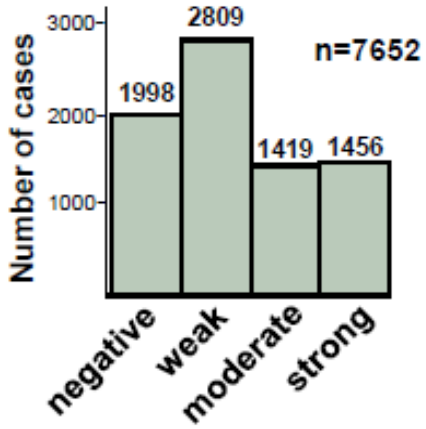



**Figure 28 Sample distribution based on *BAZ2A* expression in TMA.** Distribution of negative, weak, moderate and strong *BAZ2A* tumors as assessed by tissue microarray analysis.

Strong *BAZ2A* levels were highly associated with advanced pT stage, high Gleason grade, the presence of lymph node metastasis, high preoperative PSA level and positive surgical margin when considering all tumors or following subgrouping by *ERG* status (P<0.0001 each; **Table 4**). The time to postoperative PSA recurrence was significantly shorter in the *BAZ2A* strong

group (P<0.0001, **Figure 29**) and this finding was again observed to be independent of *ERG* status (P<0.0001, **Figure 30**). Using Cox regression multivariate analysis to determine the relative dependence of several prognostic and surgical parameters, the level of *BAZ2A* was determined to be independently predictive for the factor of PSA recurrence in multiple scenarios including various combinations of parameters (P<0.0001, **Table 5**). The independent predictive power of *BAZ2A* was further upheld following subdivision of the cases by *ERG* status (**Table 5**).

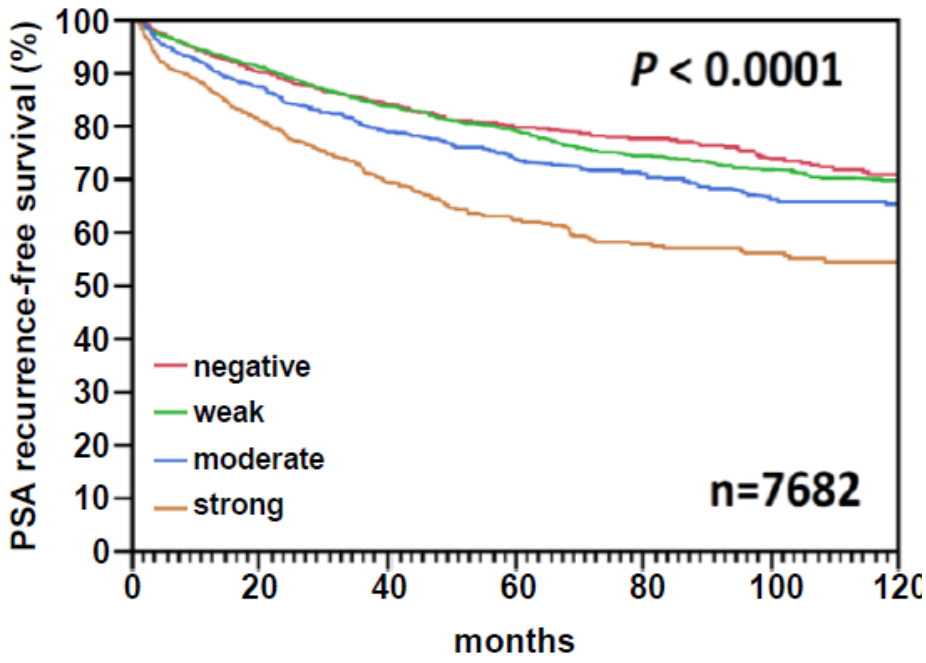



**Figure 29 PSA recurrence-free survival analysis for all prostate tumors.** Kaplan-Maier analysis of the time to postoperative PSA recurrence versus *BAZ2A* levels in all prostate tumors.

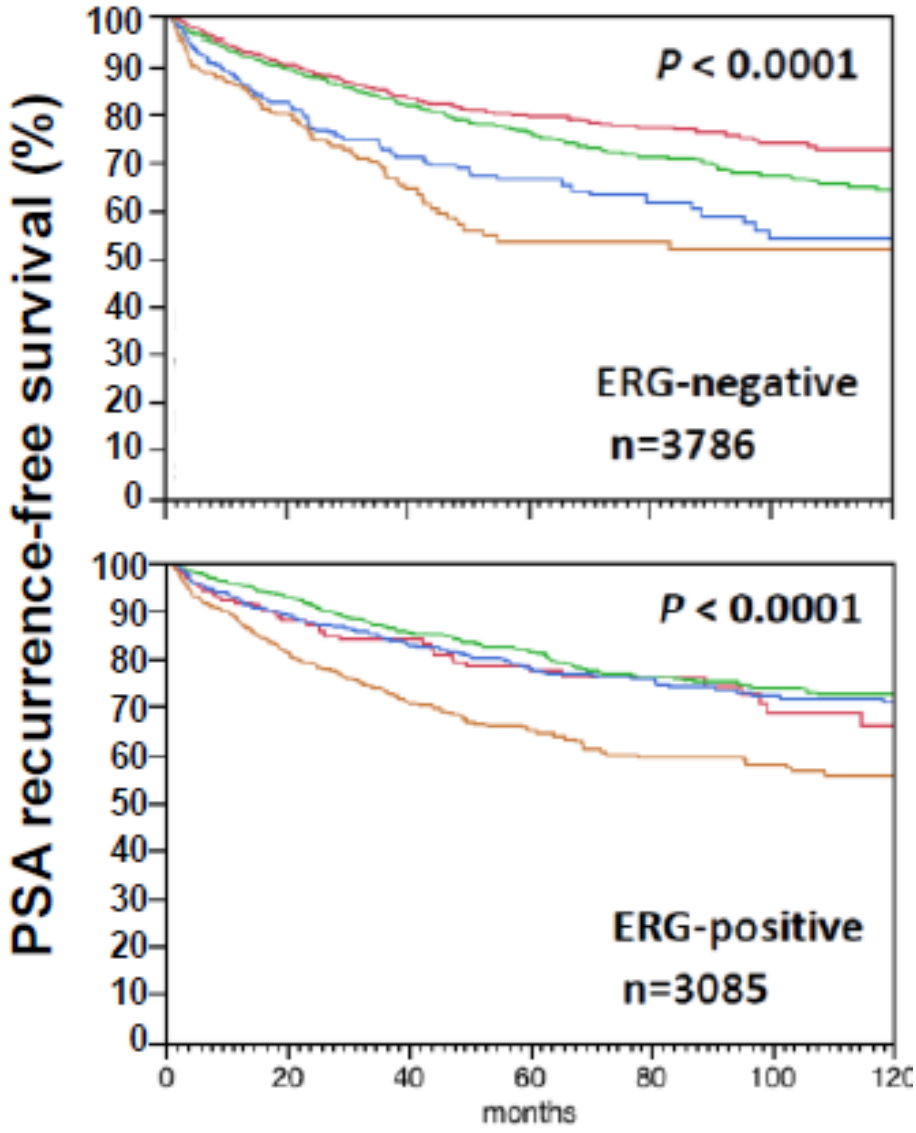



**Figure 30 PSA recurrence-free survival analysis for ERG positive and negative prostate tumors.** Kaplan-Maier analysis of the time to postoperative PSA recurrence versus *BAZ2A* levels in prostate tumors separated by *ERG* status.

a)

| Parameter | n Evaluable | IHC result | | | | *P* |
|---|---|---|---|---|---|---|
| | | Negative (%) | Weak (%) | Moderate (%) | Strong (%) | |
| **All cancers** | 7682 | 26 | 37 | 18 | 19 | |
| **Tumor stage** | | | | | | < 0.0001 |
| pT2 | 4872 | 29 | 38 | 18 | 15 | |
| pT3a | 1763 | 22 | 35 | 19 | 24 | |
| pT3b | 951 | 19 | 33 | 20 | 29 | |
| pT4 | 59 | 24 | 31 | 17 | 29 | |
| **Gleason grade** | | | | | | < 0.0001 |
| ⩽3+3 | 1791 | 32 | 43 | 15 | 11 | |
| 3+4 | 4340 | 26 | 36 | 20 | 19 | |
| 4+3 | 1163 | 20 | 33 | 19 | 29 | |
| ⩾4+4 | 341 | 21 | 25 | 21 | 34 | |
| **Lymph node metastasis** | | | | | | < 0.0001 |
| N0 | 4376 | 25 | 35 | 19 | 21 | |
| N+ | 422 | 18 | 32 | 20 | 31 | |
| **Preoperative PSA level (ng/mL)** | | | | | | < 0.0001 |
| <4 | 877 | 19 | 36 | 21 | 23 | |
| 4-10 | 4594 | 27 | 36 | 19 | 18 | |
| 10-20 | 1543 | 27 | 36 | 18 | 19 | |
| >20 | 567 | 30 | 37 | 14 | 19 | |
| **Surgical mergin** | | | | | | 0.0005 |
| Negative | 6051 | 26 | 37 | 19 | 18 | |
| Positive | 1488 | 24 | 35 | 18 | 23 | |

NOTE. Number do not always add up to 7682 in different categories because of cases with missing data

b)

| Parameter | n Evaluable | IHC result | | | | *P* |
|---|---|---|---|---|---|---|
| | | Negative (%) | Weak (%) | Moderate (%) | Strong (%) | |
| **All cancers** | 3786 | 37 | 37 | 13 | 11 | |
| **Tumor stage** | | | | | | < 0.0001 |
| pT2 | 2508 | 42 | 38 | 12 | 8 | |
| pT3a | 780 | 36 | 37 | 15 | 12 | |
| pT3b | 459 | 27 | 37 | 14 | 23 | |
| pT4 | 27 | 37 | 22 | 22 | 19 | |
| **Gleason grade** | | | | | | < 0.0001 |
| ⩽3+3 | 825 | 48 | 40 | 8 | 4 | |
| 3+4 | 2127 | 39 | 38 | 14 | 9 | |
| 4+3 | 605 | 30 | 35 | 16 | 19 | |
| ⩾4+4 | 213 | 27 | 28 | 20 | 26 | |
| **Lymph node metastasis** | | | | | | < 0.0001 |
| N0 | 2206 | 37 | 38 | 14 | 12 | |
| N+ | 200 | 26 | 30 | 19 | 27 | |
| **Preoperative PSA level (ng/mL)** | | | | | | 0.0089 |
| <4 | 367 | 32 | 38 | 14 | 16 | |
| 4-10 | 2246 | 39 | 38 | 13 | 10 | |
| 10-20 | 826 | 38 | 36 | 14 | 12 | |
| >20 | 311 | 43 | 35 | 11 | 11 | |
| **Surgical mergin** | | | | | | 0.018 |
| Negative | 2997 | 39 | 38 | 13 | 10 | |
| Positive | 720 | 38 | 37 | 12 | 14 | |

NOTE. Number do not always add up to 7682 in different categories because of cases with missing data

c)

| Parameter | n Evaluable | IHC result | | | | *P* |
|---|---|---|---|---|---|---|
| | | Negative (%) | Weak (%) | Moderate (%) | Strong (%) | |
| **All cancers** | 3085 | 9 | 35 | 26 | 30 | |
| **Tumor stage** | | | | | | < 0.0001 |
| pT2 | 1801 | 10 | 37 | 27 | 26 | |
| pT3a | 841 | 9 | 34 | 23 | 34 | |
| pT3b | 402 | 7 | 30 | 27 | 37 | |
| pT4 | 23 | 9 | 39 | 9 | 43 | |
| **Gleason grade** | | | | | | < 0.0001 |
| ⩽3+3 | 686 | 11 | 45 | 26 | 19 | |
| 3+4 | 1806 | 10 | 34 | 27 | 29 | |
| 4+3 | 465 | 6 | 29 | 23 | 42 | |
| ⩾4+4 | 104 | 6 | 17 | 23 | 54 | |
| **Lymph node metastasis** | | | | | | 0.5818 |
| N0 | 1767 | 9 | 33 | 26 | 33 | |
| N+ | 186 | 8 | 31 | 24 | 38 | |
| **Preoperative PSA level (ng/mL)** | | | | | | 0.0044 |
| <4 | 401 | 6 | 34 | 28 | 32 | |
| 4-10 | 1870 | 10 | 33 | 27 | 29 | |
| 10-20 | 559 | 7 | 38 | 25 | 30 | |
| >20 | 209 | 12 | 39 | 17 | 32 | |
| **Surgical mergin** | | | | | | 0.059 |
| Negative | 2390 | 9 | 36 | 26 | 28 | |
| Positive | 637 | 8 | 33 | 25 | 34 | |

NOTE. Number do not always add up to 7682 in different categories because of cases with missing data

**Table 4 Associations between *BAZ2A* expression and clinical outcomes.** Strong *BAZ2A* levels were highly associated with advanced pT stage, high Gleason grade, the presence of lymph node metastasis, high preoperative PSA level and positive surgical margin when considering all tumors **(a)** or following subgrouping by *ERG* status **(b and c)**.

a)

| Scenario | N evaluable | P | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Preoperative PSA-level | pT stage | cT stage | Gleason grade prostatectomy | Biopsy Gleason grade | Lymph node metastasis | Surgical margin | BAZ2A expression on TMA |
| 1 | 4095 | <0.0001 | <0.0001 | - | <0.0001 | - | <0.0001 | <0.0001 | <0.0001 |
| 2 | 6603 | <0.0001 | <0.0001 | - | <0.0001 | - | - | <0.0001 | <0.0001 |
| 3 | 6468 | <0.0001 | - | <0.0001 | <0.0001 | - | - | - | <0.0001 |
| 4 | 6367 | <0.0001 | - | <0.0001 | - | <0.0001 | - | - | <0.0001 |

b)

| Scenario | N evaluable | P | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Preoperative PSA-level | pT stage | cT stage | Gleason grade prostatectomy | Biopsy Gleason grade | Lymph node metastasis | Surgical margin | BAZ2A expression on TMA |
| 1 | 2034 | 0.0062 | <0.0001 | - | <0.0001 | - | <0.0001 | 0.0034 | 0.0199 |
| 2 | 3230 | <0.0001 | <0.0001 | - | <0.0001 | - | - | 0.0006 | 0.0006 |
| 3 | 3186 | <0.0001 | - | <0.0001 | <0.0001 | - | - | - | 0.0007 |
| 4 | 3145 | <0.0001 | - | 0.0003 | - | <0.0001 | - | - | 0.0001 |

c)

| Scenario | N evaluable | P | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Preoperative PSA-level | pT stage | cT stage | Gleason grade prostatectomy | Biopsy Gleason grade | Lymph node metastasis | Surgical margin | BAZ2A expression on TMA |
| 1 | 1673 | 0.0349 | <0.0001 | - | <0.0001 | - | 0.0077 | 0.0134 | <0.0001 |
| 2 | 2655 | 0.0008 | <0.0001 | - | <0.0001 | - | - | 0.0039 | <0.0001 |
| 3 | 2584 | <0.0001 | - | <0.0001 | <0.0001 | - | - | - | <0.0001 |
| 4 | 2537 | <0.0001 | - | <0.0001 | - | <0.0001 | - | - | <0.0001 |

**Table 5 Multivariate analysis indicating *BAZ2A* being a independent predictor of prognosis.** Cox regression multivariate analysis illustrating the relative dependence of several prognostic and surgical parameters on PSA recurrence in **a)** all cancers, **b)** *ERG* negative, and **c)** *ERG* positive prostate.

Our results demonstrate a clear link of aberrant *BAZ2A* expression with prostate cancer. Specifically, overexpression of *BAZ2A* – potentially caused by downregulation of the tumor-suppressive miR-133a - is tightly associated with a molecular subtype defined by substantial genomic instability and an aberrant genomic pattern of DNA hypermethylation (CIMP). Given these substantial implications on the biology of cancer cells, it is not surprising, that *BAZ2A* overexpression has a strong prognostic impact, which is furthermore independent of classical prognostic markers. An increasing number of key epigenetic regulatory genes, including bromodomain-containing proteins, are currently found to be dysregulated across many cancer types and represent novel targets of a new generation of cancer therapeutics. *BAZ2A* may not only thus serve as a biomarker that may help to distinguish indolent from aggressive prostate cancer, but may also qualify as a potential target for future treatment of prostate cancer.

## 4.4 Discussion

The downregulation of the tumor suppressor, miR-133a, could be partially explained by hypermethylation of its putative promoter region, and the high *BAZ2A* expression further establishes the feedback loop to maintain the hypermethylation of its promoter region (**Figure 31**). However, where is the initial hypermethylation from is still unknown. A possible explaination is that randomly induced methylation variations hit the promoter of miR-133a and then leads to the downregulation followed by the upregulation of *BAZ2A*. The upregulation of BAZ2A then changes the global methylation profile to increase the epigenetic variability and destroy epigenetic signatures in tumor cells. This may arise cancer hallmarks by creating the heterogeneous environments and phenotypes which may increase the accumulation of genetic alterations that

are advantageous to tumor cell development by natural selection. That's why the high BAZ2A expression is observed in aggressive prostate tumors showing poorer prognosis.

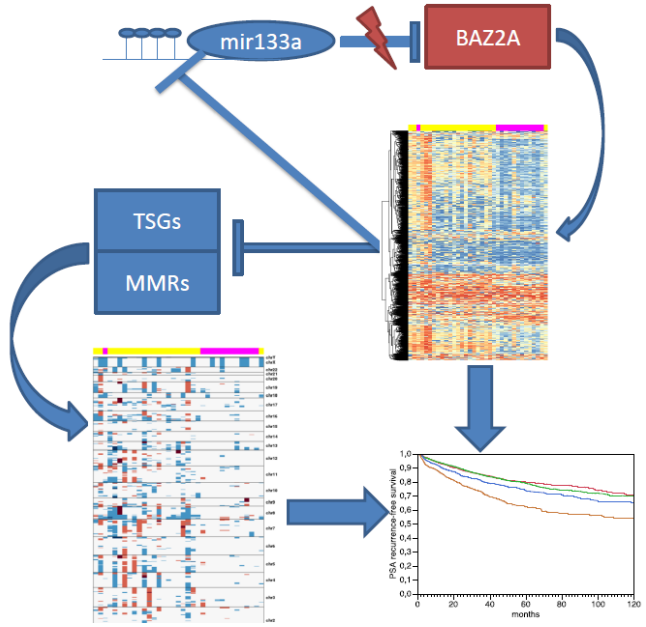



**Figure 31 A conceptional model illustrating the possible BAZ2A driven epigenetic alterations in prostate cancers.**

So far, there are only three genes, *BRAF, IDH1* and *H3.3,* discovered to associated with CIMP [107,180,206]. All these genes are affected by mutations. However, the mutation rate is low in prostate cancer. Thus, epigenetic factors might play a key role instead of mutations. It's known that microRNAs can fine tune the expression of their target genes. BAZ2A is a key gene for chromotin remodeling with stronge effects to cell survive. Therefore, the dramatic changes from mutations or structure variations might kill cells. Instead, a slight expression change modulated by microRNA might already enough to reset the chromation structure and global methylation pattern. A recent study[207] found that a long non-coding RNA, *SChLAP1*, contributes to the development of

lethal prostate cancer at least in part by antagonizing the tumor-suppressive functions of the SWI/SNF complex. Actually, the SWI/SNF complex is a nucleosome remodeling complex which have the overlapping function of BAZ2A. Unlike other known long non-coding RNAs such as HOTAIR which enhance the function of epignenetic complexes such as PRC2 and MLL [208-210], *SChLAP1*, however, impairs this key epigenetic complex with tumor suppresive function[211-216]. Thus, not only genetic mutations but also epigenetic factors can influence the epigenetic key complex and form the feedback loop to further enhancer and reprogram the epigenetic landscape in cancer.

To further investigate the molecular mechanism of how *BAZ2A* induces (epi)genetic alterations, more functional experiments should be carried out. Nevertheless, this study identified an potential epigenetic key player which has a great prognostic power to distinguish the indolent and aggressive prostate tumor.

# Chapter 5: Environmentally induced epigenetic reprogramming in mothers and their newborn children

**Note:**

Mario Bauer, Gunda Herberth, Dieter Weichenhan, Loreen Thürmann, Saskia Trump, and Kristin Junge performed experimental work, collected data and provided proband materials. The DKFZ Genomics and Proteomics Core Facility provided technical support for sequencing. Oliver Mücke, Marion Bähr, Monika Helf provided support in MassARRAY validation. Beate Fink, Anne Hain, and Melanie Nowak provided technical assistance and field work. Rolle-Kampczyk and Martin von Bergen provided urine cotinine concentrations.

## 5.1 Aim of the study

Increasing evidence has emerged that environmental exposure during the prenatal period can increase the risk to develop diseases later in life. Epigenetic mechanisms, such as changes in DNA methylation and histone modifications that together modify DNA accessibility for gene transcription in a persistent way, are discussed as potential link between early environmental exposure and later disease[217]. Prenatal exposure to tobacco smoke was described as a risk factor for a multitude of different diseases in the child, including lung diseases, obesity, and cancer[218-220]. Many studies have shown that maternal exposure to tobacco smoke induces diverse site-specific methylation changes, but the mechanistic insights derived from those epidemiological studies remain very limited[221-223]. How an important environmental stressor shapes the epigenome in healthy human individuals still remains unclear.

Here, the genome-wide, environmentally induced methylation changes

and their functional relationship with chromatin regulators in mothers and their children during pregnancy was studied for the first time at base-pair resolution. A larger validation panel of 45 mother/child pairs was explored through targeted methylome analysis followed by a broad, functional validation of the discovered epigenetic changes by RNA and protein analysis.

This integrative study provides a conceptual advance in our understanding how environmental factors act on the epigenome and suggests DNA methylation as a molecular mechanism for the long-lasting consequences of smoking during pregnancy.

## 5.2 Methods and materials

### 5.2.1 Study design

For this study samples of a prospective mother-child cohort, LINA (Lifestyle and environmental factors and their Influence on Newborns Allergy risk), were used. This cohort of 629 mother–child pairs (622 mothers and 629 children; 7 twins) were recruited between May 2006 and December 2008 in Leipzig, Germany, to investigate the pre- and postnatal influences of lifestyle and environmental factors on the immune system of the newborn and the disease risk of the child later in life. Mothers suffering from immune or infectious diseases during pregnancy were excluded from the study. Blood samples were obtained from mothers at the 36th week of gestation and cord blood at delivery[224].

During pregnancy standardized questionnaires were recorded, collecting data about smoking behavior of the parents, housing conditions, mould, traffic, noise, pets, renovation activities and personal lifestyle. Annually, starting at the child's first birthday, disease outcomes of the children were assessed via questionnaire. All questionnaires were self-administered by the parents. During annual clinical visits blood samples were obtained and body weight and

length evaluated.

Participation in the study was voluntary, and informed consent was obtained from all participants. The study was approved by the Ethics Committees of the University of Leipzig (046-2006, 160-2008).

### 5.2.2 Exposure to tobacco smoke

Exposure to environmental tobacco smoke (ETS) was recorded as smoking frequency at home (‘Did you or anybody else smoke inside your dwelling during the last 12 months?’). Answering this question as ‘(almost) daily’, ‘once a week or more’ or ‘occasionally’ was defined as exposure to ETS in the subsequent analyses and ‘never’ as no exposure to ETS in the dwelling, respectively. Furthermore, the numbers of smoked cigarettes per day in the dwelling (‘How many cigarettes per day were smoked by the mother /father /anybody else in your dwelling?’) was considered.

In addition to questionnaire data, maternal urine cotinine levels were determined to assess objective smoking metabolites[225].

### 5.2.3 Anthropometric measurements

Weight and growth development were assessed by calculating the z-score of “weight for age” and “weight for length” for each individual child in the discovery and validation panel respectively. Z scores were determined based on the WHO child growth standards using the WHO Anthro software (WHO Anthro for personal computers, version 3.2.2, 2011: Software for assessing growth and development of the world's children. Geneva: WHO, 2010) (http://www.who.int/childgrowth/software/en/).

### 5.2.4 Sample selection

**Discovery panel**. For whole genome bisulfite sequencing three smoking

mothers were selected following two criteria: a measured urine cotinine levels > 100 µg/g creatinine and a positive answer regarding smoking during pregnancy. For the non-exposed group three mothers with urine cotinine levels < 1 µg/g creatinine were selected that have not smoked or were exposed to tobacco smoke.

The age of smoking mothers (mean=26.21 years, S.D. 4.73) was similar to that of non-smoking mothers (mean=29.41 years, S.D. 6.27, p=0.518 from Student's t-test). The birth weight of the children did not differ significantly between children of smoking and non-smoking mothers (3,190 g vs. 3,270 g, p=0.789 from Student's t-test).

**Validation panel.** For validation analyses 16 smoking mothers with measured cotinine levels > 100 µg/g creatinine and/or ten and more smoked cigarettes per day during pregnancy and 29 mothers with urine cotinine levels < 1 µg/g creatinine were selected that have not smoked or were exposed to tobacco smoke. The validation panel contained the discovery panel because we wanted to include also a technical validation of the sequencing data by MassARRAY. The age of smoking mothers (mean=28.36 years, S.D. 7.20) was similar to that of non-smoking mothers (mean=31.56 years, S.D. 4.41, p=0.073 from Student's t-test) in the validation panel. The birth weight of the children did not differ significantly between children of smoking and non-smoking mothers (3,262 g vs. 3,373 g, p=0.463 from Student's t-test).

### 5.2.5 Isolation of gDNA from whole blood

Maternal blood samples were collected four weeks before birth (36th gestational week) and cord blood samples at birth. Genomic DNA from whole blood samples (peripheral blood or cord blood) was isolated using the QIAmp DNA Blood Mini Kit (Qiagen, Hilden, Germany), according to manufacturer's instruction.

### 5.2.6 Illumina WGBS Library Construction and Sequencing

Illumina Libraries were prepared using the TruSeq DNA Sample Prep Kit v2-Set A (Illumina Inc., San Diego, CA, USA) according the manufacturer's instructions. Briefly, 2 µg genomic DNA in 55 µl nuclease-free water (Ambion/Life Technologies GmbH, Darmstadt, Germany) was fragmented using a Covaris S2 ultrasonicator (Covaris, Woburn, Massachusetts, USA) and the following settings: 10% duty cycle, intensity 5, 200 cycles per burst, frequency sweeping, for 6 minutes. The fragmented DNA was end-repaired, extended with an 'A' base on the 3′ end and ligated with TruSeq paired-end indexing adapters. Then, adapter-ligated fragment libraries were treated with bisulfite using the EpiTect Bisulfite Kit (Qiagen, Hilden, Germany) following the instructions in the Illumina WGBS for Methylation Analysis Guide (Part # 15021861 Rev. B). After bisulfite conversion the fragment libraries were directly amplified using KAPA HiFi Uracil+ DNA Polymerase according to the settings for TruSeq™ DNA in the technical Data Sheet (KAPA HiFi HotStart Uracil+ Ready Mix, KR0413 - version 1.12, peqlab, Erlangen, Germany). Two 50 µl PCR reactions per sample were prepared and 14 cycles of PCR performed. Amplified fragment libraries were pooled and purified with 1x Agencourt AMPure XP beads (Beckman Coulter GmbH, Krefeld, Germany). WGBS Illumina Libraries were validated using Agilent 2100 Bioanalyzer (DNA 1000 Kit, Agilent Technologies) and Qubit flourometer (Qubit dsDNA HS Assay Kit, Invitrogen/ Life Technologies GmbH, Darmstadt, Germany).

The final libraries were clustered on the cBot (Illumina Inc., San Diego, CA, USA) using TruSeq PE Cluster Kit v3 according the manufacturer`s instructions with a final concentration of either 9 pM or 10 pM (depending on the sample) spiked with 1% PhiX control v3 and an additional dedicated PhiX

control lane. Sequencing on HiSeq2000 (101 bp paired-end) was performed using standard Illumina protocols and the 200-cycles TruSeq SBS Kit v3 (Illumina Inc., San Diego, CA, USA).

### 5.2.7 Sequencing library preparation by tagmentation used for complementation of whole genome bisulfite sequencing

Tagmentation-based whole genome bisulfite sequencing of sample LMCS00_004c and LMCS00_004m using about 20 ng genomic DNA as input was done as described previously with modifications[226]. Tagmentation adapter assembly was done with oligonucleotides Tn5mC-Apt1 and Tn5mC1.1-A1block; for the oligo replacement/gap repair step, oligonucleotide Tn5mC-ReplO1 was used. The transposome was generated using the adapter and Tn5 transposase (Epicentre via Biozym, Hessisch Oldendorf, Germany). After oligo replacement/gap repair, the DNA was bisulfite treated using the EZ methylation kit (Zymo Research, Freiburg, Germany). Sequencing libraries were prepared with primers Tn5mCP1 and Tn5mCBar5 (LMCS00_004m) and Tn5mCBar6 (LMCS00_004c), respectively with 12 PCR cycles on a LightCycler 480 (Roche Applied Science, Mannheim, Germany). These two libraries were sequenced on an Illumina HiSeq 2000 in the 101 bases paired-end mode.

### 5.2.8 Sequence alignment and cytosine methylation estimation

As described in 2.2.

### 5.2.9 DMR calling, annotation and enrichment calculation

BSmooth was used to smooth bisulfite sequencing data and call candidate DMR as described previously[227,228]. Then, calculated the average

methylation level of each DMR for each sample was calculated and a p-value was assigned to each of the DMRs using Welch's t-test. Based on the p-value ($p < 0.05$) and the level of methylation change ($\Delta$ methylation > 0.1), the DMR list was further filtered and ranked for later analysis.

To assess the functional impact of each DMR, each DMR was first annotated to the closest TSS by HOMER[229]. Furthermore, enhancers were extracted from recently published data[230], TFBS, DNAse cluster and microRNA regulatory target sites were derived from UCSC genome browser, and the distance between the center of each DMR to the center of each genomic feature was calculated. The closest genomic feature was then assigned to each DMR.

For analysis of DMR enrichment in specific genomic sites, genomic features were first extracted from UCSC genome browser, recent published paper and online databases. The percentage of total genomic CpGs for each genomic feature was calculated as a background value. Thereafter, the percentage of total hyper/hypomethylated CpGs in each genomic feature was calculated based on the DMR list. The enrichment fold change was then set as the ratio between the two percentages above. In order to test the significance of the enrichment /depletion, the CpGs from all DMRs were randomly permutated in the whole genome for 10,000 times and Fisher's exact test was used to determine the significance of the difference between the observed and simulated results.

### 5.2.10 Cellular composition estimation

Promoter methylation levels from 4 lineage markers were used to assess the proportion of each cell type in each sample: CD14 for monocytes[231], CD3D and CD3G for T cells[232] and CD19 for B cells[233,234]. The rationale behind our approach was that specific promoter regions of marker genes are fully

demethylated in the respective cell lineage, whereas they are fully methylated for all other cell types.

For each marker gene, all CpGs within the promoter region (TSS upstream 2 kb and downstream 500 bp) were extracted for each sample. Then uninformative CpGs which are lowly methylated (methylation level < 0.3) in all samples were removed. For the remaining CpGs, the average methylation level (ave_meth) was calculated in each sample. The proportion of the respective cell type was then estimated as 1-ave_meth.

For granulocytes, a cell type specific methylation signature was derived based on BRD4 promoter methylation from a genome-wide methylation analysis[235,236]. The promoter region of BRD4 is unmethylated in granulocytes (granulocytic neutrophils) and methylated in B cells and hematopoietic stem/progenitor cells (HSPC)[237]. The 7 CpG sites that are unmethylated in granulocytes and methylated in the other hematopoietic lineages were derived according to the methylation signature by Houseman and colleagues. For each of the 7 CpG sites, the methylation level around it was carefully inspected in whole-genome bisulfite sequencing data from different blood cell types.

### 5.2.11 RepliSeq based replication timing analysis

The number of DMRs binned into 1 Mb windows was correlated with genome-wide replication timing data[238]. The Repli-Seq data used in this thesis is a wavelet-smoothed, weighted average signal where high (and low) values indicate early (and late) replication during S-phase. RepliSeq replication timing data was downloaded from http://genome.ucsc.edu/ENCODE for ten different cell lines: Gm06990, Gm12801, Gm12812, Gm12813, Gm12878, HepG2, HUVEC, K562, MCF7, NHEK. We used the mean value of genomic regions that maintain similar replication timing between these different cell types, determined by low standard deviation per window.

### 5.2.12 Pathway enrichment analysis

Enrichment of KEGG pathways was determined for DMRs in either mothers or children. Only DMRs, which were significantly different from the non-smoking control group ($p < 0.05$) with a difference in the methylation level higher than 10%, were considered for analysis. The latest update (2013/01/31) of the WEB-based Gene SeT AnaLysis Toolkit (WebGestalt)[239] was used to calculate statistically significant enriched pathways. Calculation of enrichment was based on a hypergeometric test followed by a Benjamini & Hochberg multiple test adjustment. A minimum of 3 genes per pathway was required to be considered for enrichment. Enrichment was considered significant at an adjusted p-value < 0.05.

### 5.2.13 MassARRAY methylation analysis

Quantitative DNA methylation analysis of candidate DMRs was performed using Sequenom's MassARRAY platform. Briefly, genomic DNA from whole blood samples was chemically modified with sodium bisulfite using the EZ methylation kit (Zymo Research, Freiburg, Germany) according to the manufacturer's instructions. PCR primers were designed with an additional T7 promoter tag for in vivo transcription for each reverse primer, as well as a 10-mer tag on the forward primer. Bisulfite treated DNA was PCR amplified using HotStarTaq DNA Polymerase (Qiagen, Hilden, Germany) with the following cycling program: 95°C for 15 min, followed by 45 cycles of 94°C for 30 sec, 72°C for 1 min and a final elongation step at 72°C for 5 min on a LightCycler 480 (Roche Applied Science, Mannheim, Germany). The PCR product was in vitro transcribed and cleaved by RNase A using the EpiTyper T Complete Reagent Set (Sequenom, Hamburg, Germany) and subjected to MALDI-TOF mass spectrometry analysis to determine

methylation patterns as previously described[240]. DNA methylation standards (0%, 20%, 40%, 60%, 80%, and 100% methylated genomic DNA) were used to control for potential PCR bias. Note that targeted methylation analysis based on Infinium HumanMethylation450 BeadChip would not be applicable here, since the latter would cover only less than a third of the DMRs tested here.

### 5.2.14 RNA Extraction, cDNA Synthesis, and qPCR

Total RNA was prepared from fresh blood by using peqGold RNA Pure (peqlab, Erlangen, Germany), according to manufacturer's instruction. The cDNA synthesis was carried out with 5 µg of RNA by using ImProm-IITM Reverse Transcription System (Promega, Mannheim, Germany).

Gene expression was measured using the 96.96 Dynamic Array Integrated fluidic circuits (IFCs) (Fluidigm, San Francisco, CA, USA). Intron-spanning primers were designed and UPL probes selected by the Universal Probe Library Assay Design Center (http://qpcr.probefinder.com/organism.jsp). A preamplification reaction was performed by pooling all primers (final concentration, 50 nM), 5 µl of cDNA and 2x PreAmp Master Mix (Applied Biosystems/Life Technologies GmbH, Darmstadt, Germany). The cycling program consisted of 95°C for 10 min, followed by 14 cycles of 95°C for 15 sec and 60°C for 4 min on a LightCycler 480 (Roche Applied Science, Mannheim, Germany). The qPCRs of 1:5 diluted with TE buffer preamplified templates were performed following manufacture's instruction for UPL (Roche Applied Science, Mannheim, Germany) assays. Briefly, for each individual assay, a 10X Assay Mix that contained 2 µM of each forward and reverse primer, 1 µM UPL probe and 0.025% Tween-20 was prepared, and 5 µl of the mix was loaded into the assay inlets of the array. Into the sample inlets, 5 µl of the following solution was dispensed: 2.5 µl of

PreAmp sample in 1.1X of FastStart Universal Probe Master Mix (Roche Applied Science, Mannheim, Germany). The cycling program consisted of 2 min at 50°C, 10 min at 95°C, followed by 35 cycles of 95°C for 15 sec, 70°C for 5 sec, and 1 min at 60°C. All reactions were performed in triplicates.

Gene expression values were determined by using the 2-ΔΔCT method[241] with GAPD, GUSB, PGK1 and PPIA as reference genes and normalized to the lowest measured value.

### 5.2.15 Cytokine measurement

Heparinized blood samples from mother-child pairs were obtained by venipuncture and processed within six hours for further analysis. After incubating for 4 h at 37°C, samples were diluted with RPMI-1640 medium without supplements in a ratio of 1:1 and centrifuged. Cell-free supernatants were collected and stored at -80°C until subsequent analysis. Concentrations of IL-6, MCP-1 (CCL2), and TNF-α in the supernatants of whole blood samples were detected by flow cytometry using the BD CBA Human Soluble Flex Set system (BD Bioscience, Heidelberg, Germany) according to the manufacturer's instructions and as described previously[242].

In brief, cytokine specific antibody coated beads were incubated for 1 h with 25 µl of blood samples or standard solution. Thereafter, samples were incubated with the corresponding PE labeled detection antibodies for 2 h. After one washing step samples were measured by flow cytometry. Analysis of data and quantification of cytokines was performed using the FCAP ArrayTM software (Becton Dickinson, Heidelberg, Germany) on the basis of corresponding standard curves. Finally, plasma dilution factor was accounted.

## 5.3 Results

Whole genome bisulfite sequencing (WGBS) of whole blood were performed on samples from twelve individuals (maternal blood at 36th week of gestation and cord blood from their corresponding newborns) in the discovery panel at an average coverage of 38x (range: 27-50x) and thus generated 26.3 billion non-duplicate, 101 bp reads (**Table 6**). In order to rule out the blood cell-type composition induced methylation change in the tobacco smoke exposed vs. non-exposed individuals, the promoter methylation level from five lineage markers were used to assess the proportion of each cell type in each sample. It showed that the variation in response to tobacco smoke exposure is below 6% and 9% for all cell types in mothers and children (**Table 7** and **Table 8**). To exclude DMRs that are solely caused by differences in cellular blood composition between the exposed and non-exposed samples, a threshold of 10% was used for DMR calling.

Based on this DMR filter, 1981 and 1720 significant ($p < 0.05$, $\Delta$ methylation > 0.1) DMRs were identified in mothers and children, respectively (**Figure 32** and **Figure 33**). Interestingly, the ratio of hypo- vs. hypermethylated DMRs differed significantly between mothers and children (Fisher's exact test, $p < 2e\text{-}16$; **Figure 33**). While the mothers showed a dominant hypomethylation profile, children revealed hyper- and hypomethylated DMRs with a twofold higher rate of hypermethylated DMRs as their mothers. The patterns of genes associated with at least one DMR differed as well between mothers and children (Fisher's exact test, $p < 2e\text{-}16$) supporting the hypothesis that environmental modulation of the epigenome is distinct between adult and fetus. The density pattern of DMRs is highly significantly correlated with replication timing at genome-scale (**Figure 34** and **Figure 35**).

| Sample id | Gender[a] | Smoking status[b] | % of genomic CpGs covered by 1x (10x) | Average coverage |
|---|---|---|---|---|
| **children** | | | | |
| LMCS00_001c | M | 0 | 98% (96%) | 45x |
| LMCS00_002c | F | 0 | 98% (95%) | 38x |
| LMCS00_003c | F | 0 | 98% (95%) | 44x |
| LMCS00_004c | F | 1 | 98% (94%) | 27x |
| LMCS00_005c | M | 1 | 98% (92%) | 31x |
| LMCS00_006c | M | 1 | 98% (91%) | 29x |
| **mothers** | | | | |
| LMCS00_001m | F | 0 | 98% (94%) | 38x |
| LMCS00_002m | F | 0 | 98% (97%) | 50x |
| LMCS00_003m | F | 0 | 98% (96%) | 48x |
| LMCS00_004m | F | 1 | 98% (96%) | 32x |
| LMCS00_005m | F | 1 | 98% (95%) | 36x |
| LMCS00_006m | F | 1 | 98% (95%) | 36x |

a F: female; M: male

b 0: smoking or from smoking mother; 1: non-smoking or from non-smoking mother

**Table 6 Sequencing overage of study cohort.**

| Cell type | marker | Smoking group[a] | Non-smoking group[a] | p-value[b] |
|---|---|---|---|---|
| **Mothers** | | | | |
| Granulocytes | BRD4 | 0.853 (0.021) | 0.793 (0.006) | 0.10 |
| Monocytes | CD14 | 0.200 (0.026) | 0.183 (0.02) | 0.82 |
| T lymphocytes | CD3D (CD3G) | 0.180 (0.017) | 0.219 (0.025) | 0.15 |
| B lymphocytes | CD19 | 0.16 (0.01) | 0.14 (0.01) | 0.12 |
| **Children** | | | | |
| Granulocytes | BRD4 | 0.560 (0.159) | 0.633 (0.023) | 0.70 |
| Monocytes | CD14 | 0.237 (0.015) | 0.230 (0.017) | 0.64 |
| T lymphocytes | CD3D (CD3G) | 0.34 (0.087) | 0.258 (0.037) | 0.27 |
| B lymphocytes | CD19 | 0.17 (0.0053) | 0.137 (0.006) | 0.66 |

a. mean (standard deviation) methylation level in the promoter region

b. p-value was calculated by the Mann-Whitney U-test

**Table 7 Cell type distribution estimated by methylation signature ( WGBS ).**

| Cell type | marker | Smoking group[a] | Non-smoking group[a] | p-value[b] |
|---|---|---|---|---|
| **Mothers** | | | | |
| Granulocytes | BRD4 | 0.805 (0.047) | 0.791 (0.050) | 0.34 |
| Monocytes | CD14 | 0.089 (0.043) | 0.086 (0.058) | 0.64 |
| T lymphocytes | CD3D (CD3G) | 0.154 (0.097) | 0.260 (0.161) | 0.11 |
| B lymphocytes | CD19 | 0.112 (0.020) | 0.135 (0.047) | 0.96 |
| **Children** | | | | |
| Granulocytes | BRD4 | 0.631 (0.057) | 0.610 (0.048) | 0.04 |
| Monocytes | CD14 | 0.090 (0.047) | 0.099 (0.057) | 0.72 |
| T lymphocytes | CD3D (CD3G) | 0.28 (0.121) | 0.343 (0.140) | 0.32 |
| B lymphocytes | CD19 | 0.152 (0.0042) | 0.148 (0.050) | 0.66 |

a. mean (standard deviation) methylation level in the promoter region

b. p-value was calculated by the Mann-Whitney U-test

**Table 8 Cell type distribution estimated by methylation signature ( MassARRAY ).**

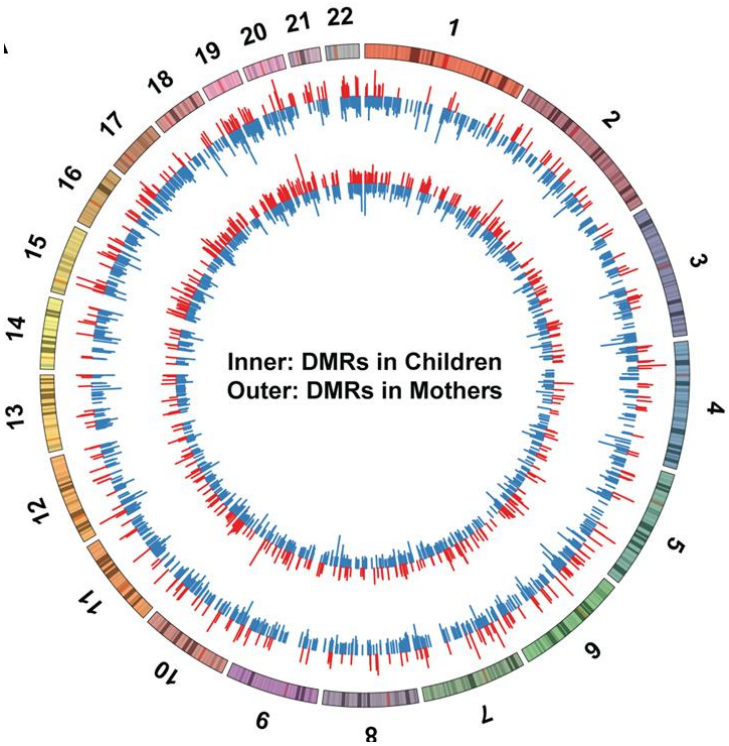



**Figure 32 Circular representation of DNA methylation levels for mothers and children.**

The height of each bar indicates the methylation change between the smoking and non-smoking group (red: hypermethylation, blue: hypomethylation).

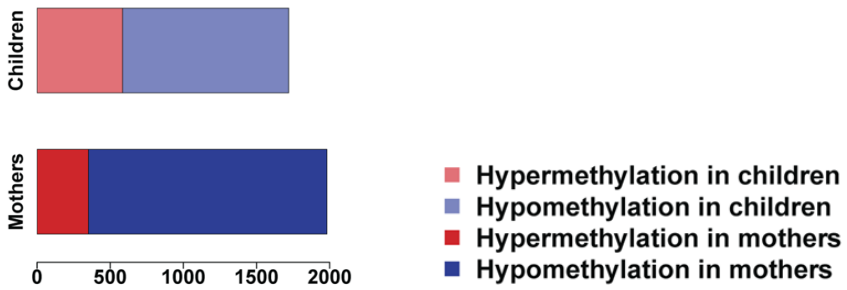



**Figure 33 Distribution of hyper/hypo methylation in children and mothers.** Bar plots represent the number of hypo- vs. hypermethylated DMRs separately for children and mothers ( ratio of these DMRs is significantly different between mothers and children, $p < 2e\text{-}16$ ).

Among different genomic features, DMRs are enriched in gene regulatory regions such as promoters, enhancers and transcription factor binding sites (**Figure 36**). Out of the total of 124 DMRs that are shared between mothers and their children (e.g. *MAPK9*, **Figure 38**), none of them are found in imprinted regions. However, there is a highly significant enrichment of DMRs in imprinted genes in children (**Figure 37**), suggesting that in the embryonic period environmental factors preferentially influence imprinted genes and that the observed epigenetic modifications in the newborn child result from de novo effects in the fetal genome rather than from a transmission from mother to child.

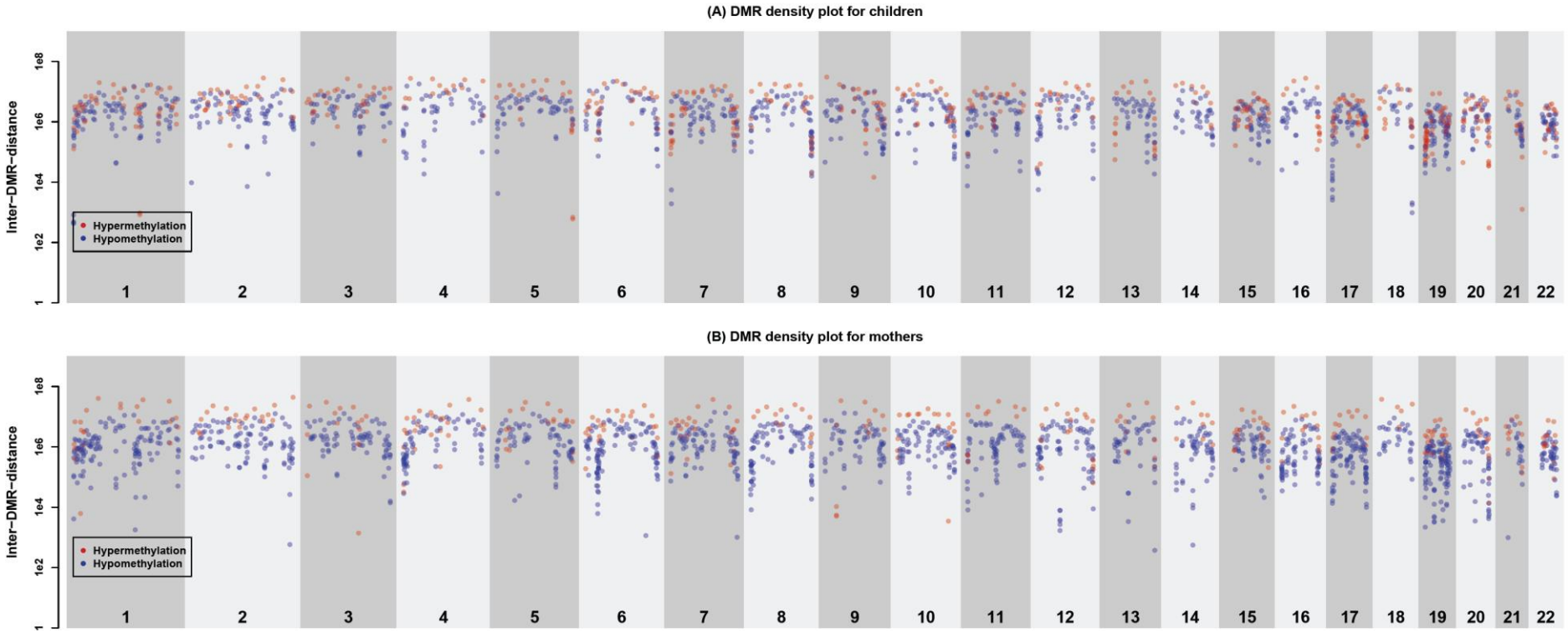



**Figure 34 Rainfall plots representing the genome-wide distribution of DMR densities in children and mothers.** Each red dot symbolizes a hypermethylated, each blue dot a hypomethylated DMR. Color shading is only used for visualization purposes. Y-axis indicates the inter-DMR distance.

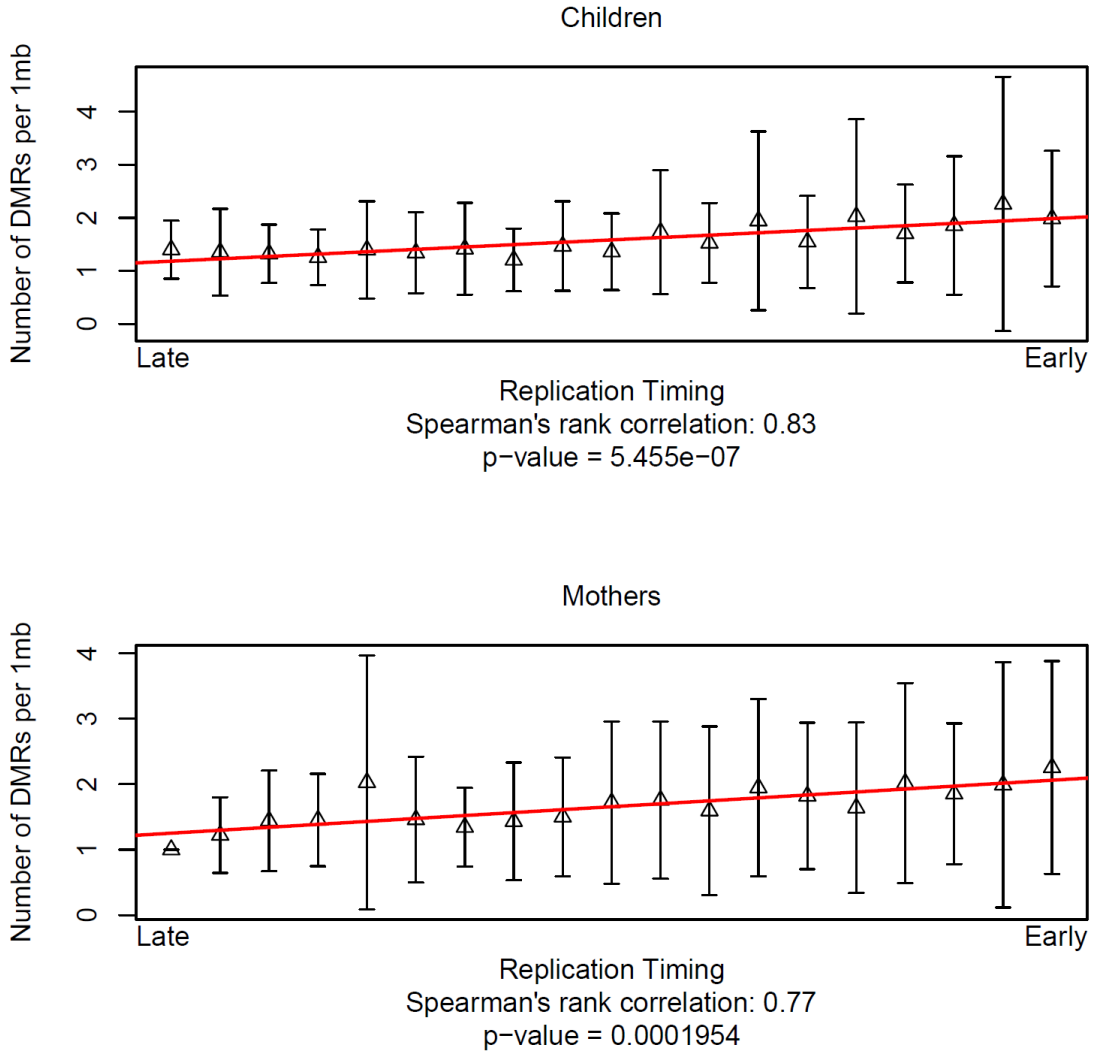



**Figure 35 Correlation between DMR density and replication timing.** Scatter plot correlating the number of DMRs (y axis) binned into 1Mb windows with genome-wide replication timing data (Repli-Seq) sorted from from “Late” to “Early” replication timing (x axis). Clearly, DMR density is inversely correlated with replication timing (Spearman’s rank correlation 0.83 (p=5.455e-7) and 0.77 (p=1.954e-4) for children and mothers, respectively). Upper (lower) panels represent genome-wide DMR distribution characteristics in children (mothers).

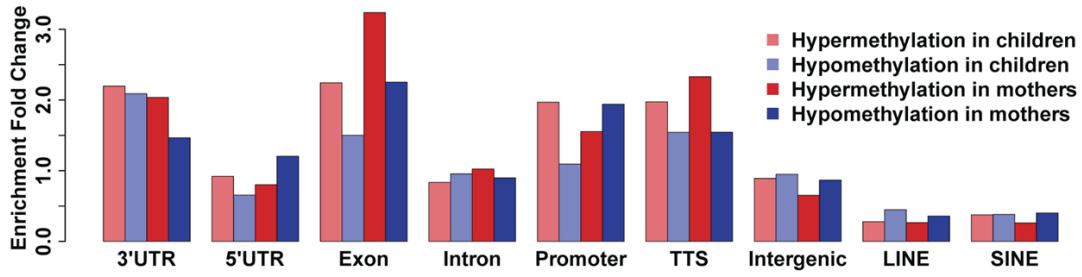



**Figure 36 Enrichment of DMRs in general genomic features for children and mothers.**

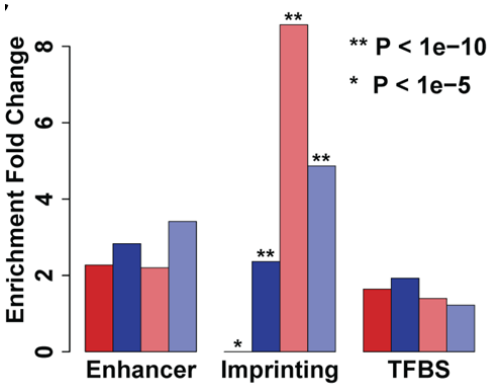
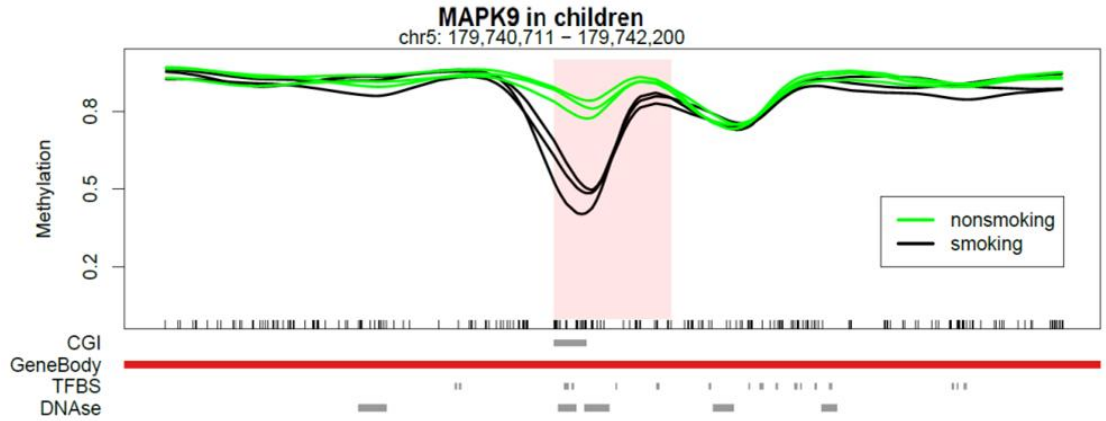



**Figure 37 Enrichment of DMRs in regulatory regions for children and mothers.**



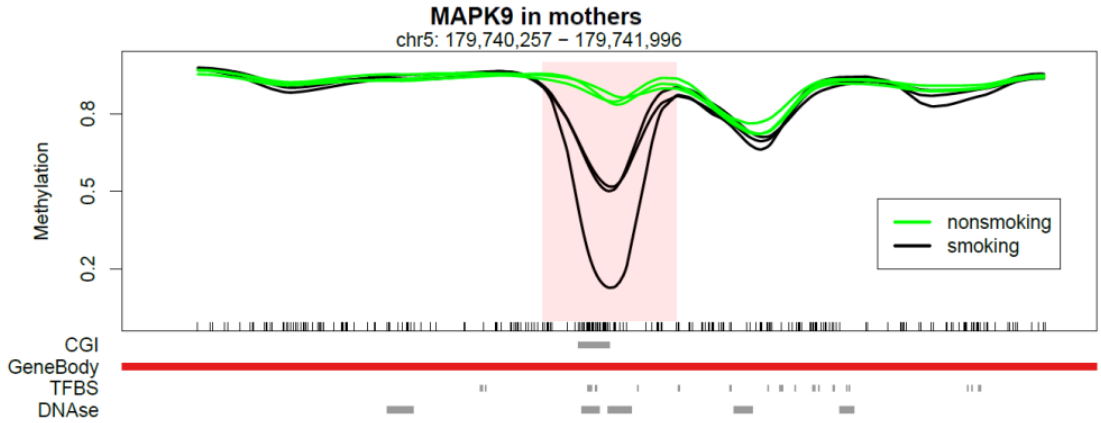



**Figure 38 DMR profiles for *MAPK9* in children and mothers.** For *MAPK9* methylation profiles are shown for the differentially methylated region (red shadowed area) discovered by DMR calling (smokers: black, non-smokers: green).

A set of 52 DMRs, linked to pathway deregulation or epigenetic reprogramming, were then validated in a total of 505 CpG sites over 90 samples (validation panel, **Table 9** and **Table 10**) by targeted mass spectrometry-based methylation analysis (MassARRAY). The correlation between MassARRAY and WGBS based methylation analysis was remarkably high across all DMRs (**Figure 39**; Pearson correlation 0.90; $p < 2e\text{-}16$). Out of the 52 DMRs the methylation difference estimated from the discovery panel could be confirmed for 30 DMRs (58%) in the validation panel. Transcriptional expression of genes related to DMRs and DNA methylation was generally weakly correlated (**Figure 40**) as reported earlier in cancer[243]. Still, 10/50 genes (20%) related to DMRs showed a significantly differential expression (**Table 9** and **Table 10**) across the smoking/non-smoking samples in the validation panel. Thereby, the correlation between DNA methylation and transcriptional response was much higher in mothers compared to their children suggesting that environmental factors acting throughout the

developmental period may induce epigenetic marks that potentially impact disease only later in life[244].

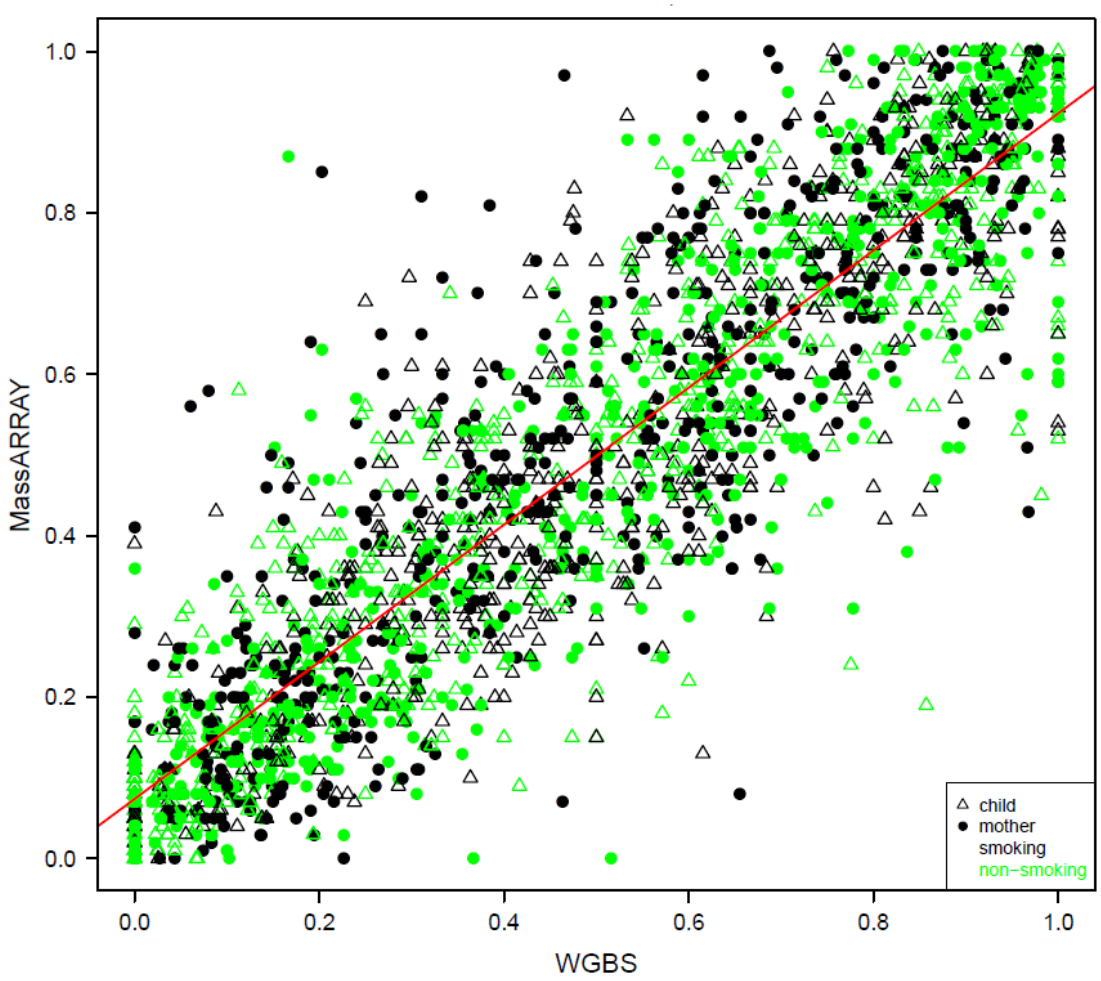



**Figure 39 Correlation of methylation changes determined by WGBS and MassARRAY for representative examples.** Overall, methylation levels of all CpGs observed by MassARRAY in the discovery panel were highly correlated with those observed by WGBS in the validation panel (Pearson coefficient R=0.9, p < 10e-16). Green: non-smokers, black: smokers.

| | | | | | | Sequencing | | | | MassARRAY | | | | | Transcription | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Gene name** | **DMR location** | | | **proximal regulatory feature/ location** | **Mother Child (M/C)** | **hypo/ hyper** | **mean meth. diff.** | **number of CpGs** | **p-value** | **mean meth. diff.** | **p-value** | **total number of CpGs in amplicon** | **CpGs used for statistics** | **CpGs with reliable Mass-ARRAY results** | **fold-change** | **p-value** |
| **PDK1** | chr2 | 173379024 | 173379341 | enhancer | C | hypo | -0.19 | 10 | 0.05 | -0.055 | 0.013 | 12 | 8 | 9 | 1.0 | $0.848_a$ |
| **MAPK9** | chr5 | 179740711 | 179742200 | intron | C | hypo | -0.25 | 44 | 0.00 | -0.073 | 0.01 | 16 | 4 | 12 | 1.6 | $0.505_b$ |
| **SLC2A1** | chr1 | 43472535 | 43473075 | enhancer | C | hypo | -0.11 | 36 | 0.01 | -0.118 | 0.002 | 30 | 3 | 14 | 1.0 | $0.909_b$ |
| **CYP2E1** | chr10 | 135343009 | 135344280 | intron | C | hypo | -0.13 | 45 | 0.01 | -0.070 | 0.039 | 8 | 6 | 6 | 2.0 | $0.306_b$ |
| **FZD10** | chr12 | 130704847 | 130705600 | TFBS | C | hypo | -0.11 | 19 | 0.01 | -0.028 | 0.078 | 10 | 3 | 6 | -1.3 | $0.702_b$ |
| **FCGR2A** | chr1 | 161423750 | 161423827 | DNAse Cluster | C | hyper | 0.19 | 3 | 0.00 | 0.039 | 0.02 | 11 | 4 | 6 | 1.5 | $0.427_b$ |
| **RUFY1** | chr5 | 178985402 | 178987269 | promoter-TSS | C | hyper | 0.16 | 77 | 0.02 | 0.035 | 0.038 | 32 | 5 | 12 | n.d. | n.d. |
| **KDM5B** | chr1 | 202777861 | 202779663 | TFBS | C | hyper | 0.12 | 38 | 0.01 | 0.04 | 0.086 | 8 | 4 | 6 | **-1.2** | **$0.056_a$** |
| **METTL24** | chr6 | 110617827 | 110618174 | intron | C | hyper | 0.18 | 11 | 0.00 | 0.080 | 0.039 | 9 | 3 | 5 | 1.6 | $0.749_b$ |
| **PPP3CA** | chr4 | 101987882 | 101988016 | intron | C | hyper | 0.16 | 6 | 0.01 | 0.069 | 0.045 | 6 | 3 | 4 | 1.4 | $0.293_a$ |
| **LMNA** | chr1 | 156093333 | 156095351 | intron | M | hypo | -0.11 | 36 | 0.02 | -0.051 | 0.053 | 10 | 1 | 8 | n.d. | n.d. |
| **MAPK9** | chr5 | 179740257 | 179742208 | intron | M | hypo | -0.31 | 44 | 0.00 | -0.116 | <0.001 | 16 | 4 | 12 | -1.1 | $0.232_b$ |
| **PLD1** | chr3 | 171495495 | 171495665 | intron | M | hypo | -0.13 | 4 | 0.01 | -0.047 | 0,018 | 4 | 3 | 4 | -1.3 | $0.753_b$ |
| **CACNA2D1** | chr7 | 82073572 | 82074006 | promoter-TSS | M | hypo | -0.07 | 16 | 0.04 | -0.032 | 0.026 | 16 | 2 | 13 | n.d. | n.d. |
| **PIK3R5** | chr17 | 8869555 | 8870436 | promoter-TSS | M | hypo | -0.09 | 36 | 0.03 | -0.050 | 0.026 | 13 | 2 | 10 | 1.5 | $0.119_a$ |
| **MAPK7** | chr17 | 19282195 | 19282738 | promoter-TSS | M | hypo | -0.14 | 30 | 0.01 | -0.168 | <0.001 | 36 | 6 | 19 | -1.0 | $0.865_b$ |
| **CACNG4** | chr17 | 64959742 | 64960474 | promoter-TSS | M | hypo | -0.09 | 22 | 0.03 | -0.044 | 0.043 | 12 | 1 | 5 | n.d. | n.d. |
| **F2RL3** | chr19 | 16998279 | 16998796 | DNAse Cluster | M | hypo | -0.10 | 8 | 0.02 | -0.095 | 0.046 | 11 | 1 | 5 | -1.6 | $0.164_a$ |
| **GABRB3** | chr15 | 26873979 | 26874471 | promoter-TSS | M | hyper | 0.13 | 46 | 0.02 | 0.022 | 0.051 | 34 | 16 | 30 | **-3.0** | **$0.067_a$** |
| **METTL24** | chr6 | 110617826 | 110618173 | intron | M | hyper | 0.15 | 11 | 0.03 | 0.152 | 0.016 | 9 | 5 | 5 | -1.1 | $0.869_a$ |

| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **FZD7** | chr2 | 202904079 | 202904886 | DNAse Cluster | C | hypo | -0.13 | 14 | 0.01 | -0.060 | 0.139 | 8 | 4 | 6 | -1.3 | $0.734_b$ |
| **PRKCB** | chr16 | 23864800 | 23864984 | intron | C | hypo | -0.24 | 3 | 0.00 | -0.051 | 0.117 | 4 | 2 | 2 | 1.6 | $0.138_a$ |
| **C18orf54** | chr18 | 51915207 | 51916404 | TFBS | C | hyper | 0.16 | 9 | 0.01 | 0.087 | 0.204 | 6 | 2 | 2 | 1.1 | $0.944_b$ |
| **CACNG4** | chr17 | 64972219 | 64972396 | intron | M | hypo | -0.10 | 6 | 0.01 | -0.029 | 0.144 | 11 | 3 | 9 | n.d. | n.d. |
| **LINC00032** | chr9 | 27205855 | 27206391 | intron | M | hyper | 0.17 | 10 | 0.00 | 0.033 | 0.133 | 9 | 5 | 5 | n.d. | n.d. |

[a] Student's t-test

[b] Mann-Whitney U test

[n.d.] not detectable

**Table 9 Correlation of DMRs identified by WGBS with differential methylation determined by MassARRAY and detected transcriptional changes.**

| | | | | | | Sequencing | | | | Transcription | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Gene name** | **DMR location** | | | **proximal regulatory feature/ location** | **Mother Child (M/C)** | **hypo/ hyper** | **mean methylation difference** | **number of CpGs** | **p-value** | **fold-change** | **p-value** |
| **KITLG** | chr12 | 88974836 | 88975118 | promoter-TS | C | hypo | -0.12 | 6 | 0.004 | 1.6 | 0.3461 [a] |
| **FHIT** | chr3 | 60942995 | 60943388 | intron | C | hypo | -0.13 | 11 | 0.005 | 1.4 | 0.5900 [b] |
| **FGF5** | chr4 | 81190228 | 81190754 | intron | C | hypo | -0.10 | 9 | 0.015 | -1.1 | 0.9415 [a] |
| **Wnt4** | chr1 | 22586826 | 22587094 | DNAse | C | hypo | -0.14 | 3 | 0.005 | n.d. | n.d. |
| **FOXA2** | chr20 | 22753930 | 22754460 | DNAse | C | hypo | -0.11 | 6 | 0.014 | n.d. | n.d. |
| **FOXO3** | chr6 | 108883163 | 108883552 | intron | C | hyper | 0.12 | 50 | 0.017 | 1.1 | 0.9209 [b] |
| **RUNX1** | chr21 | 36253761 | 36253874 | intron | C | hyper | 0.16 | 3 | 0.008 | 1.2 | 0.5048 [a] |
| **MIR657** | chr17 | 79092655 | 79092814 | intron | C | hyper | 0.20 | 7 | 0.006 | n.d. | n.d. |
| **PLB1** | chr2 | 28825272 | 28825928 | intron | C | hyper | 0.14 | 7 | 0.025 | 1.2 | 0.6480 [a] |
| **SETD9** | chr5 | 56203809 | 56204994 | promoter-TS | C | hyper | 0.12 | 31 | 0.023 | 1.1 | 0.7536 [a] |
| **CTBP2** | chr10 | 126850806 | 126851292 | DNAse | C | hyper | 0.20 | 35 | 0.003 | 1.7 | 0.1364 [a] |
| **DLG1** | chr3 | 196876335 | 196876593 | intron | C | hyper | 0.15 | 5 | 0.027 | 1.1 | 0.3421 [b] |
| **PIK3CB** | chr3 | 138565356 | 138565607 | TFBS | C | hyper | 0.12 | 7 | 0.025 | 1.0 | 0.8310 [a] |
| **PLD1** | chr3 | 171495495 | 171495665 | intron | C | hyper | 0.10 | 4 | 0.4 | 1.3 | 0.4535 [a] |
| **IL1A** | chr2 | 113541539 | 113542345 | intron | M | hypo | -0.14 | 12 | 0.009 | 1.3 | 0.7055 [a] |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **HLA-E** | chr6 | 30465492 | 30465759 | DNAse | M | hypo | -0.15 | 3 | 0.003 | n.d. | n.d. |
| **SP6** | chr17 | 45929846 | 45930672 | intron | M | hypo | -0.11 | 24 | 0.015 | n.d. | n.d. |
| **SMAD3** | chr15 | 67355626 | 67357164 | TFBS | M | hypo | -0.17 | 44 | 0.400 | 1.1 | 0.6746 [a] |
| **Wnt6** | chr2 | 219726213 | 219726546 | intron | M | hypo | -0.14 | 5 | 0.004 | n.d. | n.d. |
| **GDF7** | chr2 | 20868948 | 20872088 | exon | M | hyper | 0.22 | 177 | 0.003 | **-3.0** | **0.0536 [a]** |
| **SETD9** | chr5 | 56203809 | 56204891 | promoter-TS | M | hyper | 0.09 | 31 | 0.023 | 1.0 | 0.8142 [a] |
| **CABIN1** | chr22 | 24424783 | 24425132 | intron | M | hyper | 0.24 | 5 | 0.001 | **-1.3** | **0.0623 [b]** |
| **ZMAT3** | chr3 | 178750749 | 178751252 | intron | C | hypo | -0.28 | 10 | 0.003 | **-1.2** | **0.0762 [b]** |
| **C18orf54** | chr18 | 51915207 | 51916404 | TFBS | M | hypo | 0.17 | 9 | 0.002 | -1.5 | 0.0953 [a] |
| **CYP2E1** | chr10 | 135316185 | 135316476 | DNAse | M | hypo | -0.10 | 4 | 0.021 | **-3.4** | **0.0003 [a]** |
| **FASLG** | chr1 | 172770749 | 172770896 | DNAse | M | hypo | -0.14 | | 0.019 | **-1.9** | **0.0128 [b]** |
| **FGFR2** | chr10 | 123443190 | 123444420 | TFBS | M | hypo | -0.16 | 21 | 0.003 | **-2.5** | **0.0101 [b]** |
| **NFAT1C** | chr18 | 77292133 | 77292782 | TFBS | M | hypo | -0.2565 | | 0.0089 | **-1.5** | **0.0065 [b]** |
| **Nostrin** | chr2 | 169690829 | 169691015 | intron | M | hyper | 0.17 | 12 | 0.006 | 1.8 | 0.1438 [a] |
| **THBS1** | chr15 | 39873767 | 39874199 | intron | M | hypo | -0.12 | 12 | 0.005 | **2.3** | **0.0133 [a]** |

[a] Student's t-test

[b] Mann-Whitney U test

[n.d.] not detectable

**Table 10 Correlation of significant DMRs with differential transcription.**

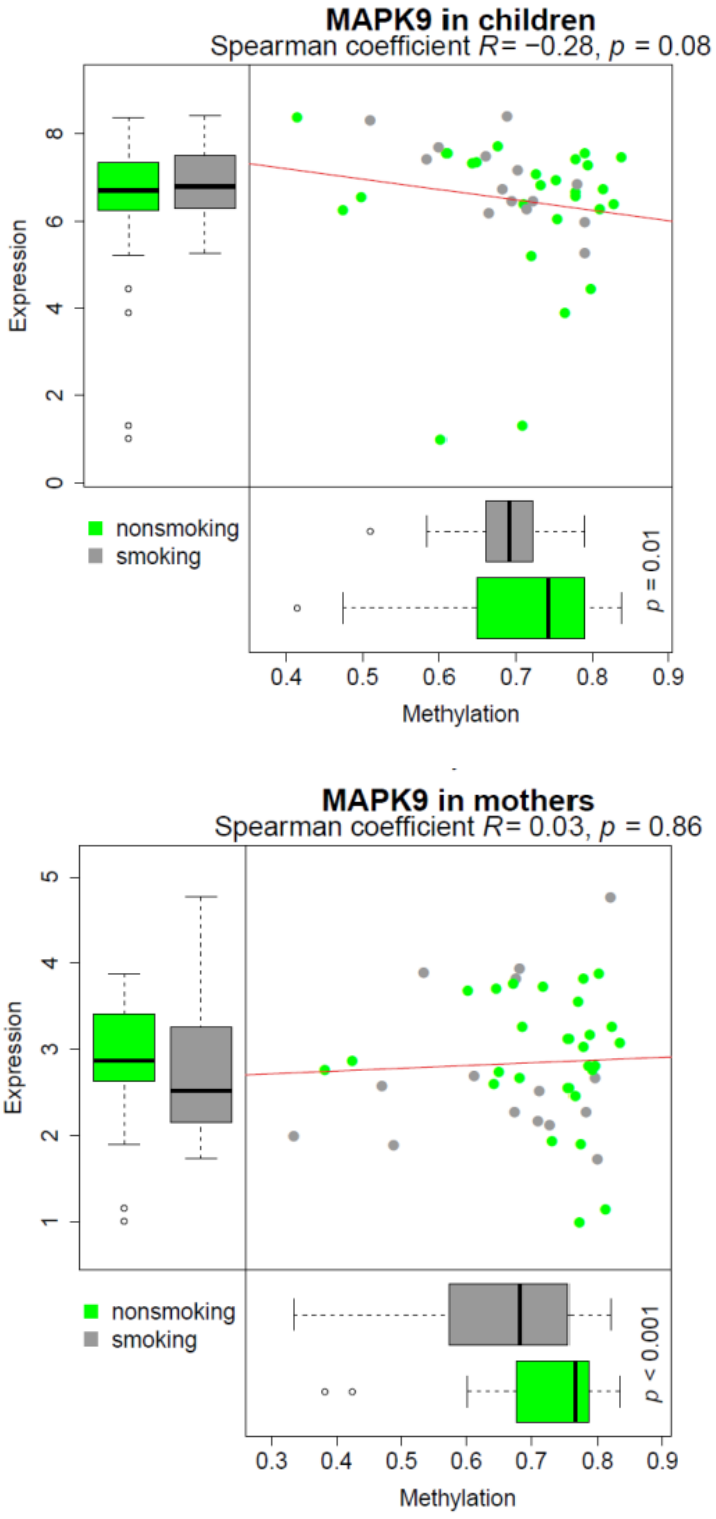



**Figure 40 correlation between methylation changes and transcription for *MAPK9*.** *MAPK9* hypomethylation is associated with a slight decrease in transcription in mothers, but

not in children. A weak correlation between methylation change and transcriptional expression is observed for *MAPK9* in children, but not in their mothers.

Pathway analysis was then performed to categorize the potential function of DMRs. Since the relevance of methylation changes in different genomic regions such as promoters, gene bodies, and enhancers, is not generally established, we considered all methylation changes attributed to a certain gene independent of its genomic location. Interestingly, only a small number of pathways, including the WNT signaling pathway, are jointly enriched in mothers and children (**Figure 41**). Aberrant WNT signaling is involved in the airway inflammatory response in healthy smokers and smokers with chronic obstructive pulmonary disease (COPD)[245] and was also described as a hallmark of many tumors, including lung cancer[246]. Thirteen differentially methylated genes identified in smoking mothers belonged to the WNT signaling pathway and 16 genes aberrantly methylated were identified in their newborn children. The striking overlap between epigenetic perturbations in this pathway in mothers and children indicates prenatal programming of impaired lung function.

The majority of affected pathways differ widely between children and mothers supporting our view that the environmental modulation of the epigenome is distinct in mothers and children (**Figure 41**). Despite the multitude of pathways enriched for DMRs in children, three functionally related groups emerged. First, signaling pathways involved in immune regulation and inflammation, among them *MAPK*, chemokine, and T cell receptor signaling. Perturbation of these pathways potentially results in an altered NFκB activation, which could be confirmed by measurements of NFκB subunit RNA expression (**Figure 42**) and blood concentrations of the NFκB target proteins IL-6, TNF-α,

and MCP-1(**Figure 43**). Those target proteins are up-regulated in children, but not in their mothers, supporting our conclusion of an increased tobacco smoke-induced inflammatory response in children compared to their mothers. Remarkably, this inflammatory phenotype is sustained until the age of one (**Figure 44**).

Second, pathways involved in metabolic dysfunction, including insulin signaling, adipocytokine and Type II diabetes mellitus pathways, were frequently encountered. These pathways include central functions of the metabolism regulating glucose homeostasis and fatty acid oxidation. PRKCB which is hypomethylated and transcriptionally upregulated in children from smoking mother, has been described to contribute to impaired insulin-signaling[247] and to participate in the regulation of glucose transport in adipocytes[248]. Smoking during pregnancy is increasingly accepted as risk factor for childhood overweight and obesity[249,250] although the underlying mechanisms remain unknown. Pathway analyses in this study support the idea that epigenetic mechanisms are involved in metabolic programming by prenatal tobacco smoke exposure. Interestingly, a continuously increasing body weight (Z-score, **Figure 45**) was observed in tobacco smoke exposed children compared to children from non-smoking mothers, suggesting a link between the epigenetic dysregulation of metabolism and an overweight phenotype.

Overlapping Pathways
Cancer
Inflammation/
Immune
Regulation
Metabolic
Dysfunction
Mothers
Children
Significance [-log10(p)]
Insulin signaling pathway
Type II diabetes mellitus
Maturity onset diabetes of the young
Adipocytokine signaling pathway
Carbohydrate digestion and absorption
VEGF signaling pathway
Natural killer cell mediated cytotoxicity
T cell receptor signaling pathway
MAPK signaling pathway
Fc gamma R-mediated phagocytosis
Cell adhesion molecules (CAMs)
Chemokine signaling pathway
mTOR signaling pathway
Cytokine-cytokine receptor interaction
B cell receptor signaling pathway
Non-small cell lung cancer
Basal cell carcinoma
Renal cell carcinoma
Hedgehog signaling pathway
Small cell lung cancer
Glioma
p53 signaling pathway
Melanoma
ErbB signaling pathway
Chronic myeloid leukemia
Glycosaminoglycan biosynthesis - keratan sulfate
Mucin type O-Glycan biosynthesis
Regulation of actin cytoskeleton
Neurotrophin signaling pathway
Melanogenesis
Hemotopoietic cell lineage
Protein processing in endoplasmatic reticulum
Osteoclast differentiation
Toxoplasmosis
Dorso-ventral axis formation
Drug metabolism -other enzymes
Glycoshingolipid biosynthesis -ganglio series
Pathways in cancer
Axon guidance
Wnt signaling pathway
Metabolic pathways
Endocytosis
Focal adhesion
N-Glycan biosynthesis
Amoebiasis
Histidine metabolism
Vitamine digestion and absorption
Neuroactive ligand-receptor interaction
Graft-versus-host-disease
Allograft rejection
Leukocyte transendothelial migration
Type I diabetes mellitus
Chagas disease (American trypanosomiasis)
African trypanosomiasis
Pancreatic secretion
Salivary secretion
ECM-receptor interaction
Other glycan degradation
Ubiquitine mediated proteolysis
Arrythmonic right ventricular cardiomyopathy (ARVC)
color key [-log10(p)]
0 2 4 6 8 10

**Figure 41 Pathway enrichment analysis.** Depicted are the results of the pathway enrichment analysis using all DMRs with a differential methylation > 10% and a significance level < 0.05. Enrichment was determined for mothers and children separately with 1671 and 1496 genes related to DMRs identified used for analysis, respectively. Three particularly interesting groups of pathways emerge in children: pathways related to metabolic dysfunction (green), inflammation/immune regulation (orange) and cancer (blue). Pathways enriched for DMRs in both mothers and children are highlighted as “overlapping pathways” (grey).

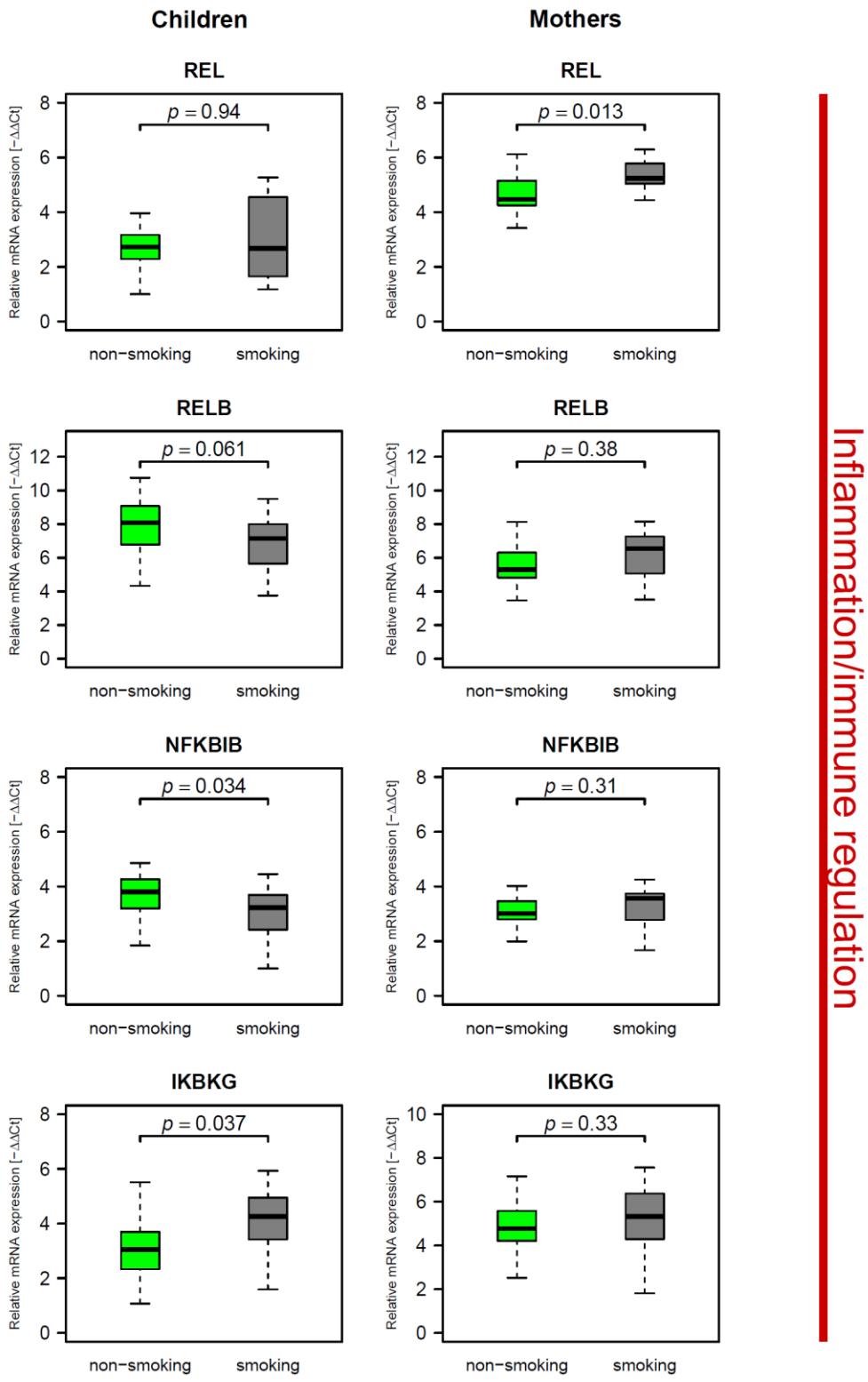



**Figure 42 Expression of mRNA of NFκB pathway genes.** mRNA expression of *REL*, *RELB*,

*NFκBIB* and *IKBKG* in the validation panel of smoking mothers (n=27, black), non-smoking mothers (n=15, green) and their children are shown. Data are represented as box plots (first and third quartile, median), the whiskers indicate ranges without outliers. P-values from Mann-Whitney U test/Student's t test.

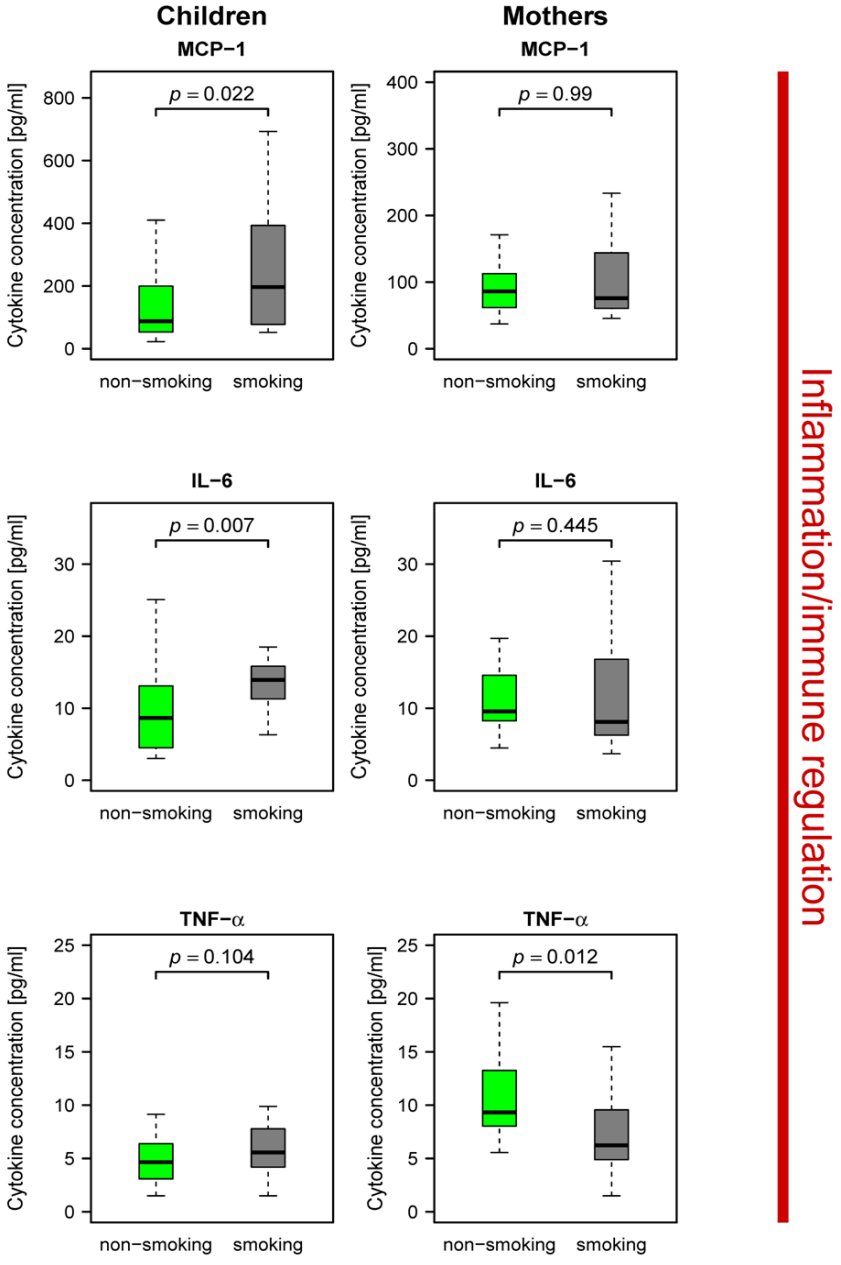



**Figure 43 Blood concentrations of inflammatory cytokines.** Concentrations of the inflammatory cytokines MCP-1, IL-6, TNF-a in the validation panel of smoking mothers (left, black, n=29), non-smoking mothers (left, green, n=16) and corresponding children (right). Data

are shown in box- plots (first and third quartile, median), the whiskers indicate ranges without outliers. P-values are from Mann-Whitney U test.

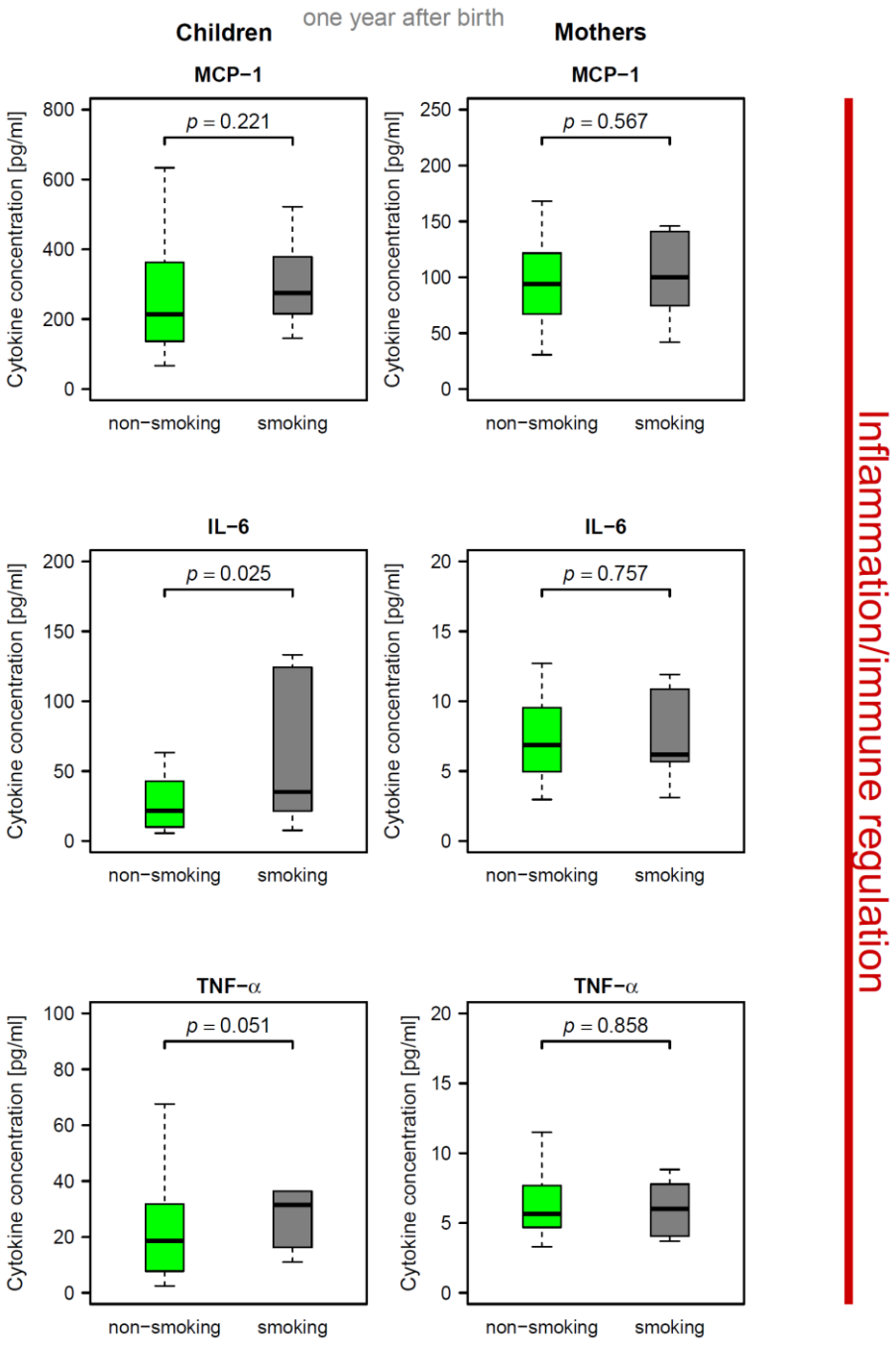
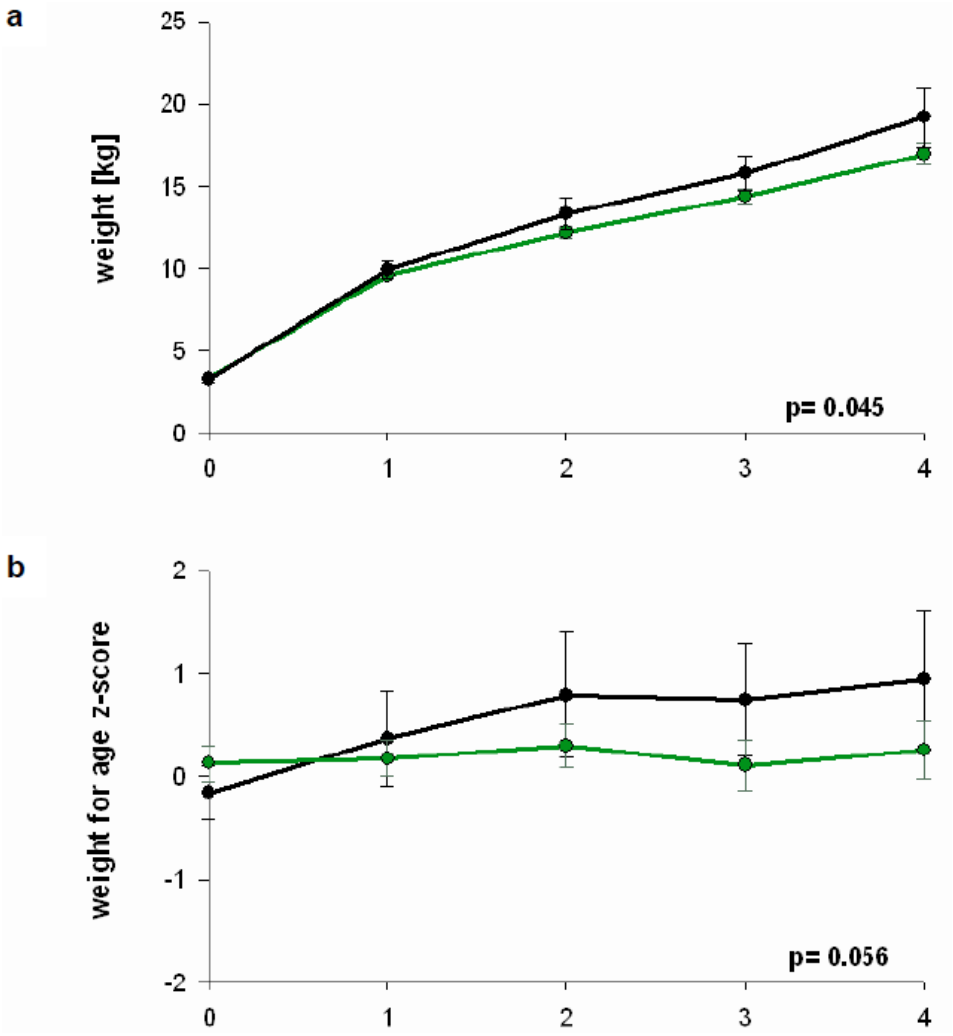



**Figure 44 Blood concentrations of inflammatory cytokines at one year after birth.** Concentrations of MCP-1, IL-6, TNF-a one year after birth in the validation panel of smoking mothers (n=28, black), non-smoking mothers (n=10, green) and their children. Data are

represented as box plots (first and third quartile, median), the whiskers indicate ranges without outliers. P-values from Student's t test of logarithmical data. Note that the inflammatory phenotype is sustained in children, but not in mothers one year after birth.



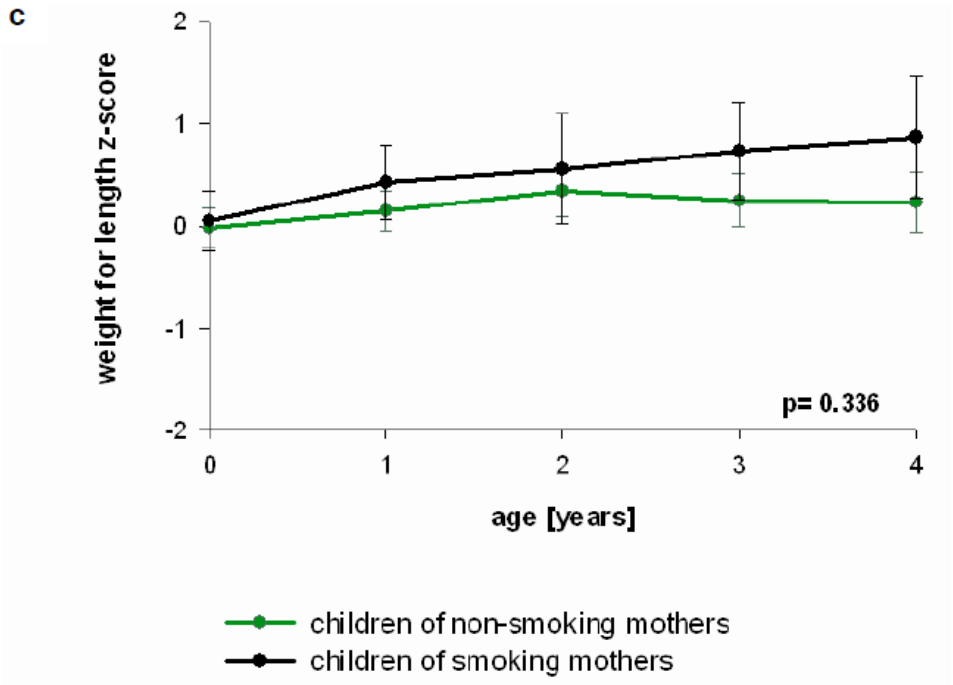



**Figure 45 Difference in the growth and weight development of children from non-smoking and smoking mothers.** Depicted are **(A)** weight, **(B)** z score for ”weight of age” and **(C)** z score for “weight of length” from birth up to the fourth year of age (mean +/- s.e.m) for both groups of children (black: from smoking mothers; green: from non-smoking mothers). Z scores were calculated using the WHO Anthro software (version 3.2.2, 2011). Data were fitted with a linear model using the LIMMA package of R and tested for statistical significance using ANOVA. Note that weight and ”weight of age” curves differ significantly, while there is no significant difference in “weight of length” between children from smoking and non-smoking mothers.

Much fewer data exist regarding the association between maternal smoking during pregnancy and type 2 diabetes (T2D) or metabolic syndrome. Animal studies showed a disruption of T2D-related pathways including an impaired pancreatic β-cell function after prenatal nicotine exposure[251,252]. Furthermore, insulin resistance, closely related to T2D, was already found related to maternal smoking during pregnancy[253]. The here described

epigenetic modification in several genes within the T2D pathway indicates maternal smoking as a risk factor for T2D.

Third, a surprisingly high number of cancer specific pathways were enriched in children. Seven out of a total of 14 cancer pathways (by KEGG-based analysis) were exclusively enriched in children, including essential signal transduction pathways related to cancer (p53, ErbB, hedgehog and WNT signaling). Thus, by epigenetically perturbing regulatory pathways, maternal smoking may also modify the cancer risk for children.

In mothers, an enrichment of DMRs was observed in cardiomyopathy-related pathways, such as cardiac muscle contraction and hypertrophic or dilated cardiomyopathy. This result is in agreement with the observation that smoking increases the risk of heart diseases[254,255]. Interestingly, the hypomethylation of *F2RL3* was detected which has been described to predict coronary heart disease[256], in smoking mothers but not in their children (**Table 9**).

| Gene name | DMR location | | | Proximal regulatory feature/ | Mother Child (M/C) | Hypo/ hyper | Mean methylation difference | Number of CpGs | p-value* | Epigenetic modifier |
|---|---|---|---|---|---|---|---|---|---|---|
| **HDAC7** | chr12 | 48197035 | 48197612 | intron | C | hyper | 0.09 | 11 | 0.0307 | repressive |
| **KDM6B** | chr17 | 7742259 | 7743410 | promoter-TSS | C | hyper | 0.09 | 11 | 0.0882 | activating** |
| **KDM5B** | chr1 | 202777861 | 202779663 | TFBS | C | hyper | 0.12 | 38 | 0.0893 | repressive |
| **SMYD3** | chr1 | 246234951 | 246236417 | intron | C | hypo | -0.10 | 23 | 0.0103 | activating |
| **MLL3** | chr7 | 152130609 | 152130970 | intron | C | hypo | -0.09 | 12 | 0.0221 | activating |
| **KAT8** | chr16 | 31127594 | 31128110 | DNAse cluster | C | hypo | -0.08 | 26 | 0.0273 | activating |
| **NSD1** | chr5 | 176555829 | 176555977 | TFBS | C | hypo | -0.11 | 4 | 0.0528 | activating |
| **SMYD2** | chr1 | 214399090 | 214399200 | DNAse cluster | C | hypo | -0.10 | 6 | 0.0240 | activating |
| **SUV39H2** | chr10 | 14954062 | 14955582 | intron | C | hypo | -0.07 | 23 | 0.0450 | repressive** |
| **DNMT1** | chr19 | 10296823 | 10298036 | intron | C | hypo | -0.06 | 26 | 0.0554 | repressive** |
| **SETDB1** | chr1 | 150897595 | 150898208 | promoter-TSS | M | hyper | 0.14 | 36 | 0.0393 | repressive |
| **KDM4A** | chr1 | 44114745 | 44115121 | promoter-TSS | M | hypo | -0.13 | 17 | 0.0040 | activating |
| **DOT1L** | chr19 | 2165738 | 2165990 | intron | M | hypo | -0.09 | 18 | 0.0098 | activating |
| **KDM4C** | chr9 | 6942611 | 6943057 | intron | M | hypo | -0.09 | 6 | 0.0299 | activating |
| **NSD1** | chr5 | 176541826 | 176542008 | enhancer | M | hypo | -0.10 | 7 | 0.0727 | activating |

* Welch's test

** deviation from pattern that repressive (activating) enzymes are hypermethylated (hypomethylated)

**Table 11 DMRs correlating to chromatin modifying enzymes in mothers and children determined by WGBS**

| Gene name | Mother/ Child (M/C) | Fold-change | p-value | | Epigenetic modifier type |
|---|---|---|---|---|---|
| **AURKC** | C | 1.30 | 0.0310 | a | activating |
| **HDAC8** | C | 2.70 | 0.0250 | a | repressive |
| **KDM5B** | C | -1.20 | 0.0560 | a | repressive |
| **SMYD2** | C | 2.00 | 0.0217 | a | activating |
| **USP16** | C | -1.30 | 0.0162 | a | |
| **AURKC** | M | 1.40 | 0.0350 | b | activating |
| **CSRP2BP** | M | -1.70 | 0.0060 | a | activating |
| **HDAC10** | M | -1.90 | 0.0220 | b | repressive |
| **HDAC8** | M | -2.20 | 0.0260 | b | repressive |
| **HDAC9** | M | -2.00 | 0.0470 | b | repressive |
| **KAT2A** | M | -1.70 | 0.0040 | b | activating |
| **KDM5B** | M | 1.30 | 0.0520 | a | repressive |
| **MLL** | M | -1.30 | 0.0070 | a | activating |
| **NCOA3** | M | -1.30 | 0.0587 | b | activating |
| **PRMT1** | M | -1.80 | 0.0004 | b | activating |
| **SETD8** | M | 1.70 | 0.0008 | a | activating |
| **SUV39H1** | M | -1.20 | 0.0606 | a | repressive |
| **SUV39H2** | M | -1.70 | 0.0052 | b | repressive |
| **UBE2A** | M | -1.30 | 0.0032 | a | |

**Table 12 Chromatin modifying enzymes showing a significant differential mRNA expression in mothers and children.**

| | Children | | Mothers | | |
|---|---|---|---|---|---|
| **Gene name** | **Fold change** | **p-value** | **Fold change** | **p-value** | |
| **ASH1L** | -1.40 | 0.0800 | -1.30 | 0.3240 | |
| **AURKA** | 1.30 | **0.0310** | -1.20 | 0.7930 | |
| **AURKB** | 1.30 | 0.2990 | 1.10 | 0.7230 | |
| **AURKC** | -1.30 | 0.2030 | 1.40 | **0.0350** | |
| **CARM1** | 1.00 | 0.7880 | -1.20 | 0.2120 | |
| **CSRP2BP** | -1.30 | 0.3760 | -1.70 | **0.0060** | |
| **DNMT1** | 1.30 | 0.4570 | -1.50 | 0.0960 | |
| **DNMT3A** | -1.00 | 0.9020 | -1.00 | 0.9370 | |
| **DNMT3b** | | n.d. | | n.d. | |
| **DNMT3L** | | n.d. | | n.d. | |
| **DOT1L** | -1.00 | 0.9790 | 1.40 | 0.1550 | |
| **DZIP3** | 1.20 | 0.4130 | -1.30 | 0.1560 | |
| **EHMT2** | 1.20 | 0.7550 | -1.20 | 0.1310 | |
| **HAT1** | 1.10 | 0.6850 | -1.20 | 0.3510 | |
| **HDAC10** | 1.40 | 0.2510 | -1.90 | **0.0220** | |
| **HDAC11** | 1.30 | 0.1590 | -1.30 | 0.1850 | |
| **HDAC3** | 1.20 | 0.4490 | 1.10 | 0.6390 | |
| **HDAC4** | 1.40 | 0.2880 | -1.50 | 0.1810 | |
| **HDAC5** | -1.20 | 0.2910 | -1.10 | 0.5290 | |
| **HDAC6** | 1.30 | 0.0940 | 1.10 | 0.5810 | |
| **HDAC7** | -1.20 | 0.4650 | -1.80 | 0.1390 | |
| **HDAC8** | 2.70 | **0.0250** | -2.20 | **0.0260** | |
| **HDAC9** | 1.30 | 0.4370 | -2.00 | **0.0470** | |
| **KAT2A** | 1.30 | 0.3920 | -1.70 | **0.0040** | |
| **KAT2B** | 1.30 | 0.5120 | -1.00 | 0.9100 | |
| **KAT5** | -1.10 | 0.5510 | -1.10 | 0.5630 | |
| **KAT6A** | -1.20 | 0.2760 | -1.00 | 0.7640 | |
| **KAT6B** | -1.80 | 0.5830 | 1.60 | 0.3760 | |
| **KAT7** | -1.10 | 0.2570 | 1.10 | 0.3100 | |
| **KAT8** | -1.10 | 0.5530 | 1.10 | 0.4510 | |
| **KDM1A** | 1.10 | 0.6230 | -1.20 | 0.2260 | |
| **KDM4A** | -1.30 | 0.6230 | -1.20 | 0.2260 | |
| **KDM4C** | 1.20 | 0.9210 | -1.40 | 0.6150 | |

| Gene name | Children | | | Mothers | | |
|---|---|---|---|---|---|---|
| | Fold change | p-value | | Fold change | p-value | |
| **KDM5B** | -1.20 | 0.0560 | a | 1.30 | **0.0520** | a |
| **KDM5C** | | n.d. | | | n.d. | |
| **KDM6B** | 1.40 | 0.4690 | a | -1.60 | 0.6550 | b |
| **MGMT** | | n.d. | | | n.d. | |
| **MLL** | -1.00 | 0.9660 | b | -1.30 | **0.0070** | a |
| **MLL3** | -1.10 | 0.4696 | b | 1.10 | 0.6553 | b |
| **MLL5** | -1.30 | 0.0614 | a | 1.10 | 0.6367 | a |
| **MYSM1** | 1.40 | 0.4059 | a | 1.20 | 0.5904 | b |
| **NCOA1** | 1.40 | 0.1340 | a | -1.30 | 0.2319 | a |
| **NCOA3** | 1.00 | 0.8574 | a | -1.30 | 0.0587 | b |
| **NCOA6** | 1.10 | 0.7088 | a | -1.20 | 0.2759 | b |
| **NEK6** | 1.20 | 0.4556 | a | -1.00 | 0.8132 | b |
| **NSD1** | 1.10 | 0.4492 | a | 1.10 | 0.6793 | a |
| **PAK1** | 1.40 | 0.1571 | a | -1.10 | 0.5995 | a |
| **PRMT1** | 1.30 | 0.0892 | a | -1.80 | **0.0004** | b |
| **RNF2** | -1.00 | 0.9891 | a | | n.d. | |
| **RNF20** | -1.10 | 0.7908 | a | -1.30 | 0.3446 | b |
| **RPS6KA3** | 1.40 | 0.1897 | a | -1.30 | 0.4158 | b |
| **RPS6KA5** | 1.10 | 0.7341 | a | -1.20 | 0.6365 | b |
| **SETD1A** | -1.20 | 0.1545 | a | 1.00 | 1.0000 | b |
| **SETD1B** | -1.00 | 0.7568 | a | 1.00 | 0.6314 | a |
| **SETD2** | -1.10 | 0.8205 | b | 1.20 | 0.2480 | b |
| **SETD3** | 1.40 | 0.2396 | a | -1.70 | 0.3721 | b |
| **SETD4** | 2.00 | 0.1333 | a | -2.60 | 0.2593 | a |
| **SETD5** | 1.20 | 0.8933 | b | 1.20 | 0.5117 | b |
| **SETD6** | 1.20 | 0.2541 | a | -1.30 | 0.1313 | a |
| **SETD7** | -1.00 | 0.5515 | b | 1.20 | 0.2533 | b |
| **SETD8** | 1.20 | 0.2669 | a | 1.70 | **0.0008** | a |
| **SETDB1** | 1.10 | 1.0000 | b | -1.10 | 0.9372 | b |
| **SETDB2** | 1.40 | 0.9774 | b | -1.20 | 0.4004 | a |
| **SMYD2** | 2.00 | **0.0217** | a | -1.10 | 0.6957 | b |
| **SMYD3** | -1.10 | 0.8445 | a | | n.d. | |
| **SUV39H1** | -1.30 | 0.3871 | b | -1.20 | 0.0606 | a |
| **SUV39H2** | 1.50 | 0.2069 | b | -1.70 | **0.0052** | b |
| **SUV420H1** | | n.d. | | | n.d. | |

| | Children | | | Mothers | | |
|---|---|---|---|---|---|---|
| **Gene name** | **Fold change** | **p-value** | | **Fold change** | **p-value** | |
| **UBE2A** | -1.10 | 0.5234 | b | -1.30 | **0.0032** | a |
| **UBE2B** | -1.00 | 0.9206 | b | 1.10 | 0.3168 | a |
| **USP16** | -1.30 | **0.0162** | a | 1.10 | 0.5201 | b |
| **USP21** | -1.30 | 0.7499 | b | -1.70 | 0.5837 | b |
| **USP22** | -1.10 | 0.3843 | a | -1.20 | 0.1687 | b |
| **WHSC1** | -1.00 | 0.8316 | b | -1.60 | 0.1633 | b |

n.d. not detectable

a Student's t-test

b Mann-Whitney U test

**Table 13 Chromatin modifying enzymes investigated by qPCR**

| **Gene name** | **DMR location** | | | **Proximal regulatory feature/ location** | **Mother Child (M/C)** | **Correlation** | **p-value** |
|---|---|---|---|---|---|---|---|
| **NCOA3*** | chr20 | 46060706 | 46060979 | enhancer | C | 0.808 | 0.052 |
| **SMYD3*** | chr1 | 246388869 | 246389547 | TFBS | M | 0.797 | 0.058 |
| **HAT1*** | chr2 | 172779525 | 172780031 | DNAse cluster | M | 0.793 | 0.060 |
| **SETDB1*** | chr1 | 150897674 | 150898138 | DNAse cluster | C | -0.774 | 0.071 |
| **HDAC4*** | chr2 | 240417580 | 240418121 | DNAse cluster | C | 0.730 | 0.099 |
| **KDM5B**** | chr1 | 202777861 | 202779663 | TFBS | M / C | −0.280 | 0.079 |

*Methylation estimated from WGBS

**Methylation estimated from MassARRAY

**Table 14 Chromatin regulators for which methylation correlates with DNMT1 expression**

Finally, to examine whether the observed changes in DNA methylation in response to tobacco smoke exposure may be linked to other epigenetic modifiers, a key set of 74 chromatin regulators (**Table 13**) were analyzed. 16/74 genes were differentially methylated in children or in mothers (**Table 12**) with only histone H3K36 methylase *NSD1* shared between mothers and children. When considering the expression of activating (histone H3/H4 acetylation, H3K4me1/me2/me3 and H3K36me1/me2 methylation) versus repressive histone marks (H3K9me2/me3, H3K27me2/me3) a striking epigenetic feature emerges: for 12/14 histone modifiers (except for *KDM6B* and *SUV39H2*) the enzymes that favor the active chromatin state were hypomethylated, while enzymes that favor repressive modifications were hypermethylated. For example, *SETDB1*, coding for the enzyme that sets the repressive H3K9me2/3 mark in euchromatin was hypermethylated, while the counteracting histone demethylases *KDM4A* and *KDM4C* that remove H3K9me2/3 were hypomethylated (**Table 11**). Likewise, *SMYD2*, which mediates the formation of the activating H3K36me2 mark, was hypomethylated and transcriptionally upregulated while the repressive H3K4 demethylase *KDM5B* was hypermethylated and transcriptionally downregulated (**Table 11**and **Table 12**). This suggested that tobacco smoke exposure results in a distinct epigenetic landscape with the potential to induce a more activated chromatin state.

It was previously shown that both *SUV39H1* as well as the H3K9me2/3 reader heterochromatin protein 1 (*HP1*) interact with *DNMT1*[257-259]; for a recent review on dependencies between histone methylation and DNA methylation[260]. Here, a highly significant transcriptional correlation between *SUV39H1* and *DNMT1* was observed in our validation panel (**Figure 46**). Furthermore, the transcription level of *DNMT1* was correlated ($p<0.1$) with the methylation level of 6/72 chromatin modifying enzymes (**Table 14**), and with the overall

methylation level of all DMRs called in the discovery panel (**Figure 47**). This leads us to propose that tobacco smoke induced differential DNA methylation is linked to deregulated histone methylation patterns, and that this might be mediated by *DNMT1*, which has been reported to be regulated by nicotine in mice[261]. As nicotine was shown to act like an *HDAC* inhibitor[262] and a number of differentially expressed histone (de)acetylases (**Table 13**) were identified, it emerges that the entire epigenetic program of smoking mothers and their children is changed on the level of DNA methylation, histone methylation and histone acetylation with the exact mechanisms still to be studied in detail.

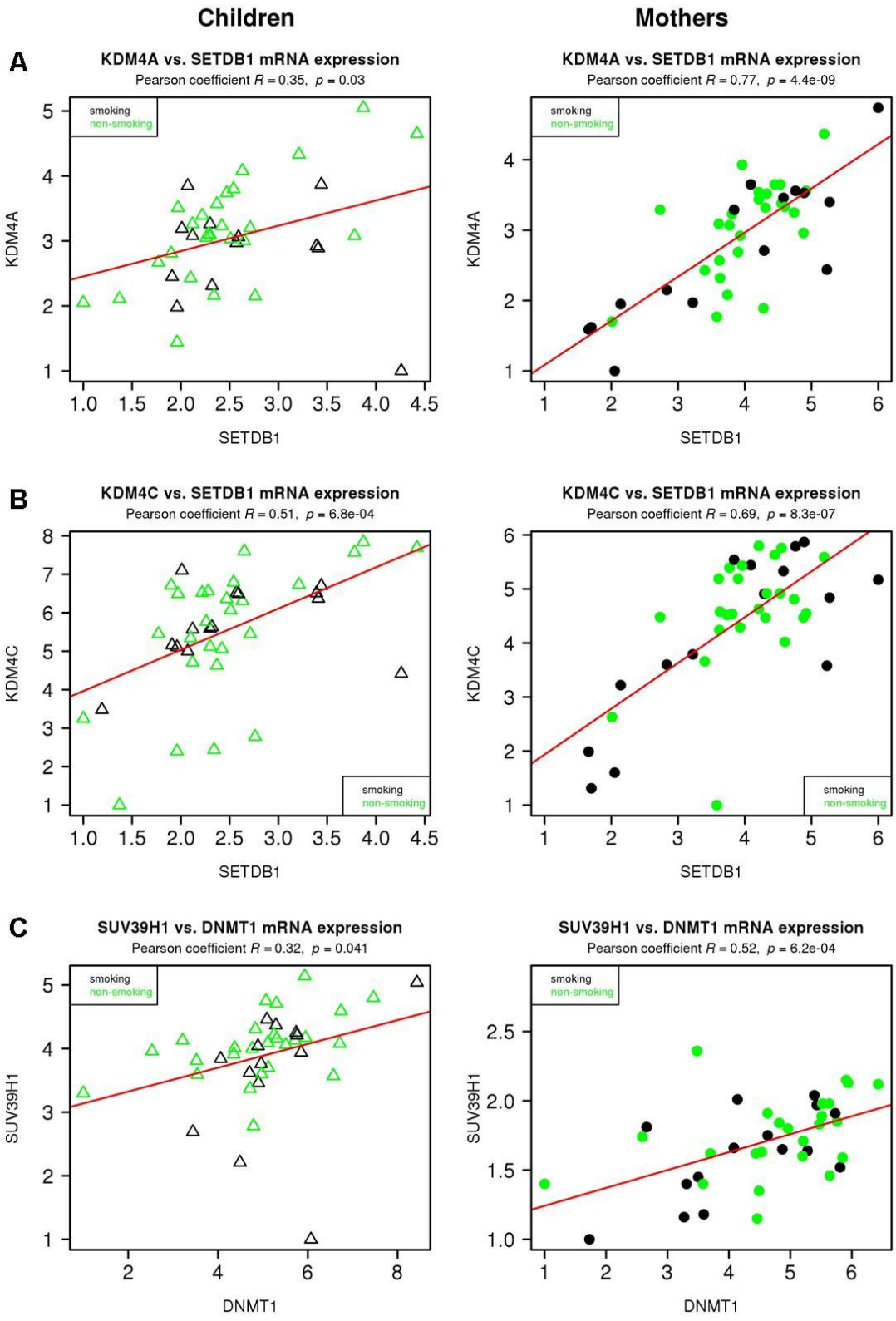



**Figure 46 Correlation of mRNA expression of different chromatin modifiers.** The enzyme *SETD1* sets the repressive H3K9me2/3 mark in euchromatin while *KDM4A* and *KDM4C*

remove this mark. Expression of *SETDB1* versus *KDM4A* and *KDM4C* mRNA, respectively, is highly correlated in mothers and slightly less correlated in children **(A,B)**. In addition, we observe a highly significant transcriptional correlation between *SUV39H1* and *DNMT1* **(C)**, two chromatin modifying enzymes which have previously been described to interact through the H3K9me2/me3 "reader" protein, *HP1* und *UHRF1*. Mothers: right panel, Children: left panel. Green: non-smoking, black: smoking.

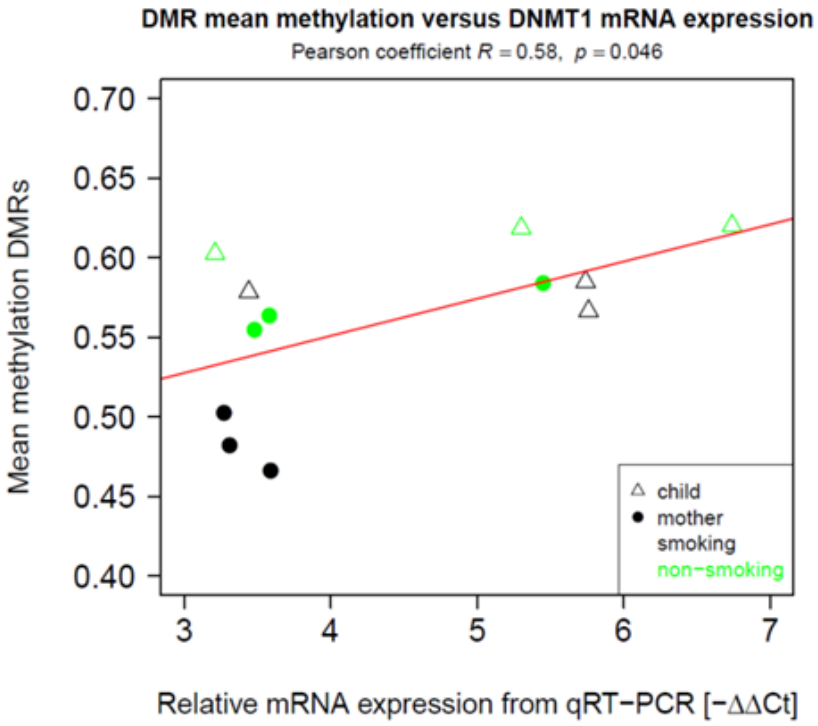



**Figure 47 Correlation between DMR mean methylation and *DNMT1* transcription.** DNMT1 is significantly correlated with the overall methylation change across all DMRs suggesting that *DNMT1* is involved in global maintenance of DNA methylation.

In summary, this study provides novel insights into the mechanisms by which an environmental stressor reprograms the epigenetic landscape in both mothers and children. The aberrant DNA methylation pattern will persist over time in the newborn child even if it is no longer exposed to smoking, since it will be faithfully copied via DNMT1 through cell divisions. Together with the

observed drastic deregulation of a number of disease related pathways in particular in children, the identified aberrant DNA methylation may act as a molecular mechanism for the long-lasting consequences of smoking during pregnancy.

## 5.4 Discussion

Transgenerational epigenetic inheritance has been recently observed in plants[263], C. elegans[264] and mice[265]. But proving epigenetic inheritance in human is difficult. Researchers must first rule out the possibility of genetic changes. Second, researchers have to show that the epigenetic effect can pass through enough generations to rule out the possibility of direct exposure. Because in a pregnant mother, three generations (1st generation: mother; 2nd generation: fetus; 3rd generation: reproductive cells in fetus) are directly exposed to the same environmental conditions at the same time. An epigenetic effect that continues into the 4th generation could be inherited and not due to direct exposure.

Although this study in the first time at a single base resolution showed a possible epigenetic mechanism to connect the maternal smoking and the influence to the next generation, the impact of this influence still needs to be validated by a longitudinal study. It will be interesting to see whether the alterations observed in newborns can be still observed after one year, five years and their later life.

As known the difficulties in such study in human beings, the limited sample size, on one hand, may decrease the power to detect subtle changes between smoking and non-smoking groups and, on the other hand, increase the false positive rate due to the variation of methylation level between each individual.

This study has revealed the association between the alteration of

methylation and the dysregulation of histone modifiers which fits the observations in C. elegans study which shows specific chromatin modifiers can induce an epigenetic memory[264]. Chip-seq on those observed dysregulated histone markers should be follow up in order to prove the functional consequence. More recently, the dysregulation of methyl group related metabolic pathways, such as folate metabolism[265], have been shown to cause the transgenerational epigenetic instability. So it makes sense to perform a systematic screenning of all possible methyl group related metabolism pathways in order to integrate metabolome into the epigenetic transgenerational inheritance machinary. In addtion, Emma Whitelaw recently proposed that RNA might be particularly involved in epigenetic inheritance[266]. Thus, combining WGBS, Chip-seq and RNA-seq will give us a more complete picture of epigenetic inheritance in different levels.

# Chapter 6: Perspectives

The processes of epigenetics have been expanded from DNA methylation and histone modifications to non-coding RNA, prion changes and polycomb mechanisms and it is likely that additional epigenetic processes will be discovered in the near future. Together with novel epigenetic mechanisms discovered by recently advanced techniques, epigenetic processes have been observed to be heavily involved not only in cancer and disease development, but also in metabolism, stem cell behavior, X chromosome inactivation, tissue regeneration, genomic imprinting, transgenerational reprogramming, memory processes and aging. However, the cause and consequences of the basic epigenetic machinery still remains a mystery. For example, what distinguishes two alleles when both have the same sequence in the same nuclear environment? Whether and how transgenerational epigenetic reprogramming occurs? What are the epigenetic marks in the germ cell which are used to maintain the totipotent genome? And how are these epigenetic marks dynamically regulated?

With the development of new techniques focused on the single cell level and the accumulation of longitudinal genome-wide epigenetic data in different populations or even different species, we will come closer to answer these fundamental questions in epigenetics.

## 6.1 Single Cell Epigenomics

Single cells are the fundamental units of life. Thus, single cell analysis will help us to better understand the fundamental biology of our life including how individual cells process information and respond to perturbations. The epigenome plays a key part in regulating the state of a single cell and makes

diversity in a population of cells.

Today, the gold standard technique for comprehensive, genome-wide analysis, is whole genome bisulfite sequencing (WGBS), which is based on ensemble measurements and requires sequencing a cell population, not a single cell. The variability between each cell is present to some degree in any cell population, and the ensemble behaviors of a population cannot represent the behaviors of any individual cell[267,268]. Stem cells, for example, including embryonic stem cells, adult stem cells and induced pluripotent stem cells are all heterogeneous populations[269,270]. Single cell amplification can target specific populations and therefore elucidate signaling pathways and networks for self-renewal and differentiation. Cancer is a heterogeneous disease and dissecting cell-to-cell variations is extremely important in understanding tumor initiation, progression, metastasis and therapeutic responses. Therefore, the current widely-used approach can just provide the distribution of DNA methylation within cells[271,272] or support models for the stochastic emergence of differential methylation[273]. The same problem exists with the ChIP-seq technique, which is used to generate genome-wide maps of histone modifications. It is impossible to know if a combination of transcription factors exists in a single individual cell.

Thus, highly sensitive methods with single cell resolution and ideally down to the single molecules level are required to accurately understand the complex intrapopulation heterogeneity and its impact on cell behavior and biological responses in cell populations which would be very revealing in the understanding of cancer evolution and stem cell development. One of the biggest challenges is to physically capture a single cell. Several approaches including micropipetting[274], FACS sorting[275] and microfluidics[276-278] already hold great promise. After capturing cells, one of whole genome amplification (WGA) strategies, named multiple displacement amplification (MDA)[279], is

used to obtain sufficient DNA for sequence analysis with potential amplification biases[280] and problems[281]. More recently, a new WGA method, named multiple annealing and looping-based amplification cycle (MALBAC), has been shown with considerable improvement on amplification fidelity[282]. Nevertheless, these methods extremely helped recent single-cell whole-genome analyses, especially in tumor evolution studies with low coverage single cell sequencing[283].

Unlike single cell genomic studies[284-286], the application of single cell approaches to epigenomic analysis has so far been limited. In single cell genomics, Helicos Biosciences has developed a high-throughput, amplification-free method for transcriptome profiling which is the single molecule sequencing digital gene expression (smsDGE)[287]. Although a recent ChIP-Seq study has shown the possibility of using very few cells and only 50 pg of input DNA[288], nobody has really been able to achieve epigenetic profiling from any single cell yet. Challenges are from both wet lab and dry lab. A major challenge in bisulfite sequencing is the up to 90% degradation of DNA when we perform the bisulfite conversion. Since the input genomic DNA in single cells is very limited, the extensive degradation makes molecular manipulations more difficult. Thus, T-WGBS and enrichment based approach might be good options because there are no harsh denaturing conditions causing severe degradation and loss of genomic DNA. For the challenges in dry lab, algorithms have been developed to tackle problems intrinsic to single cell genomics[289,290] but not for epigenomics due to the lack of real data. Thus, only proof-of-concept single cell epiegenetic analyses have been demonstrated for both DNA methylation[291,292] and histone modifications[293] so far.

In the near future, there will be a high demand for the improvement for bioinformatics[294] to study multiple individual cells to achieve statistical significance. Furthermore, interactions between cells and their extracellular

environment, need to be incorporated into experimental designs and data analyses[295].

## 6.2 Evolutionary Epigenomics

In contrast to single cell analysis, another aspect of epigenetics is to put epigenetics dynamics in the light of evolution. While the genome contains all genes, it is the epigenome that decides which are expressed. Though evolutionary genomics has focused on comparing the genomes of similar species and finding the commonalities to determine how common traits are regulated, evolutionary epigenomics provides a more in-depth look at regulatory functions.

The importance of epigenetics has long been appreciated at the molecular level. However, the role of epigenetics in evolution is a more recent focus. Epigenetic mechanisms interact with genetic and environmental factors, thus, play an important role in organism-environment interactions[296]. Epigenetic characters can be also stably transmitted across generations[297-299]. Therefore, epigenetics has now been considered in the framework of evolution and as a major force behind the evolutionary creation of new species. Indeed, epigenetic mechanisms play critical roles in phenotypic plasticity[300,301], response to environmental stressors and conservation biology[302]. Therefore, the higher level of our understandings in epigenetics, the more insights of individual and population processes at evolutionary time scales will be gained[303,304].

DNA methylation is a source of interindividual phenotypic variation and has been shown to contribute to varies phenotypic variations among individuals[305-308] such as flower shape and fruit pigmentation[309,310], mouse coat color[311,312], and traits differentiating queen and worker honeybees[313]. Thus,

DNA methylation may compensate for the decreased genetic variation in a new environment. The presence and stable transmission of an additional source of variation might be important. Therefore, it has to be incorporated into the evolutionary theory that epigenetic mechanisms mediate the increased phenotypic potential of certain genotypes.

With the current availability of vast epigenomic datasets and the prospect of even more epigenomic data coming in the near future, we will be able to compare the epigenetic signatures over different time periods for a single individual, different generations in a family, different individuals in the same population, different population in the same species and different species. All these comparisons will incredibly enhance our understanding of epigenome dynamics, which will in turn provide the power to investigate disease susceptibility and incidence, human evolution and species origins.

## 6.3 Multidisciplinary Epigenomics

The epigenetic machinery is now recognized as a fundamental mechanism in modulating the transcriptome. Thus it has been applied in many fields including not only cancer research, but also other areas of biological research. It will further continue to merge with other disciplines to assist in the explortion of biological complexity.

The brain is one of the most complex tissues in the human body, which remains one of the greatest mysteries in science and one of the greatest challenges to understand in medicine. Learning and memory are two basic functions of the brain. Thus, to understand the mechanisms of learning and memory now have become key questions that may have an essential epigenetic component. It is clear that environmental influences heavily affect the developing brain plasticity during postnatal development. By shaping

neural circuits, early environmental influences can determine structural and functional aspects of brain and behavior for the lifespan of the individual. How does the brain evolve? In particular, how (epi)genetic factors influence the brain functions under environmental selection? How to apply it to cure cognitive disease? These are questions future research is keen to answer.

In summary, if genomics is the tip of the iceberg, then epigenomics is the vastness that lies beneath.

# References:


1. Waddington, C.H. The epigenotype. 1942. *Int J Epidemiol* **41**, 10-13 (2012).

2. Liu, L., Li, Y. & Tollefsbol, T.O. Gene-environment interactions and epigenetic basis of human diseases. *Curr Issues Mol Biol* **10**, 25-36 (2008).

3. Chong, S. & Whitelaw, E. Epigenetic germline inheritance. *Curr Opin Genet Dev* **14**, 692-696 (2004).

4. Cavalli, G. & Paro, R. Epigenetic inheritance of active chromatin after removal of the main transactivator. *Science* **286**, 955-958 (1999).

5. Grewal, S.I. & Klar, A.J. Chromosomal inheritance of epigenetic states in fission yeast during mitosis and meiosis. *Cell* **86**, 95-101 (1996).

6. Rakyan, V.K., Blewitt, M.E., Druker, R., Preis, J.I. & Whitelaw, E. Metastable epialleles in mammals. *Trends Genet* **18**, 348-351 (2002).

7. Brink, R.A., Styles, E.D. & Axtell, J.D. Paramutation: directed genetic change. Paramutation occurs in somatic cells and heritably alters the functional state of a locus. *Science* **159**, 161-170 (1968).

8. Ehrlich, M., *et al.*. Amount and distribution of 5-methylcytosine in human DNA from different types of tissues of cells. *Nucleic Acids Res* **10**, 2709-2721 (1982).

9. Goll, M.G. & Bestor, T.H. Eukaryotic cytosine methyltransferases. *Annu Rev Biochem* **74**, 481-514 (2005).

10. Jeltsch, A. Molecular enzymology of mammalian DNA methyltransferases. *Curr Top Microbiol Immunol* **301**, 203-225 (2006).

11. Groth, A., Rocha, W., Verreault, A. & Almouzni, G. Chromatin challenges during DNA replication and repair. *Cell* **128**, 721-733 (2007).

12. Goll, M.G. & Bestor, T.H. Eukaryotic cytosine methyltransferases. *Annu Rev Biochem* **74**, 481-514 (2005).

13. Hermann, A., Schmitt, S. & Jeltsch, A. The human Dnmt2 has residual DNA-(cytosine-C5) methyltransferase activity. *J Biol Chem* **278**, 31717-31721 (2003).

14. Jurkowski, T.P., *et al.*. Human DNMT2 methylates tRNA(Asp) molecules using a DNA methyltransferase-like catalytic mechanism. *Rna* **14**, 1663-1670 (2008).

15. Esteller, M. Epigenetic gene silencing in cancer: the DNA hypermethylome. *Hum Mol Genet* **16 Spec No 1**, R50-R59 (2007).

16. Okano, M., Bell, D.W., Haber, D.A. & Li, E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell* **99**, 247-257 (1999).

17. Hata, K., Okano, M., Lei, H. & Li, E. Dnmt3L cooperates with the Dnmt3 family of de novo DNA methyltransferases to establish maternal imprints in mice. *Development* **129**, 1983-1993 (2002).

18. Jurkowska, R.Z., Jurkowski, T.P. & Jeltsch, A. Structure and function of mammalian DNA methyltransferases. *Chembiochem* **12**, 206-222 (2011).

19. Bird, A.P. CpG-rich islands and the function of DNA methylation. *Nature* **321**, 209-213 (1986).

20. Antequera, F. & Bird, A. Number of CpG islands and genes in human and mouse. *Proc Natl Acad Sci U S A* **90**, 11995-11999 (1993).

21. Lander, E.S., *et al.*. Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).

22. Venter, J.C., *et al.*. The sequence of the human genome. *Science* **291**, 1304-1351 (2001).

23. Razin, A. & Friedman, J. DNA methylation and its possible biological roles. *Prog Nucleic Acid Res Mol Biol* **25**, 33-52 (1981).

24. Razin, A. & Riggs, A.D. DNA methylation and gene function. *Science* **210**, 604-610 (1980).

25. Brandeis, M., *et al.*. The ontogeny of allele-specific methylation associated with imprinted genes in the mouse. *Embo J* **12**, 3669-3677 (1993).

26. Jahner, D., *et al.*. De novo methylation and expression of retroviral genomes during mouse embryogenesis. *Nature* **298**, 623-628 (1982).

27. Esteller, M. Cancer epigenomics: DNA methylomes and histone-modification maps. *Nat Rev Genet* **8**, 286-298 (2007).

28. Walsh, C.P., Chaillet, J.R. & Bestor, T.H. Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. *Nat Genet* **20**, 116-117 (1998).

29. Marmorstein, R. Protein modules that manipulate histone tails for chromatin regulation. *Nat Rev Mol Cell Biol* **2**, 422-432 (2001).

30. ALLFREY, V.G., FAULKNER, R. & MIRSKY, A.E. ACETYLATION AND METHYLATION OF HISTONES AND THEIR POSSIBLE ROLE IN THE REGULATION OF RNA SYNTHESIS. *Proc Natl Acad Sci U S A* **51**, 786-794 (1964).

31. Allfrey, V.G. & Mirsky, A.E. Structural Modifications of Histones and their Possible Role in the Regulation of RNA Synthesis. *Science* **144**, 559 (1964).

32. Kouzarides, T. Chromatin modifications and their function. *Cell* **128**, 693-705 (2007).

33. Jenuwein, T. & Allis, C.D. Translating the histone code. *Science* **293**, 1074-1080 (2001).

34. Strahl, B.D. & Allis, C.D. The language of covalent histone modifications. *Nature* **403**, 41-45 (2000).

35. Rando, O.J. & Chang, H.Y. Genome-wide views of chromatin structure. *Annu Rev Biochem* **78**, 245-271 (2009).

36. Huertas, D., Sendra, R. & Munoz, P. Chromatin dynamics coupled to DNA repair. *Epigenetics* **4**, 31-42 (2009).

37. Luco, R.F., *et al.*. Regulation of alternative splicing by histone modifications. *Science* **327**, 996-1000 (2010).

38. Li, B., Carey, M. & Workman, J.L. The role of chromatin during transcription. *Cell* **128**, 707-719 (2007).

39. Fyodorov, D.V. & Kadonaga, J.T. The many faces of chromatin remodeling: SWItching beyond transcription. *Cell* **106**, 523-525 (2001).

40. Loizou, J.I., *et al.*. Epigenetic information in chromatin: the code of entry for DNA repair. *Cell Cycle* **5**, 696-701 (2006).

41. Peterson, C.L. & Cote, J. Cellular machineries for chromosomal DNA repair. *Genes Dev* **18**, 602-616 (2004).

42. Mattick, J.S. The genetic signatures of noncoding RNAs. *PLoS Genet* **5**, e1000459 (2009).

43. Kurokawa, R., Rosenfeld, M.G. & Glass, C.K. Transcriptional regulation through noncoding RNAs and epigenetic modifications. *RNA Biol* **6**, 233-236 (2009).

44. Guttman, M., *et al.*. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**, 223-227 (2009).

45. Kapranov, P., Willingham, A.T. & Gingeras, T.R. Genome-wide transcription and the implications for genomic organization. *Nat Rev Genet* **8**, 413-423 (2007).

46. Chow, J. & Heard, E. X inactivation and the complexities of silencing a sex chromosome. *Curr Opin Cell Biol* **21**, 359-366 (2009).

47. Erwin, J.A. & Lee, J.T. New twists in X-chromosome inactivation. *Curr Opin Cell Biol* **20**, 349-355 (2008).

48. Lee, J.T., Davidow, L.S. & Warshawsky, D. Tsix, a gene antisense to Xist at the X-inactivation centre. *Nat Genet* **21**, 400-404 (1999).

49. Chaumeil, J., Le Baccon, P., Wutz, A. & Heard, E. A novel role for Xist RNA in the formation of a repressive nuclear compartment into which genes are recruited when silenced. *Genes Dev* **20**, 2223-2237 (2006).

50. Lee, J.T. & Lu, N. Targeted mutagenesis of Tsix leads to nonrandom X inactivation. *Cell* **99**, 47-57 (1999).

51. Kanellopoulou, C., *et al.*. X chromosome inactivation in the absence of Dicer. *Proc Natl Acad Sci U S A* **106**, 1122-1127 (2009).

52. Royo, H. & Cavaille, J. Non-coding RNAs in imprinted gene clusters. *Biol Cell* **100**, 149-166 (2008).

53. Lewis, A. & Reik, W. How imprinting centres work. *Cytogenet Genome Res* **113**, 81-89 (2006).

54. Feil, R., Walter, J., Allen, N.D. & Reik, W. Developmental control of allelic methylation in the imprinted mouse Igf2 and H19 genes. *Development* **120**, 2933-2943 (1994).

55. Thorvaldsen, J.L., Fedoriw, A.M., Nguyen, S. & Bartolomei, M.S. Developmental profile of H19 differentially methylated domain (DMD) deletion alleles reveals multiple roles of the DMD in regulating allelic expression and DNA methylation at the imprinted H19/Igf2 locus. *Mol Cell Biol* **26**, 1245-1258 (2006).

56. Cranston, M.J., Spinka, T.L., Elson, D.A. & Bartolomei, M.S. Elucidation of the minimal sequence required to imprint H19 transgenes. *Genomics* **73**, 98-107 (2001).

57. Shen, H. & Laird, P.W. Interplay between the cancer genome and epigenome. *Cell* **153**, 38-55 (2013).

58. Potter, V.R. Initiation and promotion in cancer formation: the importance of studies on intercellular communication. *Yale J Biol Med* **53**, 367-384 (1980).

59. Pierce, G.B. Neoplasms, differentiations and mutations. *Am J Pathol* **77**, 103-118 (1974).

60. Stehelin, D., Varmus, H.E., Bishop, J.M. & Vogt, P.K. DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA. *Nature* **260**, 170-173 (1976).

61. Bishop, J.M. Cellular oncogenes and retroviruses. *Annu Rev Biochem* **52**, 301-354 (1983).

62. Reddy, E.P., Reynolds, R.K., Santos, E. & Barbacid, M. A point mutation is responsible for the acquisition of transforming properties by the T24 human bladder carcinoma oncogene. *Nature* **300**, 149-152 (1982).

63. Cavenee, W.K., *et al.*. Expression of recessive alleles by chromosomal mechanisms in

retinoblastoma. *Nature* **305**, 779-784 (1983).

64. Weinberg, R.A. Oncogenes and tumor suppressor genes. *Trans Stud Coll Physicians Phila* **10**, 83-94 (1988).

65. Hanada, M., Delia, D., Aiello, A., Stadtmauer, E. & Reed, J.C. bcl-2 gene hypomethylation and high-level expression in B-cell chronic lymphocytic leukemia. *Blood* **82**, 1820-1828 (1993).

66. Kim, K.C. & Huang, S. Histone methyltransferases in tumor suppression. *Cancer Biol Ther* **2**, 491-499 (2003).

67. Zhang, B., Pan, X., Cobb, G.P. & Anderson, T.A. microRNAs as oncogenes and tumor suppressors. *Dev Biol* **302**, 1-12 (2007).

68. Esteller, M., *et al.*. Promoter hypermethylation of the DNA repair gene O(6)-methylguanine-DNA methyltransferase is associated with the presence of G:C to A:T transition mutations in p53 in human colorectal tumorigenesis. *Cancer Res* **61**, 4689-4692 (2001).

69. Esteller, M., *et al.*. Inactivation of the DNA repair gene O6-methylguanine-DNA methyltransferase by promoter hypermethylation is associated with G to A mutations in K-ras in colorectal tumorigenesis. *Cancer Res* **60**, 2368-2371 (2000).

70. Jones, P.A. & Baylin, S.B. The fundamental role of epigenetic events in cancer. *Nat Rev Genet* **3**, 415-428 (2002).

71. Irizarry, R.A., *et al.*. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat Genet* **41**, 178-186 (2009).

72. Ravenel, J.D., *et al.*. Loss of imprinting of insulin-like growth factor-II (IGF2) gene in distinguishing specific biologic subtypes of Wilms tumor. *J Natl Cancer Inst* **93**, 1698-1703 (2001).

73. Yuan, E., *et al.*. Genomic profiling maps loss of heterozygosity and defines the timing and stage dependence of epigenetic and genetic events in Wilms' tumors. *Mol Cancer Res* **3**, 493-502 (2005).

74. Feinberg, A.P., Ohlsson, R. & Henikoff, S. The epigenetic progenitor origin of human cancer. *Nat Rev Genet* **7**, 21-33 (2006).

75. Feinberg, A.P., Ohlsson, R. & Henikoff, S. The epigenetic progenitor origin of human cancer. *Nat Rev Genet* **7**, 21-33 (2006).

76. Fraga, M.F., *et al.*. Loss of acetylation at Lys16 and trimethylation at Lys20 of histone H4 is a common hallmark of human cancer. *Nat Genet* **37**, 391-400 (2005).

77. Tryndyak, V.P., Kovalchuk, O. & Pogribny, I.P. Loss of DNA methylation and histone H4 lysine 20 trimethylation in human breast cancer cells is associated with aberrant expression of DNA methyltransferase 1, Suv4-20h2 histone methyltransferase and methyl-binding proteins. *Cancer Biol Ther* **5**, 65-70 (2006).

78. Pogribny, I.P., *et al.*. Histone H3 lysine 9 and H4 lysine 20 trimethylation and the expression of Suv4-20h2 and Suv-39h1 histone methyltransferases in hepatocarcinogenesis induced by methyl deficiency in rats. *Carcinogenesis* **27**, 1180-1186 (2006).

79. Park, J.Y. Promoter hypermethylation in prostate cancer. *Cancer Control* **17**, 245-255 (2010).

80. Brooks, J.D., *et al.*. CG island methylation changes near the GSTP1 gene in prostatic intraepithelial neoplasia. *Cancer Epidemiol Biomarkers Prev* **7**, 531-536 (1998).

81. Li, L.C. Epigenetics of prostate cancer. *Front Biosci* **12**, 3377-3397 (2007).

82. Lee, W.H., *et al.*. Cytidine methylation of regulatory sequences near the pi-class glutathione S-transferase gene accompanies human prostatic carcinogenesis. *Proc Natl Acad Sci U S A* **91**,

11733-11737 (1994).

83. Jeronimo, C., *et al.*. A quantitative promoter methylation profile of prostate cancer. *Clin Cancer Res* **10**, 8472-8478 (2004).

84. Santourlidis, S., Florl, A., Ackermann, R., Wirtz, H.C. & Schulz, W.A. High frequency of alterations in DNA methylation in adenocarcinoma of the prostate. *Prostate* **39**, 166-174 (1999).

85. Aitchison, A., Warren, A., Neal, D. & Rabbitts, P. RASSF1A promoter methylation is frequently detected in both pre-malignant and non-malignant microdissected prostatic epithelial tissues. *Prostate* **67**, 638-644 (2007).

86. Florl, A.R., *et al.*. Coordinate hypermethylation at specific genes in prostate carcinoma precedes LINE-1 hypomethylation. *Br J Cancer* **91**, 985-994 (2004).

87. Richiardi, L., *et al.*. Promoter methylation in APC, RUNX3, and GSTP1 and mortality in prostate cancer patients. *J Clin Oncol* **27**, 3161-3168 (2009).

88. Liu, L., *et al.*. Association of tissue promoter methylation levels of APC, TGFbeta2, HOXD3 and RASSF1A with prostate cancer progression. *Int J Cancer* **129**, 2454-2462 (2011).

89. Henrique, R., *et al.*. High promoter methylation levels of APC predict poor prognosis in sextant biopsies from prostate cancer patients. *Clin Cancer Res* **13**, 6122-6129 (2007).

90. Gal-Yam, E.N., *et al.*. Frequent switching of Polycomb repressive marks and DNA hypermethylation in the PC3 prostate cancer cell line. *Proc Natl Acad Sci U S A* **105**, 12979-12984 (2008).

91. Schulz, W.A. & Hoffmann, M.J. Epigenetic mechanisms in the biology of prostate cancer. *Semin Cancer Biol* **19**, 172-180 (2009).

92. Wilson, A.S., Power, B.E. & Molloy, P.L. DNA hypomethylation and human diseases. *Biochim Biophys Acta* **1775**, 138-162 (2007).

93. Yegnasubramanian, S., *et al.*. DNA hypomethylation arises later in prostate cancer progression than CpG island hypermethylation and contributes to metastatic tumor heterogeneity. *Cancer Res* **68**, 8954-8967 (2008).

94. Poliseno, L., *et al.*. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* **465**, 1033-1038 (2010).

95. Yang, L., *et al.*. lncRNA-dependent mechanisms of androgen-receptor-regulated gene activation programs. *Nature* (2013).

96. Toyota, M., *et al.*. CpG island methylator phenotype in colorectal cancer. *Proc Natl Acad Sci U S A* **96**, 8681-8686 (1999).

97. Noushmehr, H., *et al.*. Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. *Cancer Cell* **17**, 510-522 (2010).

98. Bae, Y.K., *et al.*. Hypermethylation in histologically distinct classes of breast cancer. *Clin Cancer Res* **10**, 5998-6005 (2004).

99. Fang, F., *et al.*. Breast cancer methylomes establish an epigenomic foundation for metastasis. *Sci Transl Med* **3**, 25r-75r (2011).

100. Jing, F., *et al.*. CpG island methylator phenotype of multigene in serum of sporadic breast carcinoma. *Tumour Biol* **31**, 321-331 (2010).

101. Arai, E., *et al.*. Single-CpG-resolution methylome analysis identifies clinicopathologically aggressive CpG island methylator phenotype clear cell renal cell carcinomas. *Carcinogenesis* **33**,

1487-1493 (2012).
102. An, C., *et al.*. Prognostic significance of CpG island methylator phenotype and microsatellite instability in gastric carcinoma. *Clin Cancer Res* **11**, 656-663 (2005).
103. Etoh, T., *et al.*. Increased DNA methyltransferase 1 (DNMT1) protein expression correlates significantly with poorer tumor differentiation and frequent DNA hypermethylation of multiple CpG islands in gastric cancers. *Am J Pathol* **164**, 689-699 (2004).
104. Oue, N., *et al.*. DNA methylation of multiple genes in gastric carcinoma: association with histological type and CpG island methylator phenotype. *Cancer Sci* **94**, 901-905 (2003).
105. Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609-615 (2011).
106. Weisenberger, D.J., *et al.*. CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer. *Nat Genet* **38**, 787-793 (2006).
107. Turcan, S., *et al.*. IDH1 mutation is sufficient to establish the glioma hypermethylator phenotype. *Nature* **483**, 479-483 (2012).
108. Figueroa, M.E., *et al.*. Leukemic IDH1 and IDH2 mutations result in a hypermethylation phenotype, disrupt TET2 function, and impair hematopoietic differentiation. *Cancer Cell* **18**, 553-567 (2010).
109. Dang, L., *et al.*. Cancer-associated IDH1 mutations produce 2-hydroxyglutarate. *Nature* **462**, 739-744 (2009).
110. Kloosterhof, N.K., Bralten, L.B., Dubbink, H.J., French, P.J. & van den Bent, M.J. Isocitrate dehydrogenase-1 mutations: a fundamentally new understanding of diffuse glioma? *Lancet Oncol* **12**, 83-91 (2011).
111. Ward, P.S., *et al.*. The common feature of leukemia-associated IDH1 and IDH2 mutations is a neomorphic enzyme activity converting alpha-ketoglutarate to 2-hydroxyglutarate. *Cancer Cell* **17**, 225-234 (2010).
112. Xu, W., *et al.*. Oncometabolite 2-hydroxyglutarate is a competitive inhibitor of alpha-ketoglutarate-dependent dioxygenases. *Cancer Cell* **19**, 17-30 (2011).
113. Elsasser, S.J., Allis, C.D. & Lewis, P.W. New epigenetic drivers of cancers. *Science* **331**, 1145-1146 (2011).
114. Suva, M.L., Riggi, N. & Bernstein, B.E. Epigenetic reprogramming in cancer. *Science* **339**, 1567-1570 (2013).
115. Shen, H. & Laird, P.W. Interplay between the cancer genome and epigenome. *Cell* **153**, 38-55 (2013).
116. Plass, C., *et al.*. Mutations in regulators of the epigenome and their connections to global chromatin patterns in cancer. *Nat Rev Genet* (2013).
117. Siemiatycki, J., *et al.*. Listing occupational carcinogens. *Environ Health Perspect* **112**, 1447-1459 (2004).
118. Dubrova, Y.E., Plumb, M., Gutierrez, B., Boulton, E. & Jeffreys, A.J. Transgenerational mutation by radiation. *Nature* **405**, 37 (2000).
119. Charles, M. UNSCEAR report 2000: sources and effects of ionizing radiation. United Nations Scientific Comittee on the Effects of Atomic Radiation. *J Radiol Prot* **21**, 83-86 (2001).
120. Probst, A.V., Dunleavy, E. & Almouzni, G. Epigenetic inheritance during the cell cycle. *Nat Rev*

*Mol Cell Biol* **10**, 192-206 (2009).
121. Morgan, H.D., Sutherland, H.G., Martin, D.I. & Whitelaw, E. Epigenetic inheritance at the agouti locus in the mouse. *Nat Genet* **23**, 314-318 (1999).
122. Rakyan, V. & Whitelaw, E. Transgenerational epigenetic inheritance. *Curr Biol* **13**, R6 (2003).
123. Druker, R., Bruxner, T.J., Lehrbach, N.J. & Whitelaw, E. Complex patterns of transcription at the insertion site of a retrotransposon in the mouse. *Nucleic Acids Res* **32**, 5800-5808 (2004).
124. Duhl, D.M., Vrieling, H., Miller, K.A., Wolff, G.L. & Barsh, G.S. Neomorphic agouti mutations in obese yellow mice. *Nat Genet* **8**, 59-65 (1994).
125. Vasicek, T.J., *et al.*. Two dominant mutations in the mouse fused gene are the result of transposon insertions. *Genetics* **147**, 777-786 (1997).
126. Chan, T.L., *et al.*. Heritable germline epimutation of MSH2 in a family with hereditary nonpolyposis colorectal cancer. *Nat Genet* **38**, 1178-1183 (2006).
127. Hitchins, M.P., *et al.*. Inheritance of a cancer-associated MLH1 germ-line epimutation. *N Engl J Med* **356**, 697-705 (2007).
128. Suter, C.M., Martin, D.I. & Ward, R.L. Germline epimutation of MLH1 in individuals with multiple cancers. *Nat Genet* **36**, 497-501 (2004).
129. Baccarelli, A. & Bollati, V. Epigenetics and environmental chemicals. *Curr Opin Pediatr* **21**, 243-251 (2009).
130. Dolinoy, D.C., Huang, D. & Jirtle, R.L. Maternal nutrient supplementation counteracts bisphenol A-induced DNA hypomethylation in early development. *Proc Natl Acad Sci U S A* **104**, 13056-13061 (2007).
131. Waterland, R.A. Is epigenetics an important link between early life events and adult disease? *Horm Res* **71 Suppl 1**, 13-16 (2009).
132. Baccarelli, A., *et al.*. Exposure to particulate air pollution and risk of deep vein thrombosis. *Arch Intern Med* **168**, 920-927 (2008).
133. Brook, R.D., *et al.*. Air pollution and cardiovascular disease: a statement for healthcare professionals from the Expert Panel on Population and Prevention Science of the American Heart Association. *Circulation* **109**, 2655-2671 (2004).
134. Peters, A. Particulate matter and heart disease: evidence from epidemiological studies. *Toxicol Appl Pharmacol* **207**, 477-482 (2005).
135. Samet, J.M., Dominici, F., Curriero, F.C., Coursac, I. & Zeger, S.L. Fine particulate air pollution and mortality in 20 U.S. cities, 1987-1994. *N Engl J Med* **343**, 1742-1749 (2000).
136. Vineis, P. & Husgafvel-Pursiainen, K. Air pollution and cancer: biomarker studies in human populations. *Carcinogenesis* **26**, 1846-1855 (2005).
137. Tarantini, L., *et al.*. Effects of particulate matter on genomic DNA methylation content and iNOS promoter methylation. *Environ Health Perspect* **117**, 217-222 (2009).
138. Chahine, T., *et al.*. Particulate air pollution, oxidative stress genes, and heart rate variability in an elderly cohort. *Environ Health Perspect* **115**, 1617-1622 (2007).
139. Baccarelli, A., *et al.*. Air pollution, smoking, and plasma homocysteine. *Environ Health Perspect* **115**, 176-181 (2007).
140. Alexeeff, S.E., *et al.*. Ozone exposure, antioxidant genes, and lung function in an elderly cohort: VA normative aging study. *Occup Environ Med* **65**, 736-742 (2008).

141. Rusiecki, J.A., *et al.*. Global DNA hypomethylation is associated with high serum-persistent organic pollutants in Greenlandic Inuit. *Environ Health Perspect* **116**, 1547-1552 (2008).
142. Flom, J.D., *et al.*. Prenatal smoke exposure and genomic DNA methylation in a multiethnic birth cohort. *Cancer Epidemiol Biomarkers Prev* **20**, 2518-2523 (2011).
143. Breton, C.V., *et al.*. Prenatal tobacco smoke exposure affects global and gene-specific DNA methylation. *Am J Respir Crit Care Med* **180**, 462-467 (2009).
144. Breton, C.V., Salam, M.T. & Gilliland, F.D. Heritability and role for the environment in DNA methylation in AXL receptor tyrosine kinase. *Epigenetics* **6**, 895-898 (2011).
145. Murphy, S.K., *et al.*. Gender-specific methylation differences in relation to prenatal exposure to cigarette smoke. *Gene* **494**, 36-43 (2012).
146. Li, H. & Homer, N. A survey of sequence alignment algorithms for next-generation sequencing. *Brief Bioinform* **11**, 473-483 (2010).
147. Rakyan, V.K., *et al.*. DNA methylation profiling of the human major histocompatibility complex: a pilot study for the human epigenome project. *PLoS Biol* **2**, e405 (2004).
148. Eckhardt, F., Beck, S., Gut, I.G. & Berlin, K. Future potential of the Human Epigenome Project. *Expert Rev Mol Diagn* **4**, 609-618 (2004).
149. 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).
150. Bernstein, B.E., *et al.*. The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol* **28**, 1045-1048 (2010).
151. Wang, D., *et al.*. IMA: an R package for high-throughput analysis of Illumina's 450K Infinium methylation data. *Bioinformatics* **28**, 729-730 (2012).
152. Barfield, R.T., Kilaru, V., Smith, A.K. & Conneely, K.N. CpGassoc: an R function for analysis of DNA methylation microarray data. *Bioinformatics* **28**, 1280-1281 (2012).
153. Kilaru, V., Barfield, R.T., Schroeder, J.W., Smith, A.K. & Conneely, K.N. MethLAB: a graphical user interface package for the analysis of array-based DNA methylation data. *Epigenetics* **7**, 225-229 (2012).
154. Taiwo, O., *et al.*. Methylome analysis using MeDIP-seq with low DNA concentrations. *Nat Protoc* **7**, 617-636 (2012).
155. Brinkman, A.B., *et al.*. Whole-genome DNA methylation profiling using MethylCap-seq. *Methods* **52**, 232-236 (2010).
156. Gebhard, C., *et al.*. Genome-wide profiling of CpG methylation identifies novel targets of aberrant hypermethylation in myeloid leukemia. *Cancer Res* **66**, 6118-6128 (2006).
157. Lister, R., *et al.*. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315-322 (2009).
158. Adey, A. & Shendure, J. Ultra-low-input, tagmentation-based whole-genome bisulfite sequencing. *Genome Res* **22**, 1139-1143 (2012).
159. Krueger, F. & Andrews, S.R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571-1572 (2011).
160. Xi, Y. & Li, W. BSMAP: whole genome bisulfite sequence MAPping program. *BMC Bioinformatics* **10**, 232 (2009).
161. Hansen, K.D., Langmead, B. & Irizarry, R.A. BSmooth: from whole genome bisulfite sequencing

reads to differentially methylated regions. *Genome Biol* **13**, R83 (2012).
162. Stadler, M.B., *et al.*. DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* **480**, 490-495 (2011).
163. Burger, L., Gaidatzis, D., Schubeler, D. & Stadler, M.B. Identification of active regulatory regions from DNA methylation data. *Nucleic Acids Res* **41**, e155 (2013).
164. Zhao, Z. & Boerwinkle, E. Neighboring-nucleotide effects on single nucleotide polymorphisms: a study of 2.6 million polymorphisms across the human genome. *Genome Res* **12**, 1679-1686 (2002).
165. Liu, Y., Siegmund, K.D., Laird, P.W. & Berman, B.P. Bis-SNP: Combined DNA methylation and SNP calling for Bisulfite-seq data. *Genome Biol* **13**, R61 (2012).
166. Johnson, M.D., Mueller, M., Game, L. & Aitman, T.J. Single nucleotide analysis of cytosine methylation by whole-genome shotgun bisulfite sequencing. *Curr Protoc Mol Biol* **Chapter 21**, t21-t23 (2012).
167. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).
168. Meissner, A., *et al.*. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res* **33**, 5868-5877 (2005).
169. 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).
170. Chavez, L., *et al.*. Computational analysis of genome-wide DNA methylation during the differentiation of human embryonic stem cells along the endodermal lineage. *Genome Res* **20**, 1441-1450 (2010).
171. Bock, C., *et al.*. Quantitative comparison of genome-wide DNA methylation mapping technologies. *Nat Biotechnol* **28**, 1106-1114 (2010).
172. Chang, K., *et al.*. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* **45**, 1113-1120 (2013).
173. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* **499**, 43-49 (2013).
174. Kandoth, C., *et al.*. Integrated genomic characterization of endometrial carcinoma. *Nature* **497**, 67-73 (2013).
175. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med* **368**, 2059-2074 (2013).
176. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61-70 (2012).
177. Timp, W. & Feinberg, A.P. Cancer as a dysregulated epigenome allowing cellular growth advantage at the expense of the host. *Nat Rev Cancer* **13**, 497-510 (2013).
178. Croce, C.M. Causes and consequences of microRNA dysregulation in cancer. *Nat Rev Genet* **10**, 704-714 (2009).
179. Metzig, M., *et al.*. An RNAi screen identifies USP2 as a factor required for TNF-alpha-induced NF-kappaB signaling. *Int J Cancer* **129**, 607-618 (2011).
180. Sturm, D., *et al.*. Hotspot mutations in H3F3A and IDH1 define distinct epigenetic and biological subgroups of glioblastoma. *Cancer Cell* **22**, 425-437 (2012).
181. Schlomm, T., *et al.*. Clinical significance of p53 alterations in surgically treated prostate cancers. *Mod Pathol* **21**, 1371-1378 (2008).

182. Mirlacher, M. & Simon, R. Recipient block TMA technique. *Methods Mol Biol* **664**, 37-44 (2010).
183. Minner, S., *et al.*. High level PSMA expression is associated with early PSA recurrence in surgically treated prostate cancer. *Prostate* **71**, 281-288 (2011).
184. Grupp, K., *et al.*. Cysteine-rich secretory protein 3 overexpression is linked to a subset of PTEN-deleted ERG fusion-positive prostate cancers with early biochemical recurrence. *Mod Pathol* **26**, 733-742 (2013).
185. Muller, J., *et al.*. Loss of pSer2448-mTOR expression is linked to adverse prognosis and tumor progression in ERG-fusion-positive cancers. *Int J Cancer* **132**, 1333-1340 (2013).
186. Uchida, Y., *et al.*. MiR-133a induces apoptosis through direct regulation of GSTP1 in bladder cancer cell lines. *Urol Oncol* **31**, 115-123 (2013).
187. Ji, F., *et al.*. MicroRNA-133a, downregulated in osteosarcoma, suppresses proliferation and promotes apoptosis by targeting Bcl-xL and Mcl-1. *Bone* **56**, 220-226 (2013).
188. Tao, J., *et al.*. microRNA-133 inhibits cell proliferation, migration and invasion in prostate cancer cells by targeting the epidermal growth factor receptor. *Oncol Rep* **27**, 1967-1975 (2012).
189. Kojima, S., *et al.*. Tumour suppressors miR-1 and miR-133a target the oncogenic function of purine nucleoside phosphorylase (PNP) in prostate cancer. *Br J Cancer* **106**, 405-413 (2012).
190. Schmitz, K.M., Mayer, C., Postepska, A. & Grummt, I. Interaction of noncoding RNA with the rDNA promoter mediates recruitment of DNMT3b and silencing of rRNA genes. *Genes Dev* **24**, 2264-2269 (2010).
191. Zhou, Y. & Grummt, I. The PHD finger/bromodomain of NoRC interacts with acetylated histone H4K16 and is sufficient for rDNA silencing. *Curr Biol* **15**, 1434-1438 (2005).
192. Brase, J.C., *et al.*. TMPRSS2-ERG -specific transcriptional modulation is associated with prostate cancer biomarkers and TGF-beta signaling. *Bmc Cancer* **11**, 507 (2011).
193. Issa, J.P. CpG island methylator phenotype in cancer. *Nat Rev Cancer* **4**, 988-993 (2004).
194. Zouridis, H., *et al.*. Methylation subtypes and large-scale epigenetic alterations in gastric cancer. *Sci Transl Med* **4**, 140r-156r (2012).
195. McCabe, M.T., Lee, E.K. & Vertino, P.M. A multifactorial signature of DNA sequence and polycomb binding predicts aberrant CpG island methylation. *Cancer Res* **69**, 282-291 (2009).
196. Lee, W.H., *et al.*. Cytidine methylation of regulatory sequences near the pi-class glutathione S-transferase gene accompanies human prostatic carcinogenesis. *Proc Natl Acad Sci U S A* **91**, 11733-11737 (1994).
197. Van Neste, L., *et al.*. The epigenetic promise for prostate cancer diagnosis. *Prostate* **72**, 1248-1261 (2012).
198. Shyr, C.R., *et al.*. Tumor suppressor PAX6 functions as androgen receptor co-repressor to inhibit prostate cancer growth. *Prostate* **70**, 190-199 (2010).
199. Fraizer, G., *et al.*. Suppression of prostate tumor cell growth in vivo by WT1, the Wilms' tumor suppressor gene. *Int J Oncol* **24**, 461-471 (2004).
200. Nguyen, A.H., *et al.*. Gata3 antagonizes cancer progression in Pten-deficient prostates. *Hum Mol Genet* **22**, 2400-2410 (2013).
201. Kypta, R.M. & Waxman, J. Wnt/beta-catenin signalling in prostate cancer. *Nat Rev Urol* (2012).
202. Ostling, P., *et al.*. Systematic analysis of microRNAs targeting the androgen receptor in prostate

cancer cells. *Cancer Res* **71**, 1956-1967 (2011).
203. Kashat, M., *et al.*. Inactivation of AR and Notch-1 signaling by miR-34a attenuates prostate cancer aggressiveness. *Am J Transl Res* **4**, 432-442 (2012).
204. Shi, X.B., *et al.*. Tumor suppressive miR-124 targets androgen receptor and inhibits proliferation of prostate cancer cells. *Oncogene* (2012).
205. Postepska-Igielska, A., *et al.*. The chromatin remodelling complex NoRC safeguards genome stability by heterochromatin formation at telomeres and centromeres. *Embo Rep* (2013).
206. Weisenberger, D.J., *et al.*. CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer. *Nat Genet* **38**, 787-793 (2006).
207. Prensner, J.R., *et al.*. The long noncoding RNA SChLAP1 promotes aggressive prostate cancer and antagonizes the SWI/SNF complex. *Nat Genet* (2013).
208. Rinn, J.L., *et al.*. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **129**, 1311-1323 (2007).
209. Tsai, M.C., *et al.*. Long noncoding RNA as modular scaffold of histone modification complexes. *Science* **329**, 689-693 (2010).
210. Gupta, R.A., *et al.*. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* **464**, 1071-1076 (2010).
211. Shen, H., *et al.*. The SWI/SNF ATPase Brm is a gatekeeper of proliferative control in prostate cancer. *Cancer Res* **68**, 10154-10162 (2008).
212. Roberts, C.W. & Orkin, S.H. The SWI/SNF complex--chromatin and cancer. *Nat Rev Cancer* **4**, 133-142 (2004).
213. Reisman, D., Glaros, S. & Thompson, E.A. The SWI/SNF complex and cancer. *Oncogene* **28**, 1653-1668 (2009).
214. Jones, S., *et al.*. Frequent mutations of chromatin remodeling gene ARID1A in ovarian clear cell carcinoma. *Science* **330**, 228-231 (2010).
215. Varela, I., *et al.*. Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma. *Nature* **469**, 539-542 (2011).
216. Versteege, I., *et al.*. Truncating mutations of hSNF5/INI1 in aggressive paediatric cancer. *Nature* **394**, 203-206 (1998).
217. Martino, D.J. & Prescott, S.L. Silent mysteries: epigenetic paradigms could hold the key to conquering the epidemic of allergy and immune disease. *Allergy* **65**, 7-15 (2010).
218. Neuman, A., *et al.*. Maternal smoking in pregnancy and asthma in preschool children: a pooled analysis of eight birth cohorts. *Am J Respir Crit Care Med* **186**, 1037-1043 (2012).
219. Oken, E., Levitan, E.B. & Gillman, M.W. Maternal smoking during pregnancy and child overweight: systematic review and meta-analysis. *Int J Obes (Lond)* **32**, 201-210 (2008).
220. Hemminki, K. & Chen, B. Parental lung cancer as predictor of cancer risks in offspring: clues about multiple routes of harmful influence? *Int J Cancer* **118**, 744-748 (2006).
221. Breton, C.V., *et al.*. Prenatal tobacco smoke exposure affects global and gene-specific DNA methylation. *Am J Respir Crit Care Med* **180**, 462-467 (2009).
222. Murphy, S.K., *et al.*. Gender-specific methylation differences in relation to prenatal exposure to cigarette smoke. *Gene* **494**, 36-43 (2012).

223. Joubert, B.R., *et al.*. 450K epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy. *Environ Health Perspect* **120**, 1425-1431 (2012).
224. Hinz, D., *et al.*. Cord blood Tregs with stable FOXP3 expression are influenced by prenatal environment and associated with atopic dermatitis at the age of one year. *Allergy* **67**, 380-389 (2012).
225. Weisse, K., *et al.*. Maternal and newborn vitamin D status and its impact on food allergy development in the German LINA cohort study. *Allergy* **68**, 220-228 (2013).
226. Adey, A. & Shendure, J. Ultra-low-input, tagmentation-based whole-genome bisulfite sequencing. *Genome Res* **22**, 1139-1143 (2012).
227. Hansen, K.D., *et al.*. Increased methylation variation in epigenetic domains across cancer types. *Nat Genet* **43**, 768-775 (2011).
228. Hansen, K.D., Langmead, B. & Irizarry, R.A. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol* **13**, R83 (2012).
229. Heinz, S., *et al.*. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**, 576-589 (2010).
230. Zhu, J., *et al.*. Genome-wide chromatin state transitions associated with developmental and environmental cues. *Cell* **152**, 642-654 (2013).
231. Slaats, G.G., *et al.*. DNA methylation levels within the CD14 promoter region are lower in placentas of mothers living on a farm. *Allergy* **67**, 895-903 (2012).
232. Sehouli, J., *et al.*. Epigenetic quantification of tumor-infiltrating T-lymphocytes. *Epigenetics* **6**, 236-246 (2011).
233. Pei, L., *et al.*. Genome-wide DNA methylation analysis reveals novel epigenetic changes in chronic lymphocytic leukemia. *Epigenetics* **7**, 567-578 (2012).
234. Pascual, M., *et al.*. Epigenetic changes in B lymphocytes associated with house dust mite allergic asthma. *Epigenetics* **6**, 1131-1137 (2011).
235. Houseman, E.A., *et al.*. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* **13**, 86 (2012).
236. Hodges, E., *et al.*. Directional DNA methylation changes and complex intermediate states accompany lineage specificity in the adult hematopoietic compartment. *Mol Cell* **44**, 17-28 (2011).
237. Somm, E., *et al.*. Prenatal nicotine exposure alters early pancreatic islet and adipose tissue development with consequences on the control of body weight and glucose metabolism later in life. *Endocrinology* **149**, 6289-6299 (2008).
238. Hansen, R.S., *et al.*. Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proc Natl Acad Sci U S A* **107**, 139-144 (2010).
239. Wang, J., Duncan, D., Shi, Z. & Zhang, B. WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. *Nucleic Acids Res* **41**, W77-W83 (2013).
240. Ehrich, M., *et al.*. Cytosine methylation profiling of cancer cell lines. *Proc Natl Acad Sci U S A* **105**, 4844-4849 (2008).
241. Livak, K.J. & Schmittgen, T.D. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* **25**, 402-408 (2001).
242. Herberth, G., *et al.*. Association of neuropeptides with Th1/Th2 balance and allergic sensitization in children. *Clin Exp Allergy* **36**, 1408-1416 (2006).

243. Kulis, M., *et al.*. Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. *Nat Genet* **44**, 1236-1242 (2012).
244. Barker, D.J. The developmental origins of adult disease. *J Am Coll Nutr* **23**, 588S-595S (2004).
245. Wang, R., *et al.*. Down-regulation of the canonical Wnt beta-catenin pathway in the airway epithelium of healthy smokers and smokers with COPD. *PLoS One* **6**, e14793 (2011).
246. Ying, Y. & Tao, Q. Epigenetic disruption of the WNT/beta-catenin signaling pathway in human cancers. *Epigenetics* **4**, 307-312 (2009).
247. Tabit, C.E., *et al.*. Protein kinase C-beta contributes to impaired endothelial insulin signaling in humans with diabetes mellitus. *Circulation* **127**, 86-95 (2013).
248. Bosch, R.R., *et al.*. Inhibition of protein kinase CbetaII increases glucose uptake in 3T3-L1 adipocytes through elevated expression of glucose transporter 1 at the plasma membrane. *Mol Endocrinol* **17**, 1230-1239 (2003).
249. Oken, E., Levitan, E.B. & Gillman, M.W. Maternal smoking during pregnancy and child overweight: systematic review and meta-analysis. *Int J Obes (Lond)* **32**, 201-210 (2008).
250. Ino, T. Maternal smoking during pregnancy and offspring obesity: meta-analysis. *Pediatr Int* **52**, 94-99 (2010).
251. Somm, E., *et al.*. Prenatal nicotine exposure alters early pancreatic islet and adipose tissue development with consequences on the control of body weight and glucose metabolism later in life. *Endocrinology* **149**, 6289-6299 (2008).
252. Bruin, J.E., Kellenberger, L.D., Gerstein, H.C., Morrison, K.M. & Holloway, A.C. Fetal and neonatal nicotine exposure and postnatal glucose homeostasis: identifying critical windows of exposure. *J Endocrinol* **194**, 171-178 (2007).
253. Thiering, E., *et al.*. Prenatal and postnatal tobacco smoke exposure and development of insulin resistance in 10 year old children. *Int J Hyg Environ Health* **214**, 361-368 (2011).
254. Leone, A., Landini, L.J., Biadi, O. & Balbarini, A. Smoking and cardiovascular system: cellular features of the damage. *Curr Pharm Des* **14**, 1771-1777 (2008).
255. Balakumar, P. & Kaur, J. Is nicotine a key player or spectator in the induction and progression of cardiovascular disorders? *Pharmacol Res* **60**, 361-368 (2009).
256. Breitling, L.P., Salzmann, K., Rothenbacher, D., Burwinkel, B. & Brenner, H. Smoking, F2RL3 methylation, and prognosis in stable coronary heart disease. *Eur Heart J* **33**, 2841-2848 (2012).
257. Fuks, F., Hurd, P.J., Deplus, R. & Kouzarides, T. The DNA methyltransferases associate with HP1 and the SUV39H1 histone methyltransferase. *Nucleic Acids Res* **31**, 2305-2312 (2003).
258. Lehnertz, B., *et al.*. Suv39h-mediated histone H3 lysine 9 methylation directs DNA methylation to major satellite repeats at pericentric heterochromatin. *Curr Biol* **13**, 1192-1200 (2003).
259. Smallwood, A., Esteve, P.O., Pradhan, S. & Carey, M. Functional cooperation between HP1 and DNMT1 mediates gene silencing. *Genes Dev* **21**, 1169-1178 (2007).
260. Bergman, Y. & Cedar, H. DNA methylation dynamics in health and disease. *Nat Struct Mol Biol* **20**, 274-281 (2013).
261. Satta, R., *et al.*. Nicotine decreases DNA methyltransferase 1 expression and glutamic acid decarboxylase 67 promoter methylation in GABAergic interneurons. *Proc Natl Acad Sci U S A* **105**, 16356-16361 (2008).
262. Levine, A., *et al.*. Molecular mechanism for a gateway drug: epigenetic changes initiated by

nicotine prime gene expression by cocaine. *Sci Transl Med* **3**, 107r-109r (2011).
263. Hauser, M.T., Aufsatz, W., Jonak, C. & Luschnig, C. Transgenerational epigenetic inheritance in plants. *Biochim Biophys Acta* **1809**, 459-468 (2011).
264. Greer, E.L., *et al.*. Transgenerational epigenetic inheritance of longevity in Caenorhabditis elegans. *Nature* **479**, 365-371 (2011).
265. Padmanabhan, N., *et al.*. Mutation in folate metabolism causes epigenetic instability and transgenerational effects on development. *Cell* **155**, 81-93 (2013).
266. Daxinger, L. & Whitelaw, E. Understanding transgenerational epigenetic inheritance via the gametes in mammals. *Nat Rev Genet* **13**, 153-162 (2012).
267. Irish, J.M., Kotecha, N. & Nolan, G.P. Mapping normal and cancer cell signalling networks: towards single-cell proteomics. *Nat Rev Cancer* **6**, 146-155 (2006).
268. Graf, T. & Stadtfeld, M. Heterogeneity of embryonic and adult stem cells. *Cell Stem Cell* **3**, 480-483 (2008).
269. Takahashi, K., *et al.*. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* **131**, 861-872 (2007).
270. Chan, E.M., *et al.*. Live cell imaging distinguishes bona fide human iPS cells from partially reprogrammed cells. *Nat Biotechnol* **27**, 1033-1037 (2009).
271. Arand, J., *et al.*. In vivo control of CpG and non-CpG DNA methylation by DNA methyltransferases. *PLoS Genet* **8**, e1002750 (2012).
272. Taylor, K.H., *et al.*. Ultradeep bisulfite sequencing analysis of DNA methylation patterns in multiple gene promoters by 454 sequencing. *Cancer Res* **67**, 8511-8518 (2007).
273. Landan, G., *et al.*. Epigenetic polymorphism and the stochastic formation of differentially methylated regions in normal and cancerous tissues. *Nat Genet* **44**, 1207-1214 (2012).
274. Tang, F., *et al.*. RNA-Seq analysis to capture the transcriptome landscape of a single cell. *Nat Protoc* **5**, 516-535 (2010).
275. Bendall, S.C. & Nolan, G.P. From single cells to deep phenotypes in cancer. *Nat Biotechnol* **30**, 639-647 (2012).
276. White, A.K., *et al.*. High-throughput microfluidic single-cell RT-qPCR. *Proc Natl Acad Sci U S A* **108**, 13999-14004 (2011).
277. Wills, Q.F., *et al.*. Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments. *Nat Biotechnol* **31**, 748-752 (2013).
278. Wu, M. & Singh, A.K. Single-cell protein analysis. *Curr Opin Biotechnol* **23**, 83-88 (2012).
279. Dean, F.B., *et al.*. Comprehensive human genome amplification using multiple displacement amplification. *Proc Natl Acad Sci U S A* **99**, 5261-5266 (2002).
280. Lasken, R.S. Single-cell sequencing in its prime. *Nat Biotechnol* **31**, 211-212 (2013).
281. Bundo, M., *et al.*. A systematic evaluation of whole genome amplification of bisulfite-modified DNA. *Clin Epigenetics* **4**, 22 (2012).
282. Zong, C., Lu, S., Chapman, A.R. & Xie, X.S. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science* **338**, 1622-1626 (2012).
283. Navin, N., *et al.*. Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90-94 (2011).
284. Zhang, X., *et al.*. Investigating Evolutionary Perspective of Carcinogenesis with Single-cell Transcriptome Analysis. *Chin J Cancer* (2013).

285. Tang, F., Lao, K. & Surani, M.A. Development and applications of single-cell transcriptome analysis. *Nat Methods* **8**, S6-S11 (2011).
286. Tang, F., *et al.*. RNA-Seq analysis to capture the transcriptome landscape of a single cell. *Nat Protoc* **5**, 516-535 (2010).
287. Lipson, D., *et al.*. Quantification of the yeast transcriptome by single-molecule sequencing. *Nat Biotechnol* **27**, 652-658 (2009).
288. Goren, A., *et al.*. Chromatin profiling by directly sequencing small quantities of immunoprecipitated DNA. *Nat Methods* **7**, 47-49 (2010).
289. Harrington, E.D., Arumugam, M., Raes, J., Bork, P. & Relman, D.A. SmashCell: a software framework for the analysis of single-cell amplified genome sequences. *Bioinformatics* **26**, 2979-2980 (2010).
290. Jensen, L.J., *et al.*. STRING 8--a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res* **37**, D412-D416 (2009).
291. Kantlehner, M., *et al.*. A high-throughput DNA methylation analysis of a single cell. *Nucleic Acids Res* **39**, e44 (2011).
292. Denomme, M.M., Zhang, L. & Mann, M.R. Single oocyte bisulfite mutagenesis. *J Vis Exp* (2012).
293. Hayashi-Takanaka, Y., *et al.*. Tracking epigenetic histone modifications in single cells using Fab-based live endogenous modification labeling. *Nucleic Acids Res* **39**, 6475-6488 (2011).
294. Pop, M. & Salzberg, S.L. Bioinformatics challenges of new sequencing technology. *Trends Genet* **24**, 142-149 (2008).
295. Liu, X., *et al.*. Analysis of cell fate from single-cell gene expression profiles in C. elegans. *Cell* **139**, 623-633 (2009).
296. Angers, B., Castonguay, E. & Massicotte, R. Environmentally induced phenotypes and DNA methylation: how to deal with unpredictable conditions until the next generation and after. *Mol Ecol* **19**, 1283-1295 (2010).
297. Verhoeven, K.J., Jansen, J.J., van Dijk, P.J. & Biere, A. Stress-induced DNA methylation changes and their heritability in asexual dandelions. *New Phytol* **185**, 1108-1118 (2010).
298. Jablonka, E. & Raz, G. Transgenerational epigenetic inheritance: prevalence, mechanisms, and implications for the study of heredity and evolution. *Q Rev Biol* **84**, 131-176 (2009).
299. Johannes, F., *et al.*. Assessing the impact of transgenerational epigenetic variation on complex traits. *PLoS Genet* **5**, e1000530 (2009).
300. Bastow, R., *et al.*. Vernalization requires epigenetic silencing of FLC by histone methylation. *Nature* **427**, 164-167 (2004).
301. He, Y. & Amasino, R.M. Role of chromatin modification in flowering-time control. *Trends Plant Sci* **10**, 30-35 (2005).
302. Allendorf, F.W., Hohenlohe, P.A. & Luikart, G. Genomics and the future of conservation genetics. *Nat Rev Genet* **11**, 697-709 (2010).
303. Richards, E.J. Population epigenetics. *Curr Opin Genet Dev* **18**, 221-226 (2008).
304. Richards, C.L., Bossdorf, O. & Verhoeven, K.J. Understanding natural epigenetic variation. *New Phytol* **187**, 562-564 (2010).
305. Herrera, C.M. & Bazaga, P. Epigenetic differentiation and relationship to adaptive genetic divergence in discrete populations of the violet Viola cazorlensis. *New Phytol* **187**, 867-876 (2010).

306. Herrera, C.M. & Bazaga, P. Untangling individual variation in natural populations: ecological, genetic and epigenetic correlates of long-term inequality in herbivory. *Mol Ecol* **20**, 1675-1688 (2011).
307. Richards, C.L., *et al.*. Plasticity in salt tolerance traits allows for invasion of novel habitat by Japanese knotweed s. l. (Fallopia japonica and F.xbohemica, Polygonaceae). *Am J Bot* **95**, 931-942 (2008).
308. Paun, O., *et al.*. Stable epigenetic effects impact adaptation in allopolyploid orchids (Dactylorhiza: Orchidaceae). *Mol Biol Evol* **27**, 2465-2473 (2010).
309. Cubas, P., Vincent, C. & Coen, E. An epigenetic mutation responsible for natural variation in floral symmetry. *Nature* **401**, 157-161 (1999).
310. Manning, K., *et al.*. A naturally occurring epigenetic mutation in a gene encoding an SBP-box transcription factor inhibits tomato fruit ripening. *Nat Genet* **38**, 948-952 (2006).
311. Morgan, H.D., Sutherland, H.G., Martin, D.I. & Whitelaw, E. Epigenetic inheritance at the agouti locus in the mouse. *Nat Genet* **23**, 314-318 (1999).
312. Rakyan, V.K., *et al.*. Transgenerational inheritance of epigenetic states at the murine Axin(Fu) allele occurs after maternal and paternal transmission. *Proc Natl Acad Sci U S A* **100**, 2538-2543 (2003).
313. Kucharski, R., Maleszka, J., Foret, S. & Maleszka, R. Nutritional control of reproductive status in honeybees via DNA methylation. *Science* **319**, 1827-1830 (2008).