Technische Hochschule Wildau (FH)

Fachbereich Ingenieurswesen / Wirtschaftsingenieurswesen

Studiengang Biosystemtechnik / Bioinformatik

Statistische genomweite Analyse von MeDIP-Seq Daten zur Identifizierung differentiell methylierter Regionen

Bachelorarbeit

zur Erlangung des akademischen Grades

Bachelor of Science

vorgelegt von Jörn Dietrich

geb. am 30.06.1980

Gutachter: 1) Prof. Dr. Heike Pospisil / TH Wildau

2) Prof. Dr. Peter Beyerlein / TH Wildau

Wildau, den 29/05/2010

Danksagung

Ich möchte mich herzlich bei Prof. Dr. Pospisil und Prof. Dr. Beyerlein bedanken, die sich dazu bereit erklärt haben, meine Bachelorarbeit von Seiten der Hochschule aus zu betreuen.

Mein besonderer Dank gilt meinen Betreuern am Max-Planck-Institut für molekulare Genetik, Dr. Ralf Herwig und Lukas Chavez, die stets engagiert und mit wertvollen Ratschlägen diese Arbeit begleitet haben.

Die Probensammlung, Durchführung der MeDIP-Seq Experiment und Diskussionen zur Datenauswertung wurden zusammen mit Prof. Dr. Hans Lehrach, Dr. Dr. Michal-Ruth Schweiger sowie Dr. Christina Grimm durchgeführt. Dafür möchte ich mich bei ihnen herzlich bedanken.

Ich danke meinen Freunden und meiner Familie für die Unterstützung.

Die Arbeit wurde im Rahmen des NGFN-Plus Projektes, Modifiers of intestinal tumor formation' durchgeführt.

Inhaltsverzeichnis

Kurzzusammenfassung

In den letzten Jahren kam es zu einer rasanten Weiterentwicklung im Bereich der DNA Sequenziertechnologie. Durch das Next Generation Sequencing gibt es neue Möglichkeiten für eine beschleunigte und kostengünstige Sequenzierung kompletter Genome. Häufig fehlt jedoch eine anwendbare Experiment spezifische Datenanalysesoftware. Im Rahmen dieser Studie wurde ein neues Werkzeug für die Auswertung von Hochdurchsatzdaten aus methylierter DNA Immunpräzipitaion (MeDIP) mit anschließender Sequenzierung (MeDIP-Seq) entwickelt. Im Speziellen wurden verschiedene statistische Tests für die genomweite Identifizierung unterschiedlich methylierter Regionen eingesetzt. Die entwickelte Methode wurde auf verfügbare MeDIP-Seq Daten aus insgesamt 14 verschiedenen Darmkrebspatienten angewandt. Hierbei konnten genomische Regionen identifiziert werden, die veränderte Methylierungsmuster im Tumorgewebe im Vergleich zu Normalgewebe aufzeigen. Die identifizierten unterschiedlich methylierten genomischen Regionen ermöglichten eine anschließende Einteilung der untersuchten Proben in histopathologisch sinnvolle Gruppe mittels einer Clusteranalyse. Anschließend wurden die identifizierten Regionen auf biologisch interessante Zusammenhänge hin untersucht und annotiert. Im letzten Schritt dieser Studie wurden einige Regionen im Bereich von bekannten Biomarkern ausgewählt und beispielhaft im UCSC Genombrowser visualisiert.

Abstract

During the last years, novel DNA sequencing technologies have been developed. Next Generation Sequencing offers the opportunity to sequence entire genomes fast and cost-efficient. The biggest challenge for next generation sequencing approaches is the development of experiment specific data-analysis software. Within this study a new tool for the analysis of sequencing data derived from high-throughput Methyl-DNA Immunoprecipitation (MeDIP-Seq) experiments has been developed. In particular various statistical tests were applied in order to identify differentially methylated regions within the genome. The newly developed methods were used to analyse MeDIP-Seq data from 14 colorectal cancer patients and differentially methylated genomic regions were identified. These regions showed DNA methylation differences between

tumor and normal tissue. Based on the identified differentially methylated regions, a clustering approach was performed. The results show that the several examined samples can be classified into separated groups according to their histological origin. Subsequently, the identified regions were analyzed for biologically relevant coherency and were annotated by known biological functions. Finally, for selected regions, known biomarkers were visualized in the UCSC Genome Browser.

Schlüsselwörter: *MeDIP-Seq, DNA Methylierung, Darmkrebs*
Keywords: *MeDIP-Seq, DNA methylation, colorectal cancer*

1. Einleitung

1.1 Epigenetik

Der Begriff „Epigenetik“ definiert alle mitotisch und meiotisch vererbbaren Veränderungen der Genexpression, die nicht in der DNA selbst kodiert sind. Bislang sind drei solcher epigenetischer Systeme bekannt: RNA-assoziierte Genregulation, Histonmodifikation und DNA-Methylierung (einen Überblick liefert Egger et al., 2004).

1.2 DNA-Methylierung

Die kovalente Bindung einer Methylgruppe (CH_3) an das fünfte Kohlenstoffatom von Cytosinbasen bezeichnet man bei Säugetieren als DNA-Methylierung. Diese biochemische Modifikation kann durch mehrere Methyltransferasen (DNMTs) katalysiert (Robertson et al., 1999) werden und findet in der Regel an Cytosinen statt, die sich benachbart zu 5' Guanosin befinden (CpG Kontext). Es liegen etwa 2 bis 5 % der Cytosine im menschlichen Genom als 5-Methylcytosin vor (Paz et al., 2002) und etwa 70 % der CG Dinukleotide (CpGs) enthalten ein methyliertes Cytosin (Antequera, F. and A. Bird, 1993). CpGs sind jedoch nur mit 20 % der erwarteten statistischen Häufigkeit im menschlichen Genom vertreten. Regionen im Genom, die einen erhöhten CG-Gehalt von über 60 % aufweisen, werden als CpG-Inseln bezeichnet. Dabei handelt es sich um kurze genomische Regionen (0,5 – 2 kb), die zumeist unmethyliert vorliegen und während der Evolution stark konserviert wurden. Die Konservierung von CpG-Inseln und die Unterrepräsentation von CpGs im Genom kann dadurch erklärt werden, dass 5-Methylcytosin durch spontane Desaminierung zur Base Thymin wird, welche nicht als DNA-fremde Base erkannt wird (**Abbildung 1**). Uracil entsteht jedoch bei der spontanen Desaminierung von unmethyliertem Cytosin und wird von dem zellulären Reparatursystemen als DNA-fremde Base erkannt und repariert.

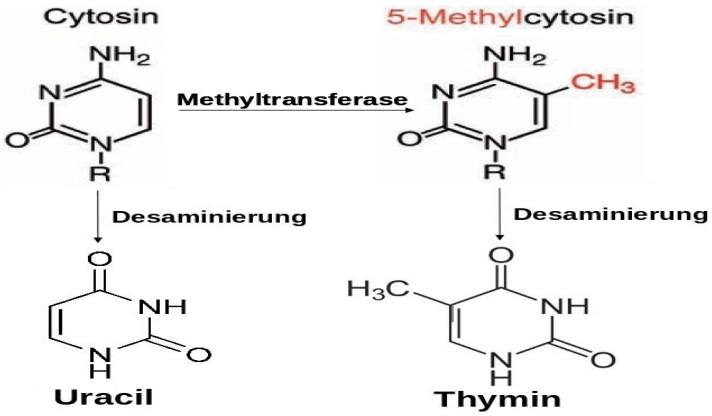

Abbildung 1 : DNA-Methylierung und spontane Desaminierung. DNA Methyltransferasen katalysieren die kovalente Bindung von Methylgruppen (rot) an der 5' Position des Cytosin. Durch spontane Desaminierung wird Cytosin zu Uracil umgewandelt. Aus 5-Methylcytosin entsteht Thymin.

DNA-Methylierung spielt in verschiedenen molekularbiologischen Prozessen eine wichtige Rolle. Hierzu gehören u.a. die Inaktivierung parasitärer DNA, wie z.B. Transposons und die Abwehr von Viren (Lorl et al., 1999), Geninaktivierung im Rahmen von Zelldifferenzierung und Entwicklung (Dean et al., 2005), Imprinting (Li et al., 1993), X-Chromosom-Inaktivierung (Chang et al., 2006) und auch die Sicherung der strukturellen Integrität von Chromosomen (Xu et al., 1999, Robertson, 2005) . Studien haben außerdem gezeigt, dass veränderte DNA-Methylierungsmuster auch im Zusammenhang mit verschiedenen humanen Krankheiten stehen, wie z.B. dem ICF-Syndrom, Rett-Syndrom, Fragile-X-Syndrom (Robertson and Wolffle, 2000) und vor allem mit verschiedenen Arten der Krebsentwicklung (Michal et al., 2006).

1.4 DNA-Methylierung im Zusammenhang mit Krebs

Früher wurde angenommen, dass Krebs auf rein genetischen Veränderungen, wie z.B. Mutationen in Tumorsuppressorgenen oder Onkogenen, sowie chromosomalen Anomalien beruht (Aaltonen et al., 1993). Nach heutigem Kenntnisstand spielen jedoch epigenetische Modifikationen, wie die DNA-Methylierung ebenso eine wichtige Rolle bei der Krebsentstehung. So weisen Tumorzellen eine chromosomale Instabilität auf, die auf einer genomweiten Abnahme der Methylierung (Hypomethylierung) beruht (Rodriguez et al., 2006). Aber auch die Zunahme der Methylierung (Hypermethylierung) konnte in

lokalen Bereichen bei Tumorzellen beobachtet werden, speziell in den Promotorregionen bekannter Gene (Hermann et al., 2003). Vor allem Tumorsuppressorgene die im Tumor inaktiv sind, weisen eine abnormale *de novo* Methylierung von CpG Inseln auf (Jones et al., 2003).

1.5. Ziel der Studie

Immer mehr Labore und Pharmaunternehmen beschäftigen sich mit der Entwicklung epigenetischer Medikamente. Um eine zielgerichtete Wirkung zu gewährleisten ist es notwendig, das Methylierungsmuster genauer zu verstehen. Aber auch für die Entdeckung neuer prädiktiver prognostischer und diagnostischer Krebsmarker ist es notwendig, eine geeignete Methode zur Bestimmung des genomweiten Methylierungsmusters zu entwickeln, welche noch dazu präzise und kostengünstig ist. Im Rahmen dieser Studie soll für solch eine neue Methode (MeDIP-Seq, siehe Kapitel 2) zur Identifizierung Genom weit differentiell methylierter Regionen (DMRs) ein bioinformatisches Werkzeug zur Auswertung der experimentellen Daten konzipiert und implementiert und anhand von verfügbaren Tumorproben getestet werden.

2. Daten und Algorithmen

2.1 Immunpräzipitation methylierter DNA

Eine Möglichkeit der genomweiten Methylierungsanalyse ist die Immunpräzipitation methylierter DNA (MeDIP). Bei dieser Methode werden methylierte DNA-Abschnitte zunächst anhand ihrer Affinität zu Methylcytosin-bindenden Proteinen oder Antikörpern angereichert. Dazu wird die genomische DNA extrahiert und mittels Ultraschall in 200 – 1000 Basenpaare lange Fragmente zerteilt. Um eine möglichst gute Bindung der Antikörper zu gewährleisten, wird die DNA denaturiert und anschließend ein Teil mittels eines Anti-5-Methylcytosin-Antikörpers immunpräzipitiert (**Abbildung 2**). Dies ermöglicht es, je nach Grad der Methylierung, DNA-Fragmente bis zu 100fach anzureichern. Der andere Teil wird nicht immunpräzipitiert und dient als Input-Probe. Die anschließende Detektion bzw. Bestimmung der DNA-Abschnitte kann mittels einer speziellen Art von Microarrays, sogenannter Tilingarrays (MeDIP-Chip), oder mittels Sequenzierung (MeDIP-Seq) vorgenommen werden.

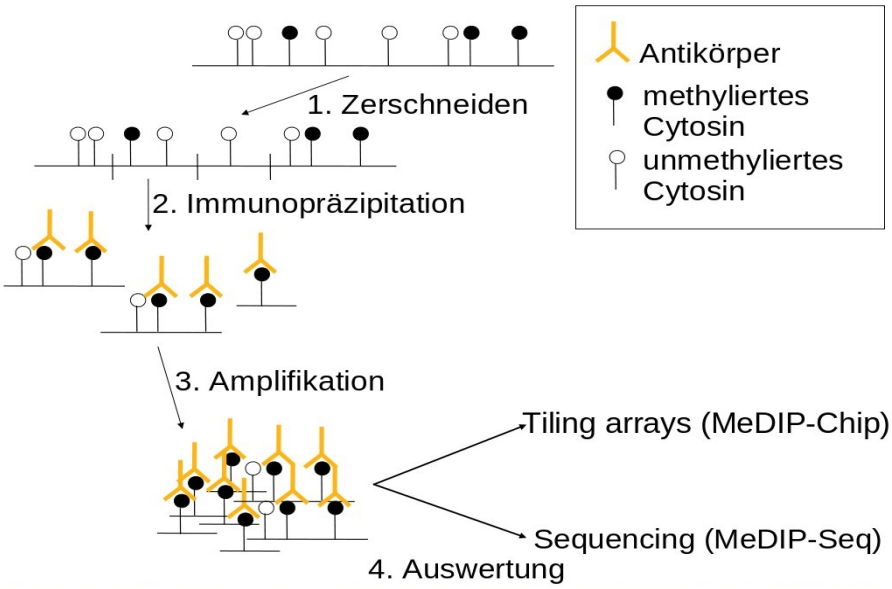

Abbildung 2: Schematische Darstellung des MeDIP Experimentes. Zunächst wird die genomische DNA fragmentiert, z.B. mittels Ultraschal. Anschließend mit einem für Methylcytosin spezifischen Antikörper gefällt und vervielfältigt. Die Identifizierung der DNA, kann anhand von Tilingarrays oder Sequenzierung vorgenommen werden.

2.2 Funktion und Anwendung von DNA-Micro- und Tilingarrays

DNA-Tilingarrays basieren auf der gleichen Microarray-Technologie, die ursprünglich für die Analyse der Regulation und Funktion von Genen entwickelt wurde. Sie ermöglicht die parallele Untersuchung von mehreren tausend Genen. Das Prinzip dieser Methode basiert auf der Hybridisierung markierter Zielmoleküle, wie z.B. cDNA oder cRNA, mit sogenannten Fängermolekülen, die auf einem Chip fixiert sind. Man unterscheidet hauptsächlich zwei verschiedene Arten von DNA-Microarrays. Zum einen cDNA Arrays, bei denen längere Nukleotidsequenzen direkt auf einen Träger, wie z.B. einer Glasplatte, aufgetragen werden. Zum anderen Oligonukleotid-Microarrays, bei denen sehr kurze Nukleotidsequenzen direkt auf dem Chip synthetisiert werden. Die typische Herangehensweise für den Vergleich der Genexpression zweier Proben, wie z.B. normales Gewebe mit Tumorgewebe, mittels Microarrays beginnt mit der Extraktion der jeweiligen mRNA. Anschließend wird diese mit reverser Transkriptase in cDNA umgewandelt und mit Fluoreszenzstoffen oder radioaktiv markiert. Nach dem Auftragen auf den Chip lagern sich die Proben an die komplementären Sequenzen an. Ungebundene Proben werden daraufhin abgewaschen. Im letzten Schritt wird der Array mit einem Laser gescannt und das Bild zur weiteren Computeranalyse gespeichert.

Nach dem gleichen Prinzip funktionieren auch Tilingarrays. Sie unterscheiden sich jedoch in der Anwendung und der Art der verwendeten Sonden. So werden Tilingarrays für die Analyse des kompletten Genoms verwendet, indem als Sonde kurze Oligonukleotide (~ 25bp) dienen, die das Genom komplementär und partiell überlappend abdecken.
Ein weit verbreitetes Beispiel für die Anwendung eines Tilingarrays ist die Chromatin Immunpräzipitation mittels Microarray Technologie (ChIP-on-chip). Dieses Verfahren wird zur Lokalisation von DNA-Bindungsstellen eines Proteins im Genom verwendet (**Abbildung 3A**). Hierzu wird ein Protein-bindendes Chromatin mittels Ultraschall zerkleinert und anschließend mit einem für das Protein spezifischen Antikörper immunpräzipitiert. Nach dem reversen Cross-linking und der Reinigung der DNA wird diese fluoreszenzmarkiert, wobei in einem Zweifarbenexperiment die Vergleichsprobe eine andere Markierung

erhält (z.B Cyanin 3) als die zu untersuchende Probe (z.B. Cyanin 5). Beide markierten Proben werden auf dem Microarray aufgetragen. Dieses Prinzip liegt auch der MeDIP-Chip Technologie zu Grunde (**Abbildung 3B**).

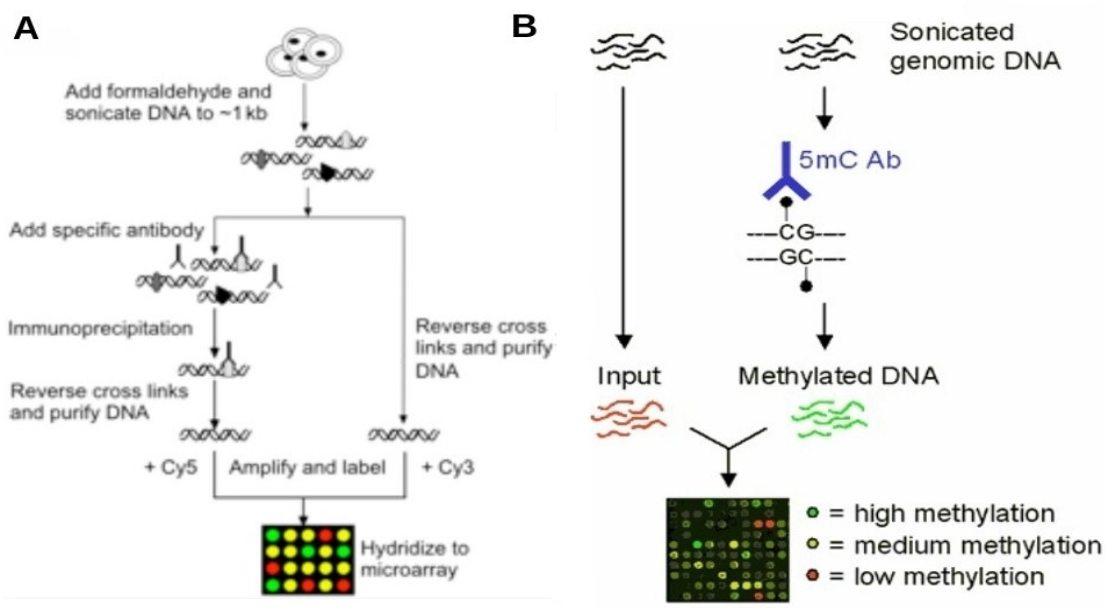

Abbildung 3: Anwendung von Microarray-Technologie: A) Schematischer Ablauf von ChIP-on-chip (Buck und Lieb, 2004). Die Protein-DNA Interaktion wird mit Formaldehyd fixiert. Anschließend wird die DNA mittels Ultraschall in 300 – 700 bp große Fragmente zerkleinert. Ein Teil der Probe wird mit spezifischen Antikörpern immunpräzipitiert. Bei einem anderen Teil wird keine Immunpräzipitation vorgenommen (Input-Probe). Es folgt die Aufhebung der Protein-DNA Fixierung (reverses Cross-linking) und die Aufreinigung. Die beiden Proben werden mit unterschiedlichen Fluoreszenzfarbstoffen markiert und auf einem Tiling-Array aufgetragen.
B) Schematischer Ablauf von MeDIP-Chip (www.epigenome-noe.net). Im ersten Schritt wird die DNA mit Ultraschall fragmentiert. Anschließend wird ein Teil mit 5-Methylcytosin spezifischen Antikörpern immunpräzipitiert. Als Input dient eine Probe ohne Immunpräzipitation. Es folgt die Markierung der beiden Proben mit unterschiedlichen Fluoreszenzfarbstoffen und die Hybridisierung auf einem Tilingarray.

2.3 Next Generation Sequencing

Das Verfahren zur Bestimmung der Reihenfolge der Basen (Adenin, Guanin, Cytosin und Thymin) in einem DNA-Molekül wird als DNA-Sequenzierung bezeichnet. Frederick Sanger entwickelte 1975 in Cambridge die Didesoxymethode, auch Kettenabruch-Synthese (Sanger et al., 1977) genannt. Prinzip dieses Verfahrens ist die Durchführung der Polymerisationsreaktion in vier getrennten Ansätzen. In jedem Reaktionsraum liegt der gleiche zu sequenzierende DNA-Einzelstrang vor und wird mit identischen markierten (Fluoreszenz oder radioaktiv markiert) Primern versehen. Primer sind kurze Oligonukleotide, die als Startpunkt für die DNA-Polymerase zur Verlängerung des komplementären DNA-Strangs dienen. Zusätzlich sind in allen Ansätzen die vier Desoxyribonukleotid-Triphosphate (dATP, dCTP, dTTP und dGTP), die zur Verlängerung der Sequenz nötig sind, enthalten. Aber jeweils nur ein bestimmtes ddNTP (Didesoxyribonukleotid). Wird ein solches modifiziertes Nukleotid eingebaut kommt es auf Grund der fehlenden 3'-Hydroxygruppe zum Kettenabbruch. Die Abbruchprodukte werden nun mittels Polyacrylamid-Gelelektrophorese der Länge nach aufgetrennt. Durch den Vergleich der vier Ansätze lässt sich die der einzelsträngigen DNA entsprechend komplementäre Sequenz ablesen. Heutige Varianten der Sanger-Methode verwenden unterschiedlich fluoreszenzmarkierte Nukleotide (Prober et al., 1987), so dass es möglich ist, die Reaktionen in einem Ansatz durchzuführen und anschließend mittels Kapillarelektrophorese aufzutrennen und zu detektieren.
Die neue Generation von Sequenziermethoden (Next Generation Sequencing), auch Hochdurchsatz-Sequenzierung genannt, beruht im Prinzip auf zwei aufeinander folgenden Ansätzen, Sequenzierung durch Hybridisierung und Ligation (Brenner et al., 2000; Shendure et al., 2005) sowie Sequenzierung durch Synthese. Sequenzierung durch Synthese kann man wiederum unterteilen in Pyrosequenzierung (Ronaghi et al., 1998) und der Methode Solexa (Bennett et al., 2005).

2.3.1 Pyrosequenzierung

Bei der Pyrosequenzierung werden einzelsträngige DNA-Template mit Primern hybridisiert. Pro Reaktionszyklus wird nur ein Nukleotid-Triphosphat hinzugegeben. Ist dieses Nukleotid komplementär zur Base der DNA wird es als Monophosphat eingebaut, wodurch Diphosphate (Pyrophosphate, PPi) freigesetzt werden. Diese werden zur Umsetzung von Adenosin-5'-phosphosulfat (APS) zu Adenosintriphosphat (ATP) durch die Sulfurylase genutzt (Ronaghi et al., 1998). Die Luciferase bildet im Anschluss aus dem ATP und Luciferin Oxiluciferin. Bei dieser Reaktion entsteht durch Chemiluminsezenz ein Lichtsignal, welches detektiert wird (**Abbildung 4**). Dieses Lichtsignal ist proportional der Menge eingebauter Nukleotid-Triphosphate. Nicht verwendete Nukleotide und ATP werden durch Apyrase am Ende jeder Reaktion abgebaut, so dass ein weiterer Zyklus folgen kann.

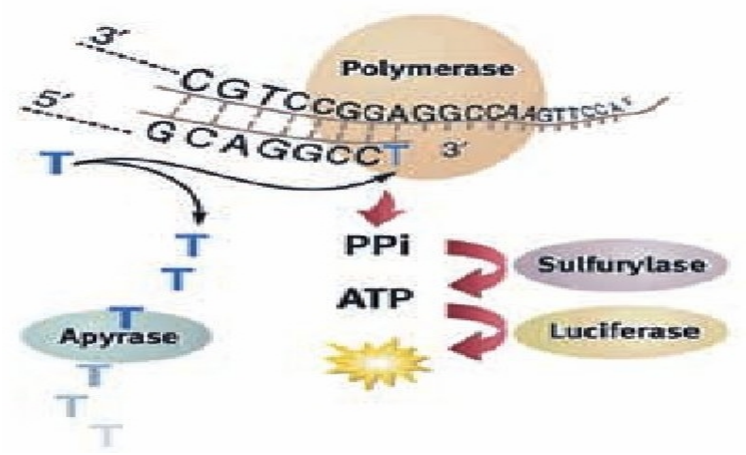

Abbildung 4: Enzymkaskade bei der Pyrosequenzierung (England und Petterson, 2005). Bei dem Einbau des komplementären Nukleotids durch die Polymerase werden Diphosphate freigesetzt. Anschließend setzt das Enzym Sulfurylase das Diphosphat, mittels Adenosin-5'-phosposulfat zu Adenosintriphosphat (ATP) um. Das ATP wird von der Luciferase für die Umsetzung von Luciferin verwendet. Bei dieser Reaktion entsteht ein messbares Lichtsignal. Das Enzym Apyrase baut nicht verwendete Nukleotide sowie Diphosphate ab.

2.3.2 Solexa

Bei der Solexa Methode werden an die fragmentierte DNA auf beiden Seiten unterschiedliche Adapter ligiert. Diese werden auf einer Oberfläche, der sogenannten *Flow cell,* immobilisiert. Die DNA-Fragmente hybridisieren mit den komplementären Primern auf der Oberfläche und bilden dabei Brücken aus. An diesen Brücken findet nun DNA-Synthese statt (Brückenamplifikation, **Abbildung 5.1a**). Anschließend folgen mehrere Zyklen aus Denaturierung, Renaturierung und Neusynthese wodurch auf einem sehr kleinen Gebiet eine hohe Dichte von gleichen DNA-Fragmenten erzeugt wird, den sogenannten klonalen Clustern. Nun folgt die Sequenzierung durch Synthese. Dazu werden schrittweise dNTPs, die je mit einem unterschiedlichen Farbstoff markiert sind, zugegeben und durch die Polymerase eingebaut. Um zu ermitteln welches dNTP eingebaut wurde, wird nach jedem Zyklus ein hochauflösendes Bild gemacht. Der Farbstoff wird eliminiert und es folgt ein neuer Zyklus.

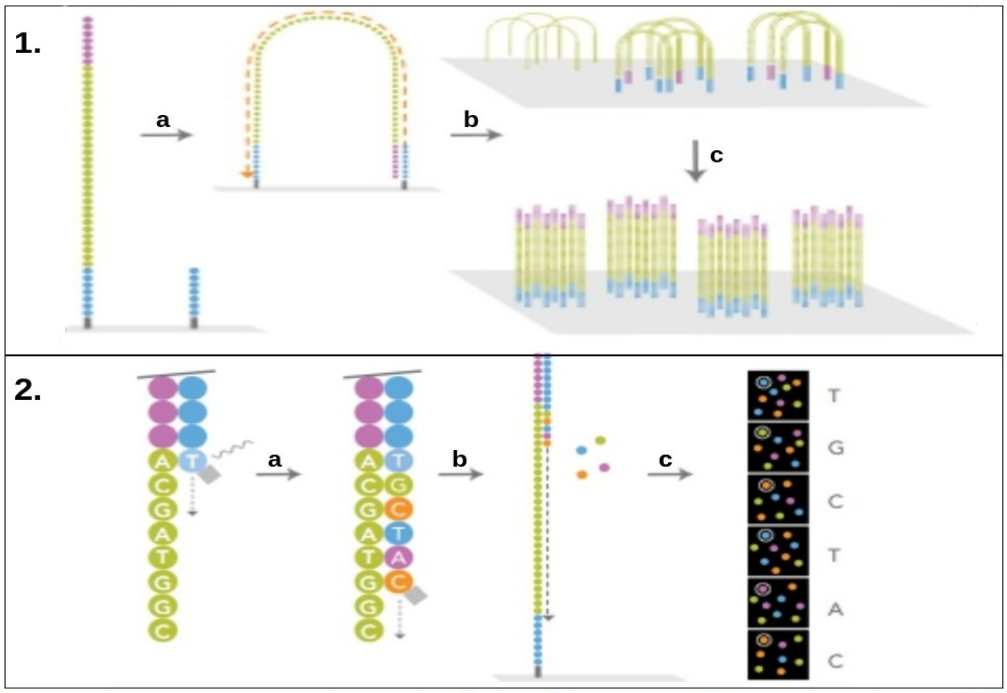

Abbildung 5: Illumina* Sequenzierung (www.Illumina.com*): 1. Brückenamplifikation Die DNA-Fragmente werden mit Adaptern an beiden Enden ligiert und über Linkermoleküle auf einer Glasoberfläche (Flow Cell) immobilisiert **(a)**. Es kommt zu einer Brückenbildung durch die Bindung von komplemetären Primern auf der Oberfläche. Anschließend wird entlang dieser Brücken die DNA synthetisiert **(b)**. Durch wiederholte Denaturierung, Renaturierung und Neusynthese entstehen Cluster identischer DNA-Fragmente **(c)**. ***2. Sequenzierung*** Die durch

die Brückenamplifikation vervielfältigten DNA-Fragmente werden mittels Sequenzierung durch Synthese bestimmt. Dazu wird ein entsprechender Primer sowie mit vier verschiedenen Fluoresenzfarbstoffen markierte dNTPs der Flow Cell hinzugegeben ***(a).*** Es erfolgt nacheinander der Einbau der komplementären Nukleotide **(b).** Nach jedem einzelnen Einbau wird eine Momentaufnahme der entsprechenden Cluster gemacht. Aus dieser Bilderfolge lassen sich die zu den DNA-Fragmenten komplementären Sequenzen ablesen **(c)**.

2.4 MEDIPS-Package

Bei der methylierten DNA Immunpräzipitation mit anschließender Sequenzierung fallen Millionen kurzer Sequenzen (36 bp) an. Diese Menge an Daten erfordert eine entsprechende Software zur weiteren Prozessierung und Auswertung. Hierzu wurde die MeDIP-Seq Datenanalyse-Software „MEDIPS“ in R angewandt (Chavez et al, 2010). Ausgangsdaten für dieses Programm sind die genomischen Koordinaten bestehend aus Chromosom, Start, Stop sowie die Strang-Information wie sie nach der Abbildung der erzeugten kurzen Sequenzen auf das Referenzgenom erhalten wurden.

2.4.1 Genomvektor

Im ersten Schritt der Datenprozessierung wird ein sogenannter Genomvektor erzeugt. Hierzu werden die kurzen Sequenzen um die Länge der ursprünglichen DNA-Fragmente (0.2-1kb) erweitert und anschließend das komplette Genom in 50 Basenpaar große Fragmente (Bins) unterteilt, denen die Anzahl der ihnen zufallenden überlappenden Sequenzen zugeordnet werden.
Die Länge (l) des Genomvektors entspricht also:

$$l = \sum_{i=1}^{N} floor\left(\frac{X_i}{b}\right) \qquad (1)$$

Wobei N die Anzahl der Chromosomen ist, X_i die Länge des i-ten Chromosoms und b die Größe der Bins. Die Funktion *floor* gibt hierbei immer die nächstliegende Ganzzahl die kleiner oder gleich einer gegebenen Zahl ist zurück. Dieser Genomvektor stellt die Rohsignale des MeDIP-Seq Experiments dar (**Abbildung 6**).

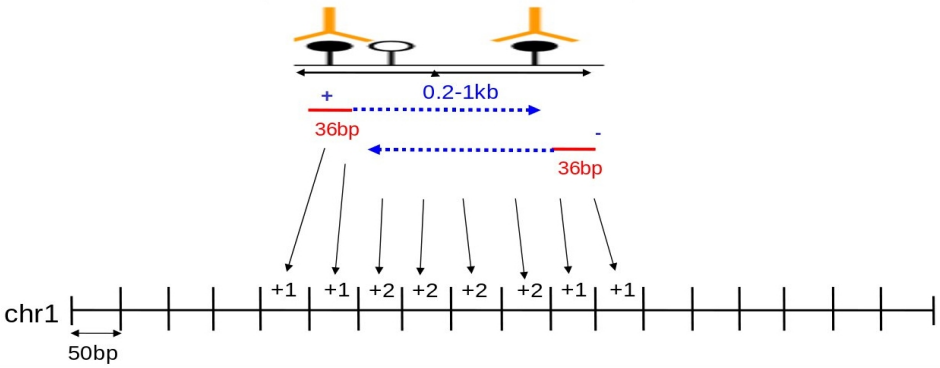

Abbildung 6: Erzeugung des Genomvektor. Die durch die Sequenzierung gewonnenen und dem Referenzgenom zugeordneten kurzen Sequenzen (rote Linien) werden entsprechend ihrer genomischen Koordinaten unter Einbeziehung der Stranginformation (Plus- bzw. Minusstrang) verlängert (gepunktete blaue Linie). Anschließend wird, entsprechend einer definierten Auflösung (50 bp Bins), die Abdeckung über jede Binposition der verlängerten Sequenzen gezählt.

2.4.2 Normalisierung

Es hat sich gezeigt, dass bei der methylierten DNA-Immunpräzipitation nicht ausschließlich methylierte DNA-Fragmente angereichert werden. Es kann auch zu geringfügigen unspezifischen Bindungen mit unmethylierten Cytosinen kommen (Pelizzola et al., 2008). Dabei spielt auch die lokale Dichte der CpGs eine Rolle. Dazu stellt das MEDIPS Paket verschiedene Möglichkeiten für die Korrektur der Rohsignale zur Verfügung, die hier aber nicht weiter erläutert werden sollen. Für die weitere Auswertung der MeDIP-Seq Daten wird auf die durch das Normalisierungsmodul generierten r*eads per million* (RPM) zurück gegriffen. Dies ermöglicht den direkten Vergleich unterschiedlicher biologischer Proben, die durch unterschiedliche Gesamtanzahlen erzeugter kurzer Sequenzen repräsentiert werden. Für jedes Bin berechnet sich der entsprechende rpm-Wert wie folgt. Sei n die Gesamtanzahl an short Reads und $x_{bin(i)}$ das Rohsignal des Bins an der Stelle i, wobei i=1,...,m und m sei gleich der Gesamtanzahl der genomischen Bins, so ist der RPM-Wert definiert als:

$$RPM_{bin_i} = \frac{x_{bin_i} * 10^6}{n} \qquad (2)$$

2.5 Statistische Auswertung zur Identifizierung differentiell methylierter Regionen

Statistische Tests haben sich bei der Analyse von Microarrays zur Bestimmung differentiell exprimierter Gene als sehr zuverlässig und sensitiv erwiesen (Herwig et al., 2007). Sie werden jedoch auch für die Analyse differentiell methylierter Regionen genutzt. Mit Hilfe statistischer Tests werden bestimmte Hypothesen anhand der Stichproben ($X=X_1,...,X_N$ und $Y=Y_1,...,Y_M$) mit einer gewissen Fehlerwahrscheinlichkeit bestätigt oder verworfen. Dazu wird als erstes eine Nullhypothese (H_0) formalisiert. Wird diese widerlegt, so gilt die Alternativhypothese (H_1).

$H_0: \mu_X - \mu_Y = 0$, die Messreihen sind im Mittel gleich.

$H_1: \mu_X - \mu_Y \neq 0$, die Messreihen haben verschiedene Mittelwerte.

Im Rahmen dieser Studie wurden für die Identifizierung differentiell methylierter Regionen der t-Test mit gleichen Varianzen sowie der Wilcoxon-Rangsummentest (Mann-Whitney-U-Test) verwendet. Bei diesen statistischen Tests können zwei Arten von Fehlern auftreten. Ein Fehler 1. Art (falsch positiv) besagt, dass die Region fälschlicherweise als differentiell methyliert deklariert wurde. Der Fehler 2. Art (falsch negativ) deklariert eine Region als nicht differentiell methyliert, obwohl ein Unterschied vorliegt. Die MeDIP-Seq Studie umfasst Millionen von genomischen Regionen, die gleichzeitig getestet werden (Multiples Testen). Dabei steigt die Wahrscheinlichkeit der falschen Schlussfolgerungen mit zunehmender Anzahl an Tests. Bei 1000 Tests werden z.B. 50 falsch positive Schlussfolgerungen erwartet, die bei Fehler 1. Art und einem Signifikanzniveau von 0.05 als signifikant deklariert sind. Daher müssen die individuellen p-Werte für multiples Testen korrigiert werden. Hierzu werden die p-Werte nach der Methode von Benjamini und Hochberg adjustiert (Benjamini und Hochberg, 1995).

2.5.1 t-Test

Zur Bewertung der Signifikanz der differenziellen Methylierung wird ein p-Wert mittels einer mathematischen Funktion berechnet. Diese mathematische Funktion oder auch Teststatistik genannt hat beim t-Test die Form (Herwig et al., 2007):

$$T\left(X_1,\ldots,X_N,Y_1,\ldots,Y_M\right)=\frac{\bar{X}-\bar{Y}}{\sqrt{\frac{(N-1)*S_X^2+(M-1)*S_Y^2}{N+M-2}}}\sqrt{\frac{N*M}{n+M}} \qquad (3)$$

wobei S_X^2 und S_Y^2 die empirischen Varianzen der Kontrollgruppe (Normalgewebe) und der behandelten Gruppe (Tumorgewebe) sind, d.h.

$$S_X^2=\frac{1}{N-1}\sum_{i=1}^{N}\left(X_i-\bar{X}\right)^2 \qquad (4)$$

bzw.

$$S_Y^2=\frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i-\bar{Y}\right)^2 \qquad (5)$$

$$\text{und } \bar{X}=\frac{1}{N}\sum_{i=1}^{N}X_i \quad , \qquad (6)$$

$$\text{und } \bar{Y}=\frac{1}{N}\sum_{i=1}^{N}Y_i \qquad (7)$$

die entsprechenden Mittelwerte.

Diese Teststatistik unterliegt der Bedingung, dass die beiden Gruppen normalverteilt sind. Man spricht von einer t-Verteilung mit M+N-2 Freiheitsgraden. Es kann zu jedem experimentellen Ergebnis ein P-Wert als Maß für die Signifikanz der Abweichung der Daten von der Nullhypothese, berechnet werden. Ein kleiner P-Wert besagt, dass die entsprechende Region unterschiedlich methyliert ist.

2.5.2 Wilcoxon-Rangsummentest

T-Tests sind parametrische Verfahren, d.h. sie setzen voraus, dass die Daten einer parametrisierbaren Wahrscheinlichkeitsverteilung folgen. Im Falle des t-Test einer Normalverteilung. Der Wilcoxon-Rangsummentest ist dahingegen ein nicht-parametrisches Verfahren, der dementsprechend eine wesentlich schwächere Verteilungsanahme benötigt (Herwig et al., 2007, Falk et al., 2007). Der Wilcoxon-Rangsummentest geht von der Annahme aus, dass es sich um unabhängige Stichproben $X_1,...,X_n$ von X und $Y_1,...,Y_m$ von Y handelt, die jedoch stetige Verteilungsfunktionen F_1 und F_2 haben, welche sich um eine Verschiebung δ unterscheiden:

$$F_1(x) = F_2(x-\delta)$$

Dementsprechend lauten die Hypothesen des Tests

$H_0: \delta = 0$, die Verteilungen der Messreihen weichen nicht von einander ab.

$H_1: \delta \neq 0$, die Verteilungen der Messreihen weichen von einander ab.

Der Test beinhaltet die Kalkulation einer Mann-Whitney-U-Statistik U, dessen Verteilung unter der Nullhypothese bekannt ist. Diese kann für kleine Stichproben berechnet werden. Bei großen Stichproben wird die Verteilung durch eine Normalverteilung approximiert. Bei jedem U Test werden die Stichproben X und Y kombiniert, der Größe nach sortiert und Ränge vergeben. Sind gleiche Werte vorhanden, so werden deren Ränge durch deren Rangmittelwert ersetzt. Anschließend werden die jeweiligen Rangsummen gebildet

$$R_1 = \sum_{i=1}^{n} R(X_i) \qquad (8)$$

und

$$R_2 = \sum_{i=1}^{m} R(Y_i) \; . \qquad (9)$$

Wobei $R(X_i)$ und $R(Y_i)$ der Rang der i-ten gepoolten und geordneten Stichprobe ist.

Dabei muss gelten:

$$R_1 + R_2 = (n+m)\frac{(n+m+1)}{2} \quad . \quad (10)$$

Nun werden für die jeweiligen Gruppen die entsprechenden Prüfgrößen U berechnet.

$$U_1 = n*m + \frac{n^2+n}{2} - R_1 \qquad (11)$$

$$U_2 = n*m + \frac{m^2+m}{2} - R_2 \qquad (12)$$

Hierbei muss gelten:

$$U_1 + U_2 = n*m \qquad (13)$$

Anschließend wird das minimale U bestimmt.

$$U = min(U_{1,} U_2) \qquad (14)$$

Wie bereits oben beschrieben, kann bei größeren Probenumfang die U Verteilung durch eine Normalverteilung approximiert werden. In diesem Fall, ist der standardisierte Wert

$$Z = \frac{U - \mu_U}{\sigma_U} \approx N(\mu = 0; \sigma^2 = 1) \qquad (15)$$

eine Standardnormalvariable, bei der

$$\mu_U = \frac{n*m}{2} \qquad (16)$$

der Mittelwert und

$$\sigma_U = \sqrt{\frac{(n*m)(n+m+1)}{12}} \qquad (17)$$

die Standardabweichung von U ist. Die Level der Signifikanz resultieren aus den Level der Signifikanz der approximierten Standardnormalverteilung $N(\mu = 0; \sigma^2 = 1)$. Daher kann die Signifikanz für das berechnete Z aus der Standardnormalverteilung gefolgert werden. Der dazugehörige P-Wert gibt nun Aufschluss, ob eine der zwei Messreihen signifikant größer ist als die andere.

2.5.3 False Discovery Rate

Wie schon in Kapitel 2.2 beschrieben, treten bei statistischen Tests zwei Arten von Fehlern auf (Tabelle 2). Wird einem durch den Test eine differenzielle Methylierung suggeriert (positiv), diese Region aber in Wirklichkeit nicht differenziell methyliert ist, so spricht man von falsch-positiv oder dem Fehler 1. Art. Suggeriert die Teststatistik einem keine differenzielle Methylierung der Region (negativ), obwohl diese differenziell methyliert ist, so bezeichnet man das als falsch-negativ bzw. Fehler 2. Art (Herwig et al., 2007).
Bei der Durchführung von m unabhängigen Hypothesentests entspricht m_0 der Anzahl der wahren Nullhypothesen und R der Anzahl der verworfenen Hypothesen. U ist die Anzahl der wahr-negativen, V der falsch-positiven, S der wahr-positiven und T die Anzahl der falsch-negativen. Die Zufallsvariable R stellt dabei die einzige beobachtete Variable dar. U,V,S und T sind jeweils unbeobachtete Zufallsvariablen.

	Nicht verworfene Hyphothesen	Verworfene Hypothesen	Total
Wahre Null-hyphothese	U	V Fehler 1.Art	m_0
Falsche Null-hypothese	T Fehler 2.Art	S	$m-m_0$
	m - R	R	m

Tabelle 1: Übersicht der Zufallsvariablen und Fehler im Bezug zu m Hypothesentests.

Der Anteil der fälschlich verworfenen Nullhypothesen wird durch die Zufallsvariable Q dargestellt

$$Q = \frac{V}{(V+S)} = \frac{V}{R} \quad . \qquad (18)$$

Dabei wird Q so definiert, dass Q = 0 ist, sofern V + S = 0 und somit keine Nullhypothesen verworfen wurden. Die False Discovery Rate Q_e ist nun definiert als der Erwartungswert von Q. Sie entspricht also dem erwarteten Anteil der fälschlicherweise verworfenen Hypothese (Benjamini und Hochberg, 1995)

$$Q_e = E(Q) = E\left\{\frac{V}{(V+S)}\right\} = E\left(\frac{V}{R}\right) \quad . \qquad (19)$$

2.5.4 Benjamini-Hochberg-Prozedur

Es seien m Hypothesen getestet $\left(H_1^0, H_2^0, \ldots, H_m^0\right)$ mit m_0 wahren Nullhypothesen und den dazugehörigen p-Werten $\left(p_1, p_2, \ldots, p_m\right)$. Nun werden die p-Werte geordnet $\left(p_{(1)} \leq p_{(2)} \leq \ldots p_{(m)}\right)$. Entsprechend verfährt man mit den Hypothesen. Anschließend werden die geordneten P-Werte P_i für die geforderte FDR q mit dem kritischen Wert $q * \frac{i}{m}$ verglichen.

Nun sei k das größte i für das gilt:

$$k = max\left\{i : p_{(i)} \leq \frac{i}{m} q\right\} \qquad (20)$$

Existiert ein k so werden alle Hypothesen H_1 bis H_k verworfen.

2.5.6 Varianzkoeffizient

Der Varianzkoeffizient (Vk) ist ein Ungleichheitsmaß. Er gibt an wie stark ein Datensatz streut. Er hat den Vorteil gegenüber der Varianz s^2

$$s^2 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)}{n} \qquad (21)$$

und der Standardabweichung s,

$$s = \sqrt{\sum_{i=1}^{n} \frac{\left(x_i - \bar{x}\right)^2}{n}}, \qquad (22)$$

dass er ein Verhältnis zum arithmetischen Mittel $\bar{X}$

$$\bar{X} = \sum_{i=1}^{n} x_i \qquad (23)$$

der Häufigkeitsverteilung widerspiegelt und dabei dimensionslos ist.
Der Varianzkoeffizient Vk wird definiert als Quotient aus der Standardabweichung s und dem arithmetischen Mittel $\bar{X}$ (Falk et al., 1995):

$$Vk = \frac{s}{\bar{x}} \qquad (24)$$

Ist die Standardabweichung größer als der Mittelwert, so ist der Varianzkoeffizient größer 1. Je näher der Varianzkoeffizient gegen 0 geht umso ähnlicher sind die entsprechenden Messwerte.

3. Material

3.1 Module zur Identifizierung von differentiell methylierten Regionen (DMRs)

Für die Identifizierung differentiell methylierter Regionen war es notwendig ein Modul (*methylProfiling*) zur implementieren. Dieses Modul *methylProfiling* ermöglicht die Erstellung von Methylierungsprofilen sowie die Identifizierung differentiell methylierter Regionen von zwei Proben. Anhand der Methylierungprofile lassen sich biologisch interessante Fragestellungen näher untersuchen. Diese könnten z.B. sein, ob der Promotor eines Tumorsuppressorgen in Tumorproben hypermethyliert ist, ob es sich um eine CpG-Insel in der Promotorregion handelt die differenziell methyliert ist, oder auch generelle Fragestellungen wie welche Regionen im Vergleich von Normal- zu Tumorproben differentiell methyliert sind.

3.1.1 Implementierung der Funktion *methylProfiling*

Die Funktion *methylProfiling* berechnet mittlere Methylierungswerte (mRPM) für genomweite Fenster oder auch für vordefinierte Fenster. In beiden Fällen sprechen wir von *regions of interest* (ROI). Der mittlere Methylierungswert (mRPM) eines Fensters bzw. einer Region wird entsprechend der folgenden Formel kalkuliert

$$mRPM_{Roi} = \frac{1}{k} \sum_{j=1}^{k_i} rpm_j \quad . \qquad (24)$$

Es sei k die Anzahl der in die entsprechende ROI fallenden RPM-Werte (RPM_j). Neben der Berechnung der $mRPM_{ROI}$-Werte für MeDIP-Daten einer Probe, ist es auch möglich, der Funktion zwei Genomvektoren als Eingabe zu geben. Hier sollen unterschiedliche methylierte Regionen zwischen zwei Proben identifiziert werden. In diesem Fall führt die Funktion die statistischen Berechnungen (Kapitel 2.2) für jede zuvor definierte ROI aus. Die Ermittlung der RPM-Werte der jeweiligen ROI basiert auf zwei unterschiedlichen Subroutinen. Zum einen die Subroutine PROFILE, welche die Berechnungen für benachbarte genomweite Fenster gleicher Größe übernimmt. Zum anderen die Subroutine ROIPROFILE, welche für die Berechnung und Auswertung für vordefinierte

Regionen dient. Solche vordefinierten Regionen können z.B. CpG-Inseln, Promotoren, Exons oder Introns sein, die der Subroutine zur Verfügung gestellt wird. Beide Subroutinen sind in C/C++ programmiert. Die Verwendung des .Call Interfaces ermöglicht die Einbindung bzw. Verwendung von C Code in R. Diese Herangehensweise war notwendig um eine beschleunigte Prozessierung zu erreichen.

3.1.2 Die Subroutine PROFILE

Die Subroutine PROFILE ermöglicht die Auswertung und Berechnung genomweiter Fenster, deren Größe durch den Parameter Fenstergröße (*framesize*) einheitlich festgelegt werden kann. Die genomischen Fenster (**Abbildung 7**) können sich dabei überschneiden (*sliding window* Ansatz). Das Ausmaß der Überschneidung kann durch den Parameter *Stepsize* (Schrittweite) beim Aufruf der Funktion spezifiziert werden.

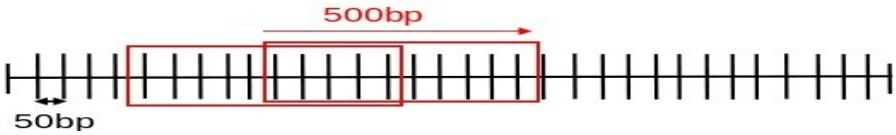

Abbildung 7: *Sliding Window*. Schematische Darstellung überlappender Fenster mit einer Fenstergröße von 500 bp, einer Überlappung bzw. Schrittweite von 250 bp über einen Genomvektor mit einer Auflösung von 50 bp.

Sofern der Parameter für die Schrittweite mit *Stepsize*=0 spezifiziert wird, werden die Fenster direkt aneinandergereiht. Vorteil der sich überschneidenden Fenster ist eine höhere Abdeckung des Genoms sowie die Vermeidung einer zu starken Glättung der Methylierungswerte.

Für jedes Fenster wird ein Subvektor (*W*) generiert mit den entsprechenden RPM-Werten aus dem Genomvektor (**Abbildung 8**). Hierfür wird zunächst ein Vektor $C=(c_1,c_2,...,c_n)$ mit den entsprechenden Chromosomenlängen (c_i) des Referenzgenoms erzeugt und an die Subroutine übergeben, wobei n die Anzahl der Chromosomen ist (z.B. n=24 für das humane Genom).

Innerhalb der Subroutine wird ein Index (p) für die RPM-Werte im Genomvektor initialisiert (p=0). Mittels einer for-Schleife wird iterativ über den Vektor C jedes einzelne Chromosomen prozessiert. Im ersten Schritt wird die Anzahl der RPM-Werte in dem sich überschneidenden Bereich (*si*) berechnet:

$$si = \frac{stepsize}{b} \quad (25)$$

Anhand der Anzahl der Werte im überschneidenden Bereich, der Chromosomenlänge c_i und der Größe der Bins (*b*) kann die Gesamtanzahl der Fenster für das jeweilige Chromosom berechnet werden.

$$nf_i = truncate\left(\frac{C_i}{b} * \frac{1}{si}\right) \quad (26)$$

Wobei die berechnete Gesamtanzahl mittels *truncate* auf eine Ganzzahl abgerundet wird. Die Anzahl der RPM-Werte in einem Fenster (*items*) berechnet sich nach der Formel:

$$items = \frac{framesize}{b} \quad (27)$$

In einem iterativen Ansatz wird für jedes Fenster mittels einer While-Schleife die Stopposition (*stop*) berechnet. Dafür wird ein Start-Wert benötigt der beim Aufruf der übergeordneten For-Schleife mit Eins initialisiert (*start=1*) und bei jeder weiteren Iteration um die *Stepsize* erhöht wird.

$$stop = start + (items * b) - 1 \quad (28)$$

Ein Speziallfall entsteht, sobald das Ende des Chromosomen erreicht wird, bzw. sobald gilt:

a) *start* > c_i: In diesem Fall kommt es zum Abbruch der While-Schleife und das nächste Chromosomen wird prozessiert.

b) *stop* > c_i: Sofern die Stopposition größer der Chromosomenlänge ist berechnet sich die Anzahl der RPM-Werte für das ensprechende Fenster nach der Formel

$$items = \frac{(c_i - start)}{b+1} \quad (29)$$

und die Stopposition wird auf die Chromosomenlänge gesetzt (stop=c_i).

Iterativ wird nun über die Anzahl der RPM-Werte ein Vektor *W* mit den entsprechenden RPM-Werte erzeugt, wobei der Index *p* jeweils erhöht wird. Anhand des Vektors wird der mittlere RPM-Wert für das jeweilige Fenster berechnet.

Im letzten Schritt wird überprüft, ob es sich um das letzte Fenster im Chromosom handelt. Solange das nicht der Fall ist wird der Index *p* um die

Anzahl der RPM-Werte des sich überschneidenden Bereiches reduziert.

$$p=p-s1$$

Ansonsten bleibt der Index unverändert.

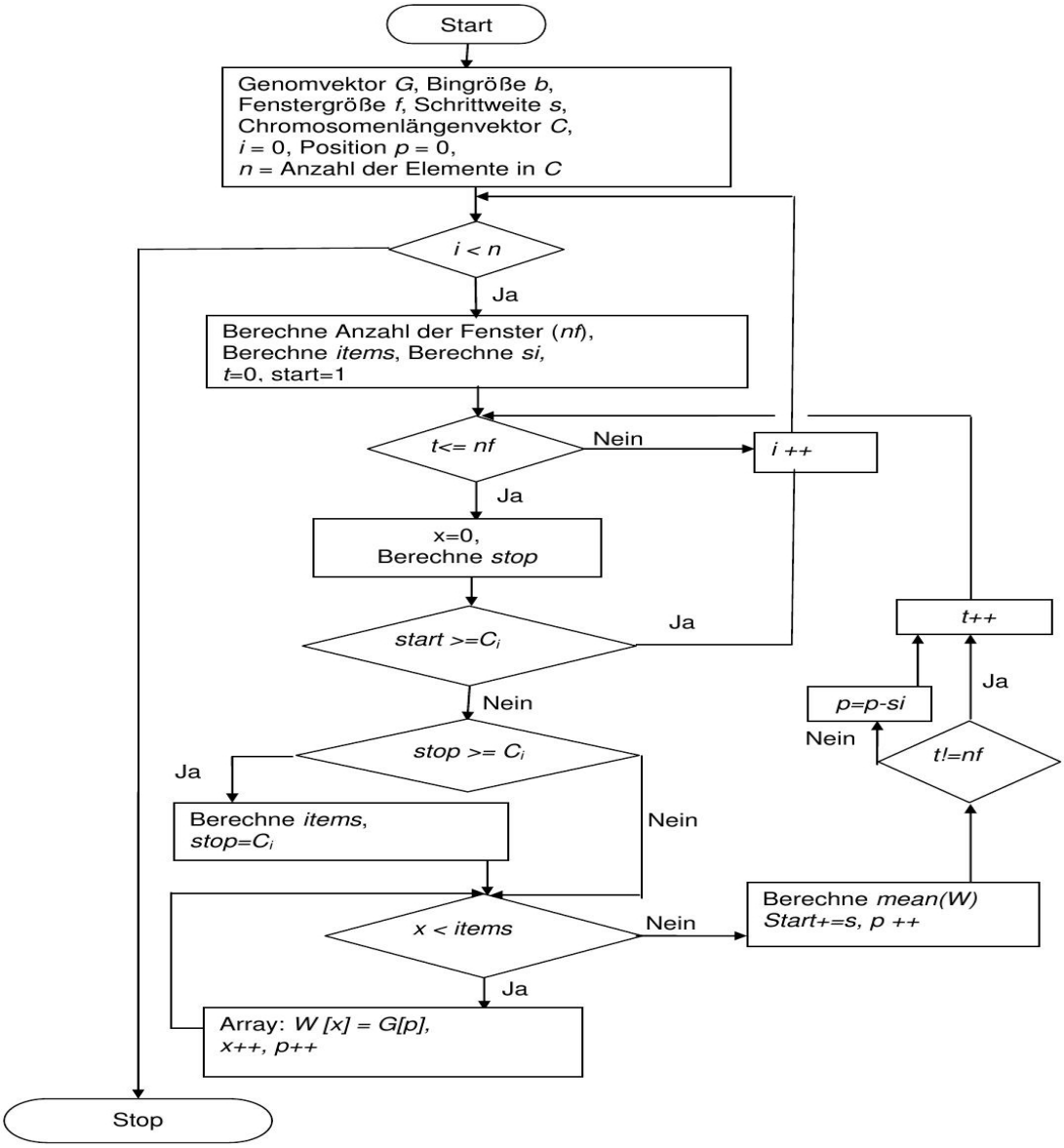

Abbildung 8: Ablaufschema der Subroutine PROFILE zur Erzeugung der Subvektoren entsprechen der Fenster. Die Subroutine Profile erwartet als Eingabe einen Genomvektor (*G*), die Binsize (*b*), einen Vektor mit den entsprechenden Chromosomenlängen (*C*), die Schrittweite (*s*) sowie die Fenstergröße (*f*). Iterativ wird für jedes Chromosom die Anzal der Fenster (*nf*) berechnet. Außerdem wird für jedes Fenster die Anzahl der in ihm enthaltenen RPM-Werte (items) berechnet. Anhand einer Zählvariable (*p*), die als Index der RPM-Werte im Genomvektor entsprechend des jeweiligen Fensters dient, wird ein Subvektor (*W*) mit den dazugehörigen RPM-Werten erzeugt. Weitere Zählvariabeln sind *i* und *x.* Start und Stop sind die jeweiligen Anfangs- und Endpunkte der Fenster.

3.1.3 Die Subroutine ROIPROFILE

Durch die Subroutine ROIPROFILE ist es möglich, gezielte vordefinierte Regionen zu analysieren, wie z.B. Promotoren oder CpG-Inseln. Als Eingabe wird hierfür eine sortierte Matrix ROI der Regionen bestehend aus Chromosomname (ROI_{Chr}), Start (ROI_{Start}) und Stop (ROI_{Stop})Position benötigt. Ebenso der entsprechende Genomvektor (*G*), sowie ein Vektor (*C)*, der die Chromosomenlängen (c_i) beinhaltet, ein Vektor (*B)*, der die Binpositionen aller Chromosomen enthält und außerdem noch die Größe der Bins (*b*) (**Abbildung 9**).

Im ersten Schritt wird ein Vektor (*BC)* erzeugt, der die kumulativen Summen, der Anzahl der Bins (bc_i) pro Chromosom *i* enthält:

$$BC = \left(bc_1, bc_1 + bc_2, bc_1 + bc_2 + bc_3, \ldots, bc_n \right) \text{ ,mit}$$

$$bc_i = ceil\left(\frac{c_i}{b} \right) . \quad (30)$$

Die Funktion *ceil* gibt dabei immer die aufgerundete Ganzzahl zurück.

Es werden nun iterativ über die Zeilen der ROI Matrix mit Hilfe der gegebenen Start (ROI_{Start}) und Stop (ROI_{Stop}) Positionen die Indizes der entsprechenden RPM-Werte im Genomvektor berechnet. Dazu wird als erstes überprüft, ob es sich bei dem Chromosom der ROI um das erste Chromosom im Genomvektor handelt. Ist dies der Fall so berechnet sich der Index der Startposition (s_1) im Genomvektor nach der Formel

$$s_1 = ceil\left(\frac{Start_{ROI}}{b} \right) \quad (31)$$

und der Index der Stopposition (s_2)

$$s_2 = ceil\left(\frac{Stop_{ROI}}{b} \right) . \quad (32)$$

Handelt es sich nicht um das erste Chromosom, so berechnen sich die Indices der Start- und Stopposition der Werte im Genomvektor wie folgt:

$$s_1 = bc_{i-1} + ceil\left(\frac{Start_{ROI}}{b} \right) \quad (33)$$

$$s_2 = bc_{i-1} + ceil\left(\frac{Stop_{ROI}}{b} \right) \quad (34)$$

wobei gilt *i=1,...,n* und n ist die Anzahl aller kumulativen Summen im *BC* Vektor.

Nun wird überprüft, ob die Binposition an der Stelle $B(s_1)$ ungleich dem Start der ROI ist. Falls dies der Fall ist, wird s_1 um eins erhöht. Für jede ROI wird ein Subvektor *(W)* mit den RPM-Werten erzeugt, die innerhalb der ROI liegen, anhand derer ein mittlerer Methylierungswert der ROI berechnet wird, oder im Falle zweier gegebener Genomvektoren ebenfalls die statistischen Tests vorgenommen werden.

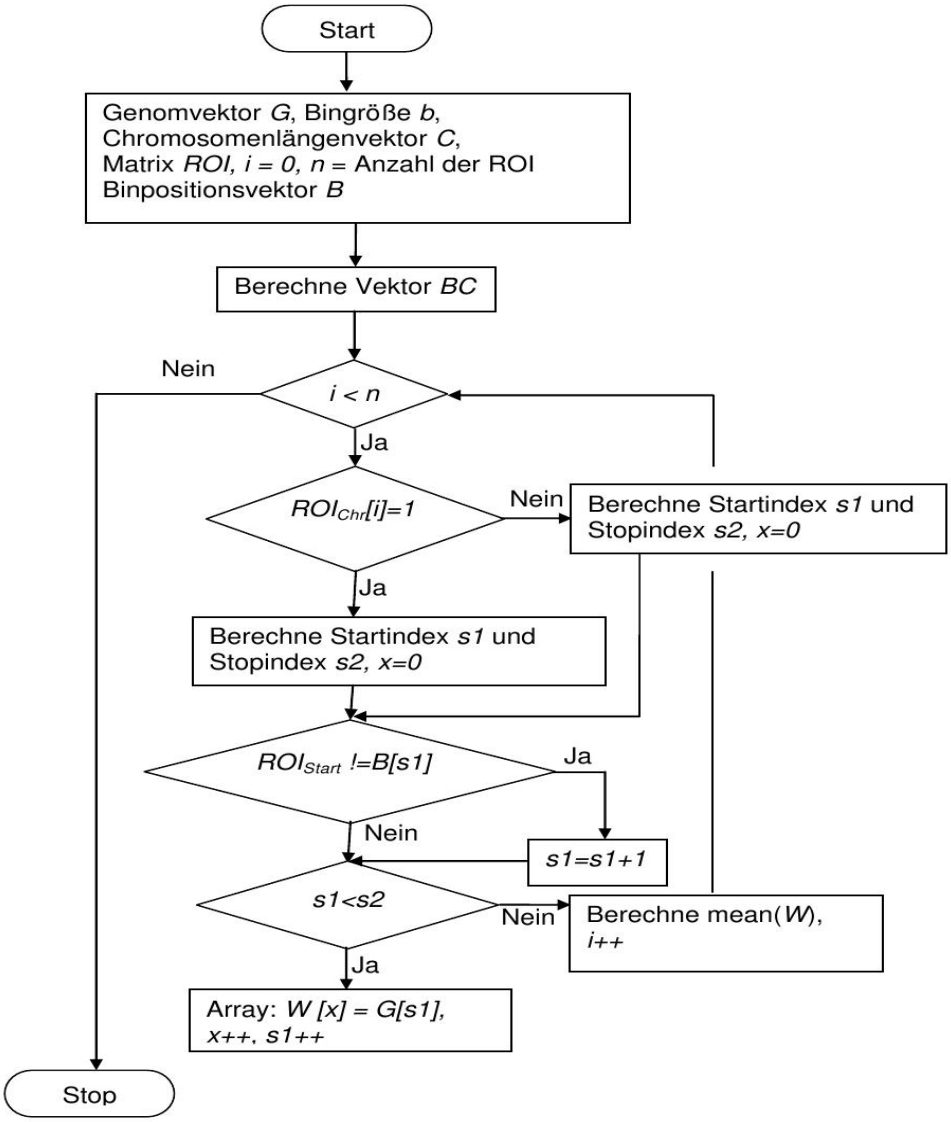

Abbildung 9: Ablaufschema der Subroutine ROIPROFILE zur Erzeugung der Subvektoren der jeweiligen ROI. Die Subroutine ROIPROFILE erwartet als Eingabe den Genomvektor (*G*), die Binsize (*b*), einen Vektor mit den Längen der im Genom enthaltenen Chromosomen (*C*), einen Vektor mit den Binpositionen (*B*) sowie die Matrix mit den ROIs (*ROI*), wobei n die Anzahl der ROIs ist. Anschließend wird für jede ROI ein Subvektor (*W*) erzeugt, der ausschließlich die RPM-Werten des Genomvektor enthält, die in die ensprechende ROI fallen. Zur Generierung werden aus den jeweiligen Start- und Stop-Positionen der ROI die Indizes (*s1, s2*) der RPM-Werte im Genomvektor berechnet. Die benötigten Zählvariabeln sind *x* und *i*.

3.2 Integration biologischer Replika

Das Modul *methylProfiling* ermöglicht den statistischen Vergleich von zwei Genomvektoren. Um z.B. zwei Gewebeproben verschiedener Patienten miteinander zu vergleichen, müssen die jeweiligen Patienten einer Gruppe zu einem Genomvektor zusammengefasst werden. Im Rahmen dieser Studie sollte jedoch auf Basis mehrerer Individuen eine Methylierungsanalyse vorgenommen werden. Diese Art der Analyse ist auch notwendig, um mögliche Subgruppen, wie z.B. aggressive und weniger aggressivere Tumorvarianten, zu klassifizieren. Daher wurde das Modul *methylProfiling* zunächst lediglich für die Berechnung der mittleren Methylierungswerte vordefinierter Regionen der einzelnen Gewebeproben der jeweiligen Patienten verwendet. Anschließend wurden die entsprechenden Werte in einer N x M Matrix zusammengefügt. Wobei N die Anzahl der genomischen Regionen und M die Anzahl der Gewebsproben ist. M setzt sich dabei zusammen aus $X=x_1,...,x_L$ und $Y=Y_1,...,Y_K$, wobei L die Anzahl der Normalgewebeproben und K die Anzahl der Tumorgewebeproben ist. Anhand dieser zwei Gruppen werden für jede Region die statistischen Tests (siehe Kapitel 2.2), Mittelwerte, Verhältnis der Mittelwerte sowie die Varianzkoeffizenten, unter Verwendung der in R implementierten Funktionen (R Development Core Team, 2009), berechnet.

4. Ergebnisse

4.1. Anwendung zur Identifikation differentiell methylierter Regionen in Tumorsamples

4.1.1 Parametersetting

Für die folgende Untersuchung standen insgesamt 28 Proben von 14 verschiedenen Darmkrebspatienten zur Verfügung, wobei 14 Proben aus Normalgewebe und 14 Proben aus Tumorgewebe entnommen wurden. Darüber hinaus lagen für jede der 28 Proben entsprechende Daten aus Input-Proben vor. Bei der Input-Probe handelt es sich um das entsprechende sequenzierte Gewebe, ohne dass daran eine Anreicherung mittels MeDIP vorgenommen wurde (**siehe Kapitel 2.1.1**). Für die Identifikation differentiell methylierter Regionen wurde das Genom in 500 Basenpaare große Fenster eingeteilt, die sich jeweils um 250 Basenpaare überschneiden. Der Genomvektor wurde auf Basis einer Bingröße von 50 Basenpaaren erzeugt. Anhand der in das jeweilige Fenster fallenden RPMs wurden die mittleren Methylierungswerte (mRPMs) für jede Probe separat gebildet. Auf Basis dieser Werte wurde entsprechend der Gewebegruppen, jeweils der t-Test sowie der Wilcoxon-Rangsummentest durchgeführt. Die daraus resultierenden P-Werte wurden unter Einbeziehung der Benjamini-Hochberg False Discovery Rate adjustiert und somit in entsprechende Q-Werte (q.t und q.w) transformiert. Es wurden ausschließlich Regionen, deren Q-Wert kleiner oder gleich 0.01 ist, berücksichtigt. Außerdem wurde der jeweilige Varianzkoeffizent berechnet. Es wurden nur Regionen mit einem Varianzkoeffizienten kleiner 1 berücksichtigt. Anhand dieser Werte wurden als lokales Filterkriterium das Verhältnis (ratio) aus mittlerem Methylierungswert der Normalgewebe und mittlerem Methylierungswert der Tumorgewebe bestimmt.

$$ratio = \frac{mRPM_{Normalgewebe}}{mRPM_{Tumorgewebe}} \qquad (35)$$

Dessen Wert sollte mehr als 1.33 oder weniger als 0.75 betragen. Das Verhältnis 1.33 ist die untere Grenze, die zu Regionen führt, die im Normalgewebe hypermethyliert sind und 0.75 die obere Grenze die zu

hypermethylierten Regionen im Tumorgewebe führt. Für jede Region wurde ebenfalls der mittlere Methylierungswert der einzelnen Input-Proben berechnet. Als weiteres lokales Filterkriterium wurden nur Regionen weiter betrachtet, wenn der mRPM-Wert des entsprechenden Gewebetyps um ein k-faches (k) größer als der mRPM-Wert der dazugehörigen Input-Proben ist. Als letztes und globales Filterkriterium wurde die Verteilung der mRPM-Werte der Input-Proben entsprechend ihres Gewebetyps bestimmt und das 95%-Quantil berechnet. Für den konkreten RPM-Wert welcher dem 95%-Quantil entspricht gilt, dass 95 % der Messwerte kleiner bzw. 5 % größer als das Quantil sind. Der entsprechende Wert dient als minimaler, globaler Grenzwert für das Hintergrundsignal.

Es wurde eine systematische Evaluation verschiedener Filterkriterien zur Identifikation statistisch signifikanter DMRs durchgeführt. Tabelle 2 listet die verschiedenen Filterkriterien auf. Für die weitere Analyse wurden die Filtereinstellungen entsprechend der Tabelle, Nummer 9 und 10 gewählt. Nummer 9 stellt dabei die Regionen dar, die eine Hypermethylierung im Tumorgewebe aufweisen und Nummer 10 eine Hypermethylierung im Normalgewebe. Zusammengenommen ergaben sich daraus 16502 teilweise überlappende Regionen die als differentiell methyliert identifiziert wurden.

Nummer	q.t	q.w	k	ratio	95%-Quantile Input	Vk	DMR
1	<=0.01	<=0.01	2	0.75>x>1.33			15383
2	<=0.01	<=0.01	1,5	0.75>x>1.33			17044
3	<=0.01	<=0.01	1,5	0.75>x>1.33		<1	16937
4	<=0.01	<=0.01	1,5	0.75>x>1.33	0,2128955 (N) ∧ 0,209282 (T)	<1	11108
5	<=0.01	<=0.01	1,5	0.75>x>1.33	0,2128955 (N) ‖ 0,209282 (T)	<1	16513
6	<=0.01	<=0.01	2	0.75>x>1.33	0,2128955 (N) ∧ 0,209282 (T)	<1	10859
7	<=0.01	<=0.01	2 (T)	0.75>x	0,209282 (T)	<1	6257
8	<=0.01	<=0.01	2 (N)	x>1.33	0,2128955 (N)	<1	8775
9	<=0.01	<=0.01	1.5 (T)	0.75>x	0,209282 (T)	<1	6585
10	<=0.01	<=0.01	1.5 (N)	x>1.33	0,2128955 (N)	<1	9917

Tabelle 2: Übersicht der Filtereinstellungen. Die q.t bzw. q.w sind die, unter Einbeziehung der False Discovery Rate, korrigierten P-Werte der jeweiligen statistischen Tests (t-Test und Wilcoxon-Rangsummentest). K ist der Faktor um den das entsprechende Signal der Tumorprobe (T) bzw. Normalgewebe (N) größer sein soll als deren lokale Inputwerte. Das Ratio gibt das Verhältnis des mittleren Methylierungswertes von Normalgewebe zu Tumorgewebe an. Das 95 %-Quantile gibt jeweils den Wert an unter den 95 % der Messwerten im Input des Normalgewebe bzw. Input des Tumorgewebe liegen. Der Varianzkoeffizient (Vk) dient als Ähnllichkeitskriterium und soll kleiner als eins sein.

4.1.2 Clusteranalyse

Clusteranalysen wurden häufig für die Gruppierung von Genen mit ähnlichen Expressionsprofilen eingesetzt. Innerhalb dieser Arbeit wurde getestet, ob Methylierungsmuster, die aus MeDIP-Seq Experimenten gewonnen wurden, zur Klassifikation von Gewebezuständen geeignet sind. Hierzu wurden zunächst differentiell methylierte Regionen (DMRs) identifiziert (siehe Kapitel 4.1). Mittels einer Clusteranalyse wurde anschließend getestet, ob die untersuchten biologischen Proben mittels der identifizierten DMRs entsprechend Ihrer Zustände in Gruppen eingeteilt werden können. Hierbei sollten genomische Regionen mit ähnlichen Methylierungsprofilen in dieselbe Gruppe eingeordnet werden und Regionen mit verschiedenen Methylierungsprofilen in verschiedene Gruppen. Die mathematische Ähnlichkeit zweier Datenpunkte wird üblicherweise durch ein Distanzmaß bzw. Ähnlichkeitsmaß gemessen, das die Methylierungszustände numerisch bewertet. Ist also $X=(x_1,...,x_P)$ der Vektor, der die Methylierung von Region X über die P Patienten beschreibt, und ist $Y=(y_1,...,y_P)$ der Vektor, der die Methylierung von Region Y über die P Patienten beschreibt, so lässt sich die Ähnlichkeit dieser Vektoren z.B. durch die Euklidische Distanz beschreiben:

$$d(X,Y)=\sqrt{\sum_{i=1}^{P}(x_i-y_i)^2} \quad (36)$$

Ähneln sich die Methylierungsprofile der Regionen über die P Versuche, so liefert dieses Maß einen kleinen Wert. Ein Clusteralgorithmus entscheidet anhand eines solchen Ähnlichkeitsmaßes, ob zwei Regionen in dieselbe Gruppe gehören.

Für die graphische Darstellung einer solchen Clusteranalyse wurde die in R (R Development Core Team, 2009) implementierte Funktion heatmap verwendet. Die einzelnen Methylierungswerte werden entsprechend ihrer Werte farbskaliert dargestellt. Darüber hinaus handelt es sich bei einer Heatmap um ein zweidimensionales Cluster. Zum einen wird über die Spalten, in unserem Falle die Patientenproben geclustert, zum anderen über die Zeilen, d.h. die differenziell methylierten Regionen. Als Distanzmaß wurde die Euklidische Distanz benutzt und als Methode für das Clustern die Default-Einstellung

complete. Diese entspricht der Complete-Linkage Methode. Hierbei werden die maximalen Abstände der Punkte betrachtet.

Wie man eindeutig anhand des Dendrogramms (**Abbildung 10**) sehen kann, werden die Patientenproben auf Basis der als differenziell methyliert identifizierten Regionen entsprechend des jeweiligen Gewebetyps gruppiert. Somit ist eine Unterscheidung zwischen Normalgewebe und Tumorgewebe auf Basis ihrer Methylierungsmuster möglich.

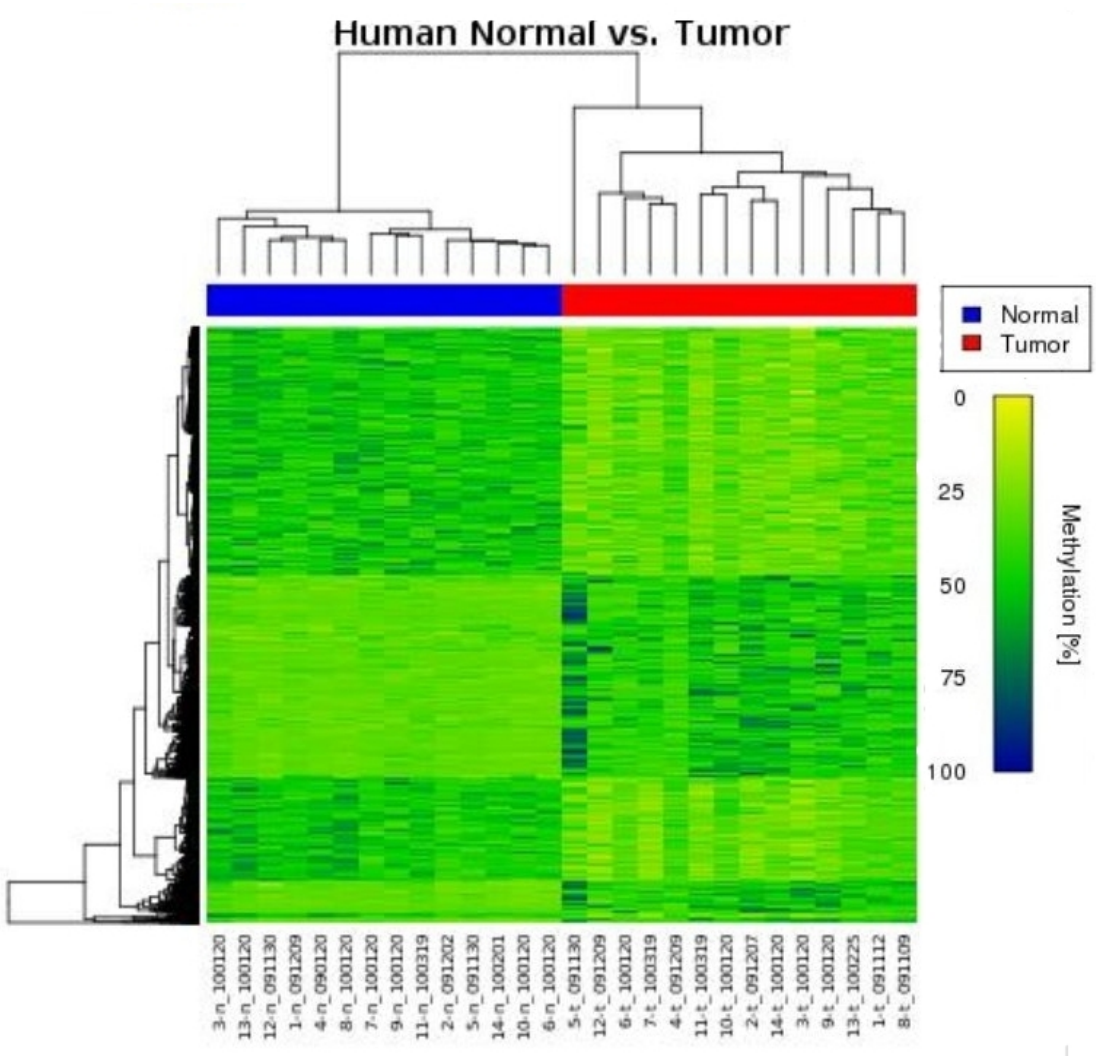

Abbildung 10: Heatmap der 16.502 differentiell methylierten Regionen. Das Dendrogramm der Patientenproben (oben) gruppiert, entsprechend der Gewebetypen, Normal (blau) und Tumor (rot). Das Dendrogramm links stellt die Gruppierung über die Regionen dar. Mittels einer Farbskalierung von Gelb nach Blau werden die mRPM-Werte dargestellt.

4.1.3 Annotation

Um einen möglicherweise funktionalen Zusammenhang zwischen identifizierten DMRs und z.B. Genexpressionsänderungen aufweisen zu können, ist es möglich diese zu annotieren. Als Annotation bezeichnet man das Verknüpfen beziehungsweise Zuordnen der Sequenzinformation, in diesem Fall der hyper- oder hypomethylierten Regionen, zu einem funktionalen Zusammenhang. Es ist dabei von besonderen Interesse, ob eine Region z.B. in proteinkodierenden Abschnitten (Exons), genregulatorischen Elementen (z.B. Bindungsstellen von Transkriptionsfaktoren) oder Promotoren fällt. Sofern durch die veränderte Methylierung ein Genbereich betroffen ist, ist es von Interesse, um welches Gen es sich handelt und welche dessen biologische Funktion ist. Aus diesem Grund wurden die Regionen mit bekannten Annotation wie Exons, Introns, Promotoren, CpG-Inseln und Genen verglichen. Da es sich bei den DMRs jedoch teilweise um überlappende Fenster handelt, wurden diese vorab zu einer Region zusammengefasst. Aus den 16.502 Regionen verblieben 8.818 sich nicht überschneidenden Regionen. Diese setzen sich zusammen aus 6.126 hypermethylierten Regionen des Normalgewebes und 2.692 hypermethylierten Regionen im Tumorgewebes. Einen Überblick über die in diesen Regionen enthaltenen Annotationen liefert nachfolgende Tabelle 3.

	Exons	Introns	Promotoren	CpG-Inseln	Intergenic	Gene
Hypermethyliert in Normalgewebe	2183	1862	635	1345	6892	1531
Hypermethyliert in Tumorgewebe	3075	1222	2117	1856	3438	1635

Tabelle 3: Übersicht der als differenziell methyliert identifizierten Annotationen

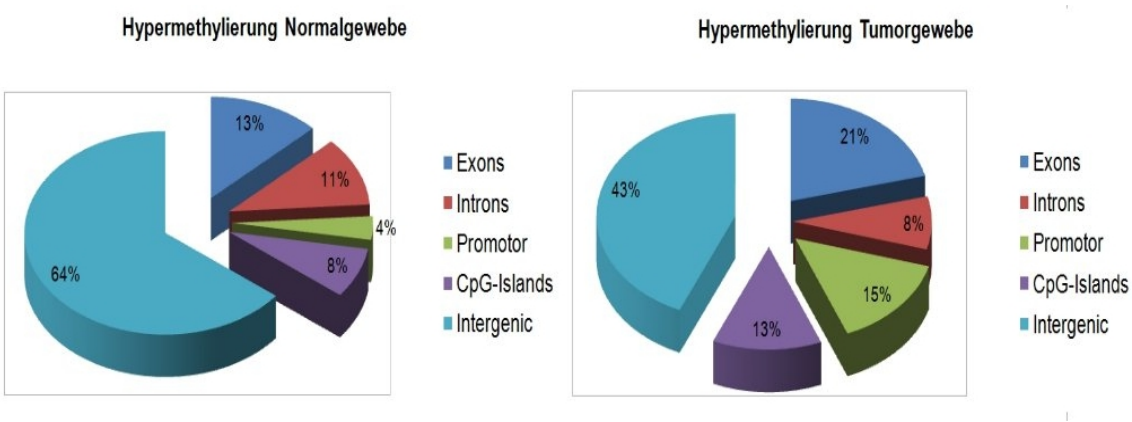

Abbildung 11: Kreisdiagramm differenziell methylierter Region. Die Darstellung zeigt die prozentuale Verteilung der Annotationen, die im Normalgewebe (links) und im Tumorgewebe (rechts) hypermethyliert sind.

Vergleicht man die Verteilung der hypermethylierten Annotationen aus Abbildung 11. des Normalgewebes mit denen des Tumorgewebes, so kann man deutlich erkennen, dass eine differenzielle Hypermethylierung des Normalgewebes hauptsächlich im intergenischen Bereich vorliegt. Nach Tabelle 3 ist die Anzahl der im Normalgewebe differenziell hypermethylierten intergenischen Regionen sogar doppelt so hoch wie die Anzahl der differenziell hypermethylierten intergenischen Regionen im Tumorgewebe. Dies entspricht auch der Beobachtung, dass die chromosomale Instabilität in Tumorzellen mit einer genomweiten Abnahme der DNA-Methylierung korreliert (Rodriguez et al., 2006). Betrachtet man die Verteilung der differenziell hypermethylierten Regionen die in den Bereich von Promotoren fallen, so zeigt sich hier der umgekehrte Effekt. Im Normalgewebe finden sich nur 635 differenziell hypermethylierte Regionen, die mit Promotoren assoziiert sind. Im Tumorgewebe hingegen wurde mehr als das dreifache an differenziell methylierten Promotoren gefunden wurde. Dies deckt sich wiederum mit der Feststellung, dass speziell bei Tumoren eine vermehrte de novo Methylierung der Promotorregionen vorliegt (Jones et al., 2003).

4.1.4 Visualisierung der DMRs

Zur visuellen Darstellung wurde der Genombrowser der *University of California Santa Cruz* (Kent et al., 2002) verwendet. Er wurde im Rahmen des *Human Genome Projects* entwickelt und beinhaltet neben zahlreichen Genomen diverse Annotationen, wie z.B. bekannte Gene oder CpG Inseln. Diese lassen sich direkt graphisch im genomischen Kontext darstellen. Zusätzlich bietet er die Möglichkeit, eigene Daten in Form sogenannter Custom Tracks hoch zu laden und zu visualisieren. Dazu ist es notwendig, die Daten in ein entsprechendes Format zu bringen. Hierzu wurden die Sequenzdaten der einzelnen Proben ins Wiggle-Format (WIG) transformiert. Die als differenziell methyliert identifizierten Regionen wurden als Browser Extensible Display (BED) Datei exportiert und hochgeladen. Dieses Format beinhaltet lediglich die Informationen Chromosom, Start und Stop und ermöglicht die Darstellung dieser Region in Form eines schwarzen Balkens im Genombrowser.

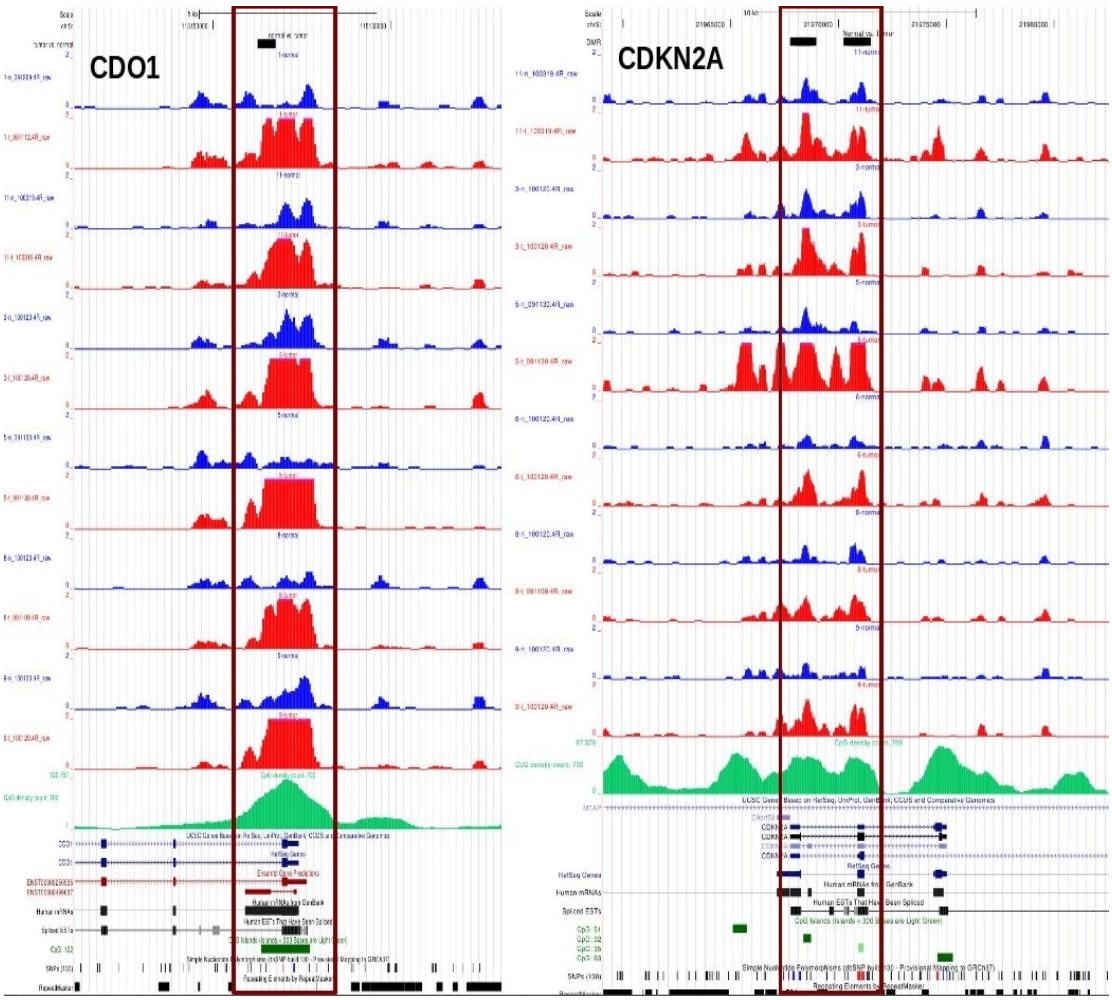

Abbildung 12: De-novo Methylierung im Promotorbereich von CDO1 und CDKN2A. Die oberen schwarzen Balken im rot umrandeten Bereich stellen die als differenziell methyliert identifizierten Regionen dar. Darunter folgen die Patientenproben. Blau stellt immer das Normalgewebe eines Patienten dar, gefolgt von seiner Probe aus dem Normalgewebe (rot). In Grün dargestellt findet sich die CpG-Dichte der Region. Darunter befinden sich die Annotation von bekannten Genen und deren Promotoren, sowie in Form von grünen Balken dargestellt CpG-Inseln.

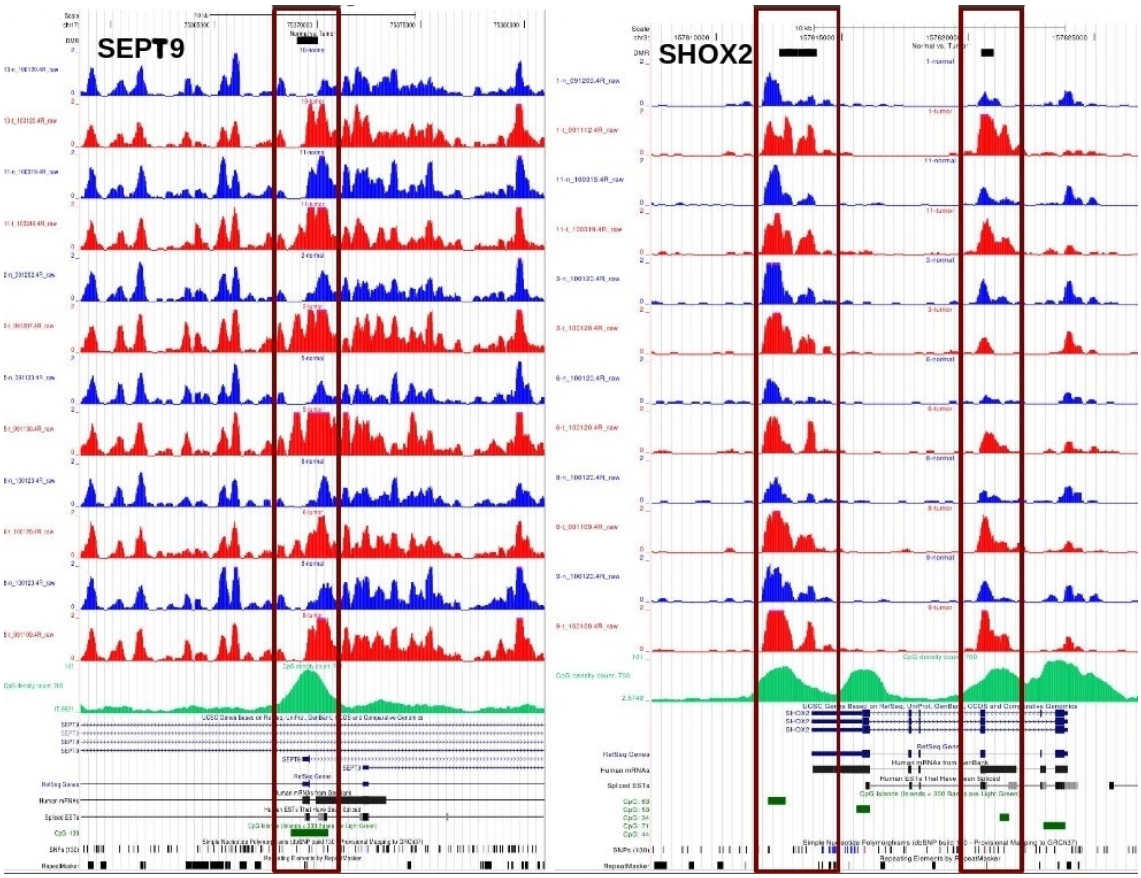

Abbildung 13: De-novo Methylierung der Promotorbereiche der Biomarker SEPT9 und SHOX2. Ebenfalls im rot umrandeten Bereich die gefundenen differenziell methylierten Regionen in Form eines schwarzen Balken dargestellt (oben). Gefolgt von den jeweiligen Proben der Patienten Normal (blau) und Tumor (rot). Darunter die CpG-Dichte (grün) sowie die Annotationen bekannter Gene und in Form grüner Balken die CpG-Inseln.

Die Abbildungen 12 und 13 zeigen eine solche Visualisierung am Beispiel der bekannten Methylierungsbiomarker CDO1, CDKN2A, SEPT9 und SHOX2. In dem Rot umrandeten Bereich sieht man oben die schwarzen Balken, die die als differenziell methylierten Regionen markieren. Darunter folgen die Methylierungswerte der einzelnen Patienten jeweils im Wechsel. Blau steht hierbei für das Normalgewebe. Darunter sieht man die entsprechende Tumorprobe des Patienten (rot). Grün dargestellt ist die CpG-Dichte in diesem Bereich. Darunter folgen noch Annotationen, wie z.B. bekannte Gene und deren Promotoren sowie in Form der grünen Balken die CpG Inseln.

Bei den hier dargestellten Genen besteht ein enger Zusammenhang mit der Entstehung verschiedener Krebsarten.

Cysteindioxigenase 1 (CDO1) wurde als starker prognostischer Biomarker für

die Therapie von Brustkrepspatienten identifiziert (Dietrich et al., in press).
Außerdem wird vermutet, dass die epigenetische Inaktivierung des Gens oder die Deletion dieser chromosomalen Region ein häufiger Mechanismus bei der Tumorentstehung von Darmkrebs ist (Staub et al., 2006). Diese Vermutung wird durch diese Studie bestätigt. Man erkennt eine klare Methylierung der CpG-Insel im Promotorbereich des Gens, was auf eine Inaktivierung schließen lässt.
Ebenfalls eine Methylierung des Promotorbereiches weist der Cyclin-abhängige Kinase-Inhibitor 2 A (CDKN2A) auf. Dabei handelt es sich um ein Tumorsuppressor-Gen, was eine wichtige Rolle bei der Regulation des Zellzyklus spielt. Es hat sich gezeigt, dass seine Inaktivierung zur Tumorentstehung zahlreicher Krebsarten, wie z.B. von Melanomen, führt. Bei Karzinomen der Kopf-Hals-Region wurde gezeigt, dass die Hypermethylierung im Promotorbereich einen möglichen Biomarker darstellt (Martone T, Gillio-Tos A, De Marco L, et al., 2007).
SEPT9 und SHOX2 wiesen ebenfalls eine signifikante Hypermethylierung im Promotorbereich der Tumorproben auf. Bei SEPT9 handelt es sich um einen Biomarker für Dickdarmkrebs (Lofton-Day, 2008) und bei SHOX2 für Lungenkrebs (Schmidt et al., in press). Diese beiden Biomarker sind die bisher einzigen Methylierungsbiomarker, die in CE markierten, zugelassenen In-vitro-Diagnostika Verwendung finden.

5.Diskussion und Zusammenfassung

DNA Methylierung ist ein wichtiger epigenetischer Mechanismus, der in vielen biologischen Prozessen eine entscheidende Rolle spielt. Frühere Studien haben gezeigt, dass DNA Methylierung eine Ursache für die Krebsentstehung (Michal et al., 2006), sowie den Verlauf der Krankheit und das Ansprechen auf eine medikamentöse Behandlung beeinflussen und prognostizieren kann.
Die in dieser Arbeit verwendete Methode zur DNA-Methylierungsanalyse basiert auf der Anwendung von methylierter DNA Immunpräzipitation (MeDIP) (Weber et al., 2005). Diese Methode ermöglicht die Anreicherung methylierter DNA-Fragmente mittels methylierungsspezifischer Antikörper. Für die eindeutige Identifizierung der DNA-Fragmente und die anschließende Bestimmung der Positionen im Referenzgenom wurde eine Sequenziertechnologie der zweiten Generation verwendet (MeDIP-Seq). Diese stellt eine sehr schnelle und kostengünstige Methode für die genomweite Analyse dar.
Bisher gab es mit dem BATMAN Algorithmus (Down et al., 2008) nur eine Methodik zur Verarbeitung von MeDIP-Seq Daten. Jedoch dauerte allein die Prozessierung des menschlichen Chromosoms 1 mit dem BATMAN Algorithmus allein drei Tage. Die in dieser Studie entwickelte neue Methode prozessiert das komplette Genome innerhalb weniger Stunden und wurde für die weitere Analyse und zur Identifizierung differentiell methylierter Regionen bei Darmkrebsproben angewandt.

Die Ergebnisse dieser Arbeit zeigen, dass die entwickelte Methode es ermöglicht, differenziell methylierte Regionen zu identifizieren. Darüber hinaus zeigte die Clusteranalyse, dass die einzelnen Proben, entsprechend ihrer Methylierungswerte der gefundenen Regionen, eine sinnvolle histopathologische Unterscheidung ergaben. Beobachtungen anderer Studien, wie z.B. die genomweite Abnahme der Methylierung (Rodriguez et al., 2006) im intergenischen Bereich sowie eine Hypermethylierung von Promotoregionen bei Tumorgewebe (Jones et al., 2003), konnten anhand der Annotierung der Regionen bestätigt werden. Darüber hinaus zeigte sich bei den identifizierten Regionen ebenfalls eine vermehrte Hypermethylierung von CpG-Inseln in den Tumorproben. Bekannte Biomarker wie SEPT9, SHOX2, CDO1 und CDKN2A

konnten ebenfalls als differenziel methyliert bestätigt werden.

Der hier vorgestellte Algoritmus basiert auf vordefinierten Fenstern. Das komplette Genom mittels kurzer, sich überschneidender Fenster zu analysieren, ist sehr rechenintensiv und somit zeitaufwändig, und äußerst speicherintensiv. Zur weiteren Optimierung der Methode wäre es erstrebenswert, einen Algorithmus zu generieren, der es ermöglicht dynamische Fenster zu generieren. Eine mögliche Herangehensweise wäre, zuerst eine grobe Analyse des Genoms vorzunehmen und auf Basis der daraus resultierenden Regionen, die von potentiellem Interesse sind, die betrachteten Fenstergrößen dynamisch zu optimiert. Darüber hinaus wurde in dieser Studie auf nicht normalisierten MeDIP-Seq Daten gearbeitet. Da jedoch bekannt ist, dass sowohl unspezifische Bindungen des Antikörpers als auch die CpG-Dichte das Methylierungssignal beeinflussen (Pelizzola et al., 2008), wäre es sinnvoll, diese Studie anhand normalisierter Werte durchzuführen.
Die Methode (der Algorithmus) sollte in weiteren Studien getestet werden und die gefundenen Ergebnisse mit unabhängigen Methoden, wie z.B. mittels quantitativer Bisulfitsequenzierung (Dietrich et al, 2009) verifiziert werden. Anhand der daraus gewonnen Ergebnisse wäre es ebenfalls möglich, eine bessere Abstimmung der Filtereinstellungen, sowie neue Ansätze für die Normalisierung der MeDIP-Seq Daten zu gewinnen.
Dazu erscheint es lohnenswert, kombinierte *in-silico* und *in-vitro* Studien zu betreiben, um mit Hilfe optimierter Filterkriterien in Zukunft ein verbessertes Verständnis der MeDIP-Seq Daten basierten Methylierungsmuster zu erhalten.

6. Literaturverzeichnis

[1] Aaltonen LA, Peltomäki P, Leach FS, Sistonen S, Pylkkänen L, Mecklin J, Järvinen H, Powell SM, Jen J, Hamilton SR, Petersen GM, Kinzler KW, Vogelstein B, De la Chapelle A (1993). Clues to the pathogenesis of familial colorectal cancer. Science 260(5109): 812-816.

[2] Antequera F, Bird A (1993). Number of CpG islands and genes in human and mouse. Proc Natl Acad Sci U S A 90(24): 11995-11999.

[3] Bachman KE, Hermane JG, Corn PG, Merlo A, Costello JF, Cavenee WK, Baylin SB, Graff JR (1999). Methylation-associated silencing of the tissue inhibitor of metalloproteinase-3 gene suggest a suppressor role in kidney, brain, and other human cancers. Cancer Res. 59(4): 798-802.

[4] Bennett ST, Barnes C, Cox A, Davies L, Brown C (2005). Toward the 1,000 dollars human genome. Pharmacogenomics 6: 373-382.

[5] Benjamini Y, Hochberg Y (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society B. 57: 289–300.

[6] Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, Luo S, McCurdy S, Foy M, Ewan M, Roth R, George D, Eletr S, Albrecht G, Vermaas E, Williams SR, Moon K, Burcham T, Pallas M, DuBridge RB, Kirchner J, Fearon K, Mao J, Corcoran K (2000). Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat. Biotechnol. 18: 630-634.

[7] Buck MJ and Lieb JD (2004). ChIP-chip: considerations for the design, analysis and application of genome-wide chromatin immunoprecipitation experiments. Genomics 83(3): 349-360.

[8] Chavez L, Jozefczuk J, Grimm C, Dietrich J, Timmermann B, Herwig R, Adjaye J (eingereicht). Computational analysis of genome-wide DNA-methylation during the dierentiation of human embryonic stem cells along the endodermal lineage

[9] Chang SC, Tucker T, Thorogood NP, Brown CJ (2006). Mechanisms of X-chromosome inactivation. Front Biosci. 11: 852-66.

[10] Dean W, Lucifero D, Santos F (2005). DNA methylation in mammalian development and disease. Birth Defects Res C Embryo Today 75(2): 98-111.

[11] Dietrich D, Krispin M, Dietrich J, Fassbender A, Lewin J, Harbeck N, Schmitt M, Eppenberger-Castori S, Vuaroqueaux V, Spyratos F, Foekens JA, Lesche R, Martens JWM, (im Druck). CDO1 Promoter Methylation is a Biomarker for Outcome Prediction of Anthracycline Treated, Estrogen Receptor-Positive, Lymph Node-Positive Breast Cancer Patients.

[12] Dietrich D, Lesche R, Tetzner R, Krispin M, Dietrich J, Haedicke W, Schuster M, Kristiansen G (2009). Analysis of DNA Methylation of Multiple Genes in Microdissected Cells From Formalin-fixed and Paraffin-embedded Tissues. J Histochem Cytochem. 57(5):477-89.

[13] Down TA, Rakyan VK, Turner DJ, Flicek P, Li H, Kulesha E, Graef S, Johnson N, Herrero J, Tomazou EM, Thorne NP, Baeckdahl L, Herberth M, Howe KL, Jackson DK, Miretti MM, Marioni JC, Birney E, Hubbard TJP, Durbin R, Tavare S, Beck A (2008). A bayesian deconvolution strategy for immunoprecipitation-based dna methylome analysis. Nat Biotechnol. 26(7):779-785.

[14] Egger G, Liang G, Aparicio A, Jones PA (2004). Epigenetics in human disease and prospects for epigenetic therapy. Nature 429(6990):457-463.

[15] Eisen MB, Brown PO (1999). DNA Arrays for Analysis of Gene Expression. Methods Enzymol 303: 179-205.

[15] Eisen M, Spellman P, Brown P, Botstein D (1998). Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences, 95: 14863-14868.

[16] England R, Pettersson M (2005). Pyro Q-CpG™: quantitative analysis of methylation in multiple CpG sites by Pyrosequencing. Nature Methods 2:1-2.

[17] Falk M, Becker R, Marohn F (1995). Angewandte Statistik mit SAS. Springer Lehrbuch, Springer, Heidelberg, 22, 73-80.

[18] Florl AR, Löwer R, Schmitz-Dräger BJ, Schulz WA (1999). DNA methylation and expression of LINE-1 and HERV-K provirus sequences in urothelial and renal cell carcinomas. Br J Cancer 80(9): 1312-1321.

[19] Herman JG, Baylin SB (2003). Gene silencing in cancer in association with promoter hypermethylation. N Engl J Med. 349: 2042-2054.

[20] Herwig R, Schuchhardt J, Chavez L, Lehrach H (2007). Analyse von Biochips: Von der Sequenz zum System, in Grundlagen der Molekularen Medizin, Springer Verlag, Berlin, 76-77, 80-81.

[21] Jones PA, Baylin SB (2002). The fundamental role of epigenetic events in cancer. Nat Rev Genet. 3(6): 415-428.

[22] Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D (2002). The Human Genome Browser at UCSC. Genome Res. 12:996-1006.

[23] Li E, Beard C, Jaenisch R (1993). Role for DNA methylation in genomic imprinting. Nature. 366(6453): 362-365.

[24] Lofton-Day C, Model F, Devos T, Tetzner R, Distler J, Schuster M, Song X, Lesche R, Liebenberg V, Ebert M, Molnar B, Grützmann R, Pilarsky C, Sledziewski A. (2008). DNA methylation biomarkers for blood-based colorectal cancer screening. Clin. Chem. 54(2):414-423.

[25] Martone T, Gillio-Tos A, De Marco L, Fiano V, Maule M, Cavalot A, Garzaro M, Merletti F, Cortesina G (2007). Association between hypermethylated tumor and paired surgical margins in head and neck squamous cell carcinomas. Clin Cancer Res. 13:5089–5094.

[26] Luczak MW, Jagodzinski PP (2006). The role of DNA methylation in cancer developement. Folia Histochem Cytobiol. 44(3):143-154.

[27] Paz MF, Avila S, Fraga MF, Pollan M, Capella G, Peinado MA, Sanchez-Cespedes M, Herman JG, Esteller M (2002). Germ-line variants in methyl-group metabolism genes and susceptibility to DNA methylation in normal tissues and human primary tumors. Cancer Res. 62(15): 4519-4524.

[28] Pelizzola M, Koga Y, Urban AE, Krauthammer M, Weissman S, Halaban R, Molinaro AA (2008). Medme: an experimental and analytical methodology for the estimation of dna methylation levels based on microarray derived medip-enrichment. Genome Res. 18(10):1652-1659.

[29] Prober JM, Trainor GL, Dam RJ, Hobbs FW, Robertson CW, Zagursky RJ, Cocuzza AJ, Jensen MA, Baumeister K (1987). A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. Science 238:336-341.

[30] R Development Core Team (2009). R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, URL http://www.R-project.org.

[31] Ronaghi M, Uhlén M, Nyrén P (1998). "A sequencing method based on real-time pyrophosphate". Science 281: 363-365.

[32] Robertson KD, Uzvolgyi E, Liang G, Talmadge C, Sumegi J, Gonzales FA, Jones PA (1999). The human DNA methyltransferases (DNMTs) 1, 3a and 3b: coordinate mRNA expression in normal tissues and overexpression in tumors. Nucleic Acids Res. 27(11):2291-2298.

[33] Robertson KD, Wolffle AP (2000). DNA methylation and human disease. Nat Rev Genet. 1:11-19.

[34] Robertson KD (2005). DNA methylation and human disease. Nat Rev Genet. 6(8):597-610.

[35] Rodriguez J, Frigola J, Vendrell E, Risques RA, Fraga MF, Morales C, Moreno V, Esteller M, Capellà G, Ribas M, Peinado MA (2006). Chromosomal instability correlates with genome-wide DNA demethylation in human primary colorectal cancers. Cancer Res. 66(17): 8462-9468.

[36] Sanger F, Coulson AR (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J. Mol. Biol. 94, 441–448.

[37] Sanger F, Nicklen S, Coulson AR (1977). DNA sequencing with chainterminating inhibitors. Proceedings of the National Academy of Sciences of the United States of America 74:5463–5467.

[38] Schena M, Shalon D, Davis RW, Brown PO (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270:467-470.

[39] Staub E, Gröne J, Mennerich D, Röpcke S, Klamann I, Hinzmann B, Castanos-Velez E, Mann B, Pilarsky C, Brümmendorf T, Weber B, Buhr HJ, Rosenthal A (2006). A genome-wide map of aberrantly expressed chromosomal islands in colorectal cancer. Mol Cancer 5: 37.

[40] Shendure, J., Porreca, G. J., Reppas, N. B., Lin, X., McCutcheon, J. P., Rosenbaum, A. M., Wang, M. D., Zhang, K., Mitra, R. D. and Church, G. M. (2005). Accurate multiplex polony sequencing of an evolved bacterial genome. Science 309: 1728-1732.

[41] Xu GL, Bestor TH, Bourc'his D, Hsieh CL, Tommerup N, Bugge M, Hulten M, Qu X, Russo JJ, Viegas- Pequignot E (1999). Chromosome instability and immunodeficiency syndrome caused by mutations in a DNA methyltransferase gene. Nature 402(6758):187-191.

[42] Weber M, Davies JJ, Wittig D, Oakeley EJ, Haase M, Lam WL, Schuebeler D (2005). Chromosome-wide and promoter-specifc analyses identify sites of differential dna methylation in normal and transformed human cells. Nat Genet. 37(8):853-862.

7. Abkürzungsverzeichnis

ATP	Adenosintriphosphat
APS	Adenosin-5’-phosphosulfat
b	Bingröße
bc	Anzahl der Bins
BC	Vektor der kumulativen Summen der Anzahl der Bins
BED	Browser Extensible Display
bp	Basenpaare
C	Vektor mit Chromosomlängen
CDKN2A	Cyclin-dependent kinase inhibitor 2A
cDNA	komplementäre Desoxyribonukleinsäure
CDO1	Cysteine dioxygenase type I
ChIP	Chromatin Immunpräzipitation
CpG	benachbarte Cytosine (C) und Guanine (G)
cRNA	komplementäre Ribonukleinsäure
dATP	2'-Desoxyadenosin-5'-triphosphat
dCTP	2'-Desoxycytidin-5'-triphosphat
ddNTP	Didesoxyribonukleosid-Triphosphate
dGTP	2'-Desoxyguanosin-5'-triphosphat
DNA	Desoxyribonukleinsäure
DNMT	DNA Methyl Transferasen
DMR	Differentiell methylierte Regionen
dNTP	2'-Desoxyribonukleosid-5'-triphosphat
dTTP	2'-Desoxythymidin-5'-triphosphat
G	Genomvektor
kb	Kilobasen
MeDIP	methylierte DNA Immunpräzipitation
mRNA	Messenger- Ribonukleinsäure
mRPM	mittlere Methylierungswerte
nf	Gesamtanzahl der Fenster eines Chromosomen
PPi	Pyrophosphat
ratio	Verhältnis
RNA	Ribonukleinsäure

ROI	Region of Interest
RPM	Methylierungswerte in reads per million
SEPT9	Septin 9
SHOX2	short stature homeobox 2
s_1	Index der Startposition
s_2	Index der Stopposition
si	Anzahl der RPM-Werte in einem sich überschneidenden Bereich
Stepsize	Schrittweit
Vk	Varianzkoeffizient
W	Subvektor
WIG	Wiggle-Format

Selbstständigkeitserklärung

Hiermit erkläre ich, die vorliegende Arbeit selbstständig ohne fremde Hilfe verfasst und nur die angegebene Literatur und Hilfsmittel verwendet zu haben.

Jörn Dietrich
Berlin, den 28. Mai 2010