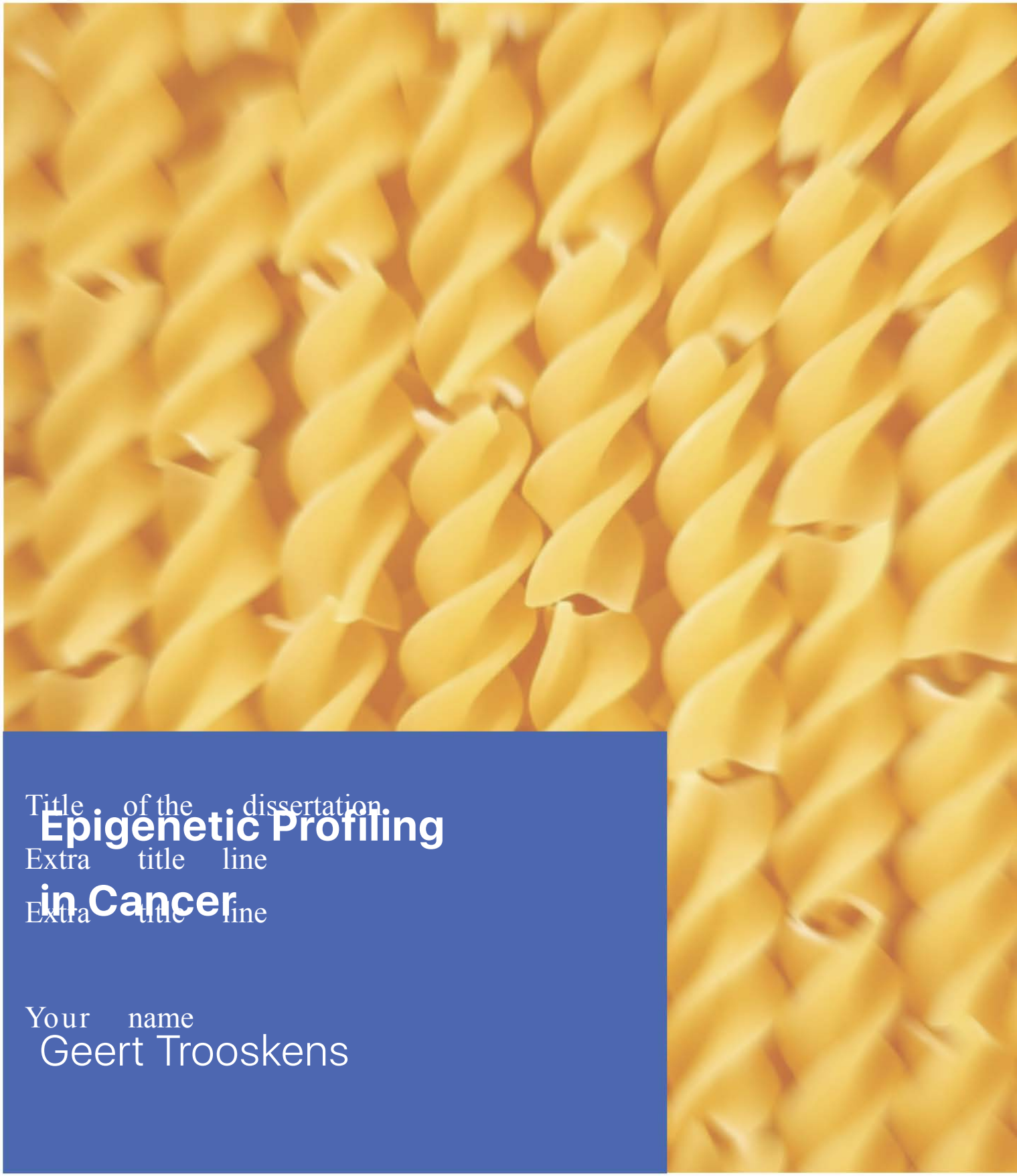
FACULTY OF
BIOSCIENCE ENGINEERING



Title of the dissertation

Epigenetic Profiling in Cancer

Extra title line

Extra title line

Your name
Geert Trooskens

GHENT
UNIVERSITY

To my girls,

Anke and Nette

"I love deadlines.
I like the whooshing sound they make as they fly by."
– Douglas Adams, *Hitchhikers Guide to the Galaxy*

Promoter: **Prof. dr. ir. Wim Van Criekinge**

Ghent University, Faculty of Bioscience Engineering,
Department of Mathematical Modelling, Statistics and
Bioinformatics, BIOBIX group

Copromoter: **Prof. dr. ir. Tim De Meyer**

Ghent University, Faculty of Bioscience Engineering,
Department of Mathematical Modelling, Statistics and
Bioinformatics, Biostatistics group (BioStat)

Dean: **Prof. dr. ir. Marc Van Meirvenne**

Rector: **Prof. dr. ir. Rik Van de Walle**

Vice-rector: **Prof. dr. Mieke Van Herreweghe**

ir. Geert Trooskens

Epigenetic Profiling in Cancer

FACULTY OF BIOSCIENCE ENGINEERING

Thesis submitted in fulfilment
of the requirements for the degree of
Doctor (PhD) of Applied Biological Sciences
Academic year 2017-2018

Epigenetisch Profileren in Kanker

Cover Illustration: Fusilli, The Italian version of DNA.

Cite As: Geert Trooskens (2017), Epigenetic Profiling in Cancer. Ghent University.

ISBN: 978-94-6357-045-9

Reprinting costs paid by MDxHealth SA.

The Promoters:

Prof. dr. ir. Wim Van Criekinge

Prof. dr. ir. tim De Meyer

The Author:

ir. Geert Trooskens

Members of the examination committee

Prof. dr. Godelieve Gheysen (Chair)
Department of Molecular biotechnology, Faculty of Bioscience Engineering, Ghent University, Belgium

Prof. dr. Tom Desmet (Secretary)
Department of Biochemical and microbial technology, Faculty of Bioscience Engineering, Ghent University, Belgium

Prof. dr. ir. Wim Van Criekinge (Promoter)
Department of Mathematical Modelling, Statistics and Bioinformatics, Faculty of Bioscience Engineering, Ghent University, Belgium

Prof. dr. ir. Tim De Meyer (Copromoter)
Department of Mathematical Modelling, Statistics and Bioinformatics, Faculty of Bioscience Engineering, Ghent University, Belgium

dr. Jan Groen
CEO, MDxHealth, Nijmegen, The Netherlands

Prof. dr. Jack A. Schalken
Center for Molecular Life Sciences, Faculty of Medical Sciences, Radboud University, Nijmegen, The Netherlands

Prof. dr. ir. Jo Vandesompele
Department of Pediatrics and medical genetics, Faculty of Medicine and Health Sciences, Ghent University, Belgium

Prof. dr. Manon Van Engeland
Department of Pathology, GROW-School for Oncology and Developmental Biology, Maastricht University Medical Center, The Netherlands

Prof. dr. Filip Van Nieuwerburgh
Department of Pharmaceutics, Faculty of Pharmaceutical Sciences, Ghent University, Belgium

Contents

Woord Vooraf

Het heeft wat voeten in de aarde gehad, maar als je dit leest is het mij gelukt om mijn boekje af te maken.

Tijdens mijn zeven jaar doctoraat heb ik epigenetica zien evolueren van een groepje gemotiveerde wetenschappers tot een gerespecteerd onderzoeksgebied. Hiermee is ook de consensus gegroeid dat milieu en levenstijl op een directe wijze met je genoom kunnen interageren.

Grote 'believer' in de epigenetica van het eerste uur is Wim Van Criekinge, mijn promoter, huidige baas en hoofd van onze vakgroep. Zonder zijn geduld en de vrijheid die hij mij gaf om mijn ding te doen zou dit waarschijnlijk nooit gelukt zijn. Ook vele dank aan mijn andere promoter; Tim de Meyer. Bij hem kon ik altijd terecht voor de meer exotische statistische vragen. Dank u Gerben, Martijn, Maté, Joachim, Jeroen, Jeroen, Sandra, Elvis, Klaas en Simon voor de goede sfeer op bureau. Alsook de collega's van NXTGNT en MDXHealth; Johan, Hendrik, An, Anna, Sarah en Ellen. Ik wil ook de mensen buiten UGent bedanken waarmee ik nauw heb samengewerkt: Malav Trivedi, Nathaniel Hodgson, en Richard Deth van Northwestern University in Boston; Linda Bosch, Beatriz Carvalho, Evert van den Broek en Gerrit Meijer van het Nationaal Kanker Instituut in Nederland; Leander Van Neste, Alexander Koch en Manon van Engeland in het Maastricht University Medical Center.

Mijn familie wil ik bedanken voor mijn (epi)genetische erfenis en hun onvoorwaardelijke steun. Niet in het allerminst mijn grootvader Jan Andriessens, Papy, die mij er elke keer aan herinnerde dat mijn doctoraat toch wel zeer belangrijk was. Mijn moeder wil ik bedanken omdat ik zonder haar waarschijnlijk zelfs mijn universteitsdiploma niet had gehaald. Merci! Tot slot nog een staande ovatie voor mijn steun en toeverlaat Anke.

Abbreviations

5-FU Capecitabine
ADT Androgen-deprivation therapy
AUC Area under the curve
CAP Capecitabine
CAPIRI Capecitabine plus irinotecan
CAPOX Capecitabine plus oxaliplatin
CI Confidence interval
CRC Colorectal cancer
ChIP Chromatin immunoprecipitation
DNA Deoxyribonucleic acid
EGFR Epidermal growth factor receptor
FF Fresh Frozen
FOLFIRI 5-FU plus Irinotecan
HDAC Histone deacetylase
HR Hazard ratio
GBM Glioblastoma multiforme
IRI Irinotecan
LDH Lactate dehydrogenase
lncRNA Long noncoding RNA
miRNA Micro RNA
MBD Methyl binding domain
MC Methylation cores
MeDip anti-5-methyl cytosine antibody immunoprecipitatie
MHM Map of the human methylome
MSI Mircosatellite instability
MSP Methylation specific PCR
MZ Monozygotic
ncRNA Noncoding RNA
NGS Next generation sequencing
OS Overall survival
PFS Progression free survival
PSA Prostate specific antigen
qMSP Quantitative methylation specific PCR
SNP Single nucleotide polymorphism
RISC RNA induced silencing complex
RNA Ribonucleic acid
RNAi RNA interference
ROC Receiver Operating Curve
RRBS Reduced representation bisulphite sequencing
rRNA Ribosomal RNA
RT Radiotherapy
TMZ Temozolomide
TSS Transcription start site
VEGF Vascular endothelial growth factor
WGBS Whole genome bisulphite sequencing

Introduction

A decade ago, the Human Genome Project completed the ambitious effort to read the 3 billion DNA letters of genetic information found in most human cells. It provided the blueprint for human life, an incredible human achievement comparable to the landing on the moon.
We should recognize that even after 10 years, we are only at the early stages of interpreting that sequence. Decades from now we will still be interpreting, and reinterpreting, it.
What we did already find however, is that when we compared the human genome to other mammals, some regions were very conserved. In other words: across tens of millions of years of evolutionary time, the sequences did not change that much at all. Highly evolutionary conserved sequences almost certainly point to important functional sequences, since these are things that life doesn't want to change because they have some vital fundamental function. Scientists thought the majority of those most conserved regions were going to be in the protein-coding genes, the parts of the genome that are the recipe for our functional proteins. It turns out, the majority of the most highly conserved regions are not in protein coding regions; they are located outside of these regions...

So what is going on here? We found out that a lot of these conserved regions are basically circuit switches, like dimmer switches for a light, that determine where, when and how much a gene gets turned on. If a protein coding gene is turned on, it is transcribed into RNA molecules and those in turn get translated into working functional proteins.
This regulation of genes is much more complicated in humans than it is in lower organisms like yeasts or worms. Our biological complexity is not so much in our gene number, it is in the complex framework of switches, dimmers, amplifiers and silencers, that regulate where, when, and how much genes get turned on.

It is this process that we call Epigenetics, and it has added a dazzling complexity to modern biology. Trying to make sense of these huge amounts of biological data, the field is rapidly acquiring the character of a data science. Trillions of data points on genes, proteins and other molecules are stored in large files and systematically studied using data-analytical approaches. Diet, lifestyle, stress factors and environmental exposures are able to alter your epigenetic switches and dimmers, effectively regulating interactions between genes and environment.
Interestingly, it was found that disruption of these epigenetic processes can lead to abnormal cellular behaviour due to altered gene functions. The initiation and progression of cancer, traditionally seen as a genetic disease, is now realized to involve epigenetic abnormalities along with genetic alterations.

During my PhD, I focused on the analysis, interpretation and visualization of epigenetic data, being both a witness and participator in this fast growing branch of biology.

Nederlandse Samenvatting

Nederlandse Titel: Epigenetisch Profileren in Kanker.

In het kader van dit doctoraat werd onderzoek verricht naar de epigenetische veranderingen in cellen en hun invloed op ziekten zoals kanker. In het eerste deel van dit onderzoek werd op genoomwijde schaal gezocht naar epigenetische biomerkers die specifiek aanwezig zijn in bepaalde kankers. In een tweede deel van dit onderzoek werd nagegaan of deze epigenetische biomerkers konden worden gebruikt in een klinische setting om een vroege diagnose of prognose te kunnen stellen.

In hoofdstuk 1 wordt een overzicht gegeven van de epigenetische mechanismen. Er wordt verder ingegaan op de rol van epigenetica in kanker en hoe deze kan gebruikt worden voor diagnose en behandeling van kanker. In de laatste sectie wordt een uitgebreid overzicht gegeven van de huidige epigenetische opsporingstechnieken, waarvan sommigen uitvoerig in dit doctoraat zijn gebruikt zijn om nieuwe biomerkers op te sporen.

Hoofdstuk 2 omvat de epigenetische onderzoeksmethoden die in het kader van dit doctoraat gebruikt en ontwikkeld zijn om DNA methylatie te bepalen in het menselijke genoom. In een eerste fase hebben we genome-wide naar epigenetische merkers gezocht door de gemethyleerde regio's eerst aan te rijken met een methylatie specifiek bindend domein en vervolgens te sequeneren. Dit laat toe om op een kost-efficiënte manier een volledig methylatie profiel te genereren van een staal. De geïdentificeerde regio's worden dan in een volgende fase gevalideerd door bisulfiet behandeling van het DNA gevolgd door amplificatie van de doelwit sequentie via PCR.

Hoofdstuk 3 is een selectie van de papers die uit dit onderzoek zijn voortgekomen. Meer specifiek is nagegaan of promotor methylatie van het WRN gen en het DCR1 gen predictief is voor de behandeling van colorectale kanker patiënten met irinotecan chemotherapie. Een ander onderzoek had als doel om de predictieve waarde van MGMT promotor methylatie te verhogen voor glioblastoma patiënten die een temozolomide chemotherapy krijgen. Hierbij werd de promotor regio gesequeneerd om zo een hoger resolutie profiel van de DNA methylatie te bekomen in vergelijking met de huidige methylatie specifieke PCR (MSP) standaard.

Part I

EPIGENETICS

1

Nature-Nurture: Genetics and Epigenetics

1.1 Situation

1.1.1 Genetics

Genetics is the study of genes, heredity, and genetic variation in living organisms. It seeks to understand the process of trait inheritance from parents to offspring.

The founder of the modern science of genetics is Gregor Mendel, a late 19th-century Czech scientist. Mendel studied 'trait inheritance,' patterns in the way traits were handed down from parents to offspring, through experiments in his garden. He observed that organisms (pea plants) inherit traits by way of discrete "units of inheritance". This term, still used today, is a somewhat ambiguous definition of what is referred to as a gene.

Trait inheritance and molecular inheritance mechanisms of genes are still a primary principle of genetics in the 21st century, but modern genetics has expanded beyond inheritance to studying the function and behaviour of genes. Gene structure and function, variation, and distribution are studied within the context of the cell, the organism (e.g. dominance) and within the context of a population. Genetics has given rise to a number of sub-fields including epigenetics and population genetics.

1.1.2 Epigenetics

Epigenetics is the study of cellular and physiological modifications that are heritable by daughter cells. These alterations are not coded in the cells

DNA but are stable, long-term alterations in the transcriptional potential of a cell. These alterations may or may not be heritable to the offspring, although the use of the term epigenetic to describe processes that are not heritable is controversial.

Unlike simple genetics based on changes to the DNA sequence (the genotype), epigenetic studies the changes in gene expression or cellular phenotype that have other causes dan the genotype.

The term Epigenetics was first used in 1942 by Conrad H. Waddington, a British developmental biologist, paleontologist, geneticist, embryologist and philosopher who laid the foundations for systems biology. He used the term as a portmanteau of the words epigenesis and genetics. The Waddingtonian equation holds that epigenesis + genetics = epigenetics, and refers in retrospect to the debate on epigenesis versus preformationism in neoclassical embryology. Whereas Waddington actualised this debate by linking epigenesis to developmental biology and preformation to genetics, thereby stressing the importance of genetic action in causal embryology. Todays epigenetics increasingly broadens biological reasoning in terms of genes only, as it expands the gene-centric view in biology by introducing a flexible and pragmatically oriented hierarchy of crucial genomic contexts. [1]

1.2 Genetics

1.2.1 Classical Genetics

Discrete inheritance and Mendel's laws

Discrete inheritance in organisms occurs by passing discrete heritable units, called genes, from parents to progeny. The first evidence came from Gregor Mendel, who studied the segregation of heritable traits in pea plants. [2] While studying the trait for flower color, Mendel observed that the flowers of each pea plant were either purple or white, but nothing in between. These discrete variations of the same gene are called alleles.

In the case of diploid species like pea plants, each individual has two copies of each gene, one copy inherited from each parent. Multiple species, including humans, have this pattern of inheritance. An organism with two copies of the same allele of a given gene is called homozygous at that gene locus, while an organism with two different alleles of a given gene is called heterozygous. The underlying particular set of alleles for a

gene in one organism is called a genotype, while the observable trait is called a phenotype. In a heterozygous genotype there is often one allele that is observable in the phenotype, this is called the dominant allele, while the other allele is called recessive as its properties are not observed in the presence of the dominant allele. Some alleles dont have (a complete) dominance over the other, resulting in an intermediate phenotype. A situation were both alleles are expressed is called co-dominance.

After sexual reproduction the offspring receives randomly one out of the two alleles from each parent. The rules of discrete inheritance and the segregation of alleles are collectively known as Mendel's first law or the Law of Segregation

Multiple Genes interactions or Epistasis

Epistasis describes how the interactions between genes can affect phenotypes. Today, scientists know that Mendel's predictions about inheritance depended on the genes he chose to study. Specifically, Mendel carefully selected seven unlinked genes that affected seven different traits. However, unlike the phenotypes that Mendel considered, the majority of phenotypes are affected by more than one gene. Most of the characteristics of organisms are (much) more complex than the characteristics that Mendel studied, and epistasis is one source of this complexity. Epistasis can occur in a variety of different ways and result in a variety of different phenotypic ratios. Beyond epistasis, gene-environment interactions like epigenetics further increase the variety of phenotypes we see around us each day.

Epistasis is currently a topic of interest in molecular and quantitative genetics: The search for loci linked to complex diseases such as diabetes, asthma, hypertension and multiple sclerosis has, to date, been less successful than for simple Mendelian disorders. Complicating factors such as an increased number of contributing loci and susceptibility alleles, incomplete penetrance, and contributing environmental effects are probably the reason why these loci are not found yet.

Epistasis is a particular cause for concern for complex traits: If the effect of one locus is altered or masked by another locus, the power to detect the first locus is likely to be reduced and the discovery of the effects at the two loci will be hindered by their interaction. Logically, further complications can be expected for traits where more than two loci are involved. [3]

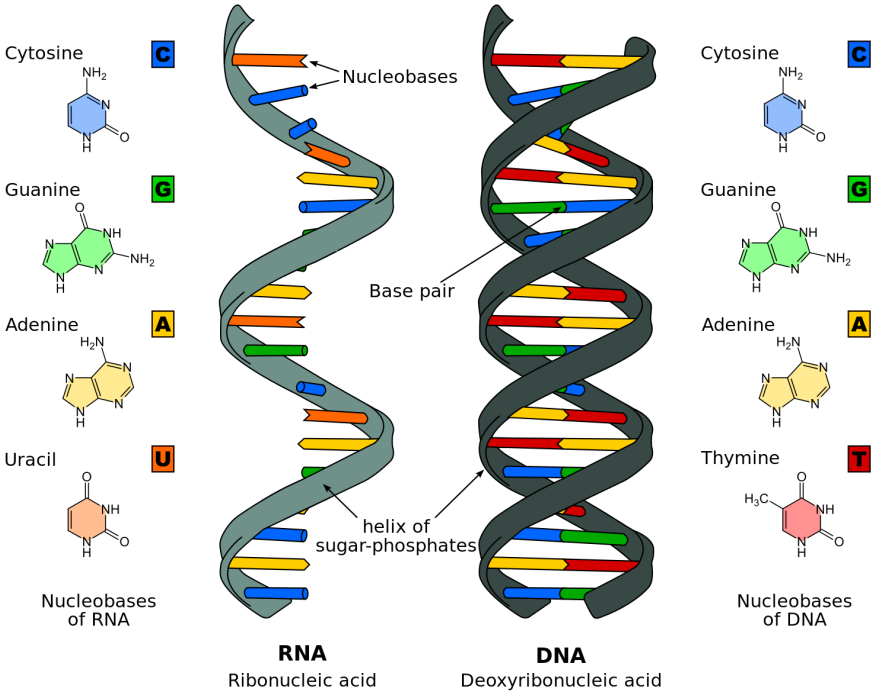


Figure 1.1: The structure of DNA and RNA.

1.2.2 Molecular Genetics

Early experiments by geneticists Hershey and Chase discovered that the transmission of traits was due to a substance called Deoxyribonucleic Acid (DNA). It was determined in 1953 by the famous discovery of Watson and Crick that the method by which DNA carry's and transmits information about the organism was through its sequence of nitrogenous bases:

The structure of DNA is a double helix where Adenine is the complement to Thymine and Guanine is complement to Cytosine (Figure 1.1). The sequence of these bases along a strand of DNA determines the function of the DNA at that location or locus. One of the primary purposes of DNA, therefore, is to provide the genetic blueprint for proteins to be made. This flow of genetic information within a biological system is also called the central dogma (Figure 1.2). The first step in this process is the transcription where a DNA Polymerase protein that binds to the DNA and transcribes a copy of the template DNA to RNA called mRNA. The sequence of the mRNA strand gets interpreted to form proteins, biological compounds that act as the building blocks of most cellular function. This process is called Translation, and occurs at ribosomal proteins. Proteins are made up of structures called Amino Acids bonded together in a sequence specified by the mRNA strand. Every 3 nucleotides (nitrogenous

The sequence of the mRNA strand gets interpreted to form proteins, biological compounds that act as the building blocks of most cellular function. This process is called Translation, and occurs at ribosomal proteins. Proteins are made up of structures called Amino Acids bonded together in a sequence specified by the mRNA strand. Every 3 nucleotides (nitrogenous base + backbone), the ribosome will add a specific Amino Acid to the protein being synthesised. These proteins ultimately mediate most cellular functions and determine most of the features that are expressed in organisms. This is the central dogma of molecular biology, stating that DNA provides the sequence that determines the proteins, and ultimately, the traits expressed by the organism.

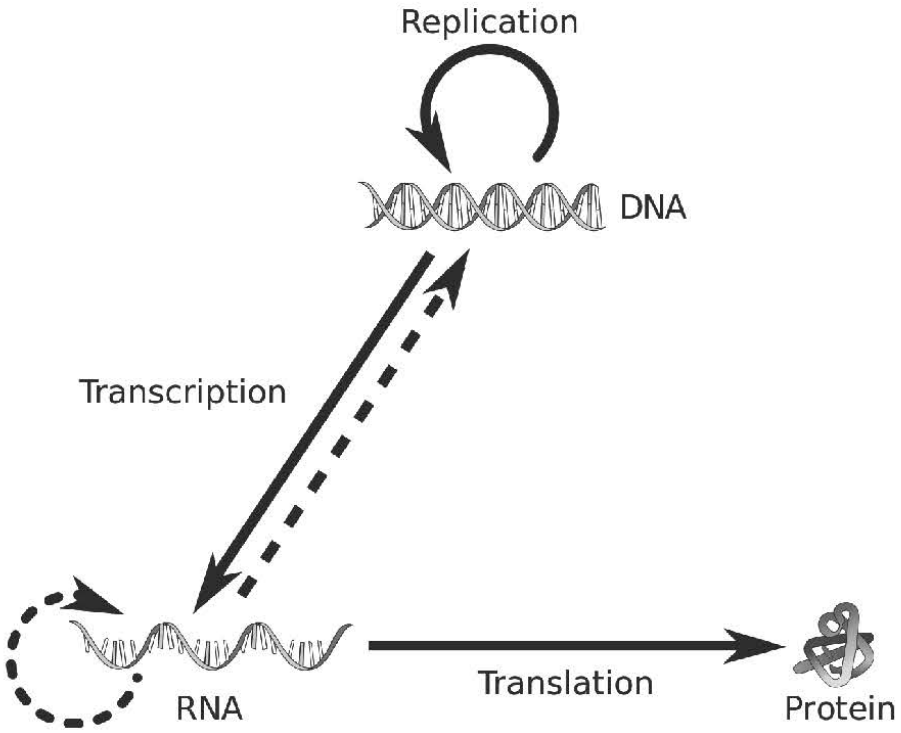


FIGURE 1.2: THE CENTRAL DOGMA OF MOLECULAR BIOLOGY: DNA IS TRANSCRIBED TO RNA AND DNA IS SUBSEQUENTLY TRANSLATED INTO (FUNCTIONAL) PROTEINS.

Figure 1.2: The central dogma of molecular biology deals with the detailed residue-by-residue transfer of sequential information. It states that such information cannot be transferred back from protein to either protein or nucleic acid. - Francis Crick [4]

Cells divide, grow and differentiate, which is the basis behind multicellular development of organisms. The blueprint of every cell, the organism's genome with contain all the DNA

4

base + backbone), the ribosome will add a specific Amino Acid to the protein being synthesized. These proteins ultimately mediate most cellular functions and determine the features that are expressed in organisms.

Cells divide, grow and differentiate, which is the basis behind multicellular development of organisms. The blueprint of every cell, the organism's genome that contains all the DNA information, must replicate every time the cell divides, so that all newly formed cells have a functional copy to operate the cell with. In most cells, the process of cell division includes a process of genome duplication and division called Mitosis.0 During this process, proteins unwind the double stranded DNA into two complementary single stranded DNA molecules, and add complementary nucleotides to each strand. This results in the synthesis of two identical strands of DNA from the original strand of DNA. After synthesis, the DNA condenses into chromosomes, each composed of two identical sister chromatids, from which after cell division one ends up into the daughter cell. The result is a daughter cell with a copy of the original genome. The protein responsible for most of the DNA replication activity (DNA Poly-

merase II) has very important implications in the field of genetics for use in PCR and DNA sequencing. Most organisms of the same species have an almost identical genetic code. In between two non-related humans, more then 99% of their DNA is identical. The basis for the less then 1% difference is the process of genetic mutation in germ line cells, which are passed on from one generation to the next. Mutations include point mutations of nucleotides, insertions, deletions or chromosomal disjunction. Point mutations that are passed on from generation to generation are called Single Nucleotide Polymorphisms. Mutations (SNPs) that aren't fatal and can be passed on from generation to generation often gets dispersed in the population. Accumulation of different random mutations that are passed on to future generations in forms such as SNPs account for part of the variation in physical appearance, biological physiology, and health between different people. These SNPs give genetic research its purpose, as many of these SNPs often result in different metabolism and susceptibility to diseases such as alzheimer or colon cancer. Laboratory techniques have been developed, and in combination with the use of computers, have allowed geneticists to undertake large-scale projects to find the genetic basis of many human diseases. [5]

1.3 Epigenetics

1.3.1 Introduction

Historically, the term "epigenetics" was used to describe a collection of diverse phenomena that could not be explained by the standard genetic principles. Over the years, numerous biological phenomena, some considered bizarre and inexplicable, have been lumped into the category of epigenetics.

Epigenetics, in a broad sense, is a bridge between the genotype and phenotype. It changes the final outcome of a genetic locus without changing the underlying DNA sequence.

For example, Monozygotic (MZ) or identical twins occur when a single egg is fertilised to form one zygote (hence, "monozygotic") which then divides into two separate embryos. Even though these two individuals contain almost identical genomes and thus share an identical genotypes, their phenotypes can differ considerably due to their environments. More specifically, epigenetics may be defined as the study of potentially stable and, ideally, heritable changes in gene expression or cellular phenotype

complex of DNA and its closely associated proteins. It provides an attractive candidate for haping the features of a cell's epigenetic landscape like the one Conrad Waddington first oposed in 1957 (Figure 1.3)

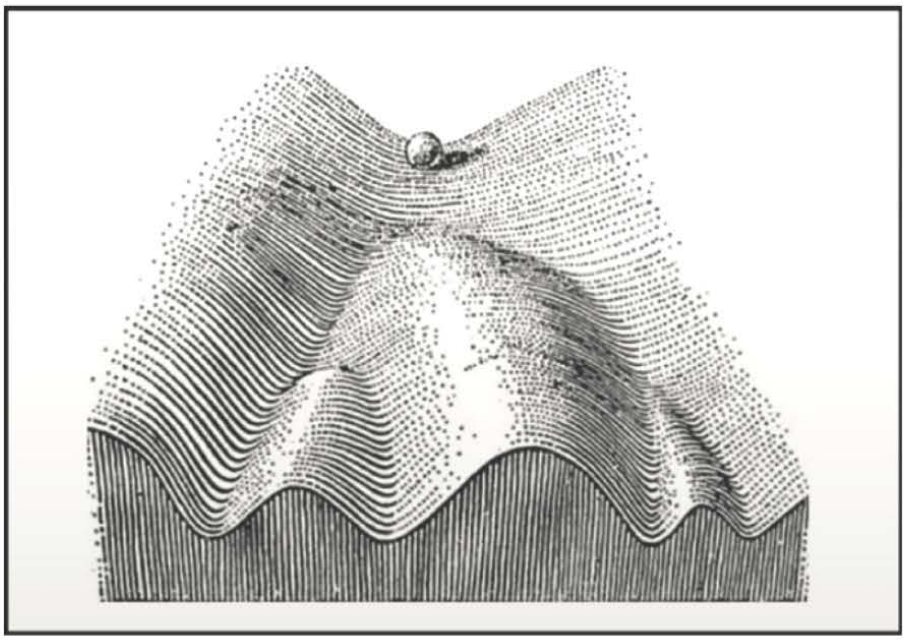


Figure 1.3: Waddingtons's classical epigenetic landscape

that occurs without inducing changes in the DNA code. [6]

1.3.2 Epigenetic Mechanisms

Much of todays epigenetic research is performed within the study of covalent and non-covalent modifications of DNA and the proteins that bind and condense the DNA (histones), especially on the mechanisms by which such modifications influence overall chromatin structure. Chromatin is the complex of DNA and its closely associated proteins. It provides an attractive candidate for shaping the features of a cells epigenetic landscape like the one Conrad Waddington first proposed in 1957 (Figure 1.3).

DNA (Hydroxy-)Methylation

DNA methylation was the first epigenetic mark to be discovered. It provides a stable, heritable, and critical component of epigenetic regulation. It is also the best characterised chemical modification of chromatin: In 5-methylcytosine, a methyl group, is attached to the 5th atom in the 6-atom ring (counting counterclockwise from the NH nitrogen at the six o'clock position, not the 2 o'clock). This methyl group distinguishes 5-methylcytosine from cytosine. 5-Methylcytosine is an epigenetic modification formed by the action of DNA methyltransferase proteins (Figure

DNA hydroxymethylation, caused by oxidation of 5-methylcytosine through the TET family of enzymes, was further discovered to be involved in switching genes on and off. DNA methylation plays an important role in normal human development and is associated with the regulation of gene expression, tumorigenesis, and other genetic and epigenetic diseases.{Goldberg:2007bq}

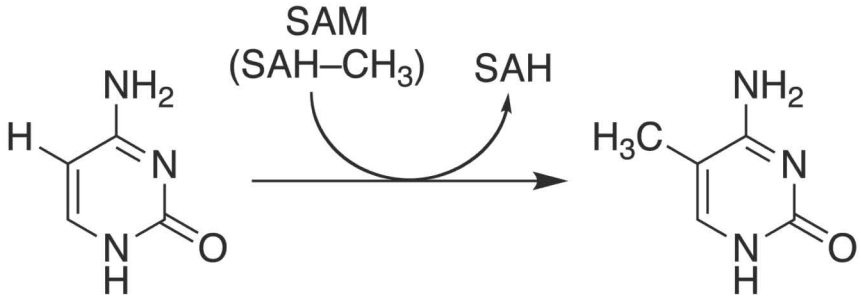


FIGURE 1.4: DNA METHYLTRANSFERASES ARE ENZYMES THAT CATALYZE THE TRANSFER OF A METHYL GROUP TO DNA. THE KNOWN DNA METHYLTRANSFERASES USE S-ADENOSYL METHIONINE (SAM) AS THE METHYL DONOR.

Figure 1.4: DNA Methyltransferases are enzymes that catalyze the transfer of a methyl group to DNA. the known DNA-methyltranfgerases use S-Adenosyl Methionine (SAM) as the methyl donor.

Maintenance methyltransferases add methyl groups to hemi-methylated DNA during DNA replication, whereas de novo DNA methyltransferases add methyl groups after DNA replication. The formation of dense heterochromatin is mediated in part by DNA methylation in combination with RNA and histone modifications resulting in silent chromatin. DNA methylation plays a crucial role in many cellular processes including gene regulations, cell differentiation, silencing of repetitive and centromeric sequences, X chromosome inactivation in female mammals and mammalian imprinting.

1.4). While spontaneous deamination of cytosine forms uracil, which is recognized and removed by DNA repair enzymes, deamination of 5-methylcytosine forms thymine. This conversion of a DNA base from cytosine (C) to thymine (T) can result in a transition mutation.

DNA hydroxymethylation, caused by oxidation of 5-methylcytosine through the TET family of enzymes, was further discovered to be involved in switching genes on and off. DNA methylation plays an important role in normal human development and is associated with the regulation of gene expression, tumorigenesis, and other genetic and epigenetic diseases. [7]

Maintenance methyltransferases add methyl groups to hemi-methylated DNA during DNA replication, whereas de novo DNA methyltransferases add methyl groups after DNA replication. The formation of dense heterochromatin is mediated in part by DNA methylation in combination with RNA and histone modifications resulting in silent chromatin. DNA methylation plays a crucial role in many cellular processes including gene regulations, cell differentiation, silencing of repetitive and centromeric sequences, X chromosome inactivation in female mammals and mammalian imprinting.

1.3.2.2 HISTONE MODIFICATIONS

Histones are highly alkaline proteins found in the nucleus of eukaryotic cells. Their function is to package and order the DNA into structural units called nucleosomes. They are the major protein components of chromatin, acting as spools around which DNA winds. Histones contain modifications that play a role in gene regulation. They can condense the DNA by winding it up around them. In order to fit the DNA enough packing into the nucleus. Each human cell has about 1.8 meters of DNA, wounded up on the histones has about 30 micrometers (0.03 mm) of

Other DNA Modifications

Next to 5-methylcytosine (5mC) and 5-hydroxymethylcytosine (5hmC), new research is continuing to reveal numerous gene-regulatory effects of covalent DNA modifications. An increasing number of studies demonstrate the importance of other cytosine modifications, such as 5-formylcytosine (5fC), and 5-carboxylcytosine (5caC). More recently, three analogous

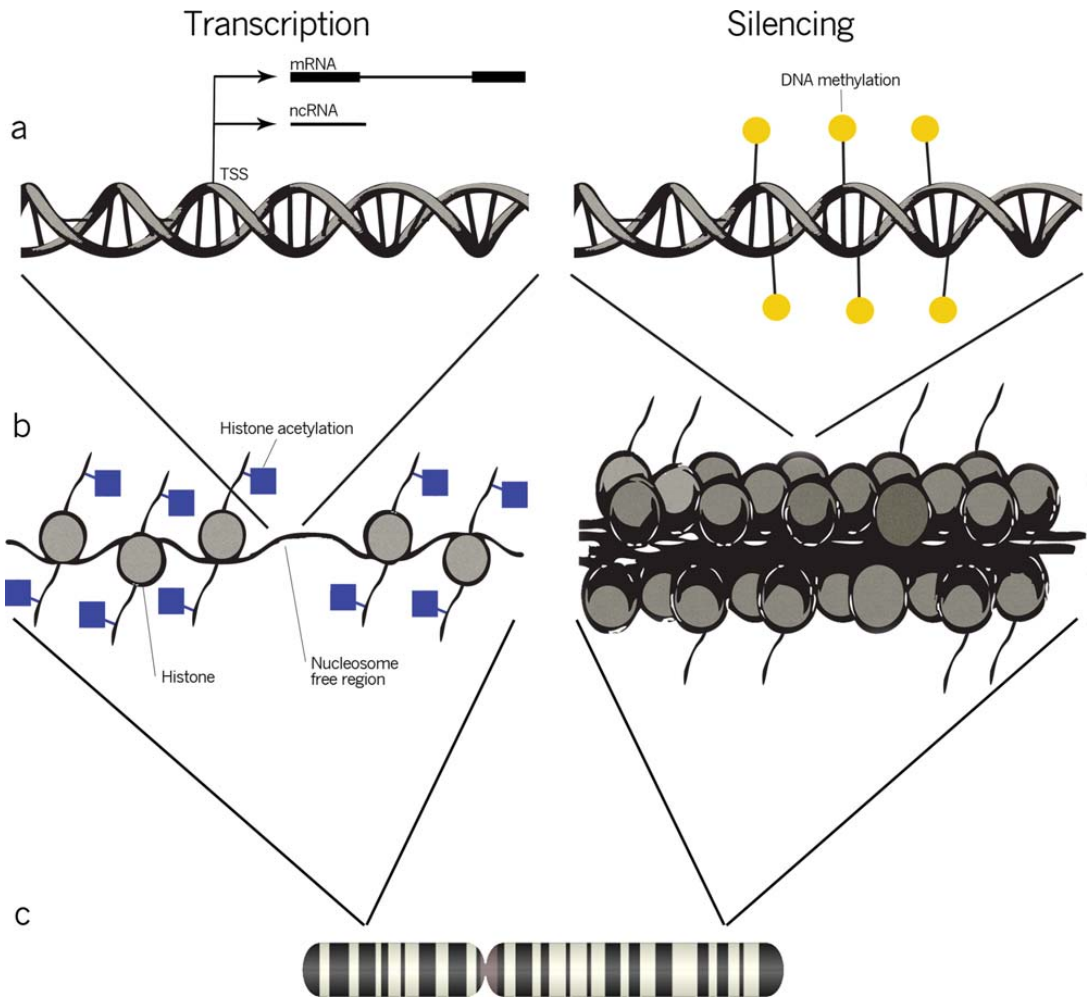


Figure 1.5: Overview of epigenetic features. Each chromosome (panel c) consists of both condensed and open chromatin regions (panel b), with different histone modifications present. Loose regions are, for example, characterized by histone lysine acetylation and the possibility of gene expression. Nucleosome (re)positioning results in nucleosome free regions, for example, at the transcription start site (TSS) (panel b), which typically associated with transcriptional silencing. Nucleosomes consist of 147-bp-long DNA stretches wrapped around an octamer of histone proteins, and are connected through linker DNA.. Promoter regions of transcriptionally silenced genes are typically densely packed without nucleosome free regions, lack histone lysine acetylation (panel b), and are often featured by DNA methylation (panel a).

modifications of thymine were found to occur in mammals and can now largely be sequenced. N6-methyladenine, better known as a RNA modification, has now been found in the DNA of multiple eukaryotes. [8] Future research needs to address the potential functions of these recently discovered DNA modifications, as they could play an important role in epigenetic regulation [9] . This diversity of DNA modifications and their potential for combinatorial interactions indicates that the epigenetic DNA code is substantially more complex than recently thought.

Histone Modifications

Histones are highly alkaline proteins found in the nucleus of eukaryotic cells. Their function is to package and order the DNA into structural units called nucleosomes. They are the major protein components of chromatin, acting as spools around which DNA winds. Histones contain modifications that play a role in gene regulation. They can condense the DNA by winding it up around them. It condenses the DNA enough that it fits into the nucleus. Each human cell has about 1.8 meters of DNA, wounded up on the histones it has about 90 micrometers (0.09 mm) of chromatin, which, when duplicated and condensed during mitosis, result in about 120 micrometers of chromosomes.

Each nucleosomal unit is formed by wrapping approximately 146 base pairs of DNA around a histone octamer core particle containing one H3-H4 tetramer and two H2A-H2B dimers. The C-terminal histone-fold domains of core histones have similar conformations that are critical for the assembly of nucleosomes by mediating histone-histone and histone-DNA interactions. In contrast, the N-terminal tails of core histones are less structured and are not essential for maintaining the integrity of nucleosomes since removal of these tails by trypsin treatment does not diminish nucleosome stability. Instead, histone tails are thought to make secondary and more flexible contacts with DNA and adjacent nucleosomes that allow for dynamic changes in the accessibility of the underlying genome. These tails are subjected to a diverse set of post-translational modifications, such as methylation, acetylation and phosphorylation, which may modulate the contacts between histones and DNA.

Because histone modifications are reversible, they can act as flexible "on/off" switches that regulate gene expression and other DNA-related processes. Moreover, since the histone tail domains are highly accessible to the nuclear environment, they provide attractive targets for signal-

activated enzymes, and may function as molecular links between signal transduction and gene expression. [10]

Telomeres

Telomeres are repeated DNA patterns that form protective caps at the ends of eukaryotic chromosomes. These DNA structures are involved in the replication and stability of DNA molecules. They are made up of a repeating DNA sequence of six nucleotide bases (TTAGGG). Small numbers of these terminal TTAGGG sequences are lost from the tips of the chromosomes with cell division. Most of this loss is compensated by the addition of TTAGGG repeats by the telomerase enzyme but overall, the telomeres get shorter with time. Shortening telomeres lead to senescence, apoptosis, or oncogenic transformation of somatic cells, affecting the health and lifespan of an individual. Shorter telomeres have been associated with increased incidence of diseases and poor survival. The rate of telomere shortening can be either increased or decreased by (epi)genetic and specific lifestyle/environmental factors. Telomere length is influenced by genetic [11] and epigenetic factors [12]. Better choice of diet and activities has great potential to reduce the rate of telomere shortening or at least prevent excessive telomere attrition, leading to delayed onset of age-associated diseases and increased lifespan.

RNA Transcripts and their Encoded Proteins

Some genes can transcribe a product that regulates the activity of that gene. RNA can recruit chromatin modifying complexes and DNA methyltransferases to specific loci during differentiation and development. Different splice forms of RNA, or formation of double-stranded RNA (RNAi) can also induce epigenetic changes which can be passed on to descendants of the cell. [13] . A large amount of the RNA in the zygote originates from the mother during oogenesis or via nurse cells, resulting in maternal effect phenotypes. A smaller quantity of sperm RNA is transmitted from the father, but there is still evidence in male mice of alterations induced by RNA [14].

MicroRNAs

Non-coding RNAs (ncRNAs) compromise the transcribed RNA fraction that lack significant protein-coding capacity . MicroRNAs (miRNAs) are

15-22 nucleotide, short, non-coding RNAs that have emerged as critical regulators of gene expression. They affect a multitude of biological processes including cell proliferation, differentiation, survival and motility.

They are a class of evolutionally conserved, single-stranded, small (approximately 19 - 23 nucleotides), endogenously expressed, and non-protein-coding RNAs that act as posttranscriptional regulators of gene expression. The biogenesis of miRNAs is a multi-step process [15]: miRNAs are initially transcribed in the cell nucleus by RNA polymerase II to form primary miRNAs with lengths ranging from 1 to 3 kb. These primary miRNAs are cleaved in the nucleus by the RNase III enzyme Drosha and a doublestranded RNA-binding protein Pasha into pre-miRNAs of approximately 70 - 100 nucleotide-long stem-loop structures. These pre-miRNAs are then transported from the nucleus to the cytoplasm by Exportin where they are further cleaved into mature double-stranded miRNA:miRNA oligonucleotides of 15 to 22 bases by the RNase-III enzyme Dicer.

After strand separation, one of the double strands becomes a mature miRNA molecule incorporated into RNA-induced silencing complex (RISC). The RISC complex functions by perfectly or imperfectly matching with its complementary target mRNA, and induces target mRNA degradation, translational inhibition or sequestration of mRNA from the translational machinery. The translational inhibition can act as both oncogenes and tumor suppressors, highlighting their importance in human cancer. Therefore, miRNA expression profiles can be used as biomarkers for the onset of disease states and it is possible to use miRNAs in gene therapy for genetic disorders as well as potential drug targets [16].

Long non-coding RNA

Long non-coding RNAs (lncRNAs) compromise a wide range of the ncRNAs that are larger then 200 nucleotides. This somewhat arbitrary limit, based on technical aspects of RNA isolation methods, distinguishes lncRNAs from small ncRNAs such as miRNAs. lncRNAs are thought to encompass nearly 30,000 different transcripts in humans, representing a major part of the non-coding transcriptome. Many lncRNAs are lowly expressed [17] and they do not tend to contain conserved motifs, requiring sensitive technologies and explaining why lncRNAs had been thought to be mostly transcriptional noise until recently. lncRNAs can regulate gene expression at both the transcriptional and the post-transcriptional level in diverse cellular contexts and biological processes [18, 19]. lncR-

NAs are subdivided into classes based on their positional relationship to protein-coding genes and different mechanisms of processing [20].
In cancer, lncRNAs can work through several mechanisms such as chromatin remodeling, chromatin interactions, as competing endogenous RNAs and natural antisense transcripts [21]. At this writing, few lncRNAs have been characterized in detail. However, it is becoming clear that lncRNAs are important regulators in cellular, developmental and disease linked processes, adding yet another layer of complexity to our understanding of genomic regulation.

Prions

The word prion, first used in 1982 by Stanley B. Prusiner, is the result of a merger between the words protein and infection, in reference to it's ability for self-propagation by transmitting its conformation to other prions. A protein fold into a specific, three-dimensional architecture that is largely determined by their sequences of amino acids. Some proteins show a degree of structural flexibility that enables them to adapt their shape to perform multiple functions. When a protein misfolds and evades normal clearance pathways, a pathogenic process can ensue in which the protein aggregates progressively into intracellular and/or extracellular deposits. The consequence is a diverse group of disorders such as mad cow disease in cattle and Creutzfeldt-Jakob disease (CJD) seen in humans. Each of these diseases entails the aggregation of particular proteins in characteristic patterns and locations [22].

2

Cancer Epigenetics

2.1 Introduction

Cancer epigenetics is the study of epigenetic modifications in cancer cells. Epigenetic mechanisms are essential for normal development and cell maintenance. Disruption of epigenetic processes or epi-mutations can lead to altered gene functions and malignant cellular transformations. The initiation and progression of cancer, traditionally seen as a genetic disease, is actually a combination of epigenetic abnormalities along with genetic alterations. Tumour growth is positively correlated with tumour suppressor genes that become epigenetically silenced and activation of oncogenes. The epigenetic mechanisms involved include DNA methylation, histone modifications, nucleosome positioning and non-coding RNAs, specifically microRNA expression. Understanding these mechanisms holds great promise for advanced cancer treatments, early detection and prevention.

2.2 Reprogramming of the Epigenome in Cancer

The well defined global epigenetic patterns present in normal cells undergo extensive distortion in cancer. These epimutations, along with widespread genetic alterations, play an important role in cancer initiation and progression [23]. The cancer epigenome is characterized by global changes in DNA methylation and histone modification patterns as well as altered expression profiles of chromatin-modifying enzymes. These epigenetic changes result in global dysregulation of gene expression profiles leading to the development and progression of disease states. Epimutations can

lead to silencing of tumor suppressor genes independently and also in conjunction with genetic mutations or deletions; thus, serving as the second hit required for cancer initiation according to the two-hit model proposed by Alfred Knudson [24]. In addition to inactivating tumor suppressor genes, these epimutations can also promote tumorigenesis by activating oncogenes. The events that lead to initiation of these epigenetic abnormalities are still not fully understood. These epigenetic alterations are, like genetic mutations, mitotically heritable, which makes them ideal targets for selection in a rapidly growing cancer cell population.

2.3 The Cancer Stem Cell Model

(Epi-)genetic changes in cancer may involve the dysregulation of hundreds of genes during tumorigenesis. The mechanism by which a tumor cell accumulates such widespread (epi)genetic abnormalities during cancer development is still not fully understood. The selective advantage of these (epi)mutations during tumor progression is possible, but it is unlikely that the multitude of epigenetic alterations that reside in a cancer epigenome occur in a random fashion and then accumulate inside the tumor due to clonal selection. A more plausible explanation would be that the accumulation of such global epigenomic abnormalities arises from initial alterations in the central epigenetic control machinery, which occur at a very early stage of neoplastic evolution. Such initiating events can predispose tumor cells to gain further epimutations during tumor progression in a fashion similar to accumulation of the genetic alterations that occurs following defects in DNA repair machinery in cancer. The cancer stem cell model suggests that the epigenetic changes, which occur in normal stem or progenitor cells, are the earliest events in cancer initiation. The idea that these initial events occur in stem cell populations is supported by the common finding that epigenetic aberrations are some of the earliest events that occur in various types of cancer and also by the discovery that normal tissues have altered progenitor cells in cancer patients. [25]. It should be noted that stem cells also acquire genetic mutations during life, which can also contribute to abnormal stem cell behaviour and subsequent cancer development [26] .

2.4 Clinical Epigenetics: Diagnosis and Therapy

2.4.1 Epigenetics as a Diagnostic Tool in the Detection and Treatment of Diseases

Clinical epigenetics is the application of molecular biology techniques that can detect alterations in DNA methylation or histone modification to diagnose or study cancer epigenetics and conceptive epigenetics, cardiovascular epigenetics, allergy, immunology, development, aging, degenerative brain disorders, pathogen interaction and metabolic illness. [27] . It is a rapidly developing field, developing epigenetic assays as powerful early diagnostic, prognostic and predictive tools.
DNA methylation techniques account for the largest share of tests compared to other epigenetic mechanisms.

DNA methylation studies of cancerous tissues aim, apart from their contribution to a better understanding of the molecular mechanisms of cancer development and progression, mostly at the determination of markers suitable for diagnosis, prognosis and therapy response (pharmaco(epi)genomics) prediction or disease follow-up. Although a tumor itself is the major source of tumor DNA, acquiring DNA through a biopsy is invasive, risky and often not possible. Fortunately, scientists have discovered that dying tumor cells release small pieces of their DNA into the bloodstream. These pieces are called cell-free circulating tumor DNA (ctDNA). This provides an opportunity to test for epigenetic biomarkers in a non- or minimally invasive manner. [28].Transrenal cfDNA can end up in urine via the bloodstream or through direct secretion from malignant tissue in the renal system and can be enriched by centrifugation of urine samples, yielding urine sediment samples [29]

Another important source of cfDNA is stool. Normal colon epithelium renews itself every 24 hours, and this process is coupled with a continuous exfoliation of cells in the intestinal lumen. A number of studies have shown that cellular exfoliation is more pronounced in colorectal cancer patients than in healthy individuals. Although colon cancers typically account for less then 1% of the total colon surface area, tumor cell DNA actually makes up between 14% and 24% of the total DNA recovered in feces. The overall 5-year survival for colorectal cancer patients is around 64%, rising to almost 90% if tumours are diagnosed early. [30]

A promising strategy for lung cancer diagnosis involves detection of

epigenetic biomarkers from sputum or bronchial lavages of patients. Here, aberrant methylation of the CDKN2A/p16 and/or MGMT promoters could be detected in DNA from sputum in 100% of patients with squamous cell lung carcinoma up to 3 years before clinical diagnosis [31].

Rapid development of next-generation sequencing technologies has enabled new genome-wide epigenetic analysis techniques. These techniques drive studies that provide novel epigenetic biomarkers for a wide range of cancers. Several DNA methylation-based biomarkers have shown promise in cancer diagnosis and prognosis, but only a few have proven their clinical usability until now. Therefore, a major challenge will be to validate candidate markers in well defined patient groups. Sufficient numbers of samples of good quality are essential for marker validation. Furthermore, the lack of standardised assays and platforms are key issues regarding inter-laboratory variation and can produce contradictory data. Currently, mainly qPCR-based MSP assays relying on bisulfite conversion of DNA as well as MSRE-based qPCR amplification are used to investigate candidate markers, which are sensitive and suitable to quantitatively address DNA methylation levels in a large number of samples. However, the field will need to develop guidelines for sample collection and standards for quality control of bisulfite conversion and MSRE digestion efficiency or DNA integrity, in order to translate their findings into clinical routine.

Epigenetic Biomarkers in Prostate Cancer

Prostate cancer (PCa) is the most commonly diagnosed cancer in men, with an estimated 250.000 new cases in the United States diagnosed each year. PCa ranks second in cancer-related mortality in men after Lung Cancer. The clinical risk factors for PCa include aging, race, and family history of PCa [32]. Currently, most patients are diagnosed after detection of elevated serum prostate-specific antigen (PSA) levels or abnormal digital rectal examination, which involves diagnostic prostate biopsy. Clinically localized PCa, which is potentially curable, is usually treated with radical prostatectomy or radiation. Patients with locally advanced or metastatic disease are initially treated with androgen-deprivation therapy (ADT). However, almost all advanced PCa cases, after a period of ADT, progress to castration-resistant PCa, an incurable stage of prostate cancer, in which approx. 90% of patients develop metastases. [33]

Current classification and prognosis of PCa relies in histological grading, staging, and baseline serum PSA levels. PCa histological grading is based on the Gleason grading system, which combines 5 simple grades (from grade 1 (most differentiated) to 5 (least differentiated)) into 9 combined grades, the so-called Gleason score (GS) or sum (ranging from 2 (1+1) to 10 (5+5)), a feature that incorporates information from the frequent morphological heterogeneity of PCa. [34]

The best characterized epigenetic alteration in PCa is promoter hypermethylation. This epigenetic modification is associated with silencing of classic tumor- suppressor genes as well as genes involved in different cellular pathways such as cell cycle, hormone response, DNA repair, signal transduction, tumor invasion and apoptosis [35]. Hypermethylation of those genes promoter methylation in PCa may correlate with pathological grade, clinical stage, and castration resistance.

Epigenetic markers that can detect cancer are well described. GSTP1 is the best characterized epigenetic biomarker for PCa. DNA methylation of GSTP1 is nearly universally present in almost all PCa cells and is absent or low in normal cells [36]. Studies indicate that more than 90% of PCa cases show aberrant promoter methylation of GSTP1 [37]. DNA methylation of GSTP1 is mostly detected using MSP in prostate tissue samples and bodily fluids, mainly blood and urine. Testing for GSTP1 is used for screening and stratification for the need of prostate biopsy [36]. GSTP1 methylation performance displays high specificity (86.8 - 100%) but low sensitivity, both in urine (18.8 - 38.9%) and serum/plasma (13.0 - 75.5%) [38–40]. This might be overcome by a multigene promoter methylation testing, and several different gene panels have been proposed, including GSTP1/ARF/CDNK2A/MGMT [38] and GSTP1/APC/RARB2/RASSF1A [39] and GSTP1/APC/RASSF1A [41].

Epigenetic Biomarkers in Breast Cancer

Breast cancer represents a complex and heterogeneous disease at the histopathological, molecular, and genetic levels with a distinct clinical outcome.It is the second most common type of cancer after lung cancer, 12% of the total (http://www.who.int/mediacentre/ factsheets/fs297/en/). In Europe, there were an estimated 71 cases of breast cancer per 100.000 adults, whereas in the United States, there were an estimated 92 cases

of breast cancer per 100.000 adults. Overall, it is the most commonly diagnosed cancer, and it's incidence has been increasing over the last years, showing a higher incidence in developed countries. [42]

It has been shown that DNA methylation is an important mechanism in the development of breast cancer. There are a large number of genes that have been found inactivated in breast cancer due to promoter methylation.

These genes are mainly tumor suppressor and other cancer- related genes. They have been found hypo- or hypermethylated in primary tumors, and some of them have been described previously as mutated in the germ lines of patients with inheritable cancers (such as CDH1, p16INK4A/ CDKN2A, RB, BRCA1). [43–46].

Epigenetic Biomarkers in Colon Cancer

Colorectal cancer (CRC) annually affects more than one million men and women worldwide and causes more than half a million deaths [47]. CRC was the third most common malignant cancer for both men and women with 250000 cases of colorectal cancer diagnosed on an annual basis in Europe only [48]. Five-year survival was 54 percent among adult Europeans diagnosed with CRC between 1995 and 1999 [49].

CRC arises as a consequence of the accumulation of genetic and epigenetic alterations in colonic epithelial cells during neoplastic transformation. In addition to genetic mutations, DNA promoter hypermethylation of tumour suppressor genes has been widely investigated. Epigenetic markers and their combinations , have been described in CRC patients with 70%-96% sensitivity and 72%-96% specificity [50–52]. There are attempts to personalise chemotherapy based on presence or absence of specific biomarkers. Combinations of genetic and epigenetic markers have also been studied, but until now, their use in clinical practice has been limited. [53]

It is known that aberrant methylation in MLH1, MGMT, or the HIC1 promoter can lead to cancer progression [54, 55]. Cancer specific DNA methylation leads to transcriptional silencing of various genes such as tumor suppressor genes and genes involved in DNA repair and apoptosis, such as CDKN2A/p16, CDKN2A/p14, and HLTF [56]. DNA hypermethylation of MLH1 occurs in $>$80% of sporadic microsatellite instability (MSI) CRC, and the restoration of MLH1 expression and function by demethylating the MLH1 promoter in MSI CRC cell lines suggests that

Table 2.1: Commercially-available DNA methylation test kits for cancer.

Gene	Biomarker Type	Cancer Type	Test Kit
NDRG4/ BMP3	Diagnostic	Colorectal	Cologuard (Exact Sciences)
VIM	Diagnostic	Colorectal	ColoSure (Labcorp)
SEPT9	Diagnostic	Colorectal	Epi proColon (Epigenomics), ColoVantage (Quest Diagnostics), RealTime mS9 (Abbott)
SHOX2/ PTGER4	Diagnostic	Lung	Epi prolung (Epigenomics)
GSTP1/ APC/ RASSF1A	Diagnostic	Prostate	ConfirmMDx (MDxHealth)
MGMT	Predictive	Glioblastoma	PredictMDx (MDxHealth), Mismatch Repair genes (MRC-Holland), PyroMark MGMT Kit (Qiagen)
TWIST1/ OTX1/ ONECUT2	Diagnostic	Bladder	AssureMDx (MDxHealth)

such aberrant methylation is a cause rather than a consequence of colorectal carcinogenesis [57]. Detection of tumor-derived DNA alterations in stool has high potential for the noninvasive detection of CRC. Studies have identified an increasing number of genes that are methylated in both tissue and fecal DNA of CRC patients, and NDRG4 and BMP3 are of particular interest. [58, 59]

Biomarkers in Glioblastoma

Biomarkers play an important role in the diagnostics of Glioblastoma multiforme (GBM) and are very important for histopathological diagnosis and their molecular classification. Patients with GBM have a poor prognosis and only 35% of them survive for more than 5 years. The current GBM treatment standards include maximal resection followed by radiotherapy with concomitant and adjuvant therapies. Despite these aggressive therapeutic regimens, the majority of patients suffer recurrence due to molecular heterogeneity of GBM. Consequently, a number of potential diagnostic, prognostic, and predictive biomarkers have been investigated. Some of them, such as IDH mutations, 1p19q deletion, MGMT promoter methylation, and EGFRvIII amplification are frequently tested in routine clinical practice. With the development of sequencing technology, detailed characterization of GBM molecular signatures has facilitated a more personalized therapeutic approach and contributed to the development of a new generation of anti-GBM therapies such as molecular inhibitors targeting growth factor receptors, vaccines, antibody-based drug conjugates, and more recently inhibitors blocking the immune checkpoints. [60] The hypermethylation in the promoter region of MGMT gene is not only an important prognostic factor for glioblastoma patients but also a predictor for the outcome of the treatment to alkylating agents [61]. MGMT encodes O6-methylguanine-DNA methyltransferase , a DNA repair enzyme which protects cells against alkylating agents like Temozolomide (TMZ) through preventing G:C to A:T gene mutations. Disorders of MGMT promoter methylation are associated with transcriptional silencing of the MGMT gene and loss of MGMT expression that results in decreased DNA repair and retention of alkyl groups, thereby allowing alkylating agents to be more effictive in patients with MGMT promoter hypermethylation. A number of clinical trials have shown that MGMT methylation corresponds to greater PFS and OS in patients who are treated with alkylating agents [62, 63]. MGMT promoter methylation status represents the best

studied and most relevant prognostic factor in GBM and has been considered as a potent predictor of response to alkylating agents.
Several studies have demonstrated that patients with tumors with methylated MGMT promoter had a survival benefit when treated with TMZ and radiotherapy, compared with those who received RT only, whereas patients with unmethylated MGMT promoter in their tumors had no survival benefit from chemotherapy, regardless of whether it was given at diagnosis together with RT or as a salvage treatment [64, 65]. Consequently, it has been suggested that elderly GBM patients eligible for either RT or TMZ should undergo MGMT promoter methylation testing prior to the clinical decision being made.

2.4.2 Epigenetic Therapy

A relatively recent development in the field of epigenetics is epigenetic therapy. Unlike genetic events, epigenetic changes can in theory be reversed by pharmacologic intervention to block enzymes that add or remove modifications from histones (writers and erasers), inhibit DNA-methyl transferases, prevent critical proteinprotein interactions among transcription factors, or block protein domains (readers) from recognising specific histone modification states.

Currently the only epigenetically directed therapies in clinical practice are inhibitors of DNA methyltransferases and histone deacetylases (HDAC). Although these drugs yield global changes in DNA methylation and histone acetylation, respectively, it remains uncertain whether the efficacy of these agents is linked to specific changes in gene expression. HDAC inhibitors have pleiotropic actions and can affect cytoplasmic as well as nuclear processes. Furthermore, both classes of agents elicit DNA damage responses and may be acting as low-intensity cytotoxic agents. [66]

Despite initial promise, there are still many questions to be answered before we can use epigenetic markers in the clinical arena on a broader scale. An important issue is that of selectivity: How can we selectively target genes and pathways with ubiquitously expressed epigenetic regulators? The answer could lie in more selective epigenetic compounds like miRNAs. A miRNA may target tens to hundreds of transcripts to control key biological processes. The biochemical reactions underpinning miRNA biogenesis and activity are relatively well defined [67]. It is however difficult to pinpoint the underlying regulation of the miRNA pathways in

vivo. miRNAs are strongly linked to key cancer-related processes, affecting nodal points in cell cycle regulation, genome integrity and stress responses, apoptosis and metastasis. In addition, genetic models have evidenced that several miRNAs act as bone fide oncogenes and that tumours may develop addiction to oncogenic miRNA overexpression, which holds a promise for the therapeutic use of miRNA inhibitors. [68]

3

Methodologies, Analytical Approaches and Visualization

3.1 Introduction

Although epigenetic features are not encoded in the DNA sequence itself, they are directly or indirectly involved in the regulation of transcription on specific loci throughout the genome. To generate a global picture of them requires high-throughput, high resolution and cost-efficient techniques. During the course of this PhD thesis, there was a spectacular evolution in Next Generation Sequencing Techniques (NGS). And this has contributed greatly to the field of epigenetics. In this chapter, I will sketch an overview of the current High Throughput methodologies and analytical approaches used in epigenetics.

3.2 Methodologies and Analytic Approaches

Based on article:
Next-generation technologies and data analytical approaches for epigenomics
Klaas Mensaert K, Simon Denil, Geert Trooskens, Wim Van Criekinge, Olivier Thas and Tim De Meyer; Environmental and molecular mutagenesis, 2014 [69]

3.2.1 Bisulfite-based methods to study DNA methylation and hydroxymethylation

Bisulfite Conversion

Treatment of DNA with sodium bisulfite results in the deamination of unmethylated cytosine residues to uracil, whereas 5-methylcytosine remains intact [70]. The introduced sequence differences can subsequently be exploited to assess the methylation status of each cytosine, using, for example, microarrays or nextgeneration sequencing applications (Figure 3.1). Unfortunately, standard bisulfite conversion does not discriminate between methylcytosine and hydroxymethylcytosine as both are nonreactive to the chemical conversion reagent. One solution is the initial oxidation of hydroxymethylcytosine to formylcytosine, which is sensitive to conversion to uracil under bisulfite conditions, upon which both the degrees of methylation and hydroxymethylation can be determined by comparison of the oxidative and the normal bisulfite sequencing results for each locus [71]. Hydroxymethylation can also be directly assessed by the TET-assisted bisulfite sequencing methodology (TAB-seq). Enzymatic glycosylation of hydroxymethylcytosine protects the latter residues against carboxylation by a recombinant TET enzyme, whereas methylcytosine (that cannot be glysosylated) is converted to carboxylcytosine. Both carboxylcytosine and basic cytosine residues, but not the glycosylated hydroxymethylcytosine, are converted to uracil upon bisulfite treatment [72]. Although these modified protocols are in principle compatible with the different bisulfite-based methods described below, it should be noted that the impact of hydroxymethylation is currently still largely ignored. Main reasons are the uncertainty about presence and nature of function, but also the fact that hydroxymethylation are typically substantially lower than methylation levels [73]. In the remainder of this subchapter, we predominantly refer to DNA methylation for simplicity, but it is clear that without adapted protocolhydroxymethylation will be assessed simultaneously.

Whole Genome, Reduced Representation, and Targeted Bisulfite Sequencing

Induced sequence differences are exploited by whole genome bisulfite sequencing (WGBS), enabling a straightforward quantification of the (hydroxy)methylation status of each individual cytosine in the genome. The

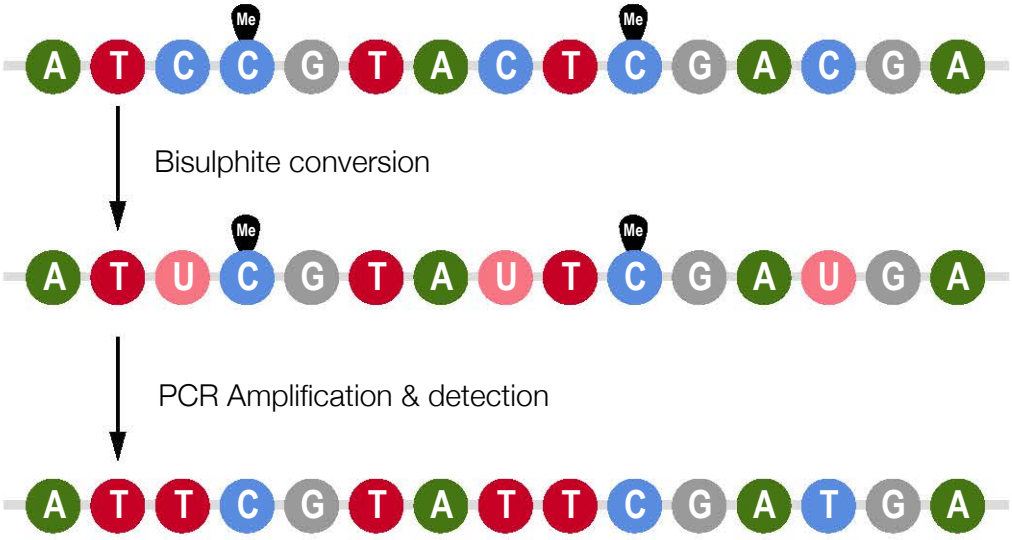


Figure 3.1: Bisulfite treatment of DNA results in the deamination of unmethylated cytosine residues to uracil, whereas 5-methylcytosine remains intact. Subsequent amplification produces T/A nucleotides on unmethylated positions and C/G nucleotides on methylated positions.

high resolution of this currently gold standard method is an advantage, as several studies have demonstrated that the methylation status of a single CpG may already be highly predictive [74, 75]. On the other hand, WGBS is expensive, particularly when taking into account the fact that the bisulfite treatment complicates the sequencing procedure [76]. For example, incomplete bisulfite conversion will yield biased methylation estimates and bisulfite treatment might result in DNA degradation, although these effects have been largely attenuated by methodological improvements. Even though current innovation toward longer reads is tempering this problem as well, the fact that a major part of the cytosine fraction is converted to (upon sequencing) thymine substantially complicates subsequent sequence alignment [76]. Finally, the sequence-dependent efficiency of PCR [77], which is indispensable for current second-generation sequencing, can also cause biased methylation estimates.

In addition to the difficulties outlined above, WGBS is generally deemed to be inefficient, as a very large part of the sequenced fragments yields no relevant information [78]. Several methods are therefore used to reduce the amount of genome to be sequenced by prior enrichment for sequences of interest. In Reduced Representation Bisulfite Sequencing

(RRBS) [79], a CpG restriction endonuclease that is DNA methylation insensitive (e.g., 30 -CCGG-50 targeting MspI) is used to generate fragments with CpG-enrichment at the ends. By ligating sequence adapters containing 5-methylcytosine before bisulfite conversion, this methodology is compatible with second-generation sequencing. Appropriate gel-based size selection of fragments before sequencing permits to generate a reproducible but reduced representation (1%) of the DNA methylome [79].

The major disadvantage of RRBS is the assays limitation to CpGs in the vicinity of the recognition site of the applied endonuclease, which is less appropriate if there are particular loci of interest [80]. Multiplex PCR of bisulfite-treated DNA before sequencing enables targeting specific loci of interest, but is profoundly complicated at a genome-scale because a large amount of multiplex-proof primers with no/few CpGs need to be designed for a template with reduced complexity.

Targeting specific loci by probes can solve this problem, for example by capture of bisulfite-treated DNA using long probes [80]. Alternatively, the solution hybrid selection methodology developed for exome sequencing can be used to capture DNA before bisulfite treatment. The latter option is deemed to perform better due to the more complex and methylation independent sequence of the targeted loci (which both substantially facilitate probe design), but requires a relatively high amount of DNA [81]. Whereas, from an experimental point of view, RRBS and WGBS do not require prior knowledge of the species genome, targeted bisulfite sequencing requires knowledge of a closely related genome for primer (PCR-based targeting) or probe (hybridization-based targeting) design.

Bisulfite Sequencing Data Analysis

Analysis of bisulfite sequencing data starts with sequence alignment followed by the identification of significantly methylated loci. Several mapping algorithms have been developed that take into account the reduced complexity and potential methylation-dependent variation of bisulfite sequencing-generated reads. The challenges associated with bisulfite sequencing alignment also imply that the presence of a reference genome is required for RRBS and WGBS.

Upon alignment and quality control (QC) (e.g., removing low-quality reads and low coverage CpGs), several methodologies for the identification of significantly (differentially) methylated regions have been pro-

posed [80,82]. Although, in general, both CpG and non-CpG methylation can be analyzed with the same tools, the term CpG will be used for consistency. For individual samples, it can be assessed whether a certain CpG is significantly methylated by comparing the observed coverage of unconverted cytosines with the (sequencing1bisulfite conversion) error rate (e.g., [83]). Experiments performed without biological replicates can only yield information about the specific samples under study and are therefore not recommended, unless biological variability is irrelevant for the research setting. In the latter case, differentially methylated CpGs between two samples can be identified using the Fisher exact test (or Chi-square test) for each CpG dinucleotide, for example [83]. To avoid multiple testing-associated problems [84], the amount of CpG variables can be reduced by imposing filters (e.g., coverage), or by combining multiple CpGs into a single variable representing a region in the genome. Several methods capable of dealing with biological/technical replicates have been developed as well. For a group of samples, several statistical distributions have been used to model methylation ratios with subsequent statistical testing, for example, beta distribution [78,85] or logistic regression models [86]. Also here, the information of different neighboring CpGs is often merged to obtain a smaller amount of less dependent variables [78, 85, 87].

Bisulfite Conversion-Dependent Arrays

Instead of sequencing, Illuminas Infinium HumanMethylation BeadChip methodology relies on the assessment of bisulfite conversion-induced single-nucleotide polymorphisms (SNPs). The most recent version, Human Methylation EPIC, assesses approximately 850,000 cytosines (predominantly in CpGs), although the effective amount of targeted cytosines is substantially lower due to aspecific probes and SNPs within the targeted regions [88, 89]. It is composed of two assay types (i.e., types I and II), whereas the previous version, HumanMethylation27 was solely based on the former type. This type I assay uses two probes per CpG locus (i.e., a methylated and an unmethylated query probe). After hybridization, but only upon perfect matching of a methylated cytosine or thymine (i.e., bisulfite converted unmethylated cytosine) at the 30 probe end, labeled nucleotides can be incorporated. As methylated and unmethylated query probes are located on different beads, methylated (m) and unmethylated (u) light signals can be detected using the same color channel, but on different locations (beads) [90]. Type II assays, on the other hand, employ

single beads coated with degenerate probes and uses two colors/channels to discriminate methylated (m) from unmethylated (u) light intensities. This is possible because differentially labeled adenine or guanine residues can be incorporated at the 30 probe end, respectively, basepairing with thymines (bisulfite converted unmethylated cytosine) or methylated (and therefore nonreactive) cytosines [90, 91]. Type II assays therefore differ from type I assays by the use of two different colors and the fact that degenerate probes rather than two different probes are used. This explains the observation that the data distributions of the resulting methylation degree estimates differ between both assay types on the HumanMethylation450 BeadChip. In addition, type II assays have been reported to perform somewhat inferior regarding reproducibility, accuracy and dynamic range [91].

Infinium Bead-Arrays: Data Analysis

Infinium data preprocessing consists of QC and subsequent filtering, data summarization and normalization. QC includes the removal of probe signals that cannot bedistinguished from the background by comparing them with control probes, but also X and Y chromosome probes, SNP containing probes, and less specific probes (implying possible cross-hybridization) can be omitted [84]. Two summary methods are routinely used and can be applied for both assay types. The b value is an intuitive statistic that is used as a proxy for the proportion of methylation (0 to 100%), whereas the M value expresses the log ratio of the methylated and unmethylated intensities.

The M value lends itself to be conveniently used in combination with existing array analysis methods, which can in part be attributed to a higher homoscedasticity of the data after log transformation [92]. Another dedicated variance-stabilizing normalization method has been implemented in the lumi package, which also includes several basic normalization algorithms [93]. Data normalization is typically performed to eliminate technical variation between samples caused by, for example, background signal and dye, GC, and CpG-bias, but also between type I and type II data [91]. After data preprocessing, standard microarray data analysis methodologies can be used, for example, LIMMA and SAM [94]. These methods employ moderated variance estimates for each locus, by sharing information over all genes, to counteract problems associated with low numbers of replicates. If larger samples are available, standard statisti-

cal methods (e.g., Wilcoxon rank-sum test, t-test, linear models) can be applied as well and these have been conveniently bundled in the Illumina Methylation Analyzer (IMA) package, with associated multiple testing correction [95]. Instead of using continuous β or M values, data may also be categorized, for example, using mixture models to define a samples qualitative methylation status for a certain probe/locus, for example [93]. Despite these complex data analysis strategies, often associated with the inherent limitations of probe-based arrays, basic Infinium data analysis is rather straightforward as it can readily yield methylation degree estimates as β values. The combination of low cost and high performance has made Infinium BeadArrays a very widely used methodology. The most important limitation, however, remains the fact that this methodology is only available for the human genome, for which it even only queries a fixed and limited portion of the DNA methylome [84].

3.2.2 Restriction Based Methodologies for Genome-Scale DNA Methylation Assessment

Whereas single-molecule sequencing allows to obtain the DNA (hydroxy)-methylome in parallel with the genome, other DNA methylation assessment methodologies mentioned higher are either limited to the human genome or require sufficient knowledge of the underlying genome sequence. When no (closely related) genome is available, fingerprinting methods can be used, for example, in ecological epigenetics studies. For these purposes, the methylation-specific amplified polymorphisms (MSAP) methoda modification of the amplified fragment length polymorphisms (AFLP) methodis widely applied. This technique makes use of the differential affinity to cytosine methylation of a pair of isoschizomers (restriction enzymes specific to the same recognition sequence). More specifically, after digestion by both restriction endonucleases separately, amplification and electrophoresis, DNA methylation is identified through the presence of isoschizomer-dependent fragment differences [96]. Different interpretation and scoring approaches, typically modifications of standard AFLP procedures, are available for the obtained banding patterns, both for multilocus (cf. population studies) and single-locus analysis [97].

Despite a low resolution and limited genome-wide character, the cost-efficiency of this methodology has resulted in a wide application, even for species with sequenced genomes [98]. Fragments indicating differential methylation between groups can subsequently be isolated and sequenced,

upon which the specific locus involved can be identified, for example [98]. Other methods making use of the (differential) affinity of restriction endonucleases and isoschizomers regarding DNA methylation [99], will not be further discussed here as their frequency of application is generally decreasing due to the advent of novel, cost-efficient, more genome-wide alternatives.

3.2.3 Enrichment-Based Epigenomics

While reduced representation and targeted bisulfite sequencing decrease costs by sequencing a limited part of the DNA methylome, an alternative strategy is to enrich for those DNA fragments that are putatively methylated. Enrichment approaches include isolation by DNA methylation-specific methyl binding domains (MBDs), antibodies or other purification strategies. This methodology is also the principal approach to obtain genome-wide information regarding histone modifications and nucleosome positioning.

Affinity Purification to Study DNA (Hydroxy)methylation

MBD, typically recombinant versions of the MBD found in, for example, the MBD2 and MCEP2 proteins, can be used to capture methylated CpGs [100]. When followed by sequencing, this methodology is known by several names: MethylCap-seq [101], MBD-seq [102], and MBDisolated genomic sequencing (MIGS) [103]. Capturing can also be followed by microarray or PCRbased analysis, and then the term methylated-CpG island recovery assay is typically used for the affinity purification step [104,105]. Although most studies have reported that these MBD domains typically do not target hydroxymethylation, for example [106], results about MCEP2 and its MBD-domain are less consistent [107, 108] indicating potential capture of hydroxymethylcytosine by MCEP2. Also for MBD3 weak affinity toward hydroxymethylcytosine has been reported [108,109].

As an alternative to MBDs, antibodies can be used to enrich for methylcytosines (methylated DNA immunoprecipitation, MeDIP, mDIP, or mCIPmethylcytosine immunoprecipitation) before microarray [110,111] or sequencing (MeDIP-seq) [112]-based analysis. More recently, antibodies targeting hydroxymethylcytosine have been used in a hMeDIP approach [113]. Although MethylCap-seq is believed to capture biologically more relevant DNA methylation due to the use of protein derived, CpG-

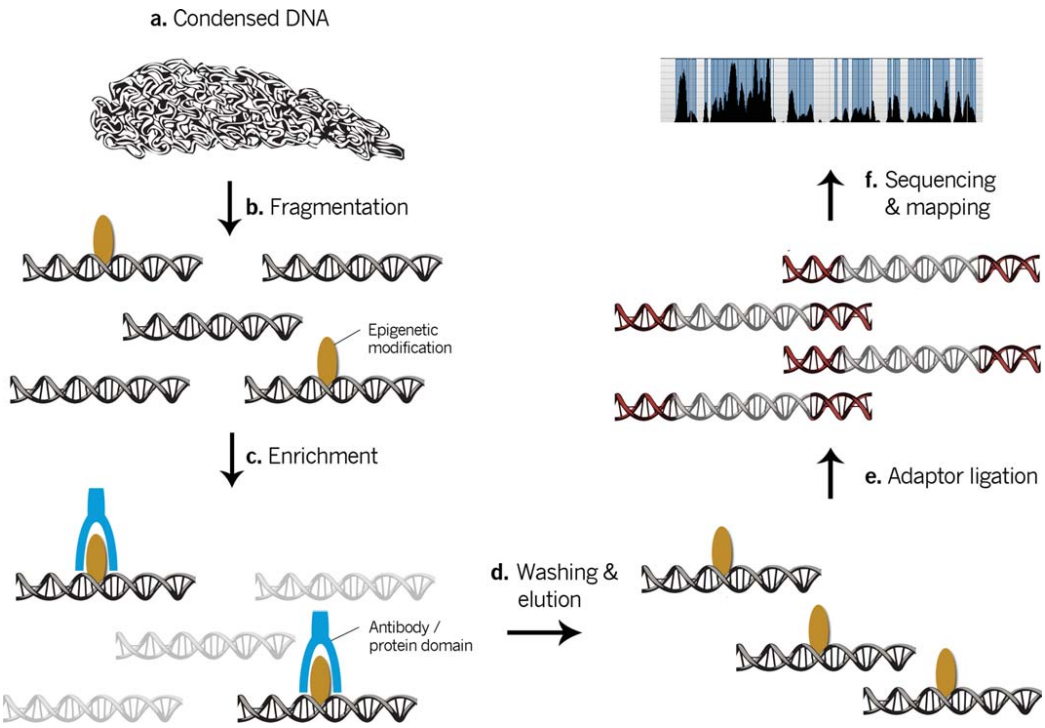


Figure 3.2: Enrichment-based sequencing. After DNA fragmentation (a, b), DNA fragments bearing the specific epigenetic modification of interest are captured using antibodies or specific protein domains (c). Unbound fragments can be washed away, whereas an elution step is required to obtain the DNA fragments of interest (d). After adaptor ligation and sequencing, sequence reads are aligned to a reference genome to identify the epigenetically modified loci (e, f).

targeting MBD, MeDIP-seq has the advantage that its affinity depends less on the CpG content of the methylated fragment, as even non-CpG methylation can be picked up [114]. However, a genuine comparison of methylCap-seq and MeDIP-seq is difficult due to the complexity of the affinity purification protocol, where for both methodologies sensitivity, specificity and overall yield may differ to a large extent between different kits, batches of antibodies/MBDs, elution step salt concentrations and sample qualities and quantities. [114–116]

Although affinity purification-based methods reduce overall cost by limiting the amount of DNA to be sequenced while maintaining a genome-wide approach, there are several drawbacks as well. First, MBD-seq and MeDIP cannot attain the base pair resolution of bisulfite sequencing. Second, target fragment CpG-density and GC-content are major biases, affecting, respectively, the efficiency of affinity purification (particularly for MethylCap-seq) [84, 117] and subsequent sequencing [117, 118]. As a consequence, several methylated regions are virtually impossible to capture and sequence by MBD-seq and/or MeDIP-seq, implying that the genome-wide character of both methods is limited.

Next-Generation Methods to Study Histone Modifications and Nucleosome Positioning

Antibodies can also be used to study histone modifications and positioning of nucleosome subsets, nowadays typically in combination with sequencing, that is, ChIPseq, although also tiling arrays are often used (see below). These experiments start with the purification of nuclei, followed by sample fragmentation, either by sonication or by micrococcal nuclease (MNase) digestion, which preferably targets linker DNA. Formaldehyde-based crosslinking reduces the chance of chromatin rearrangement during processing, but is less commonly applied to study histones as it is featured by several important side effects [119]. Upon immunoprecipitation and purification of the captured DNA, the latter is sequenced directly, or after additional preprocessing [115].

Indeed, for nucleosome positioning studies, mononucleosomal DNA fragments are selected before sequencing by the isolation of 146-bp-long DNA fragments after electrophoresis [120]. As current nucleosome positioning studies require consistent phasing across cells, it should be noted that most experiments have been performed on yeast, cell lines, or very homogeneous tissues (e.g., T-cell subtypes) [121].

While antibodies targeting specific histones, histone variants or modifications enable studying specific nucleosome subtypes [120], the application of MNase without subsequent immunoprecipitation yields profiles for the total set of nucleosomes after sequencing, that is, MNase-seq [122, 123]. Although not an enrichment-based method, an alternative strategy called nucleosome occupancy and methylome sequencing (NOMe-seq) is based on the principle that nucleosome bound DNA is protected from methylation by GpC methyltransferases (i.e., not CpG), upon which bisulfite sequencing can be used to identify nonmethylated (i.e., nucleosome bound) regions. A major advantage of NOMe-seq is that the original (typically CpG dependent) DNA methylation profile can be measured simultaneously [124].

ChIP-seq Experimental Quality Control

For the different ChIP-seq experiments, the antibody used regulates the particular epigenetic feature targeted. As the antibodys sensitivity and specificity are the main determinants of outcome quality, rigorous testing, preferably for each batch, is required. The specificity of ChIPseq is additionally compromised by several biases, caused by both immunoprecipitation (preferred enrichment for open chromatin) and sequencing. It is therefore recommended to include control samples, that is, total input control or generated by applying a control antibody targeting an irrelevant, non-nuclear antigen [125, 126]. As MNase preferably targets specific sequences [127], an appropriate control for MNase-seq is the use of input DNA free of nucleosomes [121].

Data Analysis for Enrichment-Based Sequencing Methods

Essential data analytical steps for the different enrichment-based sequencing applications include the mapping of reads, QC, data summary (peak detection), data normalization, and identification of significant (differential) enrichment. In general, standard ChIP-seq analytical tools are used for histone modifications, whereas more specific algorithms have been developed for nucleosome positioning and DNA methylation studies. Independent of the type of experiment however, the analysis starts with the alignment of the typically short sequence reads. Virtually all recent tools can cope with both single and paired-end read data, for example Bowtie2 [128], SHRiMP [129] and SOAP2 [130]. Paired-end fragments, entailing that both ends of a fragment have been sequenced, can be

more accurately mapped to a unique position. Selection of the best suited sequence read aligner for a specific task depends on general alignment quality as well as computational performance and possible memory restrictions [131].

Generic sequencing QC tools are freely available, with read positional quality scores, fraction of duplicate reads and GC-content bias as common criteria (e.g., FastQC). Lower quality sequences can be filtered, as well as duplicate fragments. The latter are different fragments of which the sequenced reads map at exactly the same locus, and are therefore likely to be PCR generated artifacts (particularly for paired-end sequencing). For single-end sequencing, loci with an unbalanced amount of reads mapped to both DNA strands (Watson and Crick) can be filtered out as well [132]. Specifically for enrichment-based methods, saturation analysis can be interesting in order to check the adequacy of the sequencing depth. Specific tools have also been developed for MethylCap-seq and MeDIP-seq, which examine, for example, the CpG coverage, CpG occurrence in reads and CpG enrichment [133, 134]. For this type of data, particularly the fragment CpG content appears to be predictive of overall quality [116].

After mapping, nucleotide level coverage data are typically transformed to count data for DNA regions, that is, data summarization. These regions can be predefined, for example, summarization per gene /promoter / equally sized bin / . . . , or extracted from the data using peak detection algorithms. For single-end sequencing data, still often used for ChIP-seq, additional preprocessing is required before peak detection to account for the typically double peaks of mapped reads (one on each strand) surrounding the enriched regions. Both peak shift adjustment and read extension are common solutions [135]. Peak calling strategies will be discussed first, followed by algorithms to test for differential enrichment between samples.'

For ChIP-seq data, a signal profile is typically constructed by counting the number of reads in a sliding window or by kernel density estimation [136–138]. Enrichment by ChIP is often nonspecific; about 60 up to 99% of reads can be off-target, thereby composing the background [132]. As an approximation, the background profile can be modeled, using either data from the ChIPped samples or a control sample [125, 139]. Statistical comparison of the putative peak with the background distribution, either by simulation (nonparametric) or by a statistical modeling approach, yields P values with associated FDRs to identify significant enrichment

(i.e., peak calling). For DNA methylation enrichment studies, background assessment and peak detection can be performed by applying the same or similar methods, for example [140, 141], or using predefined regions of interest such as gene promoters, sets of likely differentially methylated regions [78], and so forth. There exist methods that also correct for copy number variations [142]. Although several tools have been developed for the comparison of two samples, for example, for the identification of differential histone modification sites using Hidden Markov models [143], it should be noted that typically more independent samples need to be included to take biological variability into account. Importantly, this also requires normalization between these different samples, for example, by library size normalization or other algorithms assuming overall equal levels of the epigenetic mark, although the validity of this assumption is often questionable, for example [Ehrlich, 2009]. Alternatively, the principle of using noise regions to normalize ChIP samples versus total input controls, for example [125], can also be applied for normalizing multiple ChIP samples. For DNA methylation, normalization can also be based on methods for the estimation of absolute methylation degrees from the sequencing count data. These algorithms have been developed starting from the observation that the measured MethylCap-seq/MeDIP-seq intensities are a function of both absolute methylation degree and sequence CpGcontent [114, 133, 134, 144].

To evaluate differential enrichment between groups of samples, basic statistical tests such as Wilcoxon rank sum, KruskalWallis, and t-tests, can be used, for example [145, 146]. However, as datasets generated by enrichment sequencing experiments are similar to RNA-seq derived data, which are often modeled as negative binomial distributions, the use of methods designed for RNA-seq data analysis may be more reliable and powerful. These are ideally moderated methods that allow analyzing data from experiments with low numbers of replicates. Moderated methods are able to stabilize the variance estimation by sharing information between different loci [147]. For example, recent versions of EdgeR [148] and DESeq [149] provide feature variance estimates derived from both the full dataset (variance modeled as a function of the mean) and the specific locus under study. These tools are very flexible by the implementation of generalized linear models, suitable for complex, multifactorial designs. EdgeR can also be called through the Repitools package for enrichment-based epigenomics data analysis [134].

MNase-seq and ChIP-seq for nucleosome positioning aim at selec-

tively sequencing nucleosome bound fragments. After mapping, paired-end fragments are typically trimmed to obtain higher resolution profiles, and additional adjustment for sequence biases (of MNase, sequencing itself . . .) may be performed, for example, using a control sample [150]. Whereas specific loci are depleted from nucleosomes (e.g., core TSS of active genes), others are featured by either well-defined (in phase across cells) or more fuzzy positioning (heterogeneous across cells) of nucleosomes. Although broad-spectrum ChIP-seq methods can be applied for nucleosome positioning data, for example [151], several algorithms have been specifically developed for recognizing the presence of nucleosome positioning. The first step typically consists of noise removal [150, 152], upon which peak detection algorithms can be applied [150] to identify nucleosome positions. Also model-based approaches that directly incorporate the presence of noise have been developed, for example [153, 154]

3.2.4 Sequencing and PCR-based Assessment of Noncoding RNA

During a standard mRNA-seq protocol, oligo-dT-based capture is used to filter poly-A tailed RNA molecules from the total pool of RNA, which overwhelmingly consists of noninformative (nonpolyadenylated) rRNA molecules. In contrast to mRNA, ncRNA species are typically not polyadenylated, and therefore an alternative rRNA depletion strategy is required before sequencing [155]. Two commonly implemented approaches are the subtractive hybridization of rRNA molecules, based on the high degree of rRNA sequence conservation, and the removal of highly abundant RNA species through cDNA normalization (after reverse transcription). The latter procedure involves denaturation and reannealing of double stranded cDNA at an elevated temperature. As highly abundant cDNA molecules reanneal at a higher frequency, removal of double-stranded cDNA at a well-selected time-point results in relative depletion of highly abundant (r)RNA sequences [155]. Alternatively, dedicated sequencing protocols (e.g., Illuminas small RNA sequencing protocol) have been developed to specifically sequence small and therefore predominantly noncoding RNA molecules through a prior size selection, thereby implicitly eliminating most mRNA but also longer ribosomal and other overabundant RNA molecules, for example [156]. Adapted protocols are available to preferable target miRNAs, for example [157].

Depending on the sequencing procedure, either the total RNA frac-

tion (including mRNA, but rRNA depleted) will be sequenced, or solely the small RNA fraction. Total RNA-seq and standard mRNA-seq data analysis are very similar. However, after mapping with splicing-sensitive sequence aligners such as TopHat2 [158] and STAR [159] and subsequent QC using, for example, RNA-SeQC [160], total RNAseq data summary should not only consider coding regions but also collections of predicted or de novo detected ncRNAs, for example [161–163]. Subsequently, standard RNA-seq procedures for normalization, for example [164], and statistical analysis (see above), can be applied to the data. Although the loci under consideration will be far more limited than for total RNA-seq, a very similar approach can be used for small RNA-seq data analysis, for example [165].

3.2.5 Locus Specific and Validation Methods

The next-generation technologies discussed have the major advantage of being capable of assessing the epigenetic features of interest on a genome-scale. One important issue associated with genome-scale experiments though is the multiple testing problem. Although more advanced FDR control methods are being developed that particularly take into account the dependencies between variables, for example [85, 166] , experimental verification on independent samples remains the gold standard for the accurate validation of true positive results. Locus specific, low-throughput (typically PCR based) methods predating the advent of next-generation technologies are still very appropriate for validation purposes. For example, for histone modification and nucleosome positioning studies, quantitative PCR can be performed on the enriched DNA of independent samples and controls. Also for ncRNA, upon reverse transcription, quantitative PCR will yield insight into presence and differential expression of the ncRNA species of interest.

To assess/validate (differential) DNA methylation of individual loci, independent DNA samples are typically bisulfite treated first. Adapted bisulfite protocols can be used to discriminate methylcytosine from hydroxymethylcytosine (see above). Subsequently, PCR primers can be developed specifically assessing the cytosines of interest in a method called methylation-specific PCR (MSP) [167], targeting either methylated or unmethylated alleles. This very sensitive methodology cannot discriminate between very low or very high levels of methylation, for example, for heterogeneous samples where small amounts of background methylation

are often present. Therefore, real-time quantitative MSP has been developed, indicating the methylation degree for the locus under study [168]. Accurate primer design is essential for MSP to discriminate methylated from unmethylated alleles, and specialized tools have been developed, for example [169]. An alternative option is to design primers for the unmethylated (typically CpG-free) regions surrounding the locus of interest, and sequence the fragments obtained after PCR [70]. The often lower complexity of primer design, the easily verifiable specificity, the base-pair resolution and particularly the improvements (cost, labor intensity) in sequencing technology, have made the locus-specific bisulfite sequencing option very popular.

3.3 (Epi-) Genomic Data Visualization

Data visualisation is pivotal in analysing and understanding vast amounts of genomic data for different types of users ranging from researchers involved in high-throughput genome research to clinicians and a growing number of people having access to significant parts of their own genetic data. This will become even more important as sequencing prices plummet and the age of personal genomics becomes a reality.

3.3.1 Linear Genome Browsers

Linear Genome browsers are applications that provide a graphical interface to search, browse, import, export and visually analyze genomic sequence data. They use the chromosomes of a species and the genomic positions as a coordinate system for annotations generated from heterogeneous genomic data. Each distinct data set is called a track. They provide a unique, efficient and convenient analysis platform for genomics. The graphical interface helps users extract and summarise data intuitively from vast amounts of raw genomic data. They are particularly useful to zoom in and examine particular loci (e.g. a gene or gene cluster).

Resources such as genome browsers are still some of Molecular Biologists best tools. A good browser can distinguish good from poor quality genomic data sets and can show trends and patterns in the data without the need for statistical measures. These visual observations can spur questions that require more sophisticated analysis. Several browsers are available, including Entrez Genome, Ensembl and the UCSC Genome Browser. Although the amount of data available on the UCSC browser

makes it very valuable, it can be slow when attempting to browse through several data sets at various locations. Next Generation browsers, such as Anno-J, which was used for visualizing the Arabidopsis thaliana and human methylomes at nucleotide resolution [83, 170] are much more dynamic. Scrolling through the genome is very rapid and tracks can be zoomed, scaled, re-ordered and removed almost instantly. [171]

3.3.2 Circular Genome Representation

Circular Genome Representation (CGR) is used for the analysis of similarities and differences in the comparisons of genomes. CGR is an effective way to display variations in chromosome structure and, more general, any other kind of chromosomal/positional relationships between genomic loci. This data is routinely produced by sequence alignments, hybridization arrays, Next Generation sequence alignments and genotyping studies. CGR uses a circular ideogram layout to facilitate the display of relationships between pairs of positions by the use of ribbons, which encode the position, size, and orientation of related genomic elements. Circular Genome Representation software like Circos [172] is capable of displaying data as scatter, line and histogram plots, heat maps, tiles, connectors and text.

3.4 Survival analysis

Survival analysis is generally defined as a set of statistical methods for analyzing input and output variables where the outcome parameter is the time until the occurrence of an event. The event can Relapse, Progression, Death in clinical studies. The methods can also be used in other fields where the event can be marriage, divorce, the length of time people remain unemployed after a job loss or even how long it takes for the machine to break down. The time to event of interest or survival time can be measured in defined time intervals like days, weeks, or years. in the scope of this thesis, we focus on the overall survival time (OS) and progression free survival time (PFS). OS equals the time until death and PFS measures the length of time during and after the treatment of a disease, such as cancer, that a patient lives with the disease but it does not get worse.

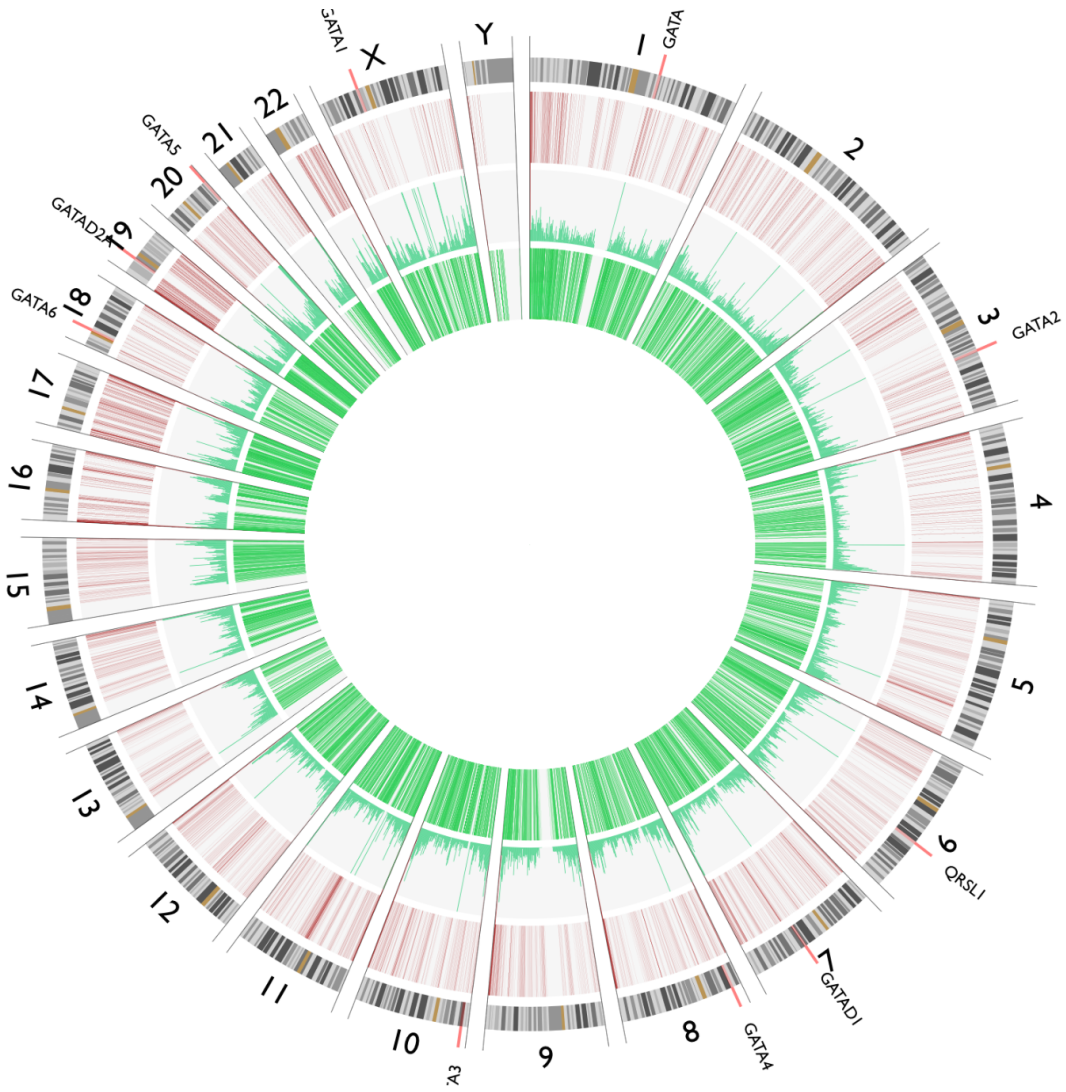


Figure 3.3: A circular genome view. Chromosome ideograms are plotted on the external ring. Three data tracks rings are plotted onto this polar coordinate system. Search results and gene lists can be plotted on the outer ring, highlighting the selected genes

3.4.1 Kaplan-Meier estimator

The Kaplan-Meier survival curve is defined as the probability of surviving in a given length of time while considering time in many small intervals. Three assumptions need to be met to perform this analysis. Firstly, we assume that at any time patients who are censored have the same survival prospects as those who continue to be followed. Secondly, we assume that the survival probabilities are the same for subjects recruited early and late in the study. Thirdly, we assume that the event happens at the time specified.
The Kaplan-Meier estimate is also called as 'product limit estimate'. It involves computing of probabilities of occurrence of event at a certain point of time. We multiply these successive probabilities by any earlier computed probabilities to get the final estimate. The survival probability at any particular time is calculated by the formula given below:

$$S_t = \frac{NumberOfSubjectsLivingAtTheStart - NumberOfSubjectsDied}{NumberOfSubjectsLivingAtTheStart}$$

Subjects who have died, dropped out, or move out are not counted as 'at risk' i.e., subjects who are lost are considered 'censored' and are not counted in the denominator. Total probability of survival till that time interval is calculated by multiplying all the probabilities of survival at all time intervals preceding that time. For example, the probability of a patient surviving two days after a tumour removal surgery can be considered to be probability of surviving the one day multiplied by the probability surviving the second day given that patient survived the first day. This second probability is called as a conditional probability. Although the probability calculated at any given interval is not very accurate because of the small number of events, the overall probability of surviving to each point is more accurate.
The graph plotted between estimated survival probabilities (on Y axis) and time past after entry into the study (on X axis) is called a survival curve. The survival curve is drawn as a step function: the proportion surviving remains unchanged between the events, even if there are some intermediate censored observations. Survival curves are very useful when comparing curves from two different groups of subject. (e.g. a cohort of patients that receive the same treatment can be divided in multiple groups based on the outcome of a biomarker) [173] (Figure 3.4)

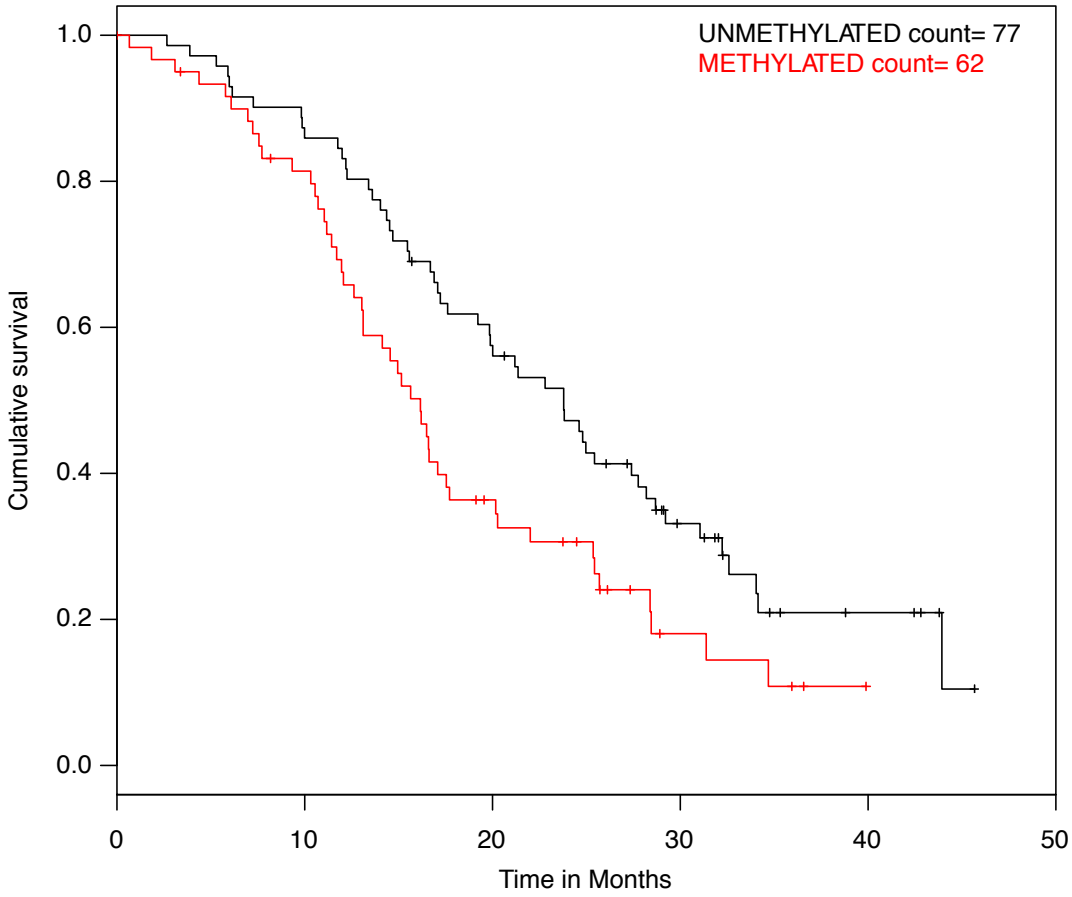


Figure 3.4: Plots of Kaplan-Meier product limit estimates of overall survival for two groups of patients receiving chemotherapy partitioned by the outcome of the promoter methylation status of the BNIP3 gene, crosses indicate censored events

3.4.2 Log-rank test

The logrank test is used to test the null hypothesis that there is no difference between the populations in the probability of an event (here a death) at any time point. The analysis is based on the times of events (here deaths). For each such time we calculate the observed number of deaths in each group (O1 and O2) and the number expected E1 and E2 if there were in reality no difference between the groups. If a survival time is censored, that individual is considered to be at risk of dying in the week of the censoring but not in subsequent weeks. The logrank test is based on the same assumptions as the Kaplan Meier survival curve. The test statistic is:

$$\text{Log-rank statistic: } = \frac{(O_1-E_1)^2}{E_1} + \frac{(O_2-E_2)^2}{E_2}$$

The significance of the log-rank statistic can be drawn by comparing the calculated value with the critical value (using chi-square table) for a degree of freedom equal to one. [174]

3.4.3 Cox Proportion Hazard Model

Cox proportion hazard model enables us to test the effect of multiple independent variables on survival times of different groups of patients, just like the multiple regression model. Hazard is defined as the probability of dying at a given time assuming that the patients have survived up to that given time. Hazard ratio is also an important term and defined as the ratio of the risk of hazard occurring at any given time in one group compared with another group at that very time. Both log-rank test and Cox proportion hazard test assume that the hazard ratio is constant over time. [175]

3.5 Conclusion and Perspectives

Over the last decade, next-generation technologies have yielded substantial insight into the epigenetic principles regulating gene expression and the epimutations associated with disease. Based on the same principles as locus-specific epigenetic analysis, which are currently still used for validation purposes, these technologies excel by their high-throughput and genome wide character.

Although currently more expensive than microarrays, it is clear that massive parallel sequencing approaches are taking the lead. Indeed, much is expected from the next generation(s) of sequencers, which envisage single molecule sequencing at an even lower cost and higher speed, and the direct measurement of, for example, DNA (hydroxy)methylation without the disadvantages of chemical conversion reactions [176, 177]. For non-model species, genomes and DNA (hydroxy)methylomes can be obtained within the same experiment. Additionally, as, in contrast to second generation sequencers, no PCR steps are required, sequence dependent (e.g., GC-content) biases will be dramatically reduced [178]. This will allow for a far more reliable quantification of epigenomic modifications, thereby also facilitating comparison of these features between different loci in the genome.

With future generations of sequencing comes the prospect of longer sequence reads [178], allowing for an easier and cost-efficient assembly of novel genomes to be used as background for subsequent epigenomic studies in non-model species. Virtually all epigenomic experiments are currently performed on populations of cells, thereby determining the average of the epigenetic feature under investigation. Longer reads also imply the possibility to evaluate the linkage of DNA (hydroxy)methylation across longer genomic distances, for example, to identify subgroups of alleles in tumors or other heterogeneous samples. In addition, progress is being made in single cell analysis by recent advances in cell selection procedures and measurement sensitivity [179], with bisulfite treatment of an individual oocyte [180] and restriction-enzyme based single-cell DNA methylation analysis [181] as typical examples, and here too long-read single-molecule sequencing might imply a major improvement. Current methods are already capable of obtaining the 147-nt-long stretches of DNA associated with nucleosomes and histone variants/modifications using ChIP-seq/MNase-seq procedures. However, methods like NOMe-seq [124] will benefit from far longer single molecule reads, as this will enable the exact mapping of a large amount of nucleosomes on an individual DNA molecule without even the need for bisulfite conversion. For ncRNAs, there will be the possibility to avoid conversion to cDNA [178], implying a decrease in time and handling required, as well as a less biased quantification (by avoiding both conversion and sequencing-associated PCR).

In conclusion, while micro-arrays and second-generation sequencing have already revolutionized the field, future generations of sequencing

hold the promise of eliminating the majority of remaining problems regarding cost, resolution, biases, and research on non-model species. However, as always, new challenges will arise, particularly with regard to data management and high-throughput (statistical) analysis of the petabyte scale data coming our way.

Part II

EPIGENETIC TOOLS

1

Epigenome-wide Profiling

From article:
A Map Of The Human Methylome
Geert Trooskens , Tim De Meyer , Veeck Jurgen, Simon Denil, Jean-Pierre Renard, Pierre Dehan, Joachim De Schrijver, Gerben Menschaert and Wim Van Criekinge

Changes in DNA methylation play a crucial role in gene regulation and disease progression. Genome wide techniques to quantify and qualify methylation are rapidly evolving, increasing the need for a straightforward comparative methylome analysis. Particularly methyl-CpG binding domain capturing based sequencing (MethylCap-Seq) is a low-cost, high-resolution technology to uncover DNA methylation in a truly genome-wide manner and is becoming increasingly popular. We applied this technique to 345 human samples from different origins and constructed a map of the human methylome by identifying the methylation-prone regions, which we attribute the name Methylation Cores (MCs). The map enables researchers to reduce the comparison problem to a discrete amount of variables. Based on this methodology, we identified 3,618,706 MCs comprising 39,6% of the genome and 53.4% of approximately 28 million human CpGs dinucleotides. We observed high enrichment of MCs in CGIs and exons, but less variability in DNA methylation compared to promoter, intronic and intergenic regions. In contrast to CpG islands (CGIs) and exon methylation that appears to be high and stable (Figure 1.1), we found clear indications that highly differential regulatory information is contained within the promoter region, but also in introns, where methylation could be involved in the regulation of gene splicing. Validation by targeted bisulfite sequencing data indicates that the map of the human

methylome encloses more then 90% of CpGs exhibiting a minimum of 10% methylation.

[Supplemental material is available for this article. The map of the human methylome is browsable in our methylome browser and downloadable at http://h2g2.ugent.be/mhm];

1.1 Introduction

DNA methylation of cytosine at the carbon-5 position is a well-studied epigenetic mark and is in humans predominantly found in a CpG-dinucleotide sequence context. When located at gene promoters, DNA methylation is generally accepted as a repressive mark [56]. By its transcriptional control function, DNA methylation plays a crucial part in tissue differentiation and aging [182], and its deregulation is often associated with disease [183]. For example, epigenetic abnormalities such as hypermethylation of tumor suppressor genes and hypomethylation of oncogenes are thought to be integral to the development of cancer [23]. Furthermore, epigenetic modifications can predict the outcome of certain treatments such as chemotherapy [62], detect cancer at early stages [184] or may be useful in patient risk stratification [185]

Since the discovery of altered DNA methylation patterns in human cancer [186], many methods have been used to assess DNA methylation. Until recently, technologies like methylation specific PCR [167] and locus-specific bisulfite sequencing [70] targeted specific sites of the genome. Typical targets include gene promoter regions or CGIs. One of the main drawbacks of these techniques is the limited applicability in genome wide experimental set-ups. Array based DNA methylation techniques [187] provide researchers with a reasonably good genomic coverage (genome-scale), but next to the inherently lower quality of hybridization methods, this methodology depends on the arbitrary choice of probes, and therefore lacks the resolution of truly genome-wide sequencing approaches.

Next-generation sequencing platforms offer several orders of magnitude higher throughput - i.e. gigabases of DNA in a single run - at reasonable costs, creating new opportunities for researchers to assess DNA methylation on a genome-wide scale. The characterization of whole genome DNA methylation patterns using bisulfite sequencing requires hunderds of Gbases of sequencing data to obtain sufficient coverage, implying practical limitations by todays standards to study and compare multiple samples. Targeted bisulfite sequencing by hybrid selection [81]

offers bisulphite sequencing of preselected regions but the loss of genome wide character is an inherent trade-off. Reduced representation bisulphite sequencng (RRBS), reduces the portion of the genome analyzed through MspI digestion and fragment size selection [188]. Alternative methods, not based on bisulphite conversion,filter out the non-methylated regions of the genome, retaining the methylated fractions, which are subsequently sequenced. A first method, anti-5-methyl cytosine antibody immunoprecipitation (MeDip) retains single stranded DNA with at least one methylated cytosine residue [110]. Marginally methylated regions can compromise a large section of the retained DNA portion, generating high background levels thus complicating the analysis. On the other hand, combining the native human methyl-CpG binding domain protein (MBD) for precipitation of multi-methylated genomic DNA fragments, with paired-end deep sequencing renders a high-resolution, whole genome methylation profile (MethylCap-Seq). A comparison study of both methods revealed that MBD sequencing can distinguish twice as many differentially methylated regions than MeDip [84]. While there have been attempts to render a map of the human methylome, this study is one of the first to render it on a genome wide level for MethylCap-Seq. Although it is virtually impossible to cover all possible methylation sites (this would require performing MethylCap-Seq on each different human tissue type and pathological variant thereof), even an incomplete map can identify those loci which are nearly always methylated or - more importantly - for which the methylation status is highly variable in the human genome.

The effects of DNA methylation depend mostly on longer stretches of CpG methylated DNA [189], implying that the lower resolution (read length) of MethylCap-Seq compared to bisulfite sequencing is a minor limitation in biological terms. This also entails that regions of CpG methylation or MCs, are an adequate unit for a methylome map. The collection of these MCs then composes the methylome. From a conceptual point of view, MCs are similar to CGIs except that for the former methylation has actually been observed while the latter are an aid to identify methylated regions. Similar to CGIs, several MCs may be present in a single gene or even a single promoter region. Also from a data-analytical point of view, a map of the human methylome will improve future analyses, since the use of the methylation status of the MCs (typically maximum tag count) ultimately reduces the epigenomic data to a set of scores for a fixed set of variables, thereby facilitating multi-sample comparisons. Furthermore, in contrast to base-pair level data-analysis, the MC methodology allows

reasonable differences in the positions corresponding to maximal CpG methylation.

The methodology outlined below uses raw MethylCap-Seq data of different samples to create the map of the human methylome, including different healthy tissues, cell lines and tumor samples (Table 1.1). Since no normalisation procedures are applied, artifacts are avoided. A Poisson background model is used to identify significantly methylated regions. However, MCs can overlap within a single methylated region due to both technical and biological reasons. The main technical issue is the fact that the resolution of the methodology is limited by the length of the sequenced fragments, which implies peak broadening and potential overlap. Biologically there are also clear indications that DNA methylation spreads to neighbouring loci [190], although this does not imply that the biological effects are the same for these loci.

Therefore, a conservative set of rules was derived that identifies adjacent MCs in a single "significantly methylated" region. First, two control sets of putative adjacent MCs were identified, i.e. a clearly positive control set for which the regions correspond with 2 MCs, and a negative control set for which this was clearly not the case. The former featured a clear decrease in CpG density between the cores, while for the latter the opposite was the case. Subsequently, decision trees using the characteristics (but not the CpG density) of the putative adjacent MCs as input data were constructed to predict the status of the control sets. As a consequence, the rules constructed by the optimal decision trees do not directly depend on the underlying CpG density and can be used to identify adjacent MCs in a genome wide manner.

After identification of the methylation prone regions, we calculated the sensitivity of our map in comparison with RRBS data from the ENCODE project [191]. We report the variation for each genomic region type and pinpoint the genomic regions that can differentiate between tissues and disease types.

Other initiatives to chart a map of the human methylome [191] [192] provide a rich resource for integrated analysis of DNA methylation. These undertakings were based on massive RRBS and Array based techniques sequencing. We provide a similar resource for the complementary Methylcap-seq technique for a large amount of samples (345) of different origins.

Table 1.1: Properties of the samples used to build the MethylCap-Seq MCs.

Tissue origin	Disease	Type	Samples
Basal cell carcinoma	cancer	primary sample	9
Bladder	cancer	cell line	6
Blood plasma	cancer	other	1
Blood plasma	normal	primary sample	1
Brain	cancer	primary sample	74
Brain	normal	primary sample	1
Breast	cancer	cell line	6
Breast	normal	primary sample	21
Cervix	cancer	cell line	13
Cervix	cancer	primary sample	52
Colon	cancer	primary sample	23
Colon	normal	primary sample	5
Head & neck	cancer	primary sample	20
Induced pluripotent stem cell	cancer	primary cell line	1
Induced pluripotent stem cell	normal	primary cell line	1
Kidney	cancer	primary sample	21
Kidney	normal	primary sample	1
Lung	cancer	primary sample	10
Melanocytes	normal	primary sample	1
Melanoma	cancer	cell line	6
Monocytes	normal	primary sample	1
Nose polyps	normal	primary sample	3
Ovaries	cancer	primary sample	20
Prostate	cancer	cell line	3
Stem cell	normal	primary sample	1
Thyroid	normal	cell line	1
White blood cells	cancer	cell line	2

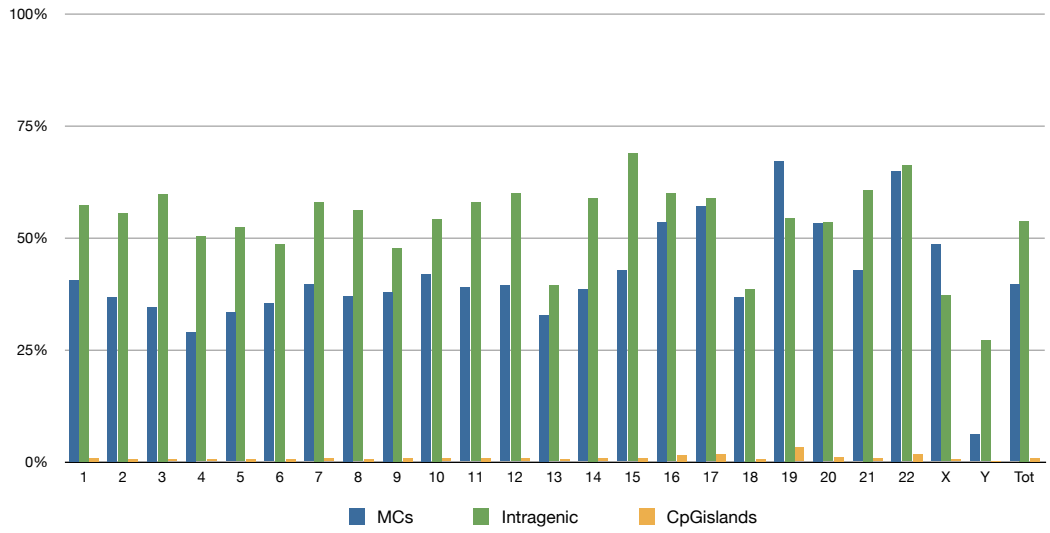


Figure 1.1: Composition of the genome. The portion of each chromosome covered by the MCs , CGIs, promoters and exons. The Data was generated on the human genome version GRCh37/hg19 with annotation from Ensembl release 56.

1.2 Results

1.2.1 Putative MC identification

First, significantly methylated regions in the summed profile are identified by imposing a minimal total coverage cut-off. To set this cut-off, the summed profile was split in different crude chromosomal regions without overlapping mapped fragments, i.e. regions separated by a region with at least one nucleotide and coverage equal to zero. 108954 of these chromosomal regions could be identified. Subsequently, those crude regions for which the corresponding DNA-sequence did not contain any CpGs were identified, 14237 in total, and the maximal coverages in those putative noise regions were listed. It was clear that the large majority of these regions were featured by very low amounts of mapped reads, with a median [IQR] of 2 [1-3]. However, as several of those putative noise regions clearly exhibited higher coverages, the 99.9th quantile of noise regions maximal coverages was used as cut-off for significant methylation. For the autosomes this yielded a cut-off of 46, for the X-and Y-chromosome this was determined individually to take into account their different frequencies, resulting in cut-offs of respectively 21 and 13.

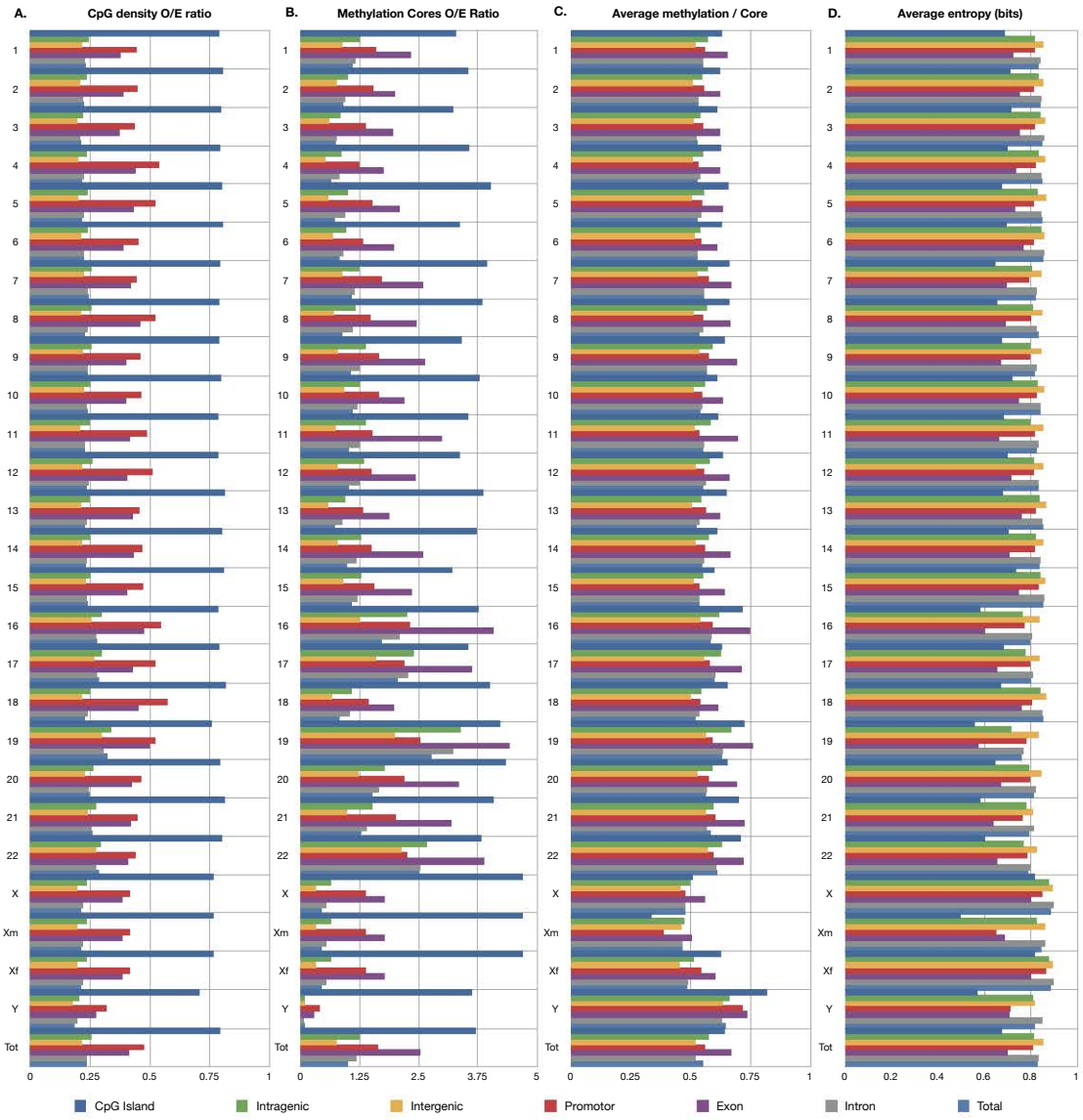


Figure 1.2: Overview of core properties per chromosome. (A) CpG density in O/E ratio. (B) The MCs O/E ratio. This is the observed MC coverage in base pairs divided by the expected MC coverage in base pairs, the expected coverage is calculated based on the total genome wide MC coverage. (C) The average methylation is the average methylation in the cores throughout the samples, where we considered a minimum of one mapped paired-end read as methylated for that sample. (D) Average information entropy is defined as the average MC binary entropy in bits. The boundaries for the promoter region were defined from -1000 to 200 relative to the TSS. CGIs were defined according to the Jones definition (Jones and Baylin 2002). Human genome version GRCh37/hg19 was used for this analysis. (Xm is the data from male samples and Xf the data from female samples).

This cut-off was imposed to the summed profile, and regions without CpGs were removed, yielding a total of 2458602 chromosomal regions. The latter regions were selected to create the MHM. Significant regions without CpG dinucleotides were typically very narrow (median 3 nt), with CpGs in the non-significant adjacent sequences (data not shown). Within the selected regions, a total of 6965691 putative Methylation Cores were identified (i.e. number of relative maxima in all significant regions).

Control set creation, decision tree based rule inference and application For each putative set of MCs and for each sliding window size w (see Materials and Methods), the minimum difference in CpG density of each of the two putative MCs (CpG position with maximal coverage) and the border in between was determined. As it is to be expected that the CpG density at the border between different true MCs is locally lower, cores with a difference in density above a certain threshold t were considered true positives, and cores with a difference in density below a certain threshold -t as true negatives. For each window w and threshold t, the ratio true positives/true negatives was determined. We expect the optimal control set to have the largest ratio, since this implies minimal noise impact (using threshold t), and an optimal matching between the window size w of the CpG density estimation and the underlying methylation degree. Using this methodology, the optimal control set had a true positive/true negative ratio of 1.24, for an optimal window of 170 nt and a threshold of 3.

With the defined setting, the number of true positives was 132149 and there were 106753 true negatives. A decision tree was created which yielded a single rule: putative MCs can be considered as such if the difference in distance between the locations of maximum coverage of the cores and the border in between is at least 64 nt. The accuracy of the model was 95.6%, with a sensitivity of 96.4% and a specificity of 94.6%. Application of this criterion on the complete dataset yields a total of 3618706 MCs in the human genome, i.e. 51.9% of the candidates. The cores are annotated as a genomic region, including the locus (CpG) associated with maximal coverage, which we define here as the MC center. The median [IQR] MC size was 281 [118;464] nt.

1.2.2 Distribution of the MCs

The MCs were annotated by their position relative to the genes in their genomic surroundings (see Materials and Methods section). The total ge-

nomic coverage of the MCs is 1,133M bp (39,63% of the genome), 660M bp intragenic (expected 609M), 472M bp intergenic (expected 524M), 29M bp in promoter regions (expected 24 M) and 65M bp overlapped with exons (expected 44M) (Figure 2). The distribution of MCs is enriched in CGI s (O/E 1.35) and exons (O/E 1.47), and to a lesser extend in promoters (O/E 1.21). The relative MC coverage differed significantly in between chromosomes. While chromosome Y consists of less than 6.15% MCs, more than 67.19% of chromosome 19 is made up out of MCs. These values are positively correlated with the amount of CpG-islands and Genes on each chromosome (Figure 1.1). The different genomic regions (intergenic, intragenic, promoter, intron,exon, CGI) have very similar CpG densitys in between chromosomes, while the relative coverage of the methylation cores shows some variability (Figure 1.2) . The O/E ratio of methylation cores is s(Figure 1.1) for CGIs, while the variability of MC coverage in between chromosomes can be attributed to the non-CGI regions. We found significant correlations between MC coverage and CGIs, promoter and exon coverage (spearman correlation of 0.952, 0.691 and 0.856 respectively) in between chromosomes.

1.2.3 Average Methylation and Variability of the MCs

To asses the variability of the MCs we calculated the binary entropy function of every MC in the 345 samples. We considered a sample methylated for a MC if we mapped back at least one MethylCap enriched read within the core boundaries. If we look at the average entropy for the MCs, we observe promoter (0.80bit), intronic (0.83bit) and intergenic regions (0.85bit) as having the lowest methylation conservation (variability). CGIs (0.67bit) and exonic regions show higher methylation conservation (0.69bit) and high average methylation levels (63,95% and 66,73% respectively). Chromosome comparison revealed that cores in CGI on the X chromosome are almost twice as likely to be methylated in females (62,7%) compared to males (33,6%) (Figure 1.2).

1.2.4 DNA methylation properties around the TSS

The MC coverage varies around the TSS. It shows high and constant in the intergenic region, lower in the promoter region from -500 to 0 , very high enrichment at the beginning of exon 1 and low coverage in

the intronic region. The average methylation of the MCs shows a minimum overlapping with the TSS while the entropy is higher intergenically compared to the intragenic regions, with a small peak on the TSS and 500 bases downstream. There is a clear difference in DNA methlyation properties between the intergenic, promoter and intragenic regions. Interestingly, the promoter is distinguishable from the intergenic region at around -750/-500 bp upstream by lower MC coverage, lower average methylation and lower entropy.The beginning of the first exon clearly has high enrichment in MC coverage,low average methylation and is located around the top of a small peak in entropy (Figure 1.4)

1.2.5 Validation with reduced representation bisulphite sequencing

We validated the map of the human methylome by comparing it with RRBS data from 174 cell lines profiled in the ENCODE project (The Encode Project Consortium. 2012). DNA Methylation in these cell lines was measured by a technique of hybrid selection and bisulphite treatment followed by massively parallel sequencing. The sensitivity of our map was measured by the amount of methylated CpGs found in the bisulphite data from the cell lines that overlap with the MCs. We used different cutoffs to consider a CpG methylated in the cell lines as seen in figure X. This sensitivity is a measure for the sensitivity of the technique (MethylCap-seq) and the completeness of the map of the human methylome. Of the CpGs showing at least 5% methylation, 85% fell within the boundaries of the methylation cores. CpGs showing at least 10% methylation gave 91.7% sensitivity and above 20% methylated we measured a sensitivity of 95.8%. Because we try to predict the methylation prone regions, only the sensitivity is a relevant measure when validating the MCs.

1.2.6 Discussion

Over the course of the last two decades, epigenetics has developed into a mature scientific field of research. Ultimately a genome-wide tissue-specific epigenetic characterization would facilitate the understanding of gene expression and differentiation regulation , and related aberrations in disease. Establishing a map of the human DNA methylome allows further epigenetic research to pinpoint the regions of interest by acting as a guide to spot the methylation prone locations. Furthermore, it allows the

gigabases of high-throughput sequencing data to be reduced to a list of discrete variables. These variables can then be analyzed and compared in similar ways as probe data or RNA-seq data. Therefore, the identification of the MCs allows us to compare methylation with other genome-wide characteristics, e.g. gene expression, SNPs and histone modifications, but also with phenotypic properties such as differentiation and disease progression.

MethylCap-Seq is one of the most sensitive high-throughput sequencing techniques to profile genome-wide methylation, providing high sensitivity while allowing analysis of multiple samples on a single run. Comparison with high-throughput bisulphite data (The Encode Project Consortium. 2012) showed that more then 90% of the CpGs that were at least 10% methylated fell within the MCs. The 15% methylated CpGs outside the MCs is partly due to the incompleteness of the map of the human methylome, but might also be partially attributed to the intrinsic capture bias of the MethylCap-Seq method. Additionally, the simple Poisson background model used to impose a threshold for "significant" methylation might be overly conservative for this kind of data. However, as no less than 39.6 % of the human genome consists of MCs, a further increase of sensitivity/completeness of the Map of the Human Methylome at a cost of decreased specificity might be not appropriate.

Our data exhibit major differences in the relative amount of methylation prone regions between chromosomes. The XY sex chromosomes have the lowest MC coverage (1% for Y and 5% for X). The MC coverage for the autosomal chromosomes varies between 7% for chomosome 13 to 32% for chromosome 19. Most of these differences can be attributed to a strong positive correlation between MC coverage and the number of CpG islands/genes. However, the low amount of MC coverage on the sex chromosomes could have additional causes.

Chromosome comparison revealed that cores in CGIs on the X chromosome are almost twice as likely to be methylated in females (62,7%) compared to males (33,6%), while the non-CGI regions remained equally methylated. The difference could be attributed to the X-chromosome inactivation, a dosage compensation mechanism that silences the majority of genes on one X chromosome in each female cell, This finding is in concordance with earlier studies [193] and [194].

Our study reveals that DNA methylation is most variable when located in gene promoters, introns and intergenic regions, while the highest and most s(Figure 1.1) methylation levels are located in the CGIs and ex-

onic regions . This indicates more qualitative effects of promoter, intron and intergenic methylation, whereas exon and CGI methylation should be more considered as quantitative features. Qualitative effects of promoter methylation on gene expression are well known, but there are also indications that the high differential intronic DNA methylation levels control the binding sites of CTCF (40-45% are located intragenically) that regulate alternative splicing [195].

Interestingly, regulation by DNA-methylation in the promoter region seems to occur from -750bp upstream into the first exon as the methylation properties start to deviate from the average intergenic properties (Figure 1.4). The lowest MC density and variability occurs between -250bp upstream and the TSS while the highest MC coverage between the TSS and 250bp downstream corresponds with an entropy peak.

It should be noted that the determination of the methylation degree and variability will be affected by the lack of normalization of the individual samples. Normalization of MethylCap-Seq data is not trivial due to the (often large) differences in global methylation degree, and we opted to avoid the introduction of noise at a cost of lower sensitivity. However, as it is expected that all loci would suffer at a similar level from this limitation, and taking into account the large amount of different samples, the relative difference between the different functional regions of a gene (or intergenic) should not be affected by the lack of normalization.

The principal advantages of MethylCap-Seq over Bisulfite sequencing comprise the experimental cost and required resources per sample. Bock et al. [84] estimated one billion reads per sample to be a viable compromise between breadth and depth of sequencing for bisulfite sequencing, while MethylCap-seq would reach sufficient coverage with 20-30 million single end 35bp reads. Thus, a genome wide DNA methylation profile requires 50 to 100 times less resources with MethylCap compared to BS-seq. Although we attained a sensitivity of 85% when compared to bisulfite sequencing, there are some disadvantages of MethylCap-Seq compared to Bisulfite sequencing. One of the pitfalls of MethylCap-Seq (and the other enrichment techniques) is that the signal is not linearly related to the actual methylation level. Several Methods were developed to correct for this bias [196]. As with most enrichment methods it is important to implement a correction step associated with CpG density. In this case, CpG dense regions will be interpreted as having a higher CpG methylation ratio as equally methylated lower CpG dense regions [102]. The MC approach does not make assumptions about this inherently linked bias

to enrichment techniques, as we only want to chart the methylome. We strongly recommend taking this bias into consideration when comparing different MCs. A second disadvantage of MethylCap-Seq, as compared to Bisulfite sequencing, is that the resolution is not on CpG base-pair level, but rather depends on the size of the sonicated DNA-fragments. Fragments determined by MethylCap-Seq to be methylated can still contain individual CpG sites which are not methylated. When using MBD sequencing, it is important to note that MBD is not able to pick up non-CpG methylation, although the latter only appears to be relevant in embryonic stem cells. It will depend on the as-yet mostly unknown importance of non-CpG methylation to assess whether this is a significant shortcoming of this technique. Although non-CpG methylation is a ubiquitous feature of human embryonic stem cell DNA, MBD is an integral part of the human DNA methylation regulation mechanism. One could assume a bigger role for CpG methylation versus non-CpG methylation, especially in tissue differentiation, disease progression and aging.

The accuracy of the MHM will increase with the amount of samples from different tissues and diseases. In addition, as the number of samples for certain tissues increases, a logical step is to profile the unique epigenetic fingerprints of distinct sample groups. Comparing the MHM with expression and other data such as histone marks will enable us to functionally annotate the MCs, helping us to understand the mechanisms involved in epigenetic regulation. This approach is a flexible methodology that can be ported to other enrichment based genome wide high-throughput methods and future third generation sequencing technologies.

1.3 Methods

1.3.1 Human Samples

A total of 345 samples were used to build the map of the human methylome (Table 1.1). gDNA was extracted from samples with the Easy DNA kit (Invitrogen K1800-01), using respectively the appropriate protocols number 4 (cancer cell lines) and number 3 (fresh frozen samples). Paraffin embedded tissue samples were pretreated with xylene followed by a classic chloroform/phenol extraction. DNA concentration was measured on a Nanodrop ND-1000 (Thermo Scientific, Wilmington, North Carolina, USA).

1.3.2 Fragmentation and MBD-capture

Fragmentation was performed on Covaris S2 with following settings: duty cycle 10%, intensity 5, 200 cycles per burst during 190 sec, to obtain fragments with an average length of 200bp. The power mode was frequency sweeping, temperature 6-8C, water level 12. Maximum 3g was loaded in 130 l TE in a microtube with AFA intensifier (Covaris, Woburn, Massachusetts, USA). For samples with less input DNA (down to 500 ng), we diluted the DNA in 1:5 diluted TE. DNA with an input of 3g was then analyzed on the Agilent 2100 (Agilent Technologies, Santa Clara, California, USA). DNA with an input lower than 3 g was concentrated in a rotary evaporator to 25 l and the fragment distribution was checked on a high sensitivity DNA chip. Methylated DNA was captured using the MethylCap kit (Diagenode AF-100-0048, Belgium). According to the manufacturers protocol, starting concentration was 200 ng. The yield was typically between 0.5 and 8 ng total captured DNA and sometimes too low to measure. Fragments were subsequently sequenced using the Illumina Genome Analyzer II. The concentrations of the fragmented and captured DNA was determined on a Fluostar Optima plate reader (BMG Labtech, Offenburg, Germany with the Quant-iT Picogreen dsDNA assay kit (Invitrogen P7589, Merelbeke, Belgium) on 480/520nm.

1.3.3 Library preparation, Amplification and Sequencing

This is a modification of the multiplexed paired end ChIP protocol (Illumina, San Diego, California, USA). We used the DNA Sample Prep Master Mix Set 1 (NEB E6040) in combination with the Multiplexing Sample Preparation Oligo Kit (96 samples, Illumina PE-400-1001). We used the total amount of fragmented DNA and followed the NEB protocols (New England BioLabs (NEB) E6040, Ipswich, Massachusetts, USA): NEBNext End Repair Module Protocol, purified on a Qiaquick PCR Purification Kit (Qiagen 28104) and eluted in 37 l EB (Elution Buffer). After applying NEBNext dA-tailing Module Protocol, we purified it with a Minelute PCR Purification Kit (Qiagen 28004) and eluted in 25l EB. NEBNext Quick Ligation Module Protocol, purify on a Minelute PCR Purification Kit (Qiagen 28004) and eluted in 30 l EB. Here we used the multiplexing sequencing adapters provided in the Multiplexing Sample Preparation Oligo Kit. Size selection of the library was done on a

2% agarose gel (Low Range Ultra agarose Biorad 161-3107). We used a 1Kb Plus ladder (Invitrogen 10787-018) and ran the gel at 120V for 2 hrs. A fragment of 300 bp +/- 50bp was excised and eluted on a Qiagen Gel Extraction Kit column (Qiagen 28704), then eluted it in 23 l EB. We followed the Illumina library amplification index protocol with the following alterations: Use 22 l DNA and perform 21 cycles. Purify on a Qiaquick PCR Purification column (Qiagen 28101) and elute in 50 l EB 1:5 diluted. Concentrate in a rotary evaporator to 10 l and put 1 l on an Agilent 2100 HS DNA chip. Determine the concentration with smear analysis on the Agilent 2100. Dilute the samples to 15 nM. Perform qPCR on the samples (1:500 diluted) and use a dilution of PhiX index3 as standard.

Calculate the exact concentration: Dilute to 10nM (endconc. 0,1 N) and pool 4 patients per lane (in total 7 lanes and 1 control lane with the PhiX index 3 control). If concentration is lower than 10 nM dilute to the lowest concentration. After denaturation with NaOH (endconc. 0,1 N), we diluted the samples to 12 pM. The paired end (PE) flow cell was prepared according to the Cluster Station User Guide (Manual). Sequencing was performed according to the GAIIx user guide performing a Multiplexed PE Run with 2 x 45 cycles.

1.3.4 Mapping

For all samples together, the paired end sequence reads are mapped using bowtie (v0.12.7) software [128]. The bowtie parameters were set to 0 mismatches in the seed (first 28 nucleotides). Only unique paired reads were retained and both fragments must be located within 400bp of each other on the human reference genome (GRCh37/hg19). The mapped PE reads allow us to give every nucleotide in the genome a coverage value. Multiple paired reads with the exact same location in one sample are discarded, as these are most likely amplified from the same sequence.

1.3.5 Background estimation

The background signal in the composite epigenomic profile arises from the non-specific capturing of unmethylated DNA-fragments. We modeled the background of this profile as a Poisson distribution, with parameter lambda estimated as the total length of mapped sequenced reads divided by the total sequenced length of the used build (3.17 billion bp,

GRCh37/hg19). Using this model, for each locus (nucleotide level) the probability under the null-hypothesis (pure background signal) can be estimated and actual signal can be defined by the minimal coverage value (intensity) for which the null- hypothesis is rejected with a certain significance level alpha (here, we used alpha = 0.05). For background estimation and MC identification, we used R version 2.11.1.

1.3.6 Methylation-core identification

Method. After filtering out the intensities designated as background, we aimed to develop a set of rules that can identify separate MCs within each significantly methylated region. Since validation data is not available, we used the underlying CpG-density as validation criterion to generate a control dataset. Subsequently, we applied decision trees on this control dataset, with putative core shape properties as input parameters, to infer the set of rules. Since CpG-density is only used to generate a control dataset, and not incorporated in the rules, the latter are CpG-density independent and therefore more broadly applicable. Importantly, we wanted to avoid an implosion of MCs, which would dramatically increase the number of variables, by selecting the most stringent set of rules (vide infra).

Filtering. In a first step of the procedure, the data -after background elimination- was further reduced by retaining only those intensity values corresponding with CpG dinucleotides and considering only those isolated genomic regions with at least two significant Cs (both strands). Subsequently, within each of the isolated genomic regions, all relative coverage value (intensities) minima and maxima were determined. Regions without local minima were instantly included in the final set of MCs since these are clearly distinct methylation sites.

Training set selection. A subset of the other regions was used to identify the set of rules based on the underlying CpG-density. The basic assumption for selecting this subset was that actual neighboring MCs are featured by a clear decrease in CpG density, whereas in a region that should certainly be considered as a single MC the opposite would be true. This training subset was determined as follows: In a first step, for each CpG dinucleotide within each chromosome, the number of nearby CpGs (CpG-density) in a sliding window with width w was determined ([i+/-round(w/2)] of the CpG at locus i), for w ranging from 5 to 1000 nt (per 5 nt). Subsequently, for each relative coverage minimum within each re-

maining significantly methylated region and value of w, the CpG-density at the relative coverage minimum was compared with the CpG-densities at the two neighboring coverage intensity maxima and the minimum difference was retained. Negative differences indicate that the CpG-density is higher at the border between two putative MCs than in (at least one of the) putative MCs, while positive differences demonstrate that the coverage value minimum also has a lower CpG-density than the neighboring putative MCs. Based on these differences a control set was selected (see results section) including both clearly distinct MCs (large positive difference) and their counterparts (large negative difference).

1.3.7 MC identification

Subsequently, decision trees were built with the package rpart (version 3.1-46) in R, using the Gini index, equal priors, and 10 fold cross-validation, to predict the status of the control set. For clarity, we use the term border for the CpG in between two putative MCs, border cove rage for the coverage value at this locus, and the term MC coverage for the maximum coverage value (at a CpG) in a specific putative MC. Input features included in the analysis were then: minimum and average MC coverage; minimum and average absolute difference between MC coverage and border coverage; minimum and average relative differences (i.e. divided by border coverage); mean and minimum relative border coverage (i.e. border coverage divided by respectively mean and minimum of both MC coverages); and finally minimum and average distances (in nt) between the border and the location of the MC coverage maxima. The rules derived from the optimal decision tree were then imposed on the original relative minima to obtain the final variables, the MCs.

1.3.8 Analysis of the genomic distribution and properties of the MCs

The MCs were annotated by their relative position to the genes in their genomic surroundings using the gene annotations from Ensembl (release 56) on the human genome assembly version GRCh37/hg19. Gene silencing through DNA methylation is mostly linked to promoter regions . Therefore we categorized the MCs by location in promoter, exonic and intronic gene regions. Gene p romoters were defined as the -1000bp to 200bp region relative to the transcription start site (TSS). The calcula-

tions were done on chromosomes 1 through 22, X and Y. We used only the ungapped sequences. Ungapped sequence lengths are calculated by summing the length of the sequenced bases only. No 'Ns' are included in the count. CpG Observed/Expected ratios are calculated according to Takai et Al. [197]. MC Observed/Expected ratios were calculated by taking the coverage of the MCs for a certain region and dividing it by the expected amount based on the average MHM coverage of the genome. The entropy for a region was calculated as the average binary entropy function (in bits) [198] of every MCs contained within these genomic regions. The binary entropy function out of the information theory has been succesfully used to quantify the sequence conservation in nucleotide and protein sequences [199]. We used the measure to describe the conservation of methylation throughout the samples, with the most variable methylation loci having the highest entropy.

In addition to categorizing the methylation cores in functional genomic regions, we plotted CGI coverage, MC coverage, average MC methylation and average binary entropy from -2000 to +2000 relative to the TSS (Figure 1.4). The plot is the average of 49506 TSS regions. We took the first exon of each gene anotated in ensembl version 56 to determine the TSS.

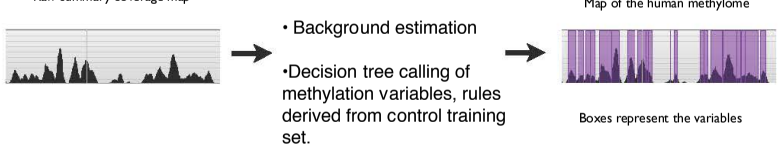
1.3.9 Data Access

Current and former methylome builds are available online at http://h2g2.ugent.be/mhm as download or browsable as separate tracks in our genome browser. The genome browser allows visual comparison with the BS-SEQ data of the cell lines from Lee EJ et al [81], CGIs, mammalian conservation and genomic variation for every genomic region. We foresee to update The Map of The Human Methylome frequently, pinning down new MCs and refining the boundaries of the known cores as more samples from different tissues and diseases are sequenced.

1.3.10 Acknowledgments

We acknowledge the support of Ghent University (Multidisciplinary Research Partnership Boinformatics: from nucleotides to networks), the Flemish Fund for Scientific Research (FWO, to TDM) and the Institute for the Promotion of Innovation in Flanders (IWT, to SD).

Since no normalization procedures are applied, artefacts are avoided.
A Poisson background model is used to identify significantly methylated regions. A conservative set of rules was derived that identifies adjacent methylation prone regions in a single region.



Results

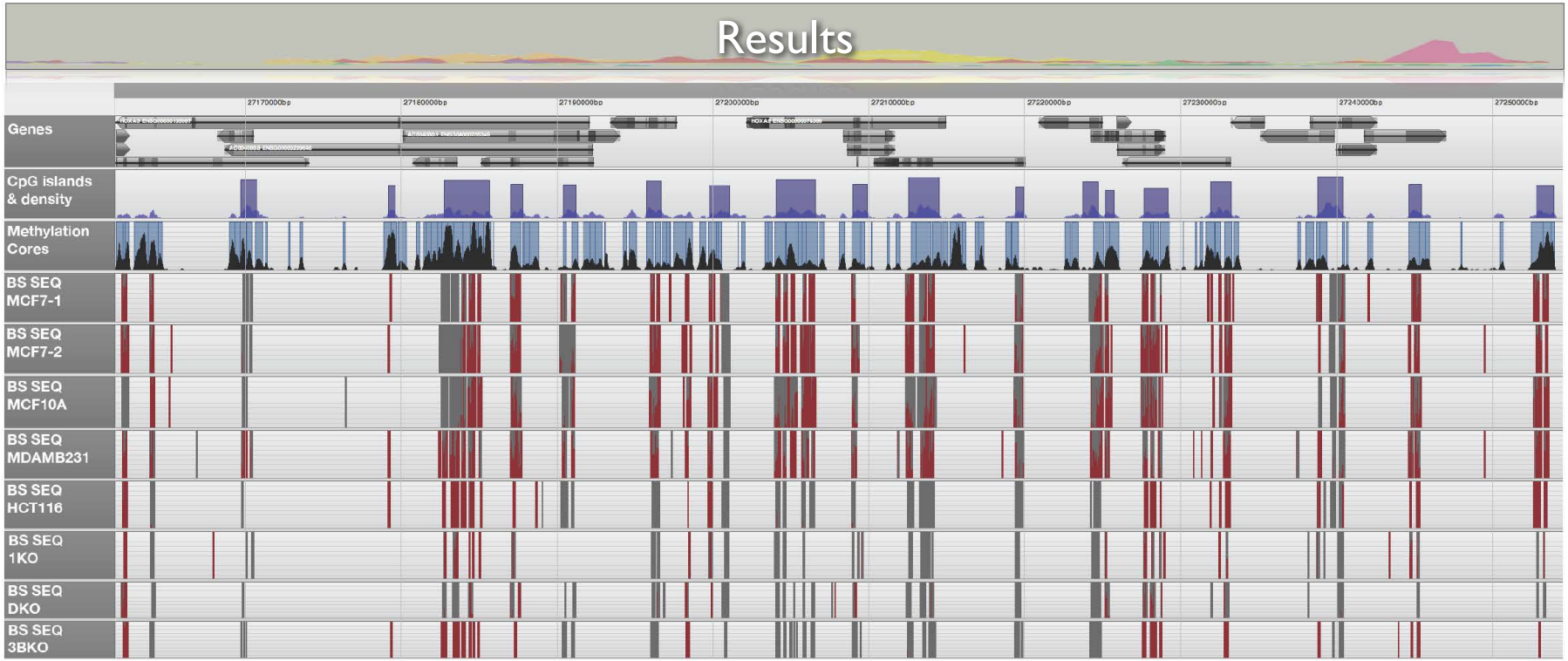


Figure 1. Visual Comparison between the MCs and targeted BS-SEQ for the colorectal cancer cell lines HCT-116, its DNMT1 knockout, DNMT & 3B double knockout, 3B knockout and the breast cancer cell lines MCF7, MCF10 and MDAMB231 from Lee et al. [81] in the HOXA gene cluster (GRCh37/hg19). The first track displays the genes in their strand-specific orientation from transcription start (5 blunt end) to gene end (3 arrowhead). The second track display the CpG density and CGIs highlighted by blue squares (2). The MCs (blue frames) are displayed underneath the latter on top of the raw coverage peaks (black). The targeted BS-SEQ of the cell lines is colored red for methylated and dark gray for not methylated in the enriched regions.

Figure 1: Visual comparison between the MCs and targeted BS-SEQ for the colorectal cancer cell lines HCT-116, its DNMT1 knockout, DNMT & 3B double knockout, 3B knockout and the breast cancer cell lines MCF7, MCF10 and MDAMB231 from Lee et al. [81] in the HOXA gene cluster (GRCh37/hg19). The first track displays the genes in their strand-specific orientation from transcription start (5 blunt end) to gene end (3 arrowhead). The second track display the CpG density and CGIs highlighted by blue squares (2). The MCs (blue frames) are displayed underneath the latter on top of the raw coverage peaks (black). The targeted BS-SEQ of the cell lines is colored red for methylated and dark gray for not methylated in the enriched regions.

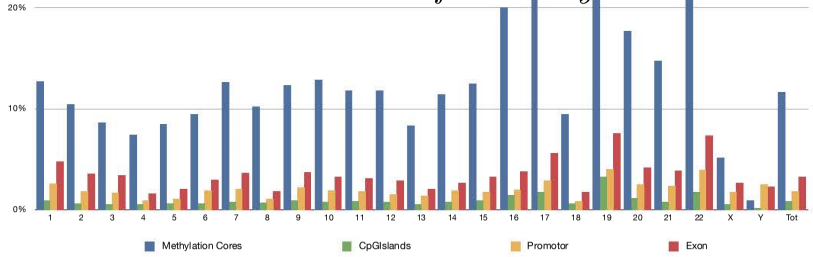


Figure 2. Composition of the genome. The portion of each chromosome covered by the MCs , CGIs, promoters and exons. The Data was generated on the human genome version GRCh37/hg19 with annotation

Conclusions

Based on this methodology, we provide a reference map of ~1.5 million methylation cores. Together they make up about 17% of the human genome and comprise 80% of the CpG islands. Validation by a different MBD kit and targeted bisulfite sequencing data indicates that the Map of the Human Methylome is approximately 95% complete.

Our data exhibit major differences in the relative amount of methylation prone regions between chromosomes. The XY sex chromosomes have the lowest MC coverage (1% for Y and 5% for X). The MC coverage for the autosomal chromosomes varies between 7% for chomosome 13 to 32% for chromosome 19. Most of these differences can be attributed to a strong positive correlation between MC coverage and the number of CpG islands/Genes. The lower amount of MC coverage on the sex chromosomes could have additional causes.

The accuracy of the methylome map will increase with more samples from different tissues and diseases. Comparing the map of the methylome with expression and other data such as histone marks will enable functional annotation of the methylation prone regions, providing a better understanding of the mechanisms involved in epigenetic regulation. This approach is a flexible methodology that can be ported to other genome wide high-throughput methods such as third generation sequencing technologies.

References

1. Lee EJ, Pei L, Srivastava G, Joshi T, Kushwaha G, Choi JH, Robertson KD, Wang X, Colbourne JK, Zhang L, et al: Targeted bisulfite sequencing by solution hybrid selection and massively parallel sequencing. Nucleic Acids Res 2011.

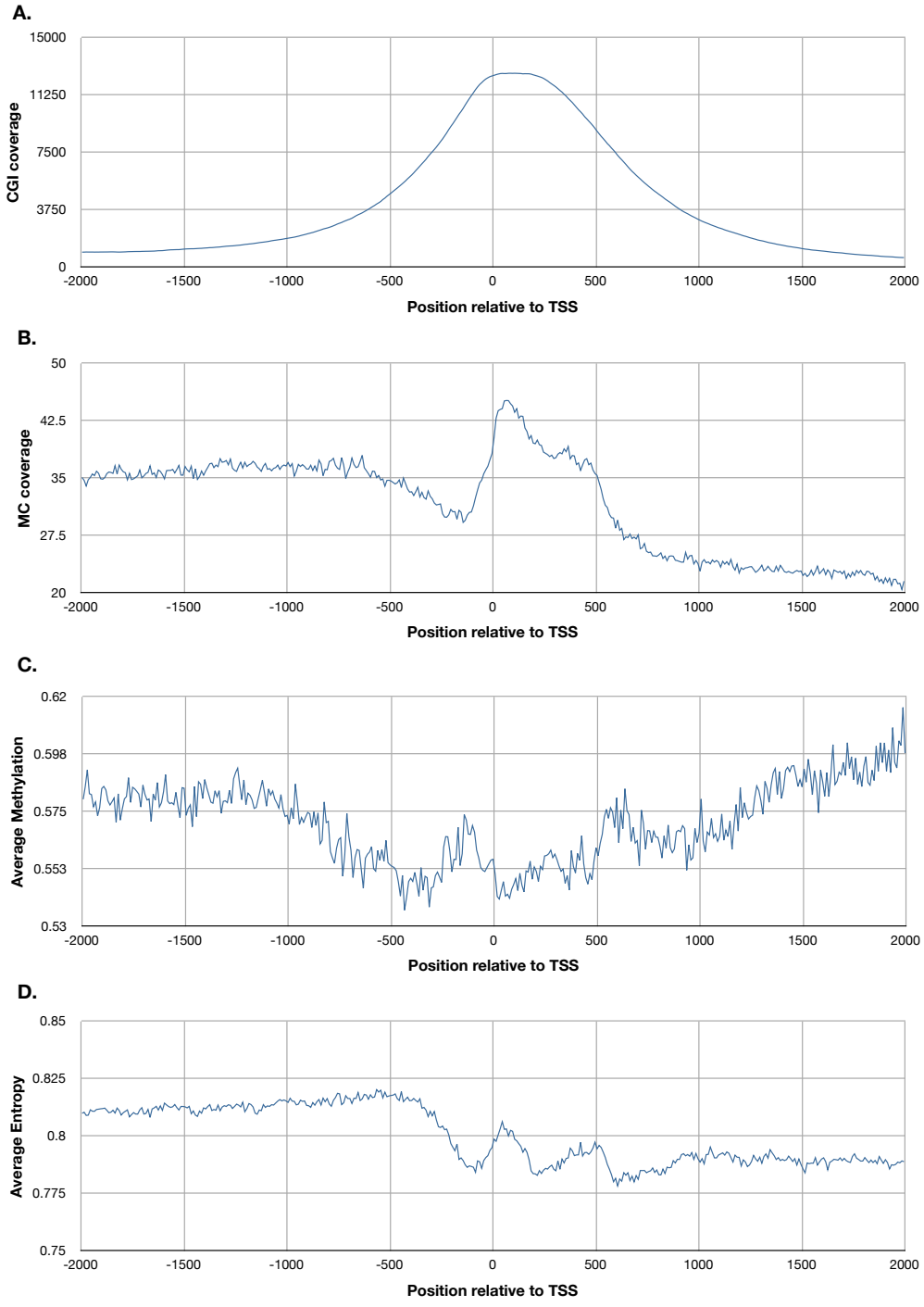


Figure 1.4: Average genome wide properties of DNA methylation relative to the transcription start site (TSS) of 49506 human genes. (A) The coverage of CpG islands shows a peak at the beginning of the first exon. (B) Coverage of the Methylation Cores indicate a steep enrichment between 0 and 500 downstream. (C) The average methylation of the MCs (D) The average entropy.

2

Primer Design for Bisulfite Treated DNA

2.1 Introduction

The polymerase chain reaction (PCR) [200], a method for making many copies of a specific DNA fragment, is one of the most widely applied tools in modern molecular biology. The development of Methylation specific PCR by Herman et Al. in 1996 [167] was a major breakthrough in speed and sensitivity for DNA methylation analysis. MSDP Primer design require a cumbersome gene by gene primer design and experimental validation. Bisulphite DNA treatment results in sequence alterations by converting unmethylated cytosines into uracils. The result is a general sequence complexity reduction as cytosines become underrepresented. Methylation specific primers of a region target methylated cytosines (M-primers containing methylated CpGs) or their unmethylated counterparts (U-primers donating TG (C $\rightarrow$ T) or CA (G $\rightarrow$ A)). When sequencing techniques became more widespread in the early 2000s due to new techniques and lower costs, Methylation unspecific Bisulphite primers became more important. These primers amplify a bisulphite treated region of the genome regardless of the methylation levels. They dont contain CpGs but allow researchers to sequence the CpGs located between the primers in the generated amplicon.

2.2 Bisulphite Primer Design

Primers for bisulphite treated DNA have a higher chance for non-specific binding on other parts of the genome because of the lowered sequence

complexity. Therefore it is very important to map the primers back onto methylated and unmethylated bisulphite version of the genome and only retain those primer pairs that only bind in the designated region. Next to specificity, its also crucial to check for primer dimers

2.3 DNA Thermodynamics

2.3.1 The nearest-neighbour model for DNA

The application of the nearest-neighbor (NN) model to nucleic acids was pioneered by Crothers and Zimm [201] and by Tinoco and coworkers [202]. The NN model for nucleic acids assumes that the stability of a given base pair depends on the identity and orientation of neighboring base pairs. For oligonucleotide duplexes, additional parameters for the initiation of duplex formation are introduced. Importantly, all other sequence-independent effects are also combined into the initiation parameter including differences between terminal and internal NNs and counterion condensation. To account for differences between duplexes with terminal AT vs. terminal GC pairs, two initiation parameters are used: initiation with terminal GC and intiation with terminal AT. An additional entropic penalty for the maintenance of the C2 symmetry of self-complementary duplexes is also included. The total ΔG37 is given by:

$$\Delta G^0(total) = \sum_i n_i \Delta G^0(i) + \Delta G^0(initw/termG \cdot C) + \Delta G^0(initw/termA \cdot T) + \Delta G^0(sym)$$

where ΔG(i) are the standard free-energy changes for the 10 possible Watson-Crick NNs, ni is the number of occurrences of each nearest neighbor i, and ΔG(sym) equals 0.43 kcal/mol if the duplex is self- complementary and zero if it is non-self-complementary. dG for different temperatures can also be calculated with the following formula:

$$\Delta G^0_T = \Delta H^0 - T\Delta S^0$$

The values can be found in table 2.1.

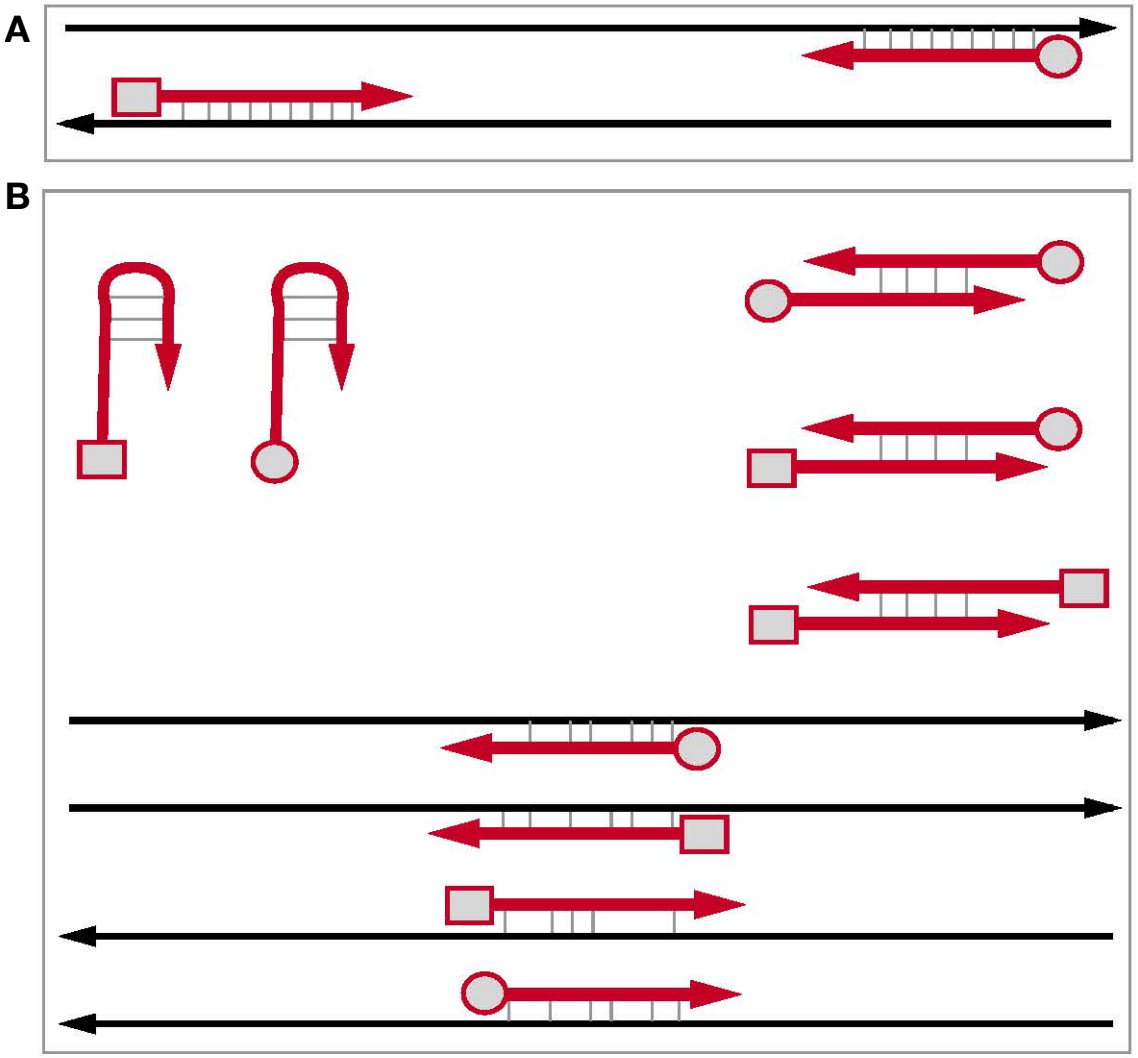


Figure 2.1: Possible binding states of primers and template: The first line is the sense strand of the template; the line below is the antisense strand of the template; the arrow with the square end is the forward primer; the arrow with the round end is the reverse primer; dashed lines indicate binding (or folding) via hydrogen bonding. (A) Desired binding interactions. High rates of binding are desired between the primers and the template priming regions. (B) Undesired binding and folding reactions. Primers can fold onto itself, dimerize with other primers, or bind to the target outside of the priming regions.

Table 2.1: Neirest-Neighbor dH,dS en dG for all possible dinucleotides

Sequence	ΔH^0 kcal/mol	ΔS^0 (cal/k.mol)	ΔG^0 (kcal/mol)
AA/TT	-7.9	-22.2	-1.00
AT/TA	-7.2	-20.4	-0.88
TA/AT	-7.2	-21.3	-0.58
CA/GT	-8.5	-22.7	-1.45
GT/CA	-8.4	-22.4	-1.44
CT/GA	-7.8	-21.0	-1.28
GA/CT	-8.2	-22.2	-1.30
CG/GC	-10.6	-27.2	-2.17
GC/CG	-9.8	-24.4	-2.24
GG/CC	-8.0	-19.9	-1.84
Init w/term G.C	0.1	-2.8	0.98
Init w/term G.C	2.3	4.1	1.03
Symmetry correction	0	-1.4	0.43

2.3.2 Predicting the melting temperature (TM)

The melting temperature of a Double DNA helix is described as the temperature at which half of the strands are in the double-helical state and half of the strands are denatured Random Coil strand. The TM for self-complementary oligonucleotide duplexes is calculated by using the equation:

$$T_M = \frac{\Delta H^0}{\Delta S^0 + R \cdot ln C_T}$$

Thus T_M can be calculated from ΔH^0 en ΔS^0 and the total oligonucleotide concentration C(T), R is the gas constant (1.987 cal/K $\cdot$ mol).

2.3.3 Salt Dependency of Oligonucleotides and Polymers

The thermodynamic salt correction for DNA is assumed to depend on the sequence length but not on the sequence composition itself. Helix formation in polymers does not formally involve an initiation parameter so the salt dependence is by default incorporated into the NN propagation terms. According to Santalucia et al. [203] the following empirical equations were derived:

$$\Delta G^0_{37}(\text{oligomer, } [Na^+]) = \Delta G^0_{37}(\text{unified oligomer, 1 M NaCl}) - 0.114 \cdot \ln[Na^+] - 0.20.$$

$$\Delta G^0_{37}(\text{polymer NN, } [Na^+]) = \Delta G^0_{37}(\text{unified NN, 1 M NaCl}) - 0.175 \cdot \ln[Na^+] - 0.20.$$

2.3.4 Thermodynamic Primer Design

A thermodynamic approach allows us to develop theoretically guided methods for predicting primer quality. This method has two significant benefits compared with commonly used ad hoc primer scoring schemes. First, they take advantage of accurate methods for assessing DNA binding and folding stability. These accurate assessments are critical because PCR relies fundamentally on DNA binding reactions. Second, a physically motivated approach reduces the number of parameters that must be chosen, and shifts the emphasis of primer selection from choosing arbitrary thresholds to specifying physically meaningful reaction conditions and primer quality criteria. Ad hoc methods for primer design compute a variety of quality scores and combine these individual metrics into a weighted sum model that produces a final score like the Primer3 software package [204]. These quality scores try to account for considerations such as primer melting temperature, thermodynamic stability of a primer at the 3 prime end, homopolymers and a variety of other criteria mostly motivated by practical experience with PCR. However, these quality metric weights can present two significant difficulties: These metrics are not always physically interpretable and they can be redundant. The thermodynamic approach calculates binding and folding energies with the formulas explained in this chapter for a variety of relevant primer species, and integrates these calculations into a final measure of PCR efficiency. In addition to predicting whether the primers will amplify a given locus, one must also evaluate the primer specificity. Specific primers will amplify only the desired locus, whereas nonspecific primers have binding sites in the background DNA that lead to undesired copying of background fragments in addition to the target locus. This approach is especially helpful for Bisulphite treated DNA. During the bisulphite conversion, the double helix loses part of its complexity by converting all the non-methylated Cytisones into Uracil and eventually into Thymines. This results into primers being more prone to aspecific binding, a lower melting temperature and higher amount of homopolymers compared to primers for genomic DNA.

Part III

EPIGENETIC BIOMARKERS IN CANCER

1

WRN promoter CpG island hypermethylation does not predict a favorable outcome for metastatic colorectal cancer patients treated with irinotecan-based therapy

From article:
WRN promoter CpG island hypermethylation does not predict more favorable outcomes for metastatic colorectal cancer patients treated with irinotecan-based therapy
Linda J.W. Bosch, Yanxin Luo, Victoria V. Lao, Petur Snaebjornsson, Geert Trooskens, Ilse Vlassenbroeck, Sandra Mongera, Weiliang Tang, Piri Welcsh, James G. Herman, Miriam Koopman, Iris D. Nagtegaal, Cornelis J.A. Punt, Wim van Criekinge, Gerrit A. Meijer, Raymond J. Monnat Jr, Beatriz Carvalho and William M. Grady; Clinical cancer research, 2016 [205]

1.1 Introduction

The current care for metastatic colorectal cancer includes, if clinically indicated, surgical resection of the primary tumor and/or liver metastases, together with chemotherapy (5- fluoruracil and oxaliplatin or irinotecan) and in some patients targeted therapy (anti-EGFR antibodies or anti-VEGF therapy). The clinical response to this regimen is variable, and it is difficult to predict who will benefit from treatment. Moreover, for most therapies, we lack accurate biomarkers to identify the optimal treatment

for individual patients. DNA repair proteins such as the Werner syndrome RECQ helicase, WRN, are promising biomarkers for predicting the response to genotoxic chemotherapy. We attempted to validate previous studies that showed WRN promoter hypermethylation predicted the response to irinotecan using an independent sample set. We did not find a clear association between aberrant WRN promoter hypermethylation and reduced WRN expression. Moreover, in contrast to earlier studies we found an inverse correlation of WRN promoter hypermethylation with survival in metastatic colorectal cancer patients treated with irinotecan. Our results highlight the need for further studies to identify biomarkers that can predict the response of colorectal cancer to standard-of-care chemotherapeutic agents including irinotecan, oxaliplatin and 5-fluorouracil.

1.2 Material and Methods

Experiments were conducted at the University of Washington in Seattle (UWSEA) and the VU University Medical Center in Amsterdam, the Netherlands (VUmc) using cell lines and patient samples.

1.2.1 cell lines and tissues

Two independent collections of cultured CRC-derived cell lines were investigated. The adenoma cell line AAC1 and CRC cell lines RKO, LoVo, SW480, LS174T, AAC1/SB10, HCT116, SW48, FET, VACO400, VACO411, VACO5 were cultured at UWSEA. The UWSEA lines were authenticated by DNA fingerprint analysis prior to use (IDEXX/Radil Bioresearch; IRB). CRC cell lines Colo205, Colo320, HCT116, HCT15, HT29, LIM1863, LS174T, LS513, RKO, SW480, and SW1398 were cultured at VUmc and authenticated by array comparative genomic hybridization (aCGH, 244 k Agilent oligonucleotide platform) at the VU University Medical Center, Amsterdam, the Netherlands. The patterns of chromosomal changes observed were in concordance to the previously described chromosomal changes in these cell lines [206]. Twenty-six fresh frozen (FF) primary CRC tissues with matched FF normal colon tissue, and 21 formalin-fixed paraffin-embedded (FFPE) normal colon tissues from cancer-free patients were collected and studied following IRB approved protocols and in accordance with the ethical regulations of the corresponding institutions (UWSEA and VUmc). The samples used at UWSEA were provided by

the Cooperative Human Tissue Network (CHTN). Collection, storage and use of patient-derived tissue and data from VUmc was performed in accordance with the Code for Proper Secondary Use of Human Tissue in The Netherlands [207].

1.2.2 Tissue samples from the CAIRO clinical trial

In the CAIRO study CRC patients with metastatic disease were randomized between sequential treatment (capecitabine (CAP) followed upon disease progression by irinotecan (IRI), then oxaliplatin plus capecitabine (CAPOX)), or combination therapy with irinotecan plus capecitabine (CAPIRI) followed by CAPOX [208]. The primary endpoint of the study was overall survival (OS). DNA was isolated from FFPE tissue of surgically resected primary tumors from 183 patients that participated in the CAIRO study. Of these 183 patients, 93 received CAPIRI as first-line therapy while 90 received first-line CAP monotherapy. From the 90 patients that received first-line CAP monotherapy, 52 received more than 2 cycles of second-line IRI. These samples were selected to match stratification factors in the original study for the subgroup of patients that underwent primary tumor resection, i.e. resection status, WHO performance status, predominant localization of metastases, previous adjuvant therapy and serum lactate dehydrogenase levels. Samples were also selected based on a high proportion of tumor cells in sections (at least 70%). A large proportion of these samples overlap with samples described in [209].

1.2.3 WRN methylation analyses

WRN methylation status was assessed by two different methylation-specific PCR (MSP) assays together with bisulfite sequencing (see Supplementary Methods for additional detail). A WRN 5 region from -31 bp to +128 relative to the transcription start site (TSS), hereafter referred to as Region 1, was analyzed by a gel-based MSP assay. Region 2, located at -410 to -331 bp upstream of the WRN TSS was analyzed with a quantitative MSP assay. Bisulfite sequencing was performed for the region -193 bp to +157bp that encompassed the TSS, and overlapped with the locations of the WRN MSP primer pairs described in Agrelo et al. [210] and an independent set of WRN MSP primers reported by Ogino et al. [211].

1.2.4 WRN expression analyses

RNA expression analyses were performed by real-time quantitative PCR assays using TaqMan Gene Expression Assays from Applied Biosystems for WRN (Hs00172155 m1), β-2 micoglobulin (B2M, Hs00984230 m1), and β-glucuronidase (GusB, Hs99999908 m1). Protein expression analyses were performed by Western blotting, using monoclonal antibodies for WRN (W0393, Sigma) and β-actin (13E5, Cell Signaling Technologies).

1.2.5 TCGA data

WRN DNA methylation (Illumina Infinium HM27 bead array; HM27) and mRNA expression (Agilent microarray) data from 223 CRC tumors from The Cancer Genome Atlas (TCGA) Colorectal Cancer project [212] were obtained via cBioPortal (http://www.cbioportal.org; data downloaded on 2 March 2014) [213]. When data from more than one probe per gene is available from the methylation assay, cBioPortal uses methylation data from the probe with the strongest negative correlation between the methylation signal and mRNA gene expression.

1.2.6 Statistical analyses

Students T-test was used to compare WRN expression levels in HCT116 and Colo205 before and after 5-aza-2-deoxycytidine (DAC) and/or trichostatin A (TSA) treatment. Pearson correlation analysis was used to measure correlation between WRN methylation and mRNA expression levels. Progression-free survival (PFS) for first-line treatment was calculated from the date of randomization to the date of first observed disease progression or death after first-line treatment. Overall survival (OS) was measured from the date of randomization to date of death due to cancer. Other causes of death were censored. The prognostic or predictive value of WRN methylation status was assessed by a Kaplan-Meier survival analysis and log-rank test. A Cox proportional hazard regression model was used to estimate Hazard Ratios (HR) and 95% Confidence Intervals (95%CI). A multivariate Cox regression model was used to assess and adjust for important prognostic variables including age, gender, serum lactate dehydrogenase (LDH), WHO performance status, previous adjuvant therapy and location of metastases. Multivariate Cox regression analysis was also used to assess and adjust for possible prognostic variables Microsatellite Instability (MSI) status, BRAF mutational status and

mucinous differentiation, for which information was available on a subset of the samples (136 out of 183) [214], [215]. Results were considered significant when p-values were ≤ 0.05.

1.3 Results

1.3.1 WRN methylation and expression status in colon cancer cell lines

In order to accurately detect and quantify WRN promoter methylation in CRC samples, we independently developed and cross-validated methylation-specific PCR (MSP) primer sets and assays in both labs (UWSEA and VUmc) for two WRN regions adjacent to and overlapping the TSS at base pair position +1: Region 1 (-31 bp to +128 bp) and Region 2 (-410 to -331 bp) (Figure 1.1A). WRN methylation status in Region 1 was assessed in 11 colon cancer cell lines (SW480, Vaco411, AAC1/SB10, Vaco400 LS174T, LoVo, HCT116, Vaco5, FET, RKO, SW48), and 1 adenoma cell line (AAC1) from UWSEA. Seven of 11 colon cancer cell lines (6%) had Region 1-methylated WRN (Figure 1.1B), while the adenoma cell line was unmethylated. There was no association between WRN Region 1 methylation and MSI and/or CpG Island Methylator Phenotype (CIMP).

WRN methylation status in Region 2 was successfully evaluated in 10 colon cancer cell lines (SW480, Vaco411, Vaco400, LS174T, LoVo, HCT116, Vaco5, FET, RKO, SW48; UWSEA), and was comparable to Region 1 methylation status within a cell line (Figure 1.1C). Bisulfite sequencing of cells lines with methylated (HCT116) or unmethylated WRN (SW480) was performed to confirm the methylation status of both regions and validate the MSP results using an orthogonal assay (Figure 1.2D).

We assessed WRN Region 2 methylation status in a second, overlapping series of colon cancer cell lines (Colo205, Colo320, HCT116, HCT15, HT29, LIM1863, LS174T, LS513, RKO, SW480, and SW1398, SW48 and Caco2; VUmc). These analyses revealed that 10 of 13 cell lines, or 77%, were WRN Region 2 methylated (Figure 1.2B).

Cell lines that carried methylated WRN expressed relatively high levels of WRN as assessed by WRN mRNA qRT-PCR (Figure 1.2A&B). There was either no or a slightly positive correlation between WRN Re-

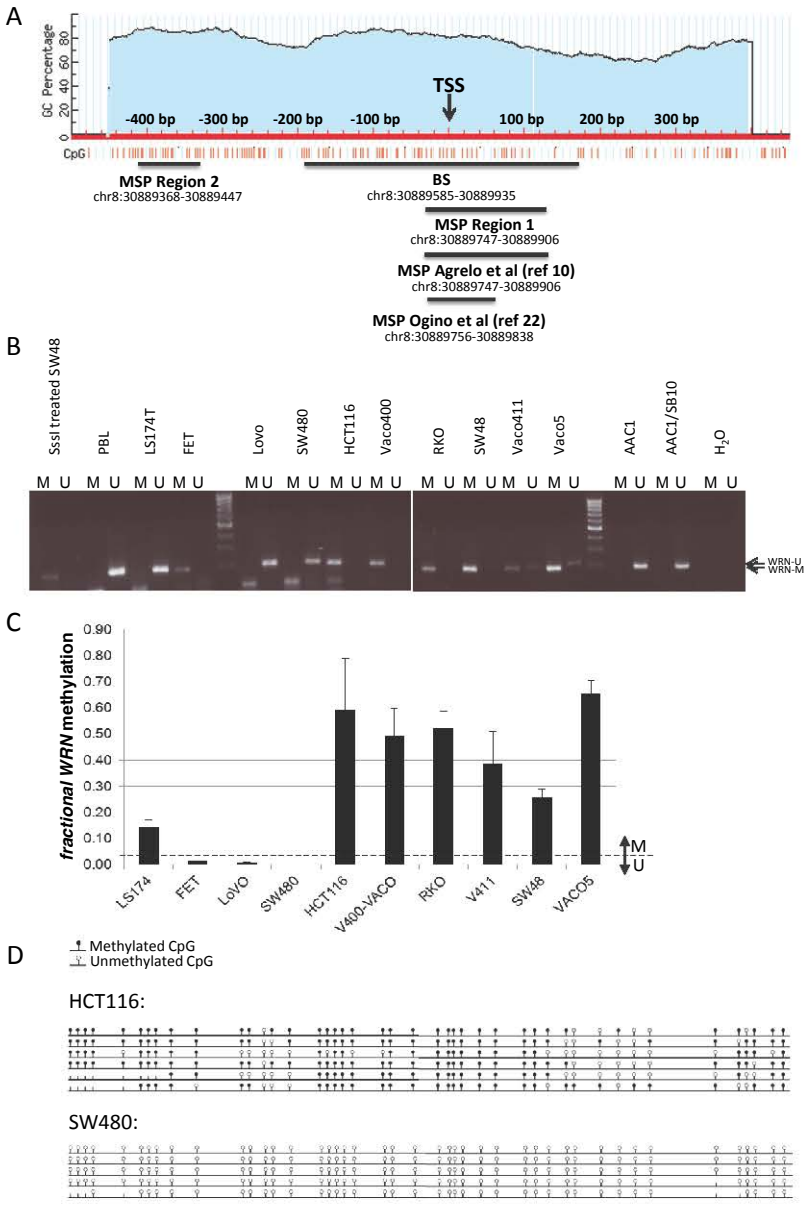


Figure 1.1: WRN promoter region methylation analysis in cell lines.
A. WRN promoter region CpG island and primer locations Genomic coordinates, CpG density and positions are shown in the top panel. Each vertical bar in the lower panel represents the presence of a CpG dinucleotide. Black horizontal bars indicate regions amplified by newly designed and validated MSP primer pairs (Region 1 and Region 2), the region amplified by the original primer pair described by Agrelo et al [210], and the region targeted for bisulfite sequencing (BS). TSS, Transcription Start Site.
B. Methylation analysis of Region 1 (see Figure 1) in the adenoma cell line AAC1 and in colon cancer cell lines. DNA from Peripheral Blood Lymphocytes (PBL) was used as an unmethylated control. H2O and DNA from SssI methylase-treated DNA from the colorectal cancer cell line SW48 were used, respectively, as no template and methylated template controls.
C. Quantitative methylation analysis of WRN promoter Region 2 in the same colon cancer cell lines shown in panel A.
D. Sodium bisulfite sequencing results of WRN gene promoter on cell lines HCT116 and SW480 in the region depicted in Figure 1. Each row represents an individual cloned allele and each circle indicates a CpG dinucleotide. Black circle=methylated CpG site; white circle=unmethylated CpG site; no circle=not determined.

Figure 2

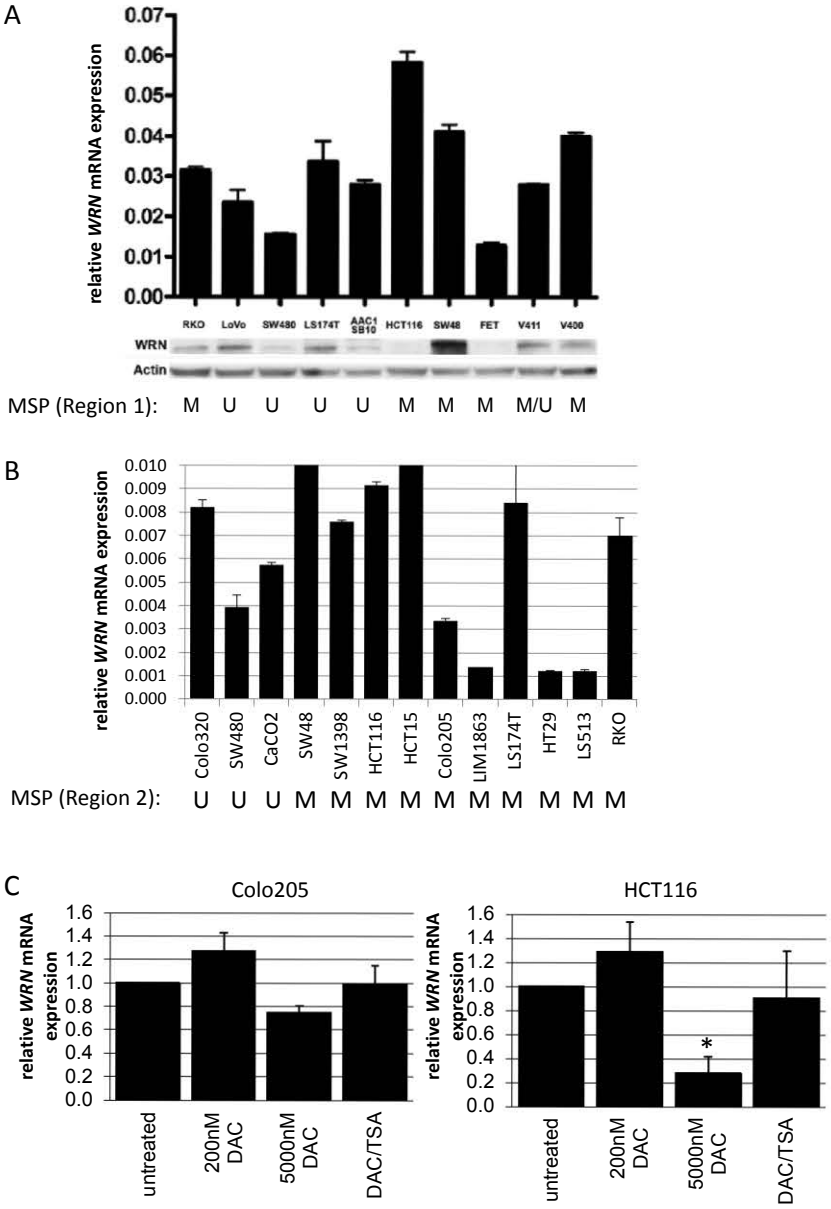


Figure 1.2: WRN expression analysis in cell lines
A. WRN mRNA (upper panel) and protein (lower paired panels) expression in CRC cell lines in relation to methylation status in WRN promoter Region 1 (lower panel). Error bars represent standard deviations across triplicate independent experiments, in which WRN mRNA was normalized to mRNA expression of the reference gene GUSB (upper panel) and, for protein expression -actin (lower panel). Methylation status of WRN promoter Region 1 is indicated below each pair of immunoblots (M = methylated; U = unmethylated).
B. WRN mRNA expression level in relation to methylation status of WRN promoter Region 2.Error bars represent standard deviations of mean expression values of two independent experiments. Methylation status of WRN promoter Region 2 is indicated below each cell line designation (M = methylated; U = unmethylated).
C. WRN mRNA expression analysis of Colo205 (left) and HCT116 (right) with and without 5-aza-2-deoxycytidine (DAC) or DAC/trichostatin A (TSA) treatment. Bars represent mean in two independent experiments, with error bars represent standard deviations. Expression was quantified relative to mRNA expression levels of B2M. *p=0.001

gion 2 methylation and expression level in two different groups of CRC cell lines: SW480, Vaco411, Vaco400 LS174T, LoVo, HCT116, Vaco5, FET, RKO, SW48 (UWSEA; Pearson correlation of 0.32, p=0.3); and Colo205, Colo320, HCT116, HCT15, HT29, LIM1863, LS174T, LS513, RKO, SW480, and SW1398, SW48 and Caco2 (VUmc; Pearson correlation of 0.68, p=0.04). Consistent with these results, treatment of the methylated CRC cell lines HCT116 and Colo205 with the demethylating agent 5-aza-2-deoxycytidine (DAC) and/or tricostatin A (TSA) either did not change or resulted in decreased WRN mRNA expression (Figure 1.2C). Western blot analysis of WRN protein expression as a function of Region 1 and 2 promoter methylation in CRC cell lines in the UWSEA collection further emphasized the lack of correlation between WRN promoter hypermethylation and mRNA and protein expression (Figure 1.2A&B).

1.3.2 WRN methylation and expression status in CRC tissues

In order to determine whether there was a more consistent relationship between WRN methylation status and expression in primary tumor samples, we analyzed WRN methylation status and expression in primary CRC samples and in adjacent normal colon mucosa. We detected Region 1 methylation in 33% (7 of 21) of primary CRCs, but in none of the paired normal mucosa samples tested (N=12). Region 2 methylation was detected in 45% (9 of 20) of primary CRCs, and in 1 of 20 matched normal mucosa samples (Figure 1.3A). Methylation status was largely concordant between the two regions: all samples that showed methylation in Region 1 were also in Region 2 methylated.

Only two cases showed an unmethylated Region 1 and a methylated Region 2. Bisulfite sequencing of a subset of these samples (8 CRCs and 2 normal mucosa samples) confirmed the results of MSP assays (data not shown). A second analysis of Region 2 methylation using an independent series of primary colorectal cancers (N=183 from the CAIRO series, see next section) and normal colon mucosa samples (N=21, VUmc) revealed WRN promoter hypermethylation in 40% (74/183) of the primary CRCs, and very low or absent WRN methylation level in normal colon mucosa.

In our first series of colon tissues, WRN mRNA expression was higher in primary CRC vs matched normal mucosa samples in 10 of 20 patients (50%), lower in 6 samples (6 of 20 or 30%) and equivalent in the remaining 4 samples (20%; Figure 1.3B). No association was observed

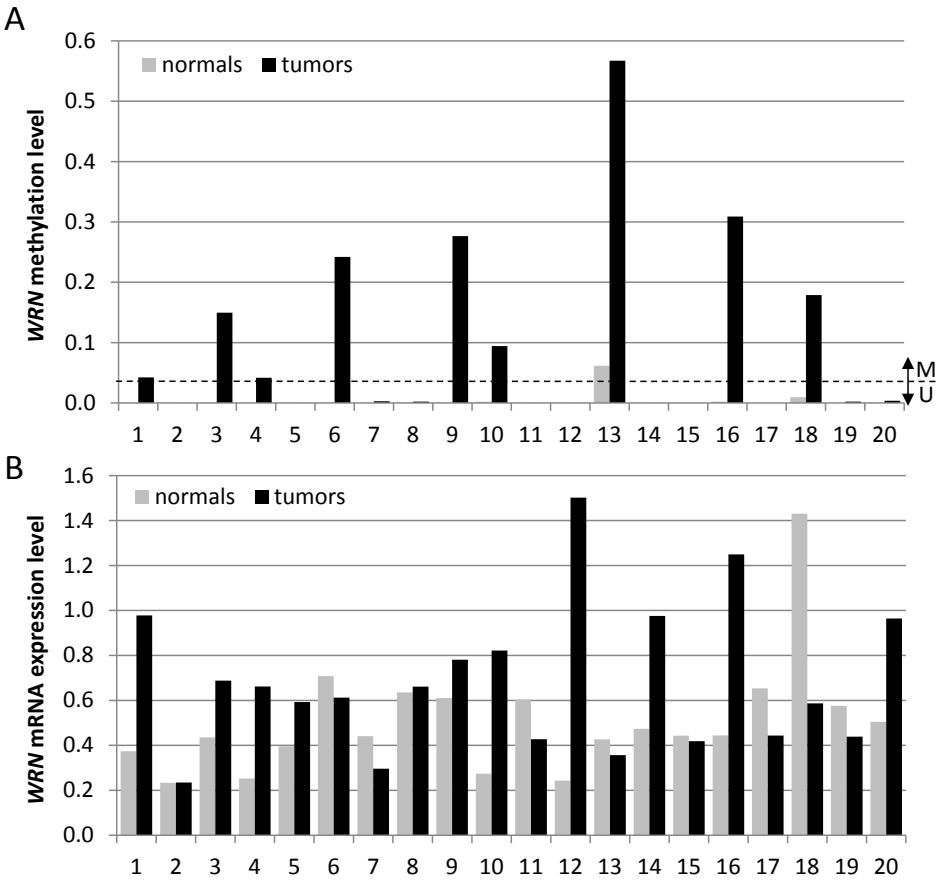


Figure 1.3: WRN promoter region methylation and expression analyses in CRC and matched normal colon tissues.
A. WRN methylation levels in CRC tumor tissues (black bars) and matched normal colon tissues (grey bars). Bars represent mean expression of duplicate measurements in one experiment. A sample was considered methylated when the Ct ratio exceeded the threshold of 0.03, which was set based on an analysis of normal colon samples (N=21), which all had values below this threshold.
B. WRN mRNA expression versus a GUSB control in the same CRC tumor (black bars) and matched normal colon (grey bars) samples shown in panel A. Bars represent mean expression of triplicate measurements in one experiment.

between WRN Region 1 or 2 hypermethylation and mRNA expression (Region 2: Pearson correlation 0.14, p=0.4). WRN protein expression could not be detected by Western blot in 10 of 20 (50%) of paired primary CRC/normal mucosa samples (data not shown). An independent assessment of WRN methylation status and mRNA expression in 223 CRCs included in the TCGA Colorectal Cancer Project (see The Cancer Genome Atlas (TCGA) database at www.cBioportal.org; [213]) did not reveal a negative correlation between WRN methylation level and mRNA expression (Pearson correlation of 0.1, p=0.03;).

1.3.3 Relationship of WRN methylation to clinical outcome

In order to determine if there is a relationship between WRN promoter hypermethylation and treatment outcomes, we assessed the correlation between WRN promoter methylation status and survival in patients who participated in the CAIRO study [208]. OS did not differ between the two treatment arms in the original study population, or in the subset included in this analysis. Patient characteristics such as age, sex, performance status, predominant localization of metastases, previous adjuvant therapy and serum lactate dehydrogenase level (LDH) were comparable between the two treatment arms in the subset included in this analysis (Supplementary Table 2). Thus we pooled patients from the two treatment arms to evaluate the association of WRN promoter methylation status and OS. The cohort of 183 patients included a total of 160 death events. The group of 109 patients with unmethylated WRN had 91 death events and the group of 74 patients with methylated WRN had 69 death events. Patients with methylated WRN CRC had shorter OS compared to patients with unmethylated WRN (median OS of 407 vs 610 days for methylated vs unmethylated WRN, respectively (HR = 1.6 (95%CI 1.2-2.2), p = 0.003; Figure 1.4A). This was observed for patients in the sequential treatment arm (median OS of 405 vs 589 days; HR = 1.5 (95%CI 1.0-2.4), p=0.05), as well as in the combination treatment arm (median OS of 410 vs 680 days for methylated vs unmethylated WRN, respectively; HR = 1.7 (95%CI 1.1-2.7), p=0.02; compare Figure 1.4B, 1.4C). However, in the sequential treatment arm, a negative effect of WRN promoter hypermethylation on outcome was observed only for patients who received irinotecan during their treatment course (n=55; median OS of 567 vs 646 days for methylated vs unmethylated WRN,

respectively; HR = 1.9 (95%CI 1.1-3.5), p=0.03; Figure 1.4D). This effect was not observed in patients who did not receive irinotecan (n=37; median OS of 320 vs 326 days for methylated vs unmethylated WRN, respectively; HR = 1.0 (95%CI 0.5-2.0), p=1.0; Figure 1.4E).

We next determined whether WRN methylation status had predictive value for irinotecan-treated outcomes by assessing the relationship between WRN methylation status and response to CAPIRI. Patients with unmethylated WRN showed significantly longer PFS when treated with CAPIRI compared to CAP alone, as was expected from the results of the original CAIRO trial [208] (median PFS of 272 vs 164 days for CAPIRI vs CAP, respectively; HR=0.48 (95%CI 0.32-0.70), p=0.0001; Figure 1.5A). However, patients with methylated WRN did not benefit from CAPIRI therapy (median PFS of 211 vs 190 days for CAPIRI vs CAP, respectively;

HR=1.1(95%CI 0.69-1.77), p=0.7; Figure 1.5B). The same trend was observed for patients receiving second-line irinotecan monotherapy in the sequential treatment arm, though the number of patients was small. Multivariate Cox regression analysis showed significant interaction effects between treatment arm and WRN methylation status, even after adjusting for potentially confounding factors including age, gender, serum LDH, WHO performance status, previous adjuvant therapy, predominant location of metastasis, MSI status, BRAF mutational status and mucinous differentiation.

1.4 Discussion

DNA repair proteins such as the RECQ helicase WRN are promising biomarkers for predicting the response to genotoxic chemotherapy. In this study, we aimed to validate the reported association between WRN promoter hypermethylation and transcriptional silencing, and determine the predictive value of WRN promoter hypermethylation for increased sensitivity to IRI-based therapy in CRC patients [210]. We developed and used two new sets of MSP PCR primers to reliably assess WRN methylation status in both CRC and normal colon tissue. Methylation status was also analyzed by bisulfite sequencing (BS) of a region overlapping the WRN TSS. Our new MSP primer pairs and BS assay covered the regions analyzed in previous reports [210, 211] (Figure 1.1A), and proved more reliable in our hands than the originally reported primer pair for WRN MSP assays [210]. Despite using these newly developed and

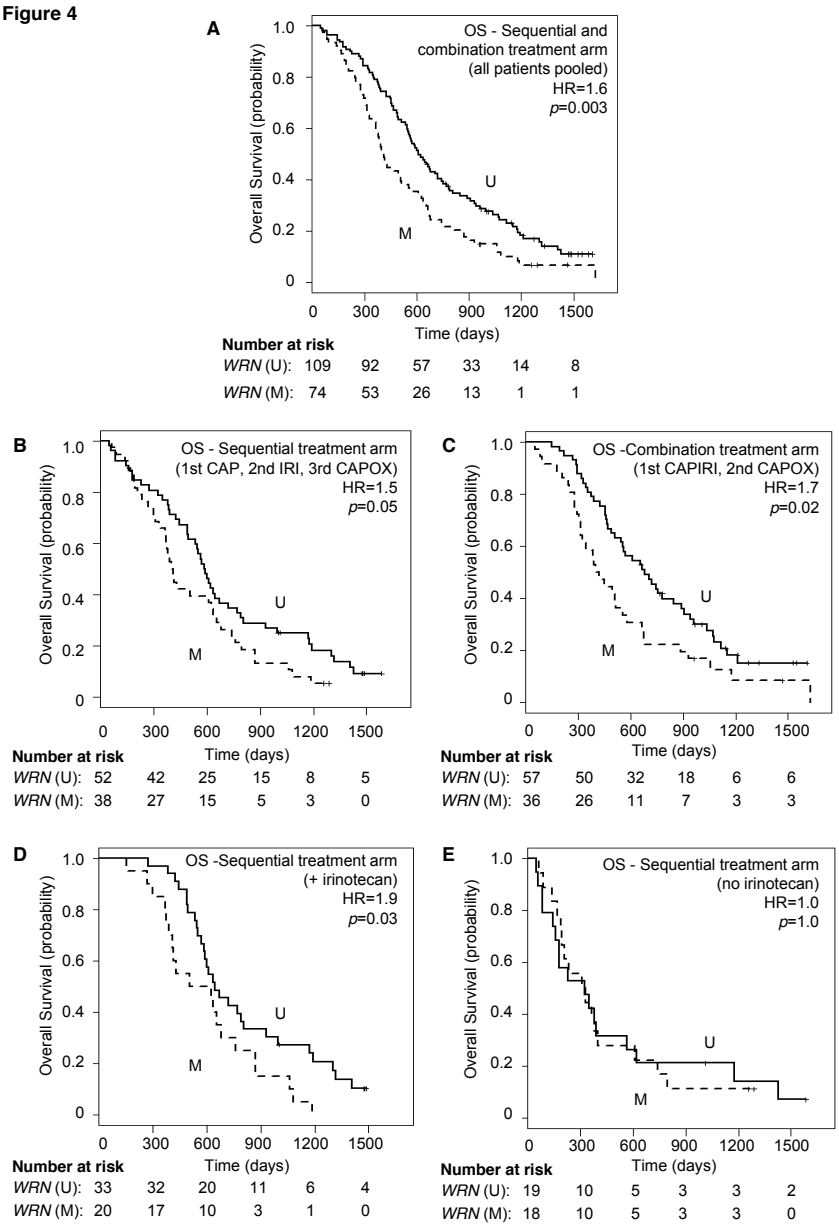


Figure 1.4: Overall survival (OS) of CRC patients with unmethylated (solid lines, U) or methylated (dashed lines, M) WRN promoter regions in response to (A) sequential and combination treatment arms combined (sequential or combined capecitabine (CAP) and Irinotecan (IRI), followed by capecitabine + oxaliplatin (CAPOX)); (B) in the sequential treatment arm alone (1st line capecitabine (CAP), 2nd line Irinotecan (IRI), 3rd line capecitabine + oxaliplatin (CAPOX)); (C) in the combination treatment arm alone (1st line capecitabine + irinotecan (CAPIRI), 2nd line capecitabine + oxaliplatin (CAPOX)); in the subset of patients who received (D) or did not receive (E) irinotecan (IRI) in the sequential treatment arm. HR = Hazard Ratio (Methylated WRN vs unmethylated WRN).

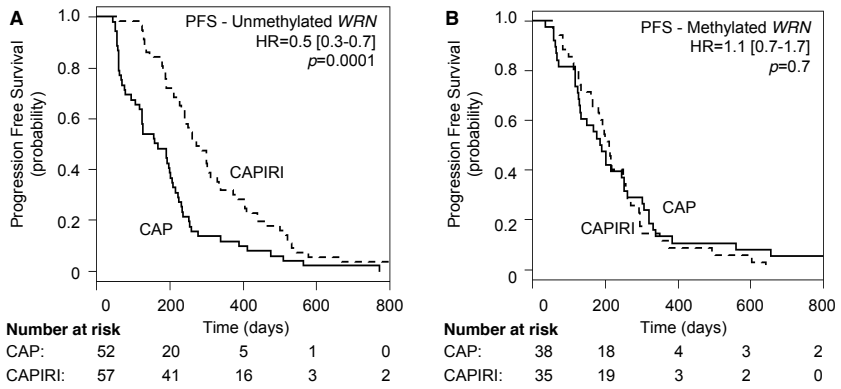


Figure 1.5: Progression-free survival in metastatic CRC patients treated with CAP (solid lines) or CAPIRI (dashed lines) as a function of WRN promoter region methylation. PFS is shown for CRCs with unmethylated (panel A) or methylated (panel B) WRN promoter regions. HR = Hazard Ratio (CAPIRI vs CAP).

well-validated methylation-specific reagents, we found no consistent association between WRN promoter hypermethylation and WRN expression at the mRNA or protein level. Moreover, we found that WRN promoter hypermethylation was associated with reduced, as opposed to the previously reported increased, OS in CRC patients with metastases who received irinotecan [210]. Progression free survival (PFS) improved only when irinotecan was added to CAP in the presence of unmethylated WRN, which was not expected from the results of the original CAIRO trial [208]. One explanation for the differing results between our study and a previous report [210] could be the use of different methylation assays. However, this is unlikely: we designed and validated new primer sets for overlapping MSP and bisulfite sequencing assays that worked reliably, and covered a 567 bp region that encompassed the TSS. These reagents reliably and accurately detected WRN promoter methylation status in both cell lines and primary tumor samples across the locations of both the originally reported [210] and an additional reported overlapping primer pair [211] (1.1A). Other possible reasons for the contrasting results in the current and previous report [210] encompass the lack of robust analytical tools in the previous report [210], together with the limited number of cell lines and the small size and nature of the clinical samples analyzed [210]. Of note, The clinicopathological details of the 88 patients reported in [210] were not described in the original report or in the reference to this cohort included in the initial report [210]. Hence, selection bias cannot be ex-

cluded. We further corroborated our finding of no consistent relationship between WRN promoter methylation level and gene expression using data on 223 CRC samples included in the TCGA Colorectal Cancer Project, where again no correlation could be identified between WRN hypermethylation and WRN transcriptional silencing [213, 216]. In order to test the association between WRN methylation status and clinical outcomes, we used material from patients enrolled in the CAIRO study (the Dutch CApecitabine, IRinotecan and Oxaliplatin (CAIRO) study [208]. Our CAIRO study cohort (n=183) was larger than the initial cohort (n=183 vs 88) and has been described in detail. The CAIRO study provided high quality clinical data, which are essential to evaluate predictive biomarkers [217] and to test the association between WRN methylation status and clinical outcomes. The CAIRO cohort also offered the opportunity to compare first-line CAP monotherapy versus CAPIRI therapy. Despite our larger well-characterized study population, we were not able to confirm the initial observation that WRN promoter hypermethylation was associated with improved outcome in irinotecan-treated metastatic CRC patients [210]. In contrast, we observed a significantly worse outcome for irinotecan-treated colorectal cancer patients with WRN-methylated tumors. This is similar to the outcome observed in an independent, well-described study [211] that used primer pairs targeting the same WRN region as the initial report [210] (see 1.1A). These observations indicate that WRN promoter hypermethylation may be useful as a biomarker, to predict a worse response to irinotecan treatment.

This effect is likely to reflect as-yet unidentified co-variables, as WRN promoter hypermethylation does not consistently alter WRN expression. WRN is a housekeeping gene that is expressed at comparatively low copy number ($\leq$1000 to 10,000 copies/cell) in many cell types [218]. The WRN promoter region includes Sp1, RCE (retinoblastoma/TP53), AP2 and MYC E-box binding sites, and there are experimental data showing that these binding sites and/or transcription factors can alter WRN transcription [219, 220]. WRN expression is also known to be cell cycle-responsive, and upregulated by cellular oncogenic transformation [221], though none of these mechanisms has been shown thus far to be WRN DNA methylation-dependent or modulated. Alternatively, WRN promoter hypermethylation has been associated to microsatellite instability, CpG island methylator phenotype, BRAF mutations and mucinous differentiation, which themselves are associated to clinical outcome in colon cancer [211, 222, 223]. Information on MSI, BRAF and mucinous dif-

ferentiation available on a sub-set of our sample set revealed that those variables did not explain the association between WRN promoter hypermethylation and clinical outcome after treatment with irinotecan- based therapy. However, the number of samples with MSI status and/or BRAF mutation was very low (n=6 and n=11, respectively), hence no hard conclusions can be drawn from these results. Future functional analyses and validation studies in large, independent and well-annotated cohorts are needed to shed light on the role of WRN promoter hypermethylation as a determinant of the response to irinotecan-based therapy.

Our study has the following limitations. First, measurements were performed on the primary tumors, while patients were treated for their metastases, which raises the question whether intra tumor heterogeneity could play a role. Although metastases can acquire additional genomic alterations, they keep most alterations present in the primary tumor [224,225]. Furthermore, DNA methylation is usually an early event in colorectal carcinogenesis, which we suspect is true for WRN methylation as well [226].

Second, we were not able to independently analyze all cell lines at both participating institutions, though note that the subset of cells analyzed by both groups gave concordant results. This strengthens our conclusion that previous findings on the negative relationship between WRN promoter methylation level and gene expression at the mRNA or protein level could not be validated. A final limitation of the current study was the use of DNA from 183 patients and tumor tissue which represented a subset of all patients in the CAIRO trial [208]. However, this selection was representative for the subgroup of patients that underwent resection of the primary tumor in terms of clinical characteristics and survival outcome (see also [209]). Furthermore, the current cohort is larger than the cohort as described in [210] (n=183 vs n=88) and was large enough to have statistical power.

In summary, we found that the methylation status of the WRN promoter region can be reliably assessed in both CRC and normal colorectal tissue using newly developed methylation-specific PCR and bisulfite sequencing assays. However, there was no consistent association between WRN promoter hypermethylation and loss of WRN expression at the mRNA or protein level in CRC cell lines or tumors. Moreover, we could not validate findings from a previous study that WRN promoter hypermethylation was associated with a better response to irinotecan-based therapy and found, instead, that WRN promoter hypermethylation was

associated with reduced OS and PFS in our well- characterized CRC patient cohort who received irinotecan-based therapy. Despite growing evidence for a role for WRN genomic alterations in CRC disease progression [227], our results indicate that WRN promoter hypermethylation does not reliably predict WRN gene expression or, as originally reported [210], improved clinical outcomes in CRC patients treated with irinotecan-based chemotherapy regimens.

2

DCR1 methylation and response to irinotecan in colorectal cancer

From:
Decoy receptor 1 (DCR1) promoter hypermethylation and response to irinotecan in metastatic colorectal cancer
Linda J.W. Bosch, Geert Trooskens, Petur Snaebjornsson, Veerle M.H. Coup, Sandra Mongera, Josien C. Haan, Susan D. Richman, Miriam Koopman, Jolien Tol, Tim de Meyer, Joost Louwagie, Luc Dehaspe, Nicole C.T. van Grieken, Bauke Ylstra, Henk M.W. Verheul, Manon van Engeland, Iris D. Nagtegaal, James G. Herman, Philip Quirke, Matthew T. Seymour, Cornelis J.A. Punt, Wim van Criekinge, Beatriz Carvalho and Gerrit A. Meijer; Oncotarget, 2017 [228]

Diversity in the biology of colorectal cancer (CRC) is associated with variable responses to standard chemotherapy. We aimed to identify and validate DNA hypermethylated genes as predictive biomarkers for irinotecan treatment of patients with metastatic CRC.

Candidate genes were selected from 389 genes involved in DNA Damage Repair by correlation analyses between gene methylation status and drug response in 32 cell lines. The discovery and initial validation set consisted of primary tumors of 185 and 166 metastatic CRC patients, respectively, from the phase III CAIRO trial. An external validation set consisted of 467 primary tumors from the phase III FOCUS study. Methylation status in tumor tissue was correlated to progression free survival (PFS) by first-line treatment regimen, containing either single-agent fluorouracil (i.e. CAP in CAIRO or 5-FU in FOCUS) or combination chemotherapy (i.e. CAP or 5-FU plus irinotecan (CAPIRI in CAIRO / FOLFIRI in

FOCUS)).

In the discovery and initial validation set, patients with methylated Decoy Receptor 1 (DCR1) tumors did not significantly benefit from CAPIRI treatment over CAP treatment (discovery set: HR=12 (95%CI 07-19, p=06), validation set: HR=0.9 (95%CI 0.6-1.4, p=0.5)), whereas patients with unmethylated DCR1 tumors did (discovery set: HR=0.4 (95%CI 0.3-0.6, p=0.00001), validation set: HR=0.5 (95%CI 0.3-0.7, p=0.0008)). These results, however, could not be validated in the external data set, where a similar effect size was found in patients with methylated and unmethylated DCR1 (methylated DCR1: HR=0.7 (95%CI 0.5-0.9, p=0.01), unmethylated DCR1: HR=0.8 (95%CI 0.6-1.2, p=0.4)).

DCR1 promoter methylation was identified and initially validated as a potential negative predictive biomarker for response to irinotecan-based therapy, but external validation could not validate these findings. These results underline the importance of extensive clinical evaluation of candidate biomarkers.

2.1 Introduction

The outcome of patients with colorectal cancer (CRC) strongly depends on tumor stage at time of diagnosis. Whereas stage I CRC patients have a 5-year overall survival of more than 90%, in stage IV CRC patients it declines to 20% or less [229]. When distant irresectable metastases develop, palliative systemic therapy is the only treatment option available to these patients. The backbone of this is 5-fluorouracil (5-FU) in combination with either oxaliplatin or irinotecan [230]. More recently, addition of targeted agents directed against vascular epithelial growth factor (VEGF) (bevacizumab) or epidermal growth factor receptor (EGFR) (cetuximab and panitumumab) has been demonstrated to give additional survival benefit [231]. Only a subset of patients benefit from these regimens, while those patients that do not, still may suffer from considerable toxicity. With the exception of KRAS/NRAS mutation status that predicts resistance to EGFR-targeted therapy [232, 233], no other biomarkers exist that adequately predict treatment response in metastatic CRC. Thus, predictive biomarkers are urgently needed to a priori identify the subset of patients that will benefit from a specific treatment.

Hypermethylated genes form a particular category of biomarkers and a number of these have been reported to predict drug response in CRC patients [210, 234], but inconsistent results for the same markers have

been reported [211, 235]. Hypermethylated genes are of particular interest, since DNA methylation is potentially reversible by DNA methyltransferase inhibitors, which could provide a way to restore expression of genes silenced by DNA hypermethylation and thus increase the sensitivity of tumor cells to the agents the gene is associated with [236]. In the present study we set out to identify and validate novel hypermethylated genes that could potentially predict response to treatment with irinotecan in patients with metastatic CRC, using material from two clinical trials, i.e. the Dutch CApecitabine, IRinotecan and Oxaliplatin (CAIRO) study [208] and the Fluorouracil, Oxaliplatin, CPT-11: Use and Sequencing (FOCUS) study from the UK [237].

2.2 Material and Methods

2.2.1 Candidate gene selection

Candidate gene selection was based on correlations between methylation of 389 genes involved in DNA Damage Repair and Response and drug response in 32 cell lines, which is described in detail in the Supplementary Information.

2.2.2 Patient sample selection

Patients selected for the current study participated in either of two phase III trials, namely the CApecitabine, IRinotecan and Oxaliplatin (CAIRO) study of the Dutch Colorectal Cancer Group (DCCG) (CKTO 2002-07, ClinicalTrials.gov; NCT00312000), and the Medical Research Council Fluorouracil, Oxaliplatin, CPT-11: Use and Sequencing (FOCUS) study (ISRCTN 79877428) under the auspices of the United Kingdom National Cancer Research Institute Colorectal Cancer Studies Group. Written informed consent was required from all patients before study entry, and included consent for translational research on tumor tissue. Details on the CAIRO and FOCUS study are provided in the Supplementary Information.

2.2.3 CAIRO biomarker populations

An initial 185 patients were selected for a discovery set of which 90 patients were treated with first-line capecitabine (CAP) and 95 were treated

with first-line capecitabine plus irinotecan (CAPIRI). The patient samples were matched according to the stratification factors in the original study (for the subgroup of patients that underwent resection of the primary tumor, since these are the patients from whom material was available to be included in this study), that is, performance status, predominant metastatic site, previous adjuvant therapy and serum lactate dehydrogenase level (LDH). In addition, only patients were included who had received at least three cycles of 1st line therapy, or two cycles when death followed due to progressive disease. A large proportion of these samples overlap with samples described in [209].

For the initial validation set, patients who had received at least three cycles of 1st line therapy or two cycles when death followed due to progressive disease were selected, with no further criteria, from the remaining patients of which tumor DNA samples were available. These comprised 166 patients, of which 78 were treated with first-line CAP and 88 were treated with first-line CAPIRI.

2.2.4 FOCUS biomarker validation population

A total of 467 tumor DNA samples from the FOCUS trial were available for the current study. These came from 331 patients treated with at least three cycles of first-line 5-FU and 136 patients treated with at least three cycles of first-line 5-FU plus irinotecan (FOLFIRI).

2.2.5 DNA isolation and methylation analysis

From formalin-fixed paraffin-embedded tissue samples from primary tumors, resected before chemotherapy, DNA was extracted as described before. [238, 239]

All methylation assays were performed blindly to information on treatment or survival outcome. The CAIRO discovery set was subjected to high-throughput LightCycler MSP assay (LightCycler 480 SYBR Green I Master kit (Roche, Vilvoorde, Belgium)). Primers were designed in promoter regions (i.e. -1000 to +200 bp relative to the transcription start site). Primers from literature were used when they experimentally passed our quality control; see supplementary table 1 for primer sequences. For the CAIRO validation set and CRC cell lines a quantitative MSP (qMSP) assay for DCR1 was used. The primers for methylated DNA were equal to the primers used for LightCycler analyses described above and were

designed at the exact location as described before [240]. The FOCUS validation set was analyzed with a qMSP assay for DCR1 as well. The primers for methylated DNA were equal to the primers used in the CAIRO discovery and validation study. All details on DNA isolation and methylation assays can be found in the Supplementary Information.

2.2.6 Cell lines

Details on the culture conditions of HCT15, HCT116, LS513, LS174T, Colo320, SW48, SW1398, HT29, Colo205, SW480 RKO, Caco2, and LIM1863 can be found in the Supplementary Information. To investigate re-expression of DCR1 after inhibition of DNA metyltransferases, HCT116 and Colo205 cells were treated with 5000 nM 5-aza-2-deoxycytidine for three days (DAC, Sigma Chemical Co., St. Louis, MO, USA).

2.2.7 RNA isolation and qRT-PCR

Details on the RNA isolation can be found in the Supplementary Information. Quantitative RT-PCR was done using TaqMan Gene Expression Assays from Applied Biosystems directed to DCR1 (Hs00182570 m1) and B2M (Hs00984230 m1). Relative expression levels were determined by calculating the Ct-ratio (Ct ratio = 2e-(Ct(DCR1)-Ct(B2M)))x1000.

2.2.8 TCGA data

DCR1 DNA methylation (Illumina Infinium HM27 bead array; HM27) and mRNA expression (Agilent array) data were obtained via cBioPortal for Cancer Genomics (http://www.cbioportal.org; [213]) on 223 CRC tumors included in The Cancer Genome Atlas (TCGA) Colorectal Cancer project. This data set was downloaded on the 14th of July 2015 from all tumors with available methylation and mRNA expression data from the Colorectal Adenocarcinoma (TCGA, Nature 2012) dataset [216].

2.2.9 Statistical analysis

PFS for first-line treatment was calculated from the date of randomization to the first observation of disease progression or death reported after first-line treatment. To test the predictive value of candidate genes, multivariate Cox proportional hazard models were built that included

the variables treatment, candidate gene and an interaction term treatment*candidate gene. For DCR1-specific analyses, we also included age, gender, WHO performance status and prior adjuvant therapy for both the CAIRO and the FOCUS samples, plus normal or abnormal LDH, and location of metastases for CAIRO. Cox proportional hazard models were used to estimate Hazard Ratios (HR) and 95% confidence intervals (CI). Kaplan-Meier analyses and log-rank tests were used to estimate survival over time. Correction for multiple testing in the discovery set was done by the Benjamini Hochberg method.

Students T-test was used for comparison of DCR1 expression levels before and after DAC treatment of HCT116. Pearson correlation analysis was used to measure correlation between DCR1 methylation and mRNA expression levels from 223 primary CRC tissue samples as provided by The Cancer Genome Atlas (TCGA) database.

Statistical analyses were performed using the computing environment R version 3.2 [241], including the packages survival (http://CRANR-projectorg/package=survival 2014) and rms (http://CRANR-projectorg/package=rms 2015). [242]

2.3 Results

2.3.1 Candidate gene selection

Correlation analyses of the DNA methylation status of 389 genes involved in DNA Damage Repair and Response with sensitivity to 118 drugs in 32 cell lines yielded 22 genes associated with topoisomerase-inhibitor related mode of action. These genes were analyzed for DNA methylation status in the discovery set (n=185). Methylation frequencies ranged from 5% to 98%, average 43%.

2.3.2 Evaluation of biomarker potential in the discovery set (CAIRO)

In concordance with the original CAIRO study, the patients of the discovery set showed significantly longer PFS when treated with CAPIRI (n=95) compared to CAP alone (n=90) (median PFS of 252 vs 182 days for CAPIRI vs CAP, respectively; HR=0.67 (95% CI 0.50-0.90, p=0.007) (figure 1A). DCR1 was methylated in 72/185 (39%) tumors. To assess

Table 1. Multivariate analysis for predictive value of candidate genes, showing p-values (size) and Hazard Ratio's (color)

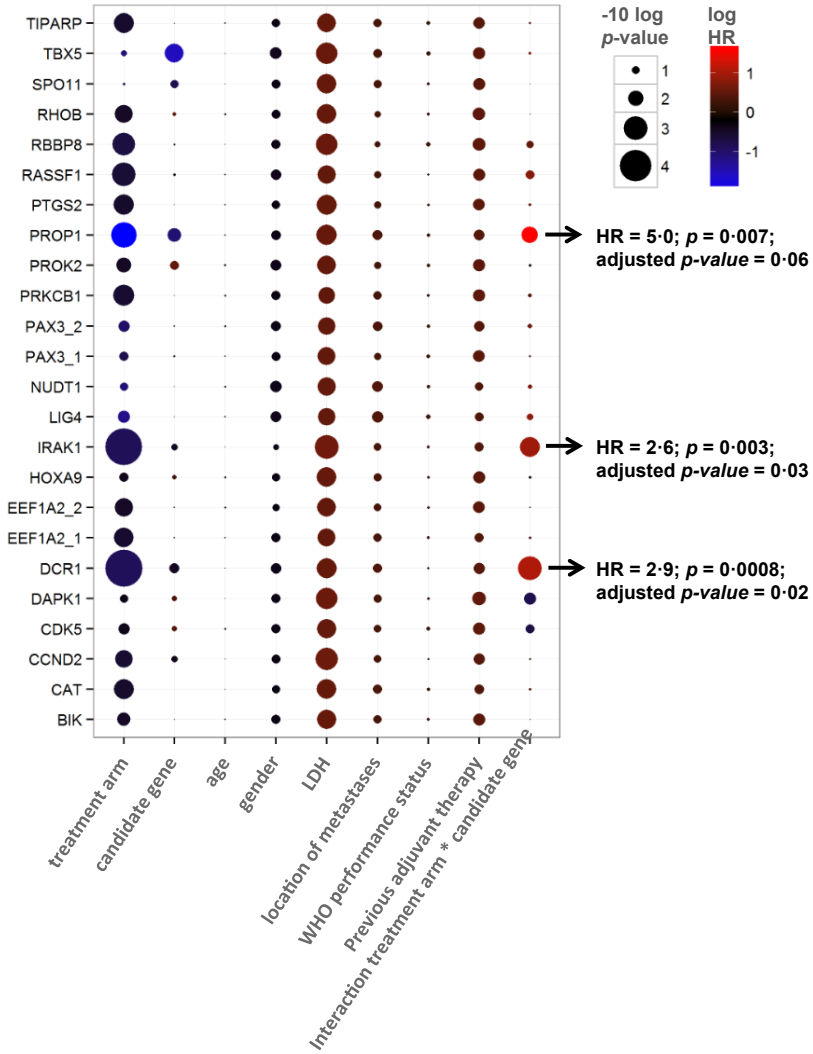


Figure 2.1: Multvariate analysis for predictive value of candidate genes, showing p-values (size) and Hazard Ratios (color)

the predictive value of each candidate gene, a multivariate survival model was generated including clinical variables, treatment arm, and an interaction term between treatment arm and candidate gene. After correcting for multiple testing, the treatment arm*candidate gene interaction remained significant for Tumor Necrosis Factor Receptor Superfamily member 10c (TNFRSF10c, also known as Decoy Receptor 1 (DCR1)) and Interleukin-1 Receptor-Associated Kinase 1 (IRAK1). This indicates that the methylation status of these candidate genes exerted an independent effect on PFS that was different in the one treatment arm compared to the other treatment arm (Figure 2.1).

Kaplan-Meyer curve analysis revealed that out of the two final candidate genes, the methylation status of DCR1 was predictive of PFS after treatment with CAPIRI, but not for PFS after treatment with CAP; patients with methylated DCR1 tumors progressed more quickly than patients with unmethylated DCR1 tumors when treated with CAPIRI (HR=2.1 (95% CI 1.3-3.3, p=0.001), but no difference was observed between patients with unmethylated or methylated DCR1 tumors when treated with CAP (HR=0.7 (95% CI 0.5-1.1, p=0.1). IRAK1 methylation was predictive of PFS after treatment with CAP and hence was not further studied . Because CAIRO was a randomized controlled trial, we were able to estimate the benefit of CAPIRI treatment over CAP treatment for patients with methylated or unmethylated DCR1 tumors by comparing PFS between the different treatment arms. Patients with methylated DCR1 (72 out of 185; 39%) did not benefit from adding irinotecan to CAP (median PFS of 192 vs 184 days for CAPIRI vs CAP, respectively; HR=1.2 (95%CI 0.7-1.9, p=0.6; figure 2.2B)). In contrast, patients with unmethylated DCR1 showed a significantly longer PFS when treated with CAPIRI compared to CAP alone (median PFS of 270 vs 178 days for CAPIRI vs CAP, respectively; HR=0.4 (95% CI 0.3-0.6, p=0.00001; figure 2.2C)).

2.3.3 Internal validation set (CAIRO)

In the second set of patients from the CAIRO study, in concordance with the original CAIRO study [208], PFS was significantly longer for patients treated with CAPIRI (n=88) compared to patients treated with CAP alone (n=78) (median PFS of 267 vs 200 days for CAPIRI vs CAP, respectively; HR=0.6 (95% CI 0.5-0.9, p=0.003; figure 2.3A)).

DCR1 was methylated in 88 out of 166 (53%) tumors. A multivariate

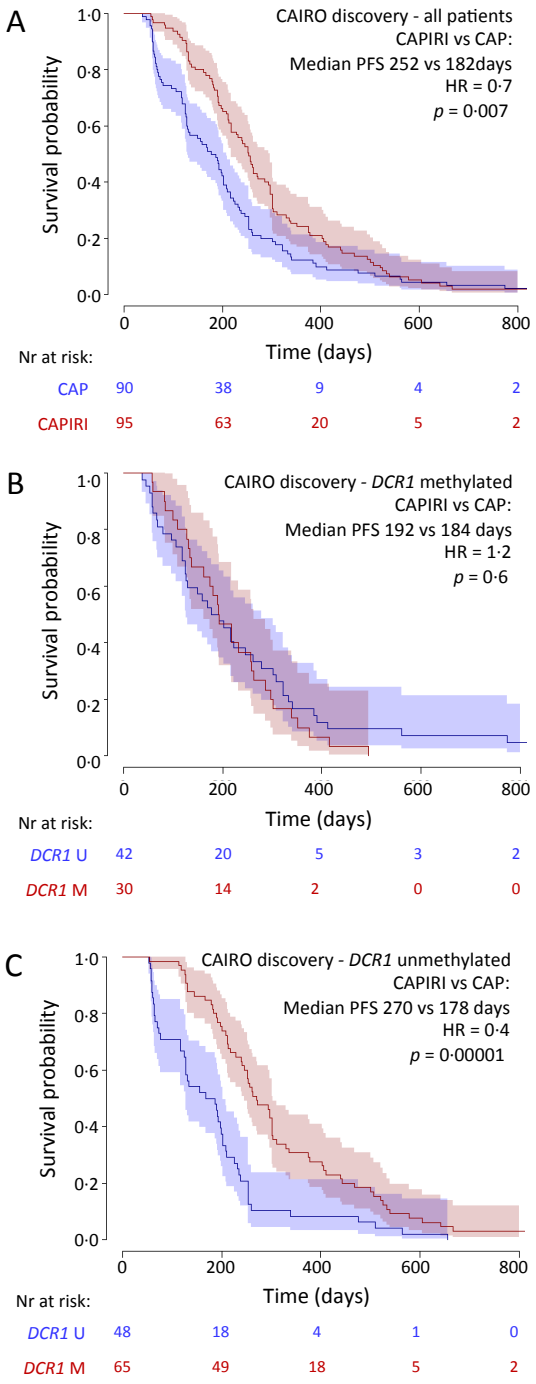


Figure 2.2: CAIRO discovery set: Progression-free survival in metastatic CRC cancer patients treated in first-line with CAP (blue line) or CAPIRI (red line) in (A) all patients from the CAIRO discovery set, in (B) patients with methylated tumor DCR1 or in (C) patients with unmethylated tumor DCR1. 95% confidence interval of the survival probability is shown by blue and red shades. HR=Hazard Ratio (CAPIRI versus CAP).

analysis, as described for the discovery set, showed a significant interaction between treatment arm and DCR1 methylation (p=0.04, table 2). Kaplan-Meyer analyses confirmed that patients with methylated DCR1 tumors did not significantly benefit from CAPIRI treatment over CAP treatment (median PFS of 267 vs 203 days for CAPIRI vs CAP, respectively; HR=0.9 (95%CI 0.6-1.4, p=0.5; figure 2.3B)), whereas patients with unmethylated DCR1 tumors did (median PFS of 261 vs 195 days for CAPIRI vs CAP, respectively; HR=0.5 (95%CI 0.3-0.7, p=0.0008) (figure 2.3C).

2.3.4 External validation set (FOCUS)

As an independent validation series, we analyzed 467 tumor samples from another randomized controlled phase III clinical trial (FOCUS).12 In this series, similar to the original FOCUS trial, PFS was significantly longer for patients treated with FOLFIRI (n=136) compared to patients treated with 5-FU alone (n=331) (median PFS of 272 vs 231 days for CAPIRI vs CAP, respectively; HR=0.8 (95%CI 0.6-1.0, p=0.02); figure 2.4A).

DCR1 was methylated in 225 out of 467 (48%) tumors. Multivariate analysis revealed that there was no significant interaction between treatment arm and DCR1 methylation status (p=0.3). Indeed, Kaplan-Meyer analyses revealed that patients with methylated or unmethylated DCR1 had a similar effect size from FOLFIRI treatment over 5-FU treatment (methylated DCR1: median PFS of 283 vs 225 days for FOLFIRI vs 5-FU, respectively; HR=0.7 (95%CI 0.5-0.9, p=0.01); figure 3C; unmethylated DCR1: median PFS of 253 vs 235 days for FOLFIRI vs 5-FU, respectively; HR=0.8 (95%CI 0.6-1.2, p=0.4) (figure 2.4B)).

2.3.5 Methylation of DCR1 is associated to decreased gene expression

The relation between DCR1 promoter hypermethylation and gene expression was investigated in vitro in a panel of 13 CRC cell lines. Ten out of 13 CRC cell lines were fully methylated for DCR1 and showed low or absent gene expression. The other three CRC cell lines were hemi-methylated and showed clearly higher gene expression levels (figure 2.5A). Treatment of two CRC cell lines, HCT116 and Colo205, with the demethylating agent DAC resulted in increased DCR1 expression (p=0.005 and p=0.08, respectively; figure 2.5B). In addition, data from The Cancer Genome At-

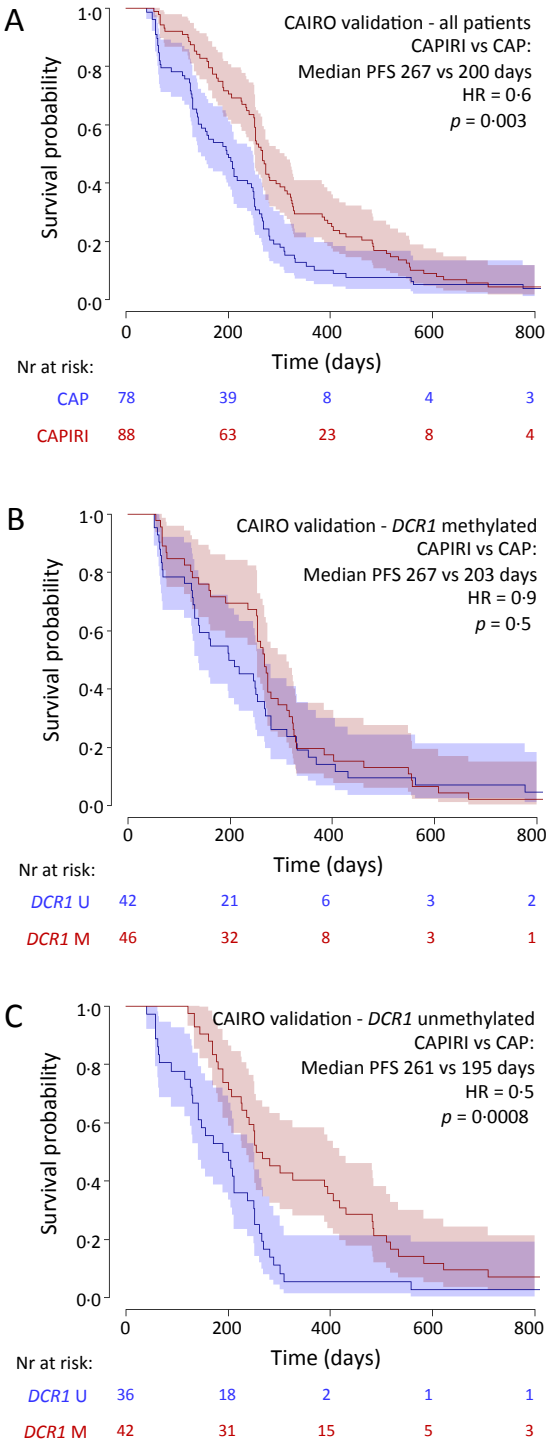


Figure 2.3: CAIRO validation set: Progression-free survival in metastatic CRC cancer patients treated in first-line with CAP (blue line) or CAPIRI (red line) in (A) all patients from the CAIRO validation set, in (B) patients with methylated tumor DCR1 or in (C) patients with unmethylated tumor DCR1. 95% confidence interval of the survival probability is shown by blue and red shades. HR=Hazard Ratio (CAPIRI versus CAP)

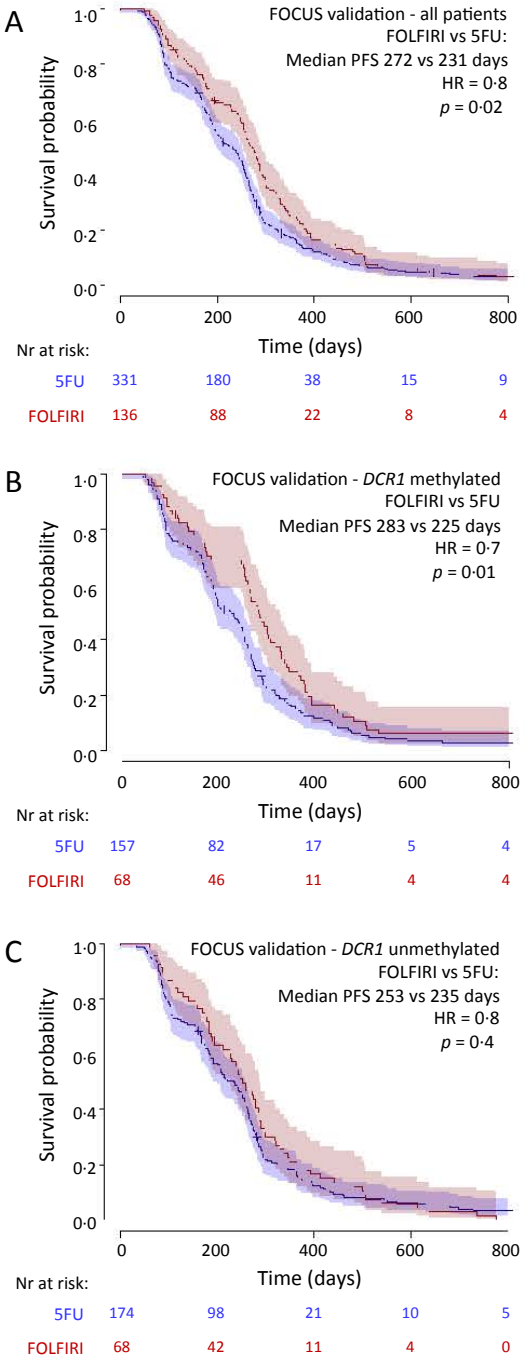


Figure 2.4: FOCUS validation set: Progression-free survival in metastatic CRC cancer patients treated in first-line with 5-FU (blue line) or FOLFIRI (red line) in (A) all patients from the FOCUS validation set, in (B) patients with methylated tumor DCR1 or in (C) patients with unmethylated tumor DCR1. 95% confidence interval of the survival probability is shown by blue and red shades. HR=Hazard Ratio (FOLFIRI versus 5-FU)

las (TCGA) database (http://cancergenome.nih.gov), including 223 CRC tumors, confirmed a negative correlation between DCR1 DNA methylation and DCR1 mRNA expression (Pearson correlation of -0.4, p=3.4E-9; figure 2.5B).

2.4 Discussion

In the present study we used a candidate gene approach to identify methylation markers for response to treatment with irinotecan-based therapy. We first made a selection of candidate genes based on in vitro findings on their function in relation to the mode of action of irinotecan, i.e. topoisomerase inhibition. We next tested for correlation of the methylation status of the candidate genes and PFS after treatment with CAPIRI therapy of metastatic CRC patients participating to the phase III CAIRO trial, [208] which identified DCR1 as a candidate marker. Because patients treated with CAP alone were used as a control group, this analysis showed DCR1 methylation as a potential negative predictive marker for response to irinotecan-based therapy. The initial finding in the discovery set could be confirmed in a second series of patients from the same CAIRO study, which indicated that the initial finding was not a stochastic statistical finding. However, validation in a second, independent series of metastatic CRC patients from the phase III FOCUS trial, [237] treated with first-line FOLFIRI or 5-FU alone, did not confirm the negative predictive value of DCR1 methylation status to irinotecan-based therapy.

Developing predictive biomarkers that reach the phase of introduction into clinical practice has proven to be highly challenging. Literature is full of proof of concept publications on potential biomarkers, but in most instances no further validation follows. The current study was carefully designed in order to overcome most common pitfalls in biomarker discovery [243, 244]; i.e. a strong biological rationale existed for the preselected candidate genes, and extensive evaluation (discovery, internal validation and external validation) was performed in a prospective-retrospective design [245] on a total of 818 archival tumor samples derived from two similar well-conducted phase III randomized clinical trials, providing the highest quality of clinical annotation. [208, 237] In addition, both clinical trials included a control group (i.e. CAP as control group for CAPIRI and 5-FU as control group for FOLFIRI), which is required to distinguish predictive from prognostic markers. Furthermore, biomarker independence was tested by including potential confounding factors in the statistical

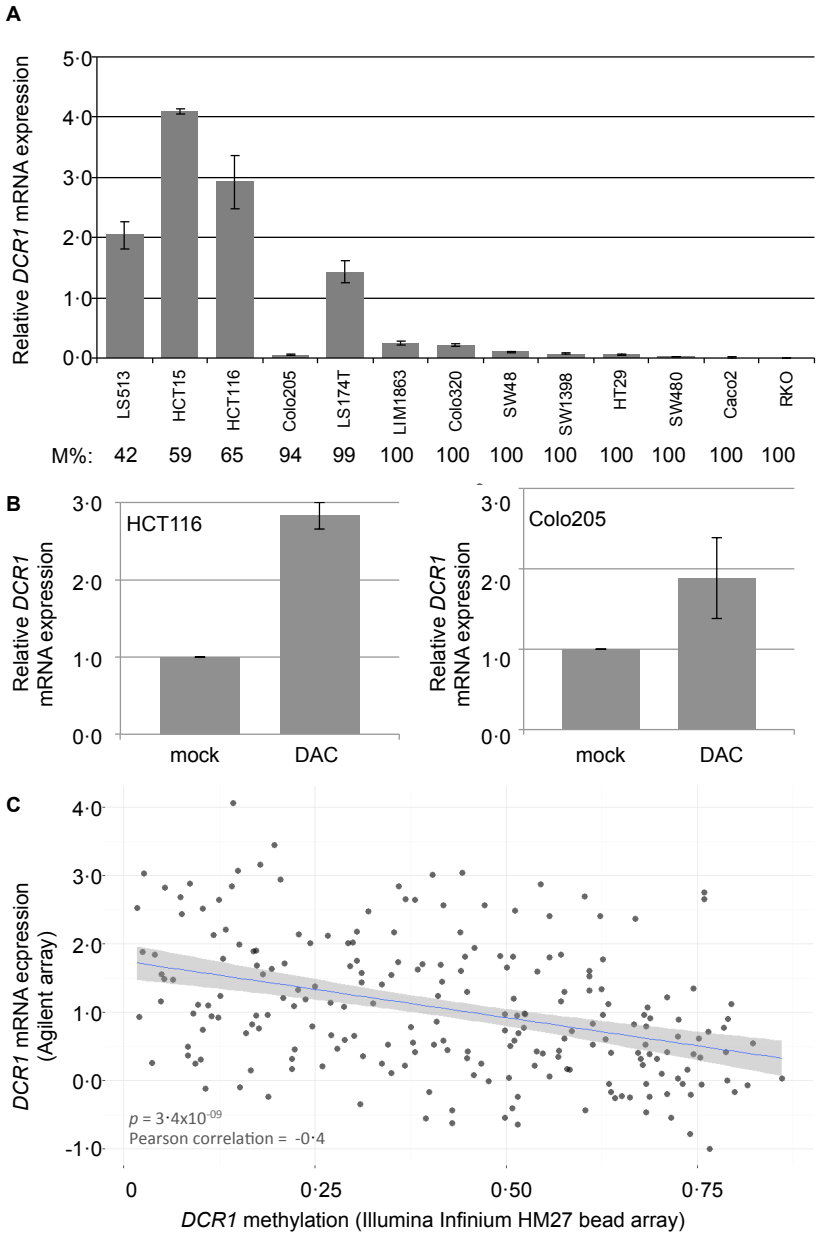


Figure 2.5: DCR1 methylation and mRNA expression levels A. DCR1 mRNA expression analysis in CRC cell lines by RT-PCR. DCR1 DNA methylation percentage as measured by qMSP (M%) is indicated below each cell line. Quantifications represent mean expression values from three independent experiments. B. DCR1 mRNA expression analysis by RT-PCR of HCT116 (left panel) and Colo205 (right panel) with and without DAC treatment (p=0.005 and p=0.08, respectively). Scatter plot including a linear regression line and 95% confdence interval, showing the correlation of DCR1 methylation levels and DCR1 mRNA expression in 223 CRC tissues from TCGA

models. Nonetheless, after initial validation in a second subsample of the CAIRO study, we were not able to validate the negative predictive value of DCR1 methylation for irinotecan-based therapy in the independent patient series from FOCUS. A lack of correlation between DCR1 methylation and DCR1 gene expression could be one of the reasons why we were not able to validate DCR1 methylation as a marker for response to irinotecan-based therapy. However, our cell-line experiments as well as analysis of a large series from the TCGA database did show a correlation between DCR1 DNA methylation and gene expression silencing. All this data together obviously raised the question whether DCR1 methylation should simply be discarded as a potential biomarker for response to irinotecan-based therapy, or whether our findings can be explained otherwise.

The two trials for instance, while they show substantial resemblances at first glance differ in a number of features related to inclusion (e.g. the performance scores leading to differences in patient characteristics) and treatment (e.g. different backbone treatment; CAP versus 5-FU). In addition, potential differences in the collection and storage of material may affect the results of analytical procedures. However, one could argue that a predictive biomarker of clinical value should be robust enough to cope with these variations. On the other hand, it is well known that standardization in sample handling and processing is critical also in the field of mRNA profiling and NGS. [246, 247]

The current study has some limitations. For example, measurements were performed on samples from the primary tumor, while patients were treated for their metastases, which provokes the question whether intra tumor heterogeneity could play a role. Although metastases can acquire additional genomic alterations, they keep most alterations present in the primary tumor. [224, 225] As DNA methylation is usually an early event in colorectal carcinogenesis, this is likely to be the case here as well. [226] Another limitation of the current study is that DCR1 methylation analyses were performed with identical primers but with different reagents in different laboratories for the three study cohorts. This could have introduced variability in test results. The proportion of patients having a positive test result was slightly different for the three cohorts indeed (39% in the discovery set, 53% in the internal validation set and 48% in the external validation set). However, because the predictive value of DCR1 methylation with regard to irinotecan-based therapy showed similar results in the two cohorts with largest relative difference in prevalence of

methylation (39% vs 53%), this variability is not likely to be the cause of the inability to validate DCR1 methylation as a predictive biomarker.

In conclusion, the present study revealed DCR1 methylation as a negative predictive marker for irinotecan-based therapy in metastatic colorectal cancer in both a discovery and an initial validation set, which could not be confirmed in an external validation data set. The present study highlights the importance of extensive evaluation of potential biomarkers. It also shows the complexity and extensiveness of systematic evaluation of a potential biomarker in order to generate more than just a proof of concept, and that a well-designed study is not a guarantee of success. Improvements in multi-team collaborations and in organizing data acquisition and biobanking in clinical trials will be necessary for efficient and successful discovery of predictive biomarkers in the future.

3

Sequencing Assay Predicting MGMT Methylation and Overall Survival in Glioblastoma Patients Receiving Chemoradiotherapy with Temozolomide

Copied from article:
Sequencing assay predicting MGMT methylation and overall survival in glioblastoma patients receiving chemoradiotherapy with temozolomide
Geert Trooskens, Annika Malmstroem, Martin Hallbeck, Peter Soderkvist, Greg Jones, Leander Van Neste, Wim Van Criekinge [Submitted]

3.1 Abstract

Glioblastoma multiforme (GBM) is the highest-grade astrocytoma and the most common and most aggressive form of brain cancer. Epigenetic silencing of MGMT is associated with longer overall survival in patients with GBM who receive radiotherapy (RT) combined with temozolomide (TMZ) chemotherapy. Methylation-specific PCR (MSP), a PCR-based technique that can sensitively detect methylated molecules in a background of unmethylated DNA is commonly used to determine the epigenetic status of the MGMT promoter. The purpose of this study was to investigate if next-generation sequencing (NGS) can be used to draw a

more detailed picture of the methylation profile compared to MSP, allowing more accurate assessment of the heterogeneity of methylation, both inter- and intra-allele. We show that NGS enhances the prognostic value of MGMT promoter methylation, offering an interesting alternative to MSP.

3.2 Introduction

Over 14.000 brain cancer-related deaths, or 2.4% of all cancer-related deaths, are reported per year in the US [248]. Glioblastoma multiforme (GBM) is the highest-grade astrocytoma and the most common and most aggressive form of brain cancer. It can occur de novo or as a secondary glioblastoma in 5 % of the cases [249]. GBM constitutes 30% of all brain tumors and patients have a 5-year survival rate lower than 17% [248].
In normal cells the O-6-methylguanine-DNA methyltransferase gene (MGMT) is responsible for repair of DNA damage caused by ionizing radiation, organic cyclic compounds and oxidative stress through DNA de-alkylation [250]. The MGMT protein removes alkyl groups from the O^6-position of guanine by an irreversible transfer of the alkyl group to a cysteine residue at its active site. The original guanine nucleotide is thereby restored and the alkylated MGMT protein send to proteasome-mediated degradation. Thus, the amount of MGMT proteins in a cell correlates directly with the cells ability to repair DNA damage [251–253] . Epigenetic silencing of MGMT by DNA methylation has been observed in various tumors [254]. When MGMT is silenced, patients showed an increased risk for developing Colorectal Cancer. [255]. Interestingly, epigenetic silencing of MGMT has been associated with longer overall survival in patients with GBM who receive radiotherapy (RT) combined with temozolomide (TMZ) chemotherapy. [61, 256]. Approximately 30% to 45% of the patients with gliomas have a methylated MGMT promoter serving as a favorable prognostic factor for chemoradiotherapy [62, 257]. The MGMT methylation status can be combined with gene expression and genomic mutations to enhance the prognostic test for GBM patients receiving RT with TMZ further. [258, 259].
With the increased utilization of next-generation sequencing (NGS), mainly in research settings, the question arises whether this technique offers benefits over PCR-based approaches. Methylation-specific PCR (MSP), a PCR-based technique that can sensitively detect methylated molecules in a background of unmethylated DNA [167], is commonly used to determine

DNA methylation. However, MSP typically results in a binary (methylated or unmethylated) call of the sample, lacking single-base methylation patterns. Earlier studies showed great variability in determining the methylation status of the MGMT promoter. Pyrosequencing and Sanger sequencing outperformed other techniques, including MSP as predictor of prognosis in GBM patients [260–262].
The purpose of this study was to investigate if ultra deep NGS (bisulphite treatment, target amplfication, followed by NGS) can be used to draw a more detailed picture of the methylation profile, allowing more accurate assessment of the heterogeneity of methylation, both inter- and intra-allele. This increased accuracy can potentially improve the prognostic value of MGMT promoter methylation. With the decreasing cost of NGS [263], this technique could be a valuable alternative to MSP, leading to increased prognostic value at the same relatively low cost.

3.3 Results

3.3.1 MSP

An average of 2986 ng of DNA (median: 2114 ng) was extracted from the samples. For 26 out of the 72 (36%) samples that received RT in combination with TMZ, the DNA yield, measured as ACTB copy numbers, was too low to assign an unambiguous methylation call using MSP according to Vlassenbroeck et al. [264]. Over the total cohort of 121 samples 40% lacked sufficient ACTB copy numbers.

3.3.2 Deep Sequencing

An average sample coverage of 659.000 reads was obtained. On average, 29% of the reads could be unambiguously mapped and passed the quality filter. The low mapping percentage is due to the strict quality filter applied (no mismatches over the entire amplicon) and a considerable amount of the synthetic control gene spiked in each sample. Two samples showed a significantly ($<$10000) lower amount of mapped reads (Sample 24020; 2765 reads, sample 24174; 640 reads). Both of them did not receive the RT + TMZ treatment thus were excluded in the overall survival analysis. The average overall methylation level was 24.1%.

3.3.3 Correlation between NGS and MSP Results

The spearman correlation between the MSP ratio and the observed methylation frequency of each individual CpG was calculated and plotted on the genomic coordinates (Figure 3.1 D). The highest correlation was found for the three middle C 's in the forward primer (po. Interestingly, only the CpG located at the most 3 '-end of the reverse primer correlated well with the MSP ratio, while the methylation frequency of the remaining two C 's in the reverse primer seemed to have limited influence on the MSP ratio. In summary, this indicates that the methylation frequency of only a few C's are really important in driving the MSP result, most notably the four most 3'-CpG 's in the forward primer and the most 3'-CpG in the reverse primer.

3.3.4 Comparing Next Generation Sequencing and MSP as a Prognostic Marker for Overall Survival

By using ultra deep bisulphite sequencing we identified two CpGs with a higher prognostic value for the 72 GBM patients receiving RT and TMZ (CPG44 with P=2.0e-05 by the log-rank test, hazard ratio 0.31; 95% confidence interval, 0.17 to 0.54 and CpG61 with P=2.9e-05 by the log-rank test, hazard ratio 0.31; 95% confidence interval, 0.18 to 0.55) compared to MSP (P=3.8e-03 by the log-rank test, hazard ratio 0.40; 95% confidence interval, 0.21 to 0.76). The proportional hazard model containing the CpG44, CpG61 and Age Variable outperformed the MSP assay and the single CpG methylation values (P=3.4e-07 by the log-rank test, hazard ratio 0.20; 95% confidence interval, 0.10 to 0.39). The ROC curve analysis for one year, two year and three year overall survival generated similar results: (1 Year OS AUCs: 0.60 MSP, 0.64 CpG44, 0.66 CpG61; 2 Year OS AUCs: 0.75 MSP, 0.78 CpG44, 0.80 CpG61; 3 Year OS AUCs: 0.68 MSP, 0.84 CpG44, 0.81 CpG61)

Survival analysis using proportional regression model of NGS methylation data

A Cox proportional Hazard model containing the all the individual CpG methylation fraction and age variables was generated to predict the Overall Survival in the patient cohort. The model outperformed the individual

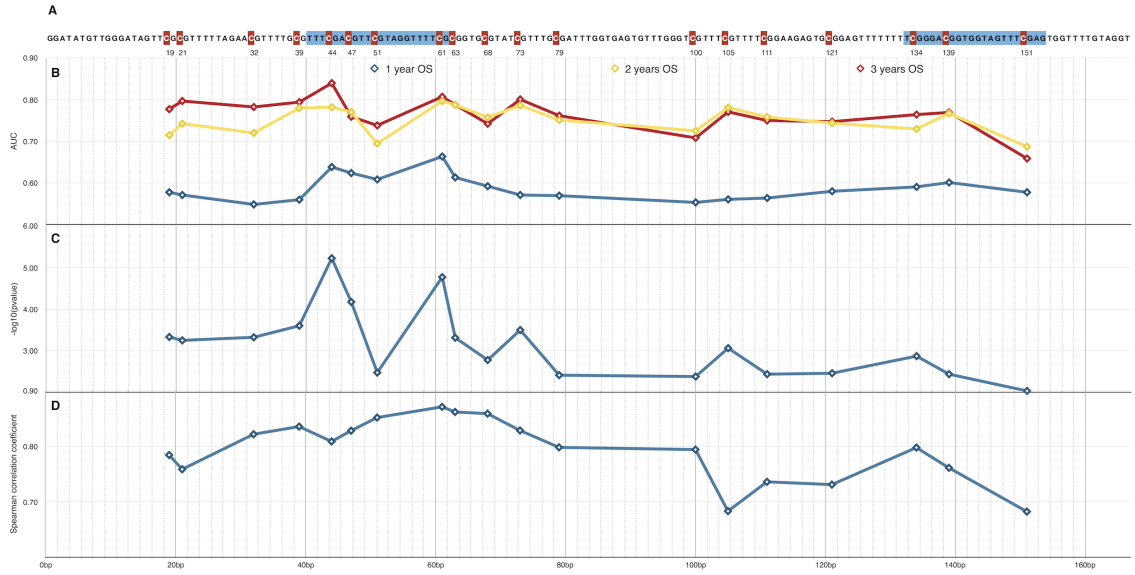


Figure 3.1: Overview of the BS Sequencing amplicon, The X-axis represent the genomic region sequenced in the MGMT gene promoter region: (A) The Sequencing amplicon. The MSP primer pair is colored blue, methylation prone cytosines in a CpG context are represented in red. (B) Survival ROC statistics: The Area under the curve scores for 1,2 and 3 year overall survival in GBM patients receiving RT + TMZ plotted for each individual CpG. A clear peak can be seen over the three years in AUC value for the CpGs at position 44 and 61 (C) The Y-axis represents -log(10) of the p-value for each individual CpG for overall survival in GBM patients receiving RT + TMZ. Higher peaks correspond to lower P-values. A difference in p-values between the CpG methylation ratios of several orders of magnitude was obeserved. (D) Spearman correlation between the MSP MGMT ratio and the individual CpG methylation levels.

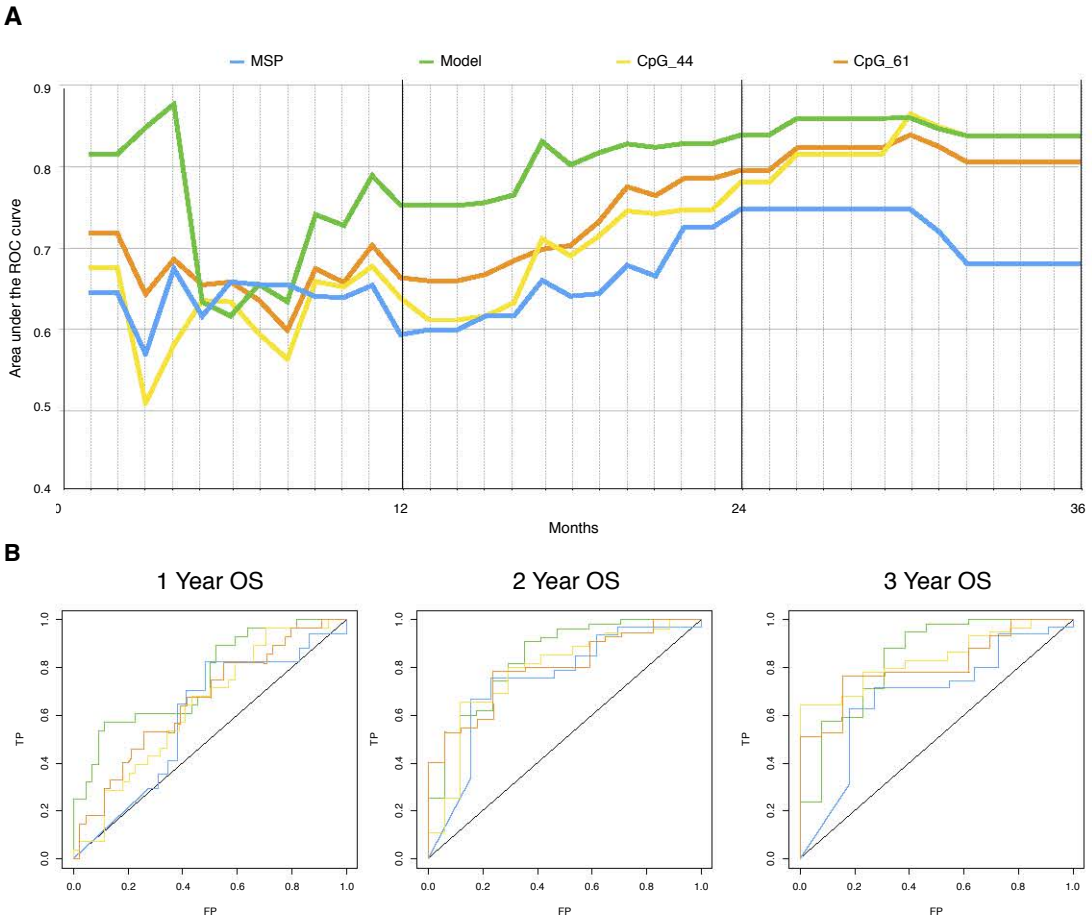


Figure 3.2: Overview of performance for predicting Overall Survival over time for the MSP assay, The proportional hazard model, CPG44 and CPG61. (A) Area under the ROC curve over time for the four variables. (B) Receiver Operating Characteristic (ROC) curve for predicting overall one year survival, two year survival and three year survival

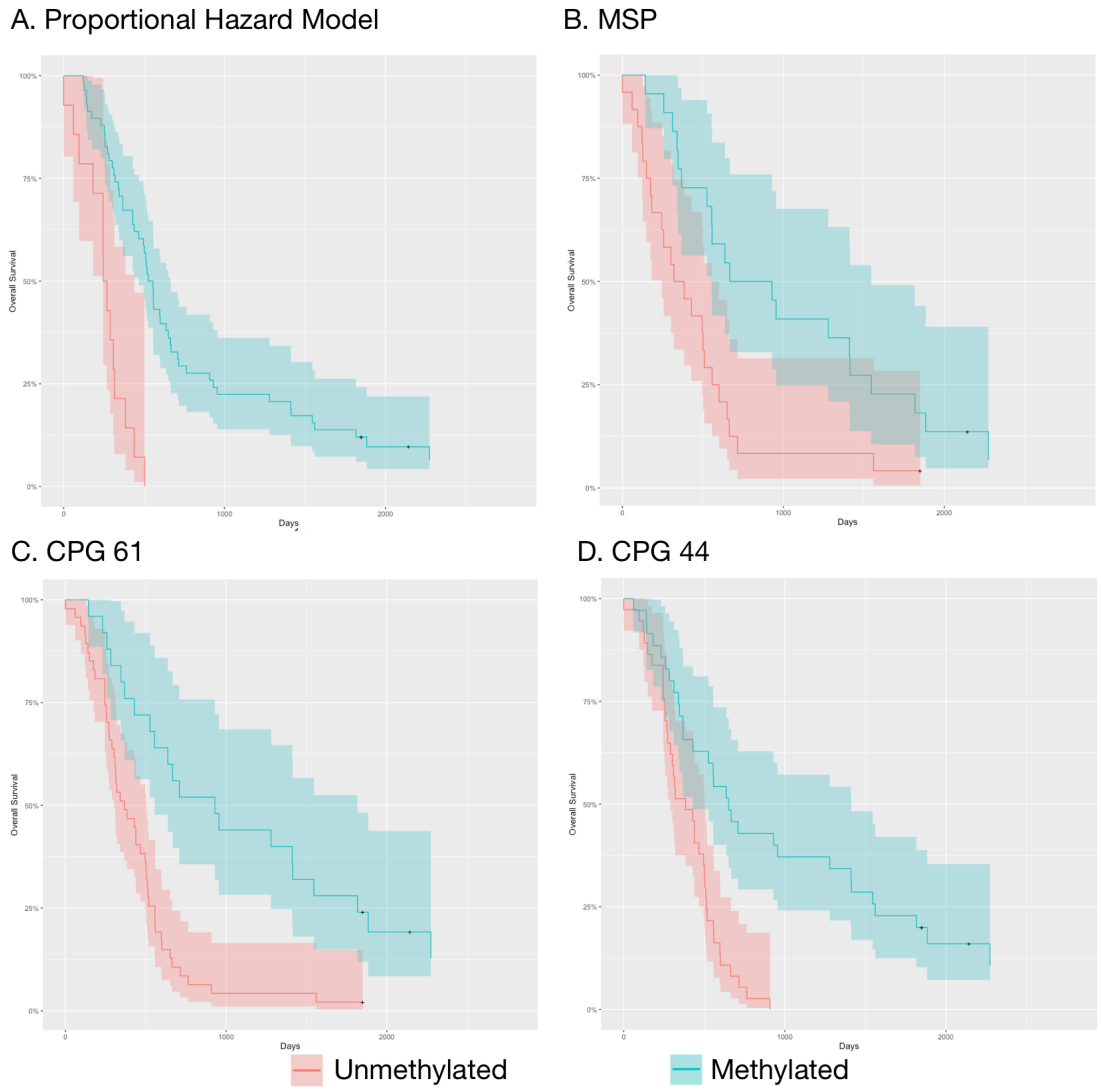


Figure 3.3: KaplanMeier Estimates of Overall Survival, According to: (A) The Proportional hazard model (P=3.45e-07) (B) MGMT Promoter Methylation Status measured by MSP (P=3.8e-03), (C) The Methylation Status of CpG 61 (P=2.9e-05), (D) The Methylation Status of CpG 44 (P=2.0e-05). The colored surface around the curves indicate the 95% upper and lower confidence interval. P values were calculated using the log-rank test. The three variables that are calculated from NGS data have a better prognostic value then the MSP assay. The proportional hazard Model that was build on all the individual CpG values does not outperform the most significant individual CpG (CpG number 44)

.

CpG methylation values and the MSP assay at the different timepoints (1 Year OS AUCs: 0.60 MSP, 0.75 Model, 0.64 CpG44, 0.66 CpG61; 2 Year OS AUCs: 0.75 MSP, 0.84 Model, 0.78 CpG44, 0.80 CpG61; 3 Year OS AUCs: 0.68 MSP, 0.84 Model, 0.84 CpG44, 0.81 CpG61) (Figure 3.3 A)

3.4 Discussion

While MSP is a technique with proven usefulness and utility to detect DNA methylation of MGMT in GBM tumor specimen, our data suggests that ultra deep NGS sequencing paints a broader picture of the DNA methylation landscape in the MGMT promoter region, providing us with high resolution quantitative measurements to increase the prognostic value for chemoradiotherapy with TMZ in GBM patients.
A challenge when comparing different techniques to measure methylation levels lies in defining a threshold to classify a sample as methylated/unmethylated. The Gel Electrophoresis-Based MSP assay [62] has an inherent binary cutoff by visualising the actual PCR product. The real-time PCR assay [264] provides us with a continuous methylation ratio variable if the DNA copy numbers of the control gene are high enough. The ultra-deep sequencing assay generates a quantitative methylation fraction for every sequenced position in the region of interest. We chose the optimal thresholds for both the MSP ratio and the individual CpG methylation levels based on the logrank test. A disadvantage is the direct influence of the cutoff on the proportional amount of patients in each group. The fraction of patients that will be classified as positive is not only a statistical question, but is part of a larger medical decision-making process based on different factors such as overall life quality, cost and the availability of alternative treatments [265]. Therefore we also compared the MSP assay to NGS with threshold independent measurements (ROC analysis) after defined periods. Regardless of the threshold, we found that some CpGs have a higher prognostic value then others within the MGMT promoter.The two most significant CpGs outperformed the MSP assay both with the treshold-dependent (logrank test) method and the treshold-independent time-dependent (ROC analysis) approach.
A relatively large portion of the MSP samples lacked a sufficient amount of ACTB gene copy numbers to assign an unambiguous methylation call using MSP, this could be due to sample handling, resulting in a subop-

timal amount of sample DNA in the PCR reaction. Interestingly, this limitations was not observed in the NGS data showing adequate coverage for all samples that received the combinational chemoradiotherapy.
A proportional hazard model incorporating the deep sequencing measurements with the patients 'age at diagnosis' variable enhances the prognostic accuracy further. Adding clinical variables to a biomarker test has been shown to notably improve the precision of the test [266]. However, caution should be exercised for overfitting when generating a model from a relatively small dataset (72 samples).

Correlation with MSP suggests that the calculated ratio is a measurement of the averages of the individual methylation levels, with the 3'end CpGs of the primers being the most influential ones showing the highest correlation. Indeed, when these CpGs are methylated, a high ratio will be obtained with MSP. However, when one or both CpGs have a low methylation frequency, it will affect primer binding. The main limitations of this study was the relatively small amount of patients in this cohort that received cobinational chemoradiotherapy with TMZ (72 samples). In order to use sequencing as an alternative for MSP, the coverage depth needs to be adequate to assure that the alleles with the low frequency's can be observed when they are present in the sample while still remaining economically feasible.

3.5 Conclusion

Ultradeep sequencing allows for more accurate high resolution quantitative measurements of individual CpG methylation levels compared to methylation specific PCR. We found two CpGs that outperformed the current MSP test in prognostic value for GBM patients receiving combinational RT with TMZ treatment. A model that combines methylation levels from NGS with clinical variables shows potential to increase the accuracy of the test further. In addition, a large portion of the MSP samples lacked a sufficient amount of ACTB copy numbers to assign an unambiguous methylation call using MSP, while this limitations was not seen in the NGS data showing adequate coverage for all samples. As next-generation sequencing becomes a routine part of health care, it shows promise as a more sensitive, accurate and reliable procedure to detect the MGMT promoter methylation status in glioblastoma multiforme patients receiving combinational chemoradiotherapy with temozolomide treatment

3.6 Materials and Methods

Samples

A total of 112 GBM samples from the Linköping University (Sweden) were analyzed using MSP and NGS. The majority of the samples were collected between 2008 and 2012, however, 10 were older, dating back to 2003. 72 received RT combined with TMZ treatment.

Sample Preparation

A total of four ten-micron formalin-fixed paraffin-embedded (FFPE) slides were obtained from all patients. The tumor areas were separated from benign tissue before DNA isolation using the phenol-chloroform method. The DNA was bisulfite treated using the EZ DNA Methylation kit (Zymo Research).

Direct, Real-Time MSP

MGMT and ACTB quantification was performed by real-time MSP assays. These consisted of parallel amplification/quantification processes using specific primer and primer/detector pairs for each analyte using the Amplifluor assay format on an ABI Prism 7900HT instrument (Applied Biosystems, Foster City, CA). The Amplifluor direct forward primers are preceded by the detection elements (underlined). Sequence details for both forward and re- verse primers are as follows: forward primer MGMT: 5 -TTCGACGTTCGTAGGTTTTCGC - 3 ; reverse primer MGMT: 5 -CTCGAAACTACCACCGTCCCGA-3 ;
forward primer ACTB: 5 -AGGGAGTATATAGGTTGG GGAAGTT-3 ;
reverse primer ACTB: 5 -AACACACAATAACAAACACAAATTCAC-3 . The MGMT target sequence is located on the sense strand of chromosome 10 between positions 131265515 and 131265629. ACTB target sequence resides on the anti-sense strand of chromosome 7 between positions 5571902 and 5571799 , based on version 37.2 of the NCBI human genome. MSP reactions were performed using 1.5 μg of input DNA as described previously [264] and ratios were calculated using the ACTB control gene.

Target Amplification and Sequencing

50 ng of all samples was used for target amplification. Flanking primes, without CpGs were designed that span the entire region of the MSP assay (5'-GGATATGTTGGGATAGTT-3', 5'-GCCTACAAAACCACTC-3', Integrated DNA Technologies, Leuven, Belgium) covering 19 CpGs (3.1). Each bisulfite deep-sequencing amplicon was generated using the FastStart High Fidelity PCR System (Roche) in a 50 μl reaction and a touchdown PCR at annealing temperatures from 60°C and 55°C (five cycles at each temperature) followed by 30 cycles at an annealing temperature of 52°C. Reactions were performed in the GENEAMP PCR system 9700 (Applied Biosystems). After amplification each amplicon was qualified for the expected length using capilair electrophoresis with the Caliper Labchip GX (HT1000 DNA chip). Amplicons were quantified with a PicoGreen assay (Quant-iT PicoGreen dsDNA Assay Kit, Invitrogen-Molecular Probes, P7589) after column purification (High Pure PCR Cleanup Micro Kit, Roche). Next, a pooled and indexed library was prepared using the TruSeq DNA Sample Prep Kit (Illumina), starting with 200 ng of DNA for the end repair reaction. After library preparation, all samples containing adaptor and index sequences were quantified and an equimolar pool was generated. Samples and the artificial amplicon were sequenced on the MiSeq in 5 runs using 24 different indexes. The MiSeq v2 Reagent Kit (Illumina), was used for paired-end sequencing with two times 251 cycles.

Mapping and Analysis

The paired end 251bp sequence reads were trimmed and aligned using the Smith-Waterman algorithm. No mismatches were allowed (in the non-CpGs) . We used the human genome build GRCh37/hg19 as a reference. A methylation percentage was obtained for every CpG within the target amplicon in all samples.

Statistical analysis

MSP ratio cutoffs and CpG methylation cutoffs were optimised based on overall survival log-rank p-value. A Cox proportional-hazards model was generated with the methylation values of the individual CpG's and the patients age variable. Overall survival curves for the MSP ratio, each CpG methylation variable and the model were estimated by the Kaplan-Meier (KM) technique and compared with use of the two-sided log-

rank test [267]. An additional methylation cutoff-independent analysis was performed: Time-dependent receiver operating characteristic (ROC) curves and corresponding area under the ROC curve values were calculated from censored survival data using the KM method [268]. Statistical analyses were performed with R, a free software environment available at http://www.r-project.org/.

4

Conclusions and Further Research

4.1 Conclusions

Epigenetics is a fine-tuning mechanism that allows the biological complexity needed in cells of higher organisms to differentiate, specialise and quickly adapt to the environment. It is now common knowledge that when epigenetics goes astray, it can trigger a wide variety of illnesses, behaviors, and other health indicators, including almost all types of cancer, cognitive dysfunction, respiratory, cardiovascular, reproductive, autoimmune, and neurobehavioral illnesses.
Promising methodologies have been developed in the last decades to accurately detect epigenetic changes. This has led to the discovery of new epigenetic biomarkers that can diagnose diseases in an early stage and predict the outcome of specific treatments.
During this PhD, the focus shifted from genome-wide epigenetic biomarker discovery to applying the newly discovered biomarkers as a diagnostic, prognostic or predictive test for patients, particularly focused on cancer. The ultimate goal is to provide personalized treatments, allowing physicians to respond earlier to a disease and reducing ineffective treatments, cutting healthcare costs and improving the quality of live for the patients.
The following are the main research contributions of this dissertation:

4.1.1 Epigenomewide DNA-Methylation Profiling Using Methyl-CpG Binding Domain Capturing Based Ssequencing

We used methyl-CpG binding domain capturing based sequencing (MethylCap-Seq) to uncover DNA methylation in a genome-wide manner. We applied this technique to 345 human samples from different tissues and diseases and constructed a map of the human methylome by identifying the methylation-prone regions. The map enables researchers to reduce the comparison problem to a discrete amount of variables. This allowed multiple research projects to look for disease and tissue specific DNA-methylation markers.

4.1.2 Validation of Epigenetic Markers Using Bisulphite Sequencing Approaches and Methylation Specific PCR (MSP)

By optimizing and automating bisulphite primer design for PCR and sequencing purposes using a thermodynamic approach, we facilitated the validation of epigenetic biomarkers by shortening the primer design process and enhancing the primer quality.

4.1.3 WRN and DCR1 Promoter Methylation and Their Response to Irinotecan in Colorectal Cancer

DNA repair proteins such as the Werner syndrome RECQ helicase, WRN, are promising biomarkers for predicting the response to genotoxic chemotherapy. We attempted to validate previous studies that showed WRN promoter hypermethylation predicted the response to irinotecan using an independent sample set. We did not find a clear association between aberrant WRN promoter hypermethylation and reduced WRN expression. Moreover, in contrast to earlier studies we found an inverse correlation of WRN promoter hypermethylation with survival in metastatic colorectal cancer patients treated with irinotecan.

DCR1 promoter methylation was identified and initially validated as a potential negative predictive biomarker for response to irinotecan-based therapy, but external validation could not validate these findings. The

results of both studies underline the importance of extensive clinical evaluation of candidate biomarkers. The fact that the follow-up studies could not validate the initial results of the biomarkers shows the complexity and extensiveness of a systematic biomarker evaluation. Most research today stops at just a proof-of-concept. Improvements in multi-team collaborations and in organizing data acquisition and biobanking in clinical trails will be necessary for the discovery and validation of future biomarkers.

4.1.4 Sequencing Assay Predicting MGMT Methylation and Overall Survival in Glioblastoma Patients Receiving Chemoradiotherapy with Temozolomide

Glioblastoma multiforme (GBM) is the highest-grade astrocytoma and the most common and most aggressive form of brain cancer. Epigenetic silencing of MGMT is associated with longer overall survival in patients with GBM who receive radiotherapy (RT) combined with temozolomide (TMZ) chemotherapy. Methylation-specific PCR (MSP), a PCR-based technique that can sensitively detect methylated molecules in a background of unmethylated DNA is commonly used to determine the epigenetic status of the MGMT promoter. Ultradeep sequencing enabled more accurate high resolution quantitative measurements of individual CpG methylation levels compared to MSP. We found two CpGs that outperformed the current MSP test in prognostic value for GBM patients receiving combinational RT with TMZ treatment. A model that combines methylation levels from NGS with clinical variables shows potential to increase the accuracy of the test further. In addition, a large portion of the MSP samples lacked a sufficient amount of ACTB copy numbers to assign an unambiguous methylation call using MSP, while this limitations was not seen in the NGS data showing adequate coverage for all samples. As next-generation sequencing becomes a routine part of health care, it shows promise as a more sensitive, accurate and reliable procedure to detect the MGMT promoter methylation status in glioblastoma multiforme patients receiving combinational chemoradiotherapy with temozolomide treatment.

4.2 Future Research

In recent years, technologies and biomedical research have offered a wide range of clinical applications for epigenetics. It is to be expected in the next decade epigenetics will overcome the limitations of traditional genetics and genomics by the development of new assays based on epigenetic biomarkers and new epigenetic drugs able to control the function of our genome. Epigenetics can provide answers to some of the most pressing unresolved questions in our understanding of personalized medicine. Epigenetic biomarkers serve in the dynamic study of diseases, and therefore, epigenetic biomarkers will serve to predict the evolution of disease and to monitor the effect of treatments on diseases. Specifically, aberrant DNA methylation of genes serves as biomarkers for screening of epigenetic diseases and many types of cancers through potentially noninvasive means such as cheek swabs, urine and blood. This allows experts to closely monitor their patients by repetitive measurements. The future of diagnosis will involve more panels that consist of multiple biomarkers for screening and prognosis of specific diseases including cancers. The panels may offer more sensitive detection and accurate prognosis of diseases, as well as the discovery of potential therapeutic targets. Many recent publications describe development of biomarker screening panels for DNA methylation as well as other epigenetic markers, including miRNA, lncRNA, and histone modifications. Some of them are already available as commercial kits (Table 2.1). These can be combined in a model with other clinical variables to improve the test performance even further [266].

The integration of classical epigenetics methods with NGS technologies has improved current assays (part 3, chapter 3) and opened up research avenues that scientists could only dream about a couple of years ago. The capacity to study epigenetic modifications on a genome-wide scale is unparalleled and has given researchers the means to answer research questions about the deeper molecular mechanisms of epigenetic regulation, the overall epigenetic profile within tissues and diseases (part 2, chapter 1), the changes in epigenetic information over time, environmental exposures on a genome-wide scale and much more.

New technology developments are on the horizon (and in some cases

already available as research tools) that include single-molecule and single-cell assays for DNA methylation, histone modification, chromosomal proteins, ncRNAs, and chromatin structure. Advances in sequencing technology are enabling the detection of modified DNA bases without bisulfite conversion. These third-generation sequencing technologies allow for the combined analysis of genetic and epigenetic features on a single run within a single day [269]. It is without question that these new technologies currently in development will have an even more profound impact on clinical epigenetics research, as they will circumvent the limitations of tissue availability and remove the problems of PCR amplification biases by allowing single cell measurements and full length, single molecule RNA and DNA detection .

This thesis highlights the importance of the extensive evaluation of potential biomarkers. It also shows the difficulties evaluating and validating potential biomarkers in a clinical setting in order to generate more than just a proof-of-concept. The whole process of developing robust biomarkers that reach the phase of introduction into clinical practice has proven to be highly challenging (e.g. part 3). Statistical overfitting, inter-lab variability, different sample handling and non-optimal patient data acquisition are just a few of the possible culprits that cause a validation study to fail. Literature is full of proof of concept publications on potential biomarkers, but for most papers, no follow-up validation is performed.

While the availability of validated protocols and easy-to-use kits has facilitated the incorporation of epigenetics research into clinical research and diagnostics, some obstacles do persist: A well known issue is the limitation by sample amount requirements for some of the epigenetics assays. Typically, tissue samples availability is limited for clinical research, and much smaller than what is often required as starting materials for most of the well established protocols. This limitation forces clinical researchers to work with more difficult protocols and drives the development of new methods that require less input material. A second critical point is the standardization in sample handling and processing to reproduce results at different time point, between different operators, at different locations and in different countries. [246, 247] Storage of clinical samples can also

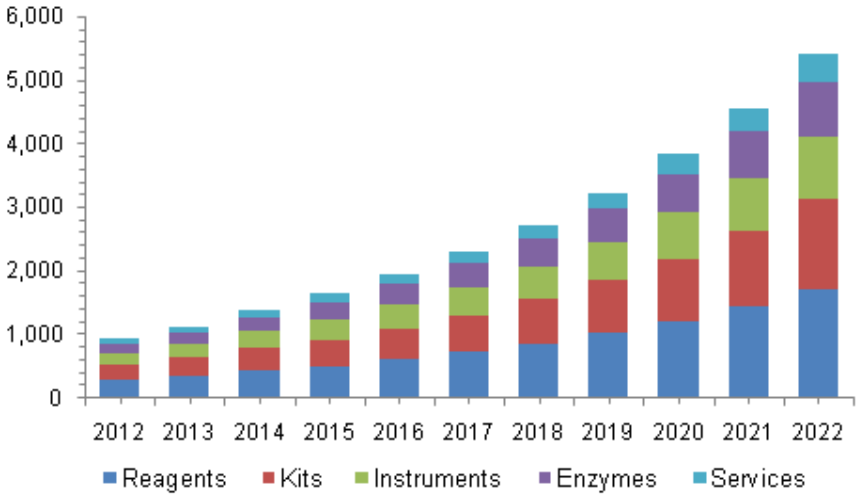


Figure 4.1: U.S. epigenetics market, by product, 2012-2020, (USD Million, according to Transparency Market Research)

be an issue, especially when fragile molecules such as RNAs need to be assayed. Improvements in multi-team collaborations and in organizing data acquisition and biobanking in clinical trials will be necessary for efficient and successful discovery of predictive biomarkers in the future.

In the light of the increasing knowledge on the role epigenetic factors play in disease, it is now becoming clear that epigenetic mechanisms could be promising therapeutic targets. Particularly taking into consideration that many of these epigenetic factors are potentially reversible. Epigenetic drugs could potentially help reverse abnormal gene expression back to pre-disease levels.
The global market for epigenetic drugs and diagnostics will be worth an estimated $5.7 billion by 2018 and is expected to reach $16.31 billion by 2022 (according to Transparency Market Research). Currently, epigenetic drugs have primarily been studied for their use in treating cancer. The increasing funding efforts have opened up new possibilities for researchers to pursue the use in other diseases like Alzheimer and asthma.

If we are to address the current increasing burden to national and private health-care systems from diseases such as metabolic diseases, obesity, cancer and neurological disorders, we need to develop a wide

array of biomarkers (diagnostic, prognostic, predictive, and for monitoring) which will contribute to improving precision medicine and helping rationalize health-care funding and resources. In the not too distant future, an amplitude of epigenetic biomarkers used in clinical diagnostics will not only help to boost the health of people, but also enhance the economic sustainability of health-care systems by avoiding side-effects, optimizing dosage and minimizing over-treatment.

Part IV
References

Bibliography

[1] Linda Van Speybroeck, Dani De Waele, and Gertrudis Van de Vijver. Theories in early embryology: close connections between epigenesis, preformationism, and self-organization. *Annals of the New York Academy of Sciences*, 981:7–49, December 2002.

[2] Anthony J F Griffiths, Jeffrey H Miller, David T Suzuki, Richard C Lewontin, and William M Gelbart. *Introduction to Genetic Analysis*. W. H. Freeman, February 2000.

[3] Heather J Cordell. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics*, 11(20):2463–2468, October 2002.

[4] F Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, August 1970.

[5] Peter M Visscher, Matthew A Brown, Mark I McCarthy, and Jian Yang. Five Years of GWAS Discovery. *The American Journal of Human Genetics*, 90(1):7–24, January 2012.

[6] R Holliday. Epigenetics: a historical overview. *Epigenetics*, 2006.

[7] Aaron D Goldberg, C David Allis, and Emily Bernstein. Epigenetics: a landscape takes shape. *Cell*, 128(4):635–638, February 2007.

[8] Ankur Jai Sood, Coby Viner, and Michael M Hoffman. DNAmod: the DNA modification database. *bioRxiv*, page 071712, 2016.

[9] Achim Breiling and Frank Lyko. Epigenetic regulatory functions of DNA modifications: 5-methylcytosine and beyond. *Epigenetics & chromatin*, 8(1):396, July 2015.

[10] Peter Cheung, C David Allis, and Paolo Sassone-Corsi. Signaling to Chromatin through Histone Modifications. *Cell*, 103(2):263–271, October 2000.

[11] Veryan Codd, Christopher P Nelson, Eva Albrecht, Massimo Mangino, Joris Deelen, Jessica L Buxton, Jouke Jan Hottenga, Krista Fischer, Tõnu Esko, Ida Surakka, Linda Broer, Dale R Nyholt, Irene Mateo Leach, Perttu Salo, Sara Hägg, Mary K Matthews, Jutta Palmen, Giuseppe D Norata, Paul F O'Reilly, Danish Saleheen, Najaf Amin, Anthony J Balmforth, Marian Beekman, Rudolf A de Boer, Stefan Böhringer, Peter S Braund, Paul R Burton, Anton J Mde Craen, Matthew Denniff, Yanbin Dong, Konstantinos Douroudis, Elena Dubinina, Johan G Eriksson, Katia Garlaschelli, Dehuang Guo, Anna-Liisa Hartikainen, Anjali K Henders, Jeanine J Houwing-Duistermaat, Laura Kananen, Lennart C Karssen, Johannes Kettunen, Norman Klopp, Vasiliki Lagou, Elisabeth M van Leeuwen, Pamela A Madden, Reedik Mägi, Patrik K E Magnusson, Satu Männistö, Mark I McCarthy, Sarah E Medland, Evelin Mihailov, Grant W Montgomery, Ben A Oostra, Aarno Palotie, Annette Peters, Helen Pollard, Anneli Pouta, Inga Prokopenko, Samuli Ripatti, Veikko Salomaa, H Eka D Suchiman, Ana M Valdes, Niek Verweij, Ana Viñuela, Xiaoling Wang, H-Erich Wichmann, Elisabeth Widen, Gonneke Willemsen, Margaret J Wright, Kai Xia, Xiangjun Xiao, Dirk J van Veldhuisen, Alberico L Catapano, Martin D Tobin, Alistair S Hall, Alexandra I F Blakemore, Wiek H van Gilst, Haidong Zhu, CARDIoGRAM consortium, Jeanette Erdmann, Muredach P Reilly, Sekar Kathiresan, Heribert Schunkert, Philippa J Talmud, Nancy L Pedersen, Markus Perola, Willem Ouwehand, Jaakko Kaprio, Nicholas G Martin, Cornelia M van Duijn, Iiris Hovatta, Christian Gieger, Andres Metspalu, Dorret I Boomsma, Marjo-Riitta Jarvelin, P Eline Slagboom, John R Thompson, Tim D Spector, Pim van der Harst, and Nilesh J Samani. Identification of seven loci affecting mean telomere length and their association with disease. *Nature Genetics*, 45(4):422–427, April 2013.

[12] Masood A Shammas. Telomeres, lifestyle, cancer, and aging. *Current opinion in clinical nutrition and metabolic care*, 14(1):28–34, January 2011.

[13] Ursula Muñoz-Najar and John M Sedivy. Epigenetic Control of Aging. *dx.doi.org*, 14(2):241–259, December 2010.

[14] Virginia Hughes. Sperm RNA carries marks of trauma. *Nature*, 508(7496):296–297, April 2014.

[15] Minju Ha and V Narry Kim. Regulation of microRNA biogenesis. *Nature Reviews Molecular Cell Biology*, 15(8):509–524, August 2014.

[16] A G Bader and P Lammers. The therapeutic potential of microRNAs. *Innovations in Pharmaceutical . . .*, 2011.

[17] Thomas Derrien, Rory Johnson, Giovanni Bussotti, Andrea Tanzer, Sarah Djebali, Hagen Tilgner, Gregory Guernec, David Martin, Angelika Merkel, David G Knowles, Julien Lagarde, Lavanya Veeravalli, Xiaoan Ruan, Yijun Ruan, Timo Lassmann, Piero Carninci, James B Brown, Leonard Lipovich, Jose M Gonzalez, Mark Thomas, Carrie A Davis, Ramin Shiekhattar, Thomas R Gingeras, Tim J Hubbard, Cedric Notredame, Jennifer Harrow, and Roderic Guigó. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome research*, 22(9):1775–1789, September 2012.

[18] Ling-Ling Chen. Linking Long Noncoding RNA Localization and Function. *Trends in biochemical sciences*, 41(9):761–772, September 2016.

[19] Jeffrey J Quinn and Howard Y Chang. Unique features of long noncoding RNA biogenesis and function. *Nature reviews. Genetics*, 17(1):47–62, January 2016.

[20] Huang Wu, Li Yang, and Ling-Ling Chen. The Diversity of Long Noncoding RNAs and Their Generation. *Trends in genetics : TIG*, 33(8):540–552, August 2017.

[21] Yiwen Fang and Melissa J Fullwood. Roles, Functions, and Mechanisms of Long Non-coding RNAs in Cancer. *Genomics, proteomics & bioinformatics*, 14(1):42–54, February 2016.

[22] Mathias Jucker and Lary C Walker. Self-propagation of pathogenic protein aggregates in neurodegenerative diseases. *Nature*, 501(7465):45–51, September 2013.

[23] Peter A Jones and Stephen B Baylin. The fundamental role of epigenetic events in cancer. *Nature reviews. Genetics*, 3(6):415–428, June 2002.

[24] A G Knudson. Mutation and cancer: statistical study of retinoblastoma. *Proceedings of the National Academy of Sciences of the United States of America*, 68(4):820–823, April 1971.

[25] Sharma, Shikhar, Kelly, Theresa K, and Jones, Peter A. Epigenetics in cancer. *Carcinogenesis*, 31(1):27–36, January 2010.

[26] Francis Blokzijl, Joep de Ligt, Myrthe Jager, Valentina Sasselli, Sophie Roerink, Nobuo Sasaki, Meritxell Huch, Sander Boymans, Ewart Kuijk, Pjotr Prins, Isaac J Nijman, Inigo Martincorena, Michal Mokry, Caroline L Wiegerinck, Sabine Middendorp, Toshiro Sato, Gerald Schwank, Edward E S Nieuwenhuis, Monique M A Verstegen, Luc J W van der Laan, Jeroen de Jonge, Jan N M IJzermans, Robert G Vries, Marc van de Wetering, Michael R Stratton, Hans Clevers, Edwin Cuppen, and Ruben van Boxtel. Tissue-specific mutation accumulation in human adult stem cells during life. *Nature*, 538(7624):260–264, October 2016.

[27] Mira Jakovcevski and Schahram Akbarian. Epigenetic mechanisms in neurological disease. *Nature medicine*, 18(8):1194–1204, August 2012.

[28] Chetan Bettegowda, Mark Sausen, Rebecca J Leary, Isaac Kinde, Yuxuan Wang, Nishant Agrawal, Bjarne R Bartlett, Hao Wang, Brandon Luber, Rhoda M Alani, Emmanuel S Antonarakis, Nilofer S Azad, Alberto Bardelli, Henry Brem, John L Cameron, Clarence C Lee, Leslie A Fecher, Gary L Gallia, Peter Gibbs, Dung Le, Robert L Giuntoli, Michael Goggins, Michael D Hogarty, Matthias Holdhoff, Seung-Mo Hong, Yuchen Jiao, Hartmut H Juhl, Jenny J Kim, Giulia Siravegna, Daniel A Laheru, Calogero Lauricella, Michael Lim, Evan J Lipson, Suely Kazue Nagahashi Marie, George J Netto, Kelly S Oliner, Alessandro Olivi, Louise Olsson, Gregory J Riggins, Andrea Sartore-Bianchi, Kerstin Schmidt, Ie-Ming Shih, Sueli Mieko Oba-Shinjo, Salvatore Siena, Dan Theodorescu, Jeanne Tie, Timothy T Harkins, Silvio Veronese, Tian-Li Wang, Jon D Weingart, Christopher L Wolfgang, Laura D Wood, Dongmei Xing, Ralph H Hruban, Jian Wu, Peter J Allen, C Max

Schmidt, Michael A Choti, Victor E Velculescu, Kenneth W Kinzler, Bert Vogelstein, Nickolas Papadopoulos, and Luis A Diaz. Detection of circulating tumor DNA in early- and late-stage human malignancies. *Science translational medicine*, 6(224):224ra24–224ra24, February 2014.

[29] Paul Cairns. Gene methylation and early detection of genitourinary cancer: the road ahead. *Nature reviews. Cancer*, 7(7):531–543, July 2007.

[30] Daniele Calistri, Claudia Rengucci, Andrea Casadei Gardini, Giovanni Luca Frassineti, Emanuela Scarpi, Wainer Zoli, Fabio Falcini, Rosella Silvestrini, and Dino Amadori. Fecal DNA for noninvasive diagnosis of colorectal cancer in immunochemical fecal occult blood test-positive individuals. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology*, 19(10):2647–2654, October 2010.

[31] Paul P Anglim, Todd A Alonzo, and Ite A Laird-Offringa. DNA methylation-based biomarkers for early detection of non-small cell lung cancer: an update. *Molecular cancer*, 7(1):81, 2008.

[32] Richard M Hoffman. Screening for Prostate Cancer. *The New England journal of medicine*, 365(21):2013–2019, November 2011.

[33] M M Shen and C Abate-Shen. Molecular genetics of prostate cancer: new prospects for old challenges. *Genes & development*, 24(18):1967–2000, September 2010.

[34] A EpsteinJI and ELL AminMB. *The 2014 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma: Definition of Grading Patterns and Proposal for a New Grading System.* American Journal of Surgical Pathology, 2005.

[35] Manel Esteller. Epigenetics in Cancer. *dx.doi.org*, 358(11):1148–1159, June 2009.

[36] Rui Henrique and Carmen Jerónimo. Molecular Detection of Prostate Cancer: A Role for GSTP1 Hypermethylation. *European urology*, 46(5):660–669, November 2004.

[37] C Jerónimo, H Usadel, R Henrique, J Oliveira, C Lopes, W G Nelson, and D Sidransky. Quantitation of GSTP1 Methylation in Non-neoplastic Prostatic Tissue and Organ-Confined Prostate Adenocarcinoma. *Journal of the . . .*, 93(22):1747–1752, November 2001.

[38] M O Hoque, O Topaloglu, and S Begum. Quantitative Methylation-Specific Polymerase Chain Reaction Gene Patterns in Urine Sediment Distinguish Prostate Cancer Patients From Control Subjects: Journal of Clinical Oncology: Vol 23, No 27. *Journal of Clinical . . .*, 2005.

[39] M Rouprêt, V Hupertan, D R Yates, and JWF Catto. Molecular detection of localized prostate cancer using quantitative methylation-specific PCR on urinary cells obtained following prostate massage. *Clinical Cancer . . .*, 2007.

[40] Paul Cairns, Manel Esteller, James G Herman, Mark Schoenberg, Carmen Jerónimo, Montserrat Sanchez-Cespedes, Nan-Haw Chow, Marc Grasso, Li Wu, William B Westra, and David Sidransky. Molecular Detection of Prostate Cancer in Urine by GSTP1 Hypermethylation. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 7(9):2727–2730, September 2001.

[41] Leander Van Neste, Alan W Partin, Grant D Stewart, Jonathan I Epstein, David J Harrison, and Wim Van Criekinge. Risk score predicts high-grade prostate cancer in DNA-methylation positive, histopathologically negative biopsies. *The Prostate*, 76(12):1078–1087, September 2016.

[42] M H Forouzanfar, K J Foreman, and A M Delossantos. Breast and cervical cancer in 187 countries between 1980 and 2010: a systematic analysis. *The lancet*, 2011.

[43] A Dobrovic and D Simpfendorfer. Methylation of the BRCA1 gene in sporadic breast cancer. *Cancer Research*, 57(16):3347–3350, August 1997.

[44] H Werner and I Bruchim. IGF-1 and BRCA1 signalling pathways in familial cancer. *The lancet oncology*, 2012.

[45] William M Grady, Joseph Willis, Parry J Guilford, Anita K Dunbier, Tumi T Toro, Henry Lynch, Georgia Wiesner, Kelly Ferguson, Charis Eng, Jae-Gahb Park, Seong-Jin Kim, and Sanford Markowitz. Methylation of the CDH1 promoter as the second genetic hit in hereditary diffuse gastric cancer. *Nature Genetics*, 26(1):16–17, September 2000.

[46] H Berman, J Zhang, and Y G Crawford. Genetic and Epigenetic Changes in Mammary Epithelial Cells Identify a Subpopulation of Cells Involved in Early Carcinogenesis. *Cold Spring Harbor ...*, 2005.

[47] Rebecca Siegel, Deepa Naishadham, and Ahmedin Jemal. Cancer statistics, 2013. *CA: A Cancer Journal for Clinicians*, 63(1):11–30, January 2013.

[48] R Labianca, B Nordlinger, and G D Beretta. Primary colon cancer: ESMO Clinical Practice Guidelines for diagnosis, adjuvant treatment and follow-up. *Annals of ...*, 2010.

[49] F Berrino, R De Angelis, M Sant, S Rosso, and M B Lasota. Survival for eight major cancers and all cancers combined for European adults diagnosed in 1995–99: results of the EUROCARE-4 study. *The lancet oncology*, 2007.

[50] Linda J W Bosch, Beatriz Carvalho, Remond J A Fijneman, Connie R Jimenez, Herbert M Pinedo, Manon van Engeland, and Gerrit A Meijer. Molecular Tests for Colorectal Cancer Screening. *Clinical Colorectal Cancer*, 10(1):8–23, March 2011.

[51] Takeshi Nagasaka, Noriaki Tanaka, Harry M Cullings, Dong-Sheng Sun, Hiromi Sasamoto, Takuyuki Uchida, Minoru Koi, Naoshi Nishida, Yoshio Naomoto, C Richard Boland, Nagahide Matsubara, and Ajay Goel. Analysis of Fecal DNA Methylation to Detect Gastrointestinal Neoplasia. *Journal of the ...*, 101(18):1244–1258, September 2009.

[52] Wai K Leung, Ka-Fai To, Ellen P S Man, Michael W Y Chan, Aric J Hui, Simon S M Ng, James Y W Lau, and Joseph J Y Sung. Detection of Hypermethylated DNA or Cyclooxygenase-2 Messenger RNA in Fecal Samples of Patients With Colorectal Cancer or

Polyps. *The American Journal of Gastroenterology*, 102(5):1070–1076, May 2007.

[53] Gustaw Lech. Colorectal cancer tumour markers and biomarkers: Recent therapeutic advances. *World Journal of Gastroenterology*, 22(5):1745, 2016.

[54] *Biallelic inactivation of hMLH1 by epigenetic gene silencing, a novel mechanism causing human MSI cancers*, 1998.

[55] Z Petko, M Ghiassi, A Shuber, and J Gorham. Aberrantly methylated CDKN2A, MGMT, and MLH1 in colon polyps and in fecal DNA from patients with colorectal polyps. *Clinical Cancer ...*, 2005.

[56] James G Herman and Stephen B Baylin. Gene silencing in cancer in association with promoter hypermethylation. *The New England journal of medicine*, 349(21):2042–2054, November 2003.

[57] M Esteller, R LeVine, S B Baylin, and L H Ellenson. MLH1 promoter hypermethylation is associated with the microsatellite instability phenotype in sporadic endometrial carcinomas. *Oncogene*, 1998.

[58] Myoung Sook Kim, Juna Lee, and David Sidransky. DNA methylation markers in colorectal cancer. *Cancer and Metastasis Reviews*, 29(1):181–206, 2010.

[59] Aga Syed Sameer and Saniya Nissar. Understanding Epigenetics: an Alternative Mechanism of Colorectal Carcinogenesis. *Current Colorectal Cancer Reports*, 12(3):113–122, 2016.

[60] Wojciech Szopa, Thomas A Burley, Gabriela Kramer-Marek, and Wojciech Kaspera. Diagnostic and Therapeutic Biomarkers in Glioblastoma: Current Status and Future Perspectives. *BioMed research international*, 2017(2):8013575–13, 2017.

[61] Monika E Hegi, Annie-Claire Diserens, Sophie Godard, Pierre-Yves Dietrich, Luca Regli, Sandrine Ostermann, Philippe Otten, Guy Van Melle, Nicolas de Tribolet, and Roger Stupp. Clinical Trial Substantiates the Predictive Value of O-6-Methylguanine-DNA Methyltransferase Promoter Methylation in Glioblastoma Patients Treated

with Temozolomide. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 10(6):1871–1874, March 2004.

[62] Monika E Hegi, Annie-Claire Diserens, Thierry Gorlia, Marie-France Hamou, Nicolas de Tribolet, Michael Weller, Johan M Kros, Johannes A Hainfellner, Warren Mason, Luigi Mariani, Jacoline E C Bromberg, Peter Hau, René O Mirimanoff, J Gregory Cairncross, Robert C Janzer, and Roger Stupp. MGMT gene silencing and benefit from temozolomide in glioblastoma. *The New England journal of medicine*, 352(10):997–1003, March 2005.

[63] M Esteller, J Garcia-Foncillas, E Andion, S N Goodman, O F Hidalgo, V Vanaclocha, S B Baylin, and J G Herman. Inactivation of the DNA-repair gene MGMT and the clinical response of gliomas to alkylating agents. *The New England journal of medicine*, 343(19):1350–1354, November 2000.

[64] Guido Reifenberger, Bettina Hentschel, Jörg Felsberg, Gabriele Schackert, Matthias Simon, Oliver Schnell, Manfred Westphal, Wolfgang Wick, Torsten Pietsch, Markus Loeffler, Michael Weller, and for the German Glioma Network. Predictive impact of MGMT promoter methylation in glioblastoma of the elderly. *International Journal of Cancer*, 131(6):1342–1350, January 2012.

[65] Terri S Armstrong, Jeffrey S Wefel, Meihua Wang, Mark R Gilbert, Minhee Won, Andrew Bottomley, Tito R Mendoza, Corneel Coens, Maria Werner-Wasik, David G Brachman, Ali K Choucair, and Minesh Mehta. Net Clinical Benefit Analysis of Radiation Therapy Oncology Group 0525: A Phase III Trial Comparing Conventional Adjuvant Temozolomide With Dose-Intensive Temozolomide in Patients With Newly Diagnosed Glioblastoma. *Journal of Clinical Oncology*, 31(32):4076–4084, November 2013.

[66] Relja Popovic and Jonathan D Licht. Emerging epigenetic targets and therapies in cancer medicine. *Cancer Discovery*, 2(5):405–413, May 2012.

[67] Jacek Krol, Inga Loedige, and Witold Filipowicz. The widespread regulation of microRNA biogenesis, function and decay. *Nature reviews. Genetics*, 11(9):597–610, September 2010.

[68] Martin D Jansson and Anders H Lund. MicroRNA and cancer. *Molecular oncology*, 6(6):590–610, December 2012.

[69] Klaas Mensaert, Simon Denil, Geert Trooskens, Wim Van Criekinge, Olivier Thas, and Tim De Meyer. Next-generation technologies and data analytical approaches for epigenomics. *Environmental and molecular mutagenesis*, 55(3):155–170, April 2014.

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

[71] Michael J Booth, Miguel R Branco, Gabriella Ficz, David Oxley, Felix Krueger, Wolf Reik, and Shankar Balasubramanian. Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science (New York, N.Y.)*, 336(6083):934–937, May 2012.

[72] Miao Yu, Gary C Hon, Keith E Szulwach, Chun-Xiao Song, Liang Zhang, Audrey Kim, Xuekun Li, Qing Dai, Yin Shen, Beomseok Park, Jung-Hyun Min, Peng Jin, Bing Ren, and Chuan He. Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell*, 149(6):1368–1380, June 2012.

[73] Gerd P Pfeifer, Swati Kadam, and Seung-Gi Jin. 5-hydroxymethylcytosine and its potential roles in development and cancer. *Epigenetics & chromatin*, 6(1):10, 2013.

[74] Kevin Gaston and Mike Fried. CpG methylation has differential effects on the binding of YY1 and ETS proteins to the bi-directional promoter of the Surf-1 and Surf-2 genes. *Nucleic acids research*, 23(6):901–909, 1995.

[75] Rainer Claus, David M Lucas, Stephan Stilgenbauer, Amy S Ruppert, Lianbo Yu, Manuela Zucknick, Daniel Mertens, Andreas Bühler, Christopher C Oakes, Richard A Larson, Neil E Kay, Diane F Jelinek, Thomas J Kipps, Laura Z Rassenti, John G Gribben, Hartmut Döhner, Nyla A Heerema, Guido Marcucci, Christoph Plass,

and John C Byrd. Quantitative DNA methylation analysis identifies a single CpG dinucleotide important for ZAP-70 expression and predictive of prognosis in chronic lymphocytic leukemia. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 30(20):2483–2491, July 2012.

[76] Russell P Darst, Carolina E Pardo, Lingbao Ai, Kevin D Brown, and Michael P Kladde. Bisulfite sequencing of DNA. *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.]*, Chapter 7:Unit 7.9.1–17, July 2010.

[77] P M Warnecke, C Stirzaker, J R Melki, D S Millar, C L Paul, and S J Clark. Detection and measurement of PCR bias in quantitative methylation analysis of bisulphite-treated DNA. *Nucleic acids research*, 25(21):4422–4426, November 1997.

[78] Michael J Ziller, Hongcang Gu, Fabian Müller, Julie Donaghey, Linus T-Y Tsai, Oliver Kohlbacher, Philip L De Jager, Evan D Rosen, David A Bennett, Bradley E Bernstein, Andreas Gnirke, and Alexander Meissner. Charting a dynamic DNA methylation landscape of the human genome. *Nature*, 500(7463):477–481, August 2013.

[79] A Meissner. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic acids research*, 33(18):5868–5877, October 2005.

[80] Jie Deng, Robert Shoemaker, Bin Xie, Athurva Gore, Emily M LeProust, Jessica Antosiewicz-Bourget, Dieter Egli, Nimet Maherali, In-Hyun Park, Junying Yu, George Q Daley, Kevin Eggan, Konrad Hochedlinger, James Thomson, Wei Wang, Yuan Gao, and Kun Zhang. Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. *Nature biotechnology*, 27(4):353–360, April 2009.

[81] Eun-Joon Lee, Lirong Pei, Gyan Srivastava, Trupti Joshi, Garima Kushwaha, Jeong-Hyeon Choi, Keith D Robertson, Xinguo Wang, John K Colbourne, Lu Zhang, Gary P Schroth, Dong Xu, Kun Zhang, and Huidong Shi. Targeted bisulfite sequencing by solution hybrid selection and massively parallel sequencing. *Nucleic acids research*, 39(19):e127, October 2011.

[82] F Krueger and S R Andrews. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics (Oxford, England)*, 27(11):1571–1572, May 2011.

[83] Ryan Lister, Mattia Pelizzola, Robert H Dowen, R David Hawkins, Gary Hon, Julian Tonti-Filippini, Joseph R Nery, Leonard Lee, Zhen Ye, Que-Minh Ngo, Lee Edsall, Jessica Antosiewicz-Bourget, Ron Stewart, Victor Ruotti, A Harvey Millar, James A Thomson, Bing Ren, and Joseph R Ecker. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 462(7271):315–322, November 2009.

[84] Christoph Bock, Eleni M Tomazou, Arie B Brinkman, Fabian Müller, Femke Simmer, Hongcang Gu, Natalie Jäger, Andreas Gnirke, Hendrik G Stunnenberg, and Alexander Meissner. Quantitative comparison of genome-wide DNA methylation mapping technologies. *Nature biotechnology*, 28(10):1106–1114, October 2010.

[85] Katja Hebestreit, Martin Dugas, and Hans-Ulrich Klein. Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. *Bioinformatics (Oxford, England)*, 29(13):1647–1653, July 2013.

[86] Akalin, A, Kormaksson, M, and Li, S. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome . . .*, 13(10):R87, 2012.

[87] Kasper D Hansen, Benjamin Langmead, and Rafael A Irizarry. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biology*, 13(10):R83, 2012.

[88] E Magda Price, Allison M Cotton, Lucia L Lam, Pau Farré, Eldon Emberly, Carolyn J Brown, Wendy P Robinson, and Michael S Kobor. Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. *Epigenetics & chromatin*, 6(1):4, March 2013.

[89] Ruth Pidsley, Elena Zotenko, Timothy J Peters, Mitchell G Lawrence, Gail P Risbridger, Peter Molloy, Susan Van Djik, Beverly Muhlhausler, Clare Stirzaker, and Susan J Clark. Critical evaluation

of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biology*, 17(1):208, October 2016.

[90] Marina Bibikova, Bret Barnes, Chan Tsan, Vincent Ho, Brandy Klotzle, Jennie M Le, David Delano, Lu Zhang, Gary P Schroth, Kevin L Gunderson, Jian-Bing Fan, and Richard Shen. High density DNA methylation array with single CpG site resolution. *Genomics*, 98(4):288–295, October 2011.

[91] Sarah Dedeurwaerder, Matthieu Defrance, Emilie Calonne, Hélène Denis, Christos Sotiriou, and François Fuks. Evaluation of the Infinium Methylation 450K technology. *dx.doi.org*, 3(6):771–784, November 2011.

[92] Pan Du, Xiao Zhang, Chiang-Ching Huang, Nadereh Jafari, Warren A Kibbe, Lifang Hou, and Simon M Lin. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC bioinformatics*, 11(1):587, November 2010.

[93] Pan Du, Warren A Kibbe, and Simon M Lin. lumi: a pipeline for processing Illumina microarray. *Bioinformatics (Oxford, England)*, 24(13):1547–1548, July 2008.

[94] Virginia Goss Tusher, Robert Tibshirani, and Gilbert Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9):5116–5121, April 2001.

[95] D Wang, L Yan, Q Hu, L E Sucheston, M J Higgins, C B Ambrosone, C S Johnson, D J Smiraglia, and S Liu. IMA: an R package for high-throughput analysis of Illumina's 450K Infinium methylation data. *Bioinformatics (Oxford, England)*, 28(5):729–730, February 2012.

[96] L Z Xiong, C G Xu, M A Saghai Maroof, and Qifa Zhang. Patterns of cytosine methylation in an elite rice hybrid and its parental lines, detected by a methylation-sensitive amplification polymorphism technique. *Molecular and General Genetics MGG*, 261(3):439–446, 1999.

[97] Benjamin Schulz, R Lutz Eckstein, and Walter Durka. Scoring and analysis of methylation-sensitive amplification polymorphisms for epigenetic population studies. *Molecular ecology resources*, 13(4):642–653, July 2013.

[98] Gertrud Lund, Linda Andersson, Massimiliano Lauria, Marie Lindholm, Mario F Fraga, Ana Villar-Garea, Esteban Ballestar, Manel Esteller, and Silvio Zaina. DNA methylation polymorphisms precede any histological sign of atherosclerosis in mice lacking apolipoprotein E. *The Journal of biological chemistry*, 279(28):29147–29154, July 2004.

[99] Axel Schumacher, Philipp Kapranov, Zachary Kaminsky, James Flanagan, Abbas Assadzadeh, Patrick Yau, Carl Virtanen, Neil Winegarden, Jill Cheng, Thomas Gingeras, and Arturas Petronis. Microarray-based DNA methylation profiling: technology and applications. *Nucleic acids research*, 34(2):528–542, January 2006.

[100] Esteban Ballestar, Maria F Paz, Laura Valle, Susan Wei, Mario F Fraga, Jesus Espada, Juan Cruz Cigudosa, Tim Hui-Ming Huang, and Manel Esteller. Methyl-CpG binding proteins identify novel sites of epigenetic inactivation in human cancer. *The EMBO journal*, 22(23):6335–6345, December 2003.

[101] Arie B Brinkman, Femke Simmer, Kelong Ma, Anita Kaan, Jingde Zhu, and Hendrik G Stunnenberg. Whole-genome DNA methylation profiling using MethylCap-seq. *Methods*, 52(3):232–236, 2010.

[102] Ning Li, Mingzhi Ye, Yingrui Li, Zhixiang Yan, Lee M Butcher, Jihua Sun, Xu Han, Quan Chen, Xiuqing zhang, and Jun Wang. Whole genome DNA methylation analysis based on high throughput sequencing technology. *Methods*, 52(3):203–212, November 2010.

[103] David Serre, Byron H Lee, and Angela H Ting. MBD-isolated Genome Sequencing provides a high-throughput and comprehensive survey of DNA methylation in the human genome. *Nucleic acids research*, 38(2):391–399, January 2010.

[104] T Rauch. MIRA-Assisted Microarray Analysis, a New Technology for the Determination of DNA Methylation Patterns, Identifies Frequent Methylation of Homeodomain-Containing Genes in Lung Cancer Cells. *Cancer Research*, 66(16):7939–7947, 2006.

[105] Tibor Rauch and Gerd P Pfeifer. Methylated-CpG island recovery assay: a new technique for the rapid detection of methylated-CpG islands in cancer. *Laboratory investigation; a journal of technical methods and pathology*, 85(9):1172–1180, September 2005.

[106] S G Jin, S Kadam, and G P Pfeifer. Examination of the specificity of DNA methylation profiling techniques towards 5-methylcytosine and 5-hydroxymethylcytosine. *Nucleic acids research*, 38(11):e125–e125, June 2010.

[107] V Valinluck. Oxidative damage to methyl-CpG sequences inhibits the binding of the methyl-CpG binding domain (MBD) of methyl-CpG binding protein 2 (MeCP2). *Nucleic acids research*, 32(14):4100–4108, August 2004.

[108] Marian Mellén, Pinar Ayata, Scott Dewell, Skirmantas Kriaucionis, and Nathaniel Heintz. MeCP2 Binds to 5hmC Enriched within Active Genes and Accessible Chromatin in the Nervous System. *Cell*, 151(7):1417–1430, December 2012.

[109] Ozlem Yildirim, Ruowang Li, Jui-Hung Hung, Poshen B Chen, Xianjun Dong, Ly-Sha Ee, Zhiping Weng, Oliver J Rando, and Thomas G Fazzio. Mbd3/NURD Complex Regulates Expression of 5-Hydroxymethylcytosine Marked Genes in Embryonic Stem Cells. *Cell*, 147(7):1498–1510, December 2011.

[110] Michael Weber, Jonathan J Davies, David Wittig, Edward J Oakeley, Michael Haase, Wan L Lam, and Dirk Schübeler. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nature Genetics*, 37(8):853–862, August 2005.

[111] Ian M Wilson, Jonathan J Davies, Michael Weber, Carolyn J Brown, Carlos E Alvarez, Calum MacAulay, Dirk Schübeler, and Wan L Lam. Epigenomics: mapping the methylome. *Cell Cycle*, 5(2):155–158, January 2006.

[112] Filipe Jacinto, Esteban Ballestar, and Manel Esteller. Methyl-DNA immunoprecipitation (MeDIP): Hunting down the DNA methylome. *BioTechniques*, 44(1):35–43, January 2008.

[113] S G Jin, X Wu, A X Li, and G P Pfeifer. Genomic mapping of 5-hydroxymethylcytosine in the human brain. *Nucleic acids research*, 39(12):5015–5024, July 2011.

[114] R Alan Harris, Ting Wang, Cristian Coarfa, Raman P Nagarajan, Chibo Hong, Sara L Downey, Brett E Johnson, Shaun D Fouse, Allen Delaney, Yongjun Zhao, Adam Olshen, Tracy Ballinger, Xin Zhou, Kevin J Forsberg, Junchen Gu, Lorigail Echipare, Henriette O'Geen, Ryan Lister, Mattia Pelizzola, Yuanxin Xi, Charles B Epstein, Bradley E Bernstein, R David Hawkins, Bing Ren, Wen-Yu Chung, Hongcang Gu, Christoph Bock, Andreas Gnirke, Michael Q Zhang, David Haussler, Joseph R Ecker, Wei Li, Peggy J Farnham, Robert A Waterland, Alexander Meissner, Marco A Marra, Martin Hirst, Aleksandar Milosavljevic, and Joseph F Costello. Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nature biotechnology*, 28(10):1097–1105, October 2010.

[115] Peter J Park. ChIP–seq: advantages and challenges of a maturing technology. *Nature reviews. Genetics*, 10(10):669–680, September 2009.

[116] Tim De Meyer, Evi Mampaey, Michaël Vlemmix, Simon Denil, Geert Trooskens, Jean-Pierre Renard, Sarah De Keulenaer, Pierre Dehan, Gerben Menschaert, and Wim Van Criekinge. Quality evaluation of methyl binding domain based kits for enrichment DNA-methylation sequencing. *PloS one*, 8(3):e59068, 2013.

[117] Mark D Robinson, Clare Stirzaker, Aaron L Statham, Marcel W Coolen, Jenny Z Song, Shalima S Nair, Dario Strbenac, Terence P Speed, and Susan J Clark. Evaluation of affinity-based genome-wide DNA methylation data: effects of CpG density, amplification bias, and copy number variation. *Genome research*, 20(12):1719–1729, December 2010.

[118] Y Benjamini and T P Speed. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic acids research*, 40(10):e72–e72, May 2012.

[119] Partha M Das, Kavitha Ramachandran, Jane vanWert, and Rakesh Singal. Chromatin immunoprecipitation assay. *BioTechniques*, 37(6):961–969, December 2004.

[120] Cizhong Jiang and B Franklin Pugh. Nucleosome positioning and gene regulation: advances through genomics. *Nature reviews. Genetics*, 10(3):161–172, March 2009.

[121] Anton Valouev, Steven M Johnson, Scott D Boyd, Cheryl L Smith, Andrew Z Fire, and Arend Sidow. Determinants of nucleosome organization in primary human cells. *Nature*, 474(7352):516–520, May 2011.

[122] Dustin E Schones, Kairong Cui, Suresh Cuddapah, Tae-Young Roh, Artem Barski, Zhibin Wang, Gang Wei, and Keji Zhao. Dynamic Regulation of Nucleosome Positioning in the Human Genome. *Cell*, 132(5):887–898, March 2008.

[123] Daniel J Gaffney, Graham McVicker, Athma A Pai, Yvonne N Fondufe-Mittendorf, Noah Lewellen, Katelyn Michelini, Jonathan Widom, Yoav Gilad, and Jonathan K Pritchard. Controls of Nucleosome Positioning in the Human Genome. *PLOS Genet*, 8(11):e1003036, November 2012.

[124] T K Kelly, Y Liu, F D Lay, G Liang, B P Berman, and P A Jones. Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. *Genome research*, 22(12):2497–2506, December 2012.

[125] Joel Rozowsky, Ghia Euskirchen, Raymond K Auerbach, Zhengdong D Zhang, Theodore Gibson, Robert Bjornson, Nicholas Carriero, Michael Snyder, and Mark B Gerstein. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nature biotechnology*, 27(1):66–75, January 2009.

[126] S G Landt, G K Marinov, A Kundaje, P Kheradpour, F Pauli, S Batzoglou, B E Bernstein, P Bickel, J B Brown, P Cayting, Y Chen, G DeSalvo, C Epstein, K I Fisher-Aylor, G Euskirchen, M Gerstein, J Gertz, A J Hartemink, M M Hoffman, V R Iyer, Y L Jung, S Karmakar, M Kellis, P V Kharchenko, Q Li, T Liu, X S Liu, L Ma, A Milosavljevic, R M Myers, P J Park, M J Pazin, M D Perry, D Raha, T E Reddy, J Rozowsky, N Shoresh, A Sidow, M Slattery, J A Stamatoyannopoulos, M Y Tolstorukov, K P White, S Xi, P J Farnham, J D Lieb, B J Wold, and M Snyder. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome research*, 22(9):1813–1831, September 2012.

[127] Colin Dingwall, George P Lomonossoff, and Ronald A Laskey. High sequence specificity of micrococcal nuclease. *Nucleic acids research*, 9(12):2659–2674, 1981.

[128] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9(4):357–359, April 2012.

[129] Stephen M Rumble, Phil Lacroute, Adrian V Dalca, Marc Fiume, Arend Sidow, and Michael Brudno. SHRiMP: Accurate Mapping of Short Color-space Reads. *PLoS computational biology*, 5(5):e1000386, May 2009.

[130] Ruiqiang Li, Chang Yu, Yingrui Li, Tak-Wah Lam, Siu-Ming Yiu, Karsten Kristiansen, and Jun Wang. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics (Oxford, England)*, 25(15):1966–1967, August 2009.

[131] Cole Trapnell and Steven L Salzberg. How to map billions of short reads onto genomes. *Nature biotechnology*, 27(5):455–457, May 2009.

[132] Shirley Pepke, Barbara Wold, and Ali Mortazavi. Computation for ChIP-seq and RNA-seq studies. *Nature methods*, 6(11 Suppl):S22–32, November 2009.

[133] L Chavez, J Jozefczuk, C Grimm, J Dietrich, B Timmermann, H Lehrach, R Herwig, and J Adjaye. Computational analysis of genome-wide DNA methylation during the differentiation of human embryonic stem cells along the endodermal lineage. *Genome research*, 20(10):1441–1450, October 2010.

[134] A L Statham, D Strbenac, M W Coolen, C Stirzaker, S J Clark, and M D Robinson. Repitools: an R package for the analysis of enrichment-based epigenomic data. *Bioinformatics (Oxford, England)*, 26(13):1662–1663, June 2010.

[135] Gordon Robertson, Martin Hirst, Matthew Bainbridge, Misha Bilenky, Yongjun Zhao, Thomas Zeng, Ghia Euskirchen, Bridget Bernier, Richard Varhol, Allen Delaney, Nina Thiessen, Obi L Griffith, Ann He, Marco Marra, Michael Snyder, and Steven Jones. Genome-wide profiles of STAT1 DNA association using chromatin

immunoprecipitation and massively parallel sequencing. *Nature Methods*, 4(8):651–657, 2007.

[136] A P Boyle, J Guinney, G E Crawford, and T S Furey. F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics (Oxford, England)*, 24(21):2537–2538, October 2008.

[137] Anton Valouev, David S Johnson, Andreas Sundquist, Catherine Medina, Elizabeth Anton, Serafim Batzoglou, Richard M Myers, and Arend Sidow. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nature methods*, 5(9):829–834, August 2008.

[138] Desmond S Lun, Ashley Sherrid, Brian Weiner, David R Sherman, and James E Galagan. A blind deconvolution approach to high-resolution mapping of transcription factor binding sites from ChIP-seq data. *Genome Biology*, 10(12):R142, 2009.

[139] A P Fejes, G Robertson, M Bilenky, R Varhol, M Bainbridge, and S J M Jones. FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics (Oxford, England)*, 24(15):1729–1730, July 2008.

[140] Xun Lan, Christopher Adams, Mark Landers, Miroslav Dudas, Daniel Krissinger, George Marnellos, Russell Bonneville, Maoxiong Xu, Junbai Wang, Tim H M Huang, Gavin Meredith, and Victor X Jin. High Resolution Detection and Analysis of CpG Dinucleotides Methylation Using MBD-Seq Technology. *PloS one*, 6(7):e22226, July 2011.

[141] Anneleen Decock, Maté Ongenaert, Jasmien Hoebeeck, Katleen De Preter, Gert Van Peer, Wim Van Criekinge, Ruth Ladenstein, Johannes H Schulte, Rosa Noguera, Raymond L Stallings, An Van Damme, Geneviève Laureys, Joëlle Vermeulen, Tom Van Maerken, Frank Speleman, and Jo Vandesompele. Genome-wide promoter methylation analysis in neuroblastoma identifies prognostic methylation biomarkers. *Genome Biology*, 13(10):R95, 2012.

[142] M D Robinson, D Strbenac, C Stirzaker, A L Statham, J Song, T P Speed, and S J Clark. Copy-number-aware differential anal-

ysis of quantitative DNA sequencing data. *Genome research*, 22(12):2489–2496, December 2012.

[143] H Xu, C L Wei, F Lin, and W K Sung. An HMM approach to genome-wide identification of differential histone modification sites from ChIP-seq data. *Bioinformatics (Oxford, England)*, 24(20):2344–2349, October 2008.

[144] Thomas A Down, Vardhman K Rakyan, Daniel J Turner, Paul Flicek, Heng Li, Eugene Kulesha, Stefan Gräf, Nathan Johnson, Javier Herrero, Eleni M Tomazou, Natalie P Thorne, Liselotte Bäckdahl, Marlis Herberth, Kevin L Howe, David K Jackson, Marcos M Miretti, John C Marioni, Ewan Birney, Tim J P Hubbard, Richard Durbin, Simon Tavaré, and Stephan Beck. A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nature biotechnology*, 26(7):779–785, July 2008.

[145] Benjamin Rodriguez, Hok-Hei Tam, David Frankhouser, Michael Trimarchi, Mark Murphy, Chris Kuo, Deval Parikh, Bryan Ball, Sebastian Schwind, John Curfman, William Blum, Guido Marcucci, Pearlly Yan, and Ralf Bundschuh. A scalable, flexible workflow for MethylCap-seq data analysis. In *2011 IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS)*, pages 1–4. IEEE, 2011.

[146] Rui-Lan Huang, Fei Gu, Nameer B Kirma, Jianhua Ruan, Chun-Liang Chen, Hui-Chen Wang, Yu-Ping Liao, Cheng-Chang Chang, Mu-Hsien Yu, Jay M Pilrose, Ian M Thompson, Hsuan-Cheng Huang, Tim Hui-Ming Huang, Hung-Cheng Lai, and Kenneth P Nephew. Comprehensive methylome analysis of ovarian tumors reveals hedgehog signaling pathway regulators as prognostic DNA methylation biomarkers. *Epigenetics*, 8(6):624–634, October 2014.

[147] M D Robinson and G K Smyth. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics (Oxford, England)*, 23(21):2881–2887, November 2007.

[148] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, 26(1):139–140, January 2010.

[149] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106, 2010.

[150] O Flores and M Orozco. nucleR: a package for non-parametric nucleosome positioning. *Bioinformatics (Oxford, England)*, 27(15):2149–2150, July 2011.

[151] Peter Humburg, Chris A Helliwell, David Bulger, and Glenn Stone. ChIPseqR: analysis of ChIP-seq experiments. *BMC bioinformatics*, 12(1):39, 2011.

[152] Yong Zhang, Hyunjin Shin, Jun S Song, Ying Lei, and X Shirley Liu. Identifying Positioned Nucleosomes with Epigenetic Marks in Human from ChIP-Seq. *BMC genomics*, 9(1):1, November 2008.

[153] J Becker, C Yau, J M Hancock, and C C Holmes. NucleoFinder: a statistical approach for the detection of nucleosome positions. *Bioinformatics (Oxford, England)*, 29(6):711–716, March 2013.

[154] Sangsoon Woo, Xuekui Zhang, Renan Sauteraud, François Robert, and Raphael Gottardo. PING 2.0: an R/Bioconductor package for nucleosome positioning using next-generation sequencing data. *Bioinformatics (Oxford, England)*, 29(16):2049–2050, August 2013.

[155] Victoria A VanderNoot, Stanley A Langevin, Owen D Solberg, Pamela D Lane, Deanna J Curtis, Zachary W Bent, Kelly P Williams, Kamlesh D Patel, Joseph S Schoeniger, Steven S Branda, and Todd W Lane. cDNA normalization by hydroxyapatite chromatography to enrich transcriptome diversity in RNA-seq applications. *BioTechniques*, 53(6), December 2012.

[156] Eulàlia Martí, Lorena Pantano, Mónica Bañez-Coronel, Franc Llorens, Elena Miñones-Moyano, Sílvia Porta, Lauro Sumoy, Isidre Ferrer, and Xavier Estivill. A myriad of miRNA variants in control and Huntington's disease brain regions detected by massively parallel sequencing. *Nucleic acids research*, 38(20):7219–7235, November 2010.

[157] Lihong Weng, Xiwei Wu, Hanlin Gao, Bing Mu, Xuejun Li, Jin-Hui Wang, Chao Guo, Jennifer M Jin, Zhuo Chen, Maricela Covarrubias, Yate-Ching Yuan, Lawrence M Weiss, and Huiqing Wu. Mi-

croRNA profiling of clear cell renal cell carcinoma by whole-genome small RNA deep sequencing of paired frozen and formalin-fixed, paraffin-embedded tissue specimens. *The Journal of pathology*, 222(1):n/a–n/a, May 2010.

[158] Daehwan Kim, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L Salzberg. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, 14(4):R36, 2013.

[159] A Dobin, C A Davis, F Schlesinger, J Drenkow, C Zaleski, S Jha, P Batut, M Chaisson, and T R Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*, 29(1):15–21, December 2012.

[160] David S DeLuca, Joshua Z Levin, Andrey Sivachenko, Timothy Fennell, Marc-Danie Nazaire, Chris Williams, Michael Reich, Wendy Winckler, and Gad Getz. RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics (Oxford, England)*, 28(11):1530–1532, June 2012.

[161] C Liu. NONCODE: an integrated knowledge database of non-coding RNAs. *Nucleic acids research*, 33(Database issue):D112–D115, December 2004.

[162] Marie-Josée Cros, Antoine de Monte, Jérôme Mariette, Philippe Bardou, Benjamin Grenier-Boley, Daniel Gautheret, Hélène Touzet, and Christine Gaspin. RNAspace.org: An integrated environment for the prediction, annotation, and analysis of ncRNA. *RNA (New York, N.Y.)*, 17(11):1947–1956, November 2011.

[163] Sarah Djebali, Carrie A Davis, Angelika Merkel, Alex Dobin, Timo Lassmann, Ali Mortazavi, Andrea Tanzer, Julien Lagarde, Wei Lin, Felix Schlesinger, Chenghai Xue, Georgi K Marinov, Jainab Khatun, Brian A Williams, Chris Zaleski, Joel Rozowsky, Maik Röder, Felix Kokocinski, Rehab F Abdelhamid, Tyler Alioto, Igor Antoshechkin, Michael T Baer, Nadav S Bar, Philippe Batut, Kimberly Bell, Ian Bell, Sudipto Chakrabortty, Xian Chen, Jacqueline Chrast, Joao Curado, Thomas Derrien, Jorg Drenkow, Erica Dumais, Jacqueline Dumais, Radha Duttagupta, Emilie Falconnet, Meagan Fastuca,

Kata Fejes-Toth, Pedro Ferreira, Sylvain Foissac, Melissa J Fullwood, Hui Gao, David Gonzalez, Assaf Gordon, Harsha Gunawardena, Cedric Howald, Sonali Jha, Rory Johnson, Philipp Kapranov, Brandon King, Colin Kingswood, Oscar J Luo, Eddie Park, Kimberly Persaud, Jonathan B Preall, Paolo Ribeca, Brian Risk, Daniel Robyr, Michael Sammeth, Lorian Schaffer, Lei-Hoon See, Atif Shahab, Jorgen Skancke, Ana Maria Suzuki, Hazuki Takahashi, Hagen Tilgner, Diane Trout, Nathalie Walters, Huaien Wang, John Wrobel, Yanbao Yu, Xiaoan Ruan, Yoshihide Hayashizaki, Jennifer Harrow, Mark Gerstein, Tim Hubbard, Alexandre Reymond, Stylianos E Antonarakis, Gregory Hannon, Morgan C Giddings, Yijun Ruan, Barbara Wold, Piero Carninci, Roderic Guigó, and Thomas R Gingeras. Landscape of transcription in human cells. *Nature*, 489(7414):101–108, September 2012.

[164] Mark D Robinson and Alicia Oshlack. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3):R25, 2010.

[165] Sören Müller, Lukas Rycak, Peter Winter, Günter Kahl, Ina Koch, and Björn Rotter. omiRas: a Web server for differential expression analysis of miRNAs derived from small RNA-Seq data. *Bioinformatics (Oxford, England)*, 29(20):2651–2652, October 2013.

[166] H Q Wang, L K Tuominen, and C J Tsai. SLIM: a sliding linear model for estimating the proportion of true null hypotheses in datasets with dependence structures. *Bioinformatics (Oxford, England)*, 27(2):225–231, January 2011.

[167] *Methylation-specific PCR: a novel PCR assay for methylation status of CpG islands*, 1996.

[168] Y M Lo, I H Wong, J Zhang, M S Tein, M H Ng, and N M Hjelm. Quantitative analysis of aberrant p16 methylation using real-time quantitative methylation-specific polymerase chain reaction. *Cancer Research*, 59(16):3899–3903, August 1999.

[169] L C Li and R Dahiya. MethPrimer: designing primers for methylation PCRs. *Bioinformatics (Oxford, England)*, 18(11):1427–1431, November 2002.

[170] Ryan Lister, Ronan C O'Malley, Julian Tonti-Filippini, Brian D Gregory, Charles C Berry, A Harvey Millar, and Joseph R Ecker. Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis. *Cell*, 133(3):523–536, May 2008.

[171] R David Hawkins, Gary C Hon, and Bing Ren. Next-generation genomics: an integrative approach. *Nature reviews. Genetics*, 11(7):476–486, July 2010.

[172] Martin Krzywinski, Jacqueline Schein, Inanç Birol, Joseph Connors, Randy Gascoyne, Doug Horsman, Steven J Jones, and Marco A Marra. Circos: an information aesthetic for comparative genomics. *Genome research*, 19(9):1639–1645, September 2009.

[173] Jugal Kishore, ManishKumar Goel, and Pardeep Khanna. Understanding survival analysis: Kaplan-Meier estimate. *International Journal of Ayurveda Research*, 1(4):274, 2010.

[174] J M Bland. The logrank test. *BMJ*, 328(7447):1073–1073, May 2004.

[175] Xian Liu. The Cox Proportional Hazard Regression Model and Advances. In *Survival Analysis*, pages 144–200. John Wiley & Sons, Ltd, Chichester, UK, July 2012.

[176] Benjamin A Flusberg, Dale R Webster, Jessica H Lee, Kevin J Travers, Eric C Olivares, Tyson A Clark, Jonas Korlach, and Stephen W Turner. Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nature methods*, 7(6):461–465, May 2010.

[177] J Schreiber, Z L Wescoe, R Abu-Shumays, J T Vivian, B Baatar, K Karplus, and M Akeson. Error rates for nanopore discrimination among cytosine, methylcytosine, and hydroxymethylcytosine along individual DNA strands. *Proceedings of the National Academy of Sciences*, 110(47):18910–18915, November 2013.

[178] Eric E Schadt, Steve Turner, and Andrew Kasarskis. A window into third-generation sequencing. *Human Molecular Genetics*, 19(R2):R227–40, October 2010.

[179] Ehud Shapiro, Tamir Biezuner, and Sten Linnarsson. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature reviews. Genetics*, 14(9):618–630, July 2013.

[180] Michelle M Denomme, Liyue Zhang, and Mellissa R W Mann. Single Oocyte Bisulfite Mutagenesis. *Journal of Visualized Experiments*, 14(64), 2012.

[181] Martin Kantlehner, Roland Kirchner, Petra Hartmann, Joachim W Ellwart, Marianna Alunni-Fabbroni, and Axel Schumacher. A high-throughput DNA methylation analysis of a single cell. *Nucleic acids research*, 39(7):e44–e44, April 2011.

[182] Louise Laurent, Eleanor Wong, Guoliang Li, Tien Huynh, Aristotelis Tsirigos, Chin Thing Ong, Hwee Meng Low, Ken Wing Kin Sung, Isidore Rigoutsos, Jeanne Loring, and Chia-Lin Wei. Dynamic changes in the human methylome during differentiation. *Genome research*, 20(3):320–331, March 2010.

[183] Keith D Robertson. DNA methylation and human disease. *Nature reviews. Genetics*, 6(8):597–610, August 2005.

[184] C Jerónimo, H Usadel, R Henrique, J Oliveira, C Lopes, W G Nelson, and D Sidransky. Quantitation of GSTP1 methylation in non-neoplastic prostatic tissue and organ-confined prostate adenocarcinoma. *Journal of the National Cancer Institute*, 93(22):1747–1752, November 2001.

[185] Mark L Gonzalgo, Christian P Pavlovich, Shing M Lee, and William G Nelson. Prostate cancer detection by GSTP1 methylation analysis of postbiopsy urine specimens. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 9(7):2673–2677, July 2003.

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

[187] Péter Adorján, Jürgen Distler, Evelyne Lipscher, Fabian Model, Jürgen Müller, Cécile Pelet, Aron Braun, Andrea R Florl, David Gütig, Gabi Grabs, André Howe, Mischo Kursar, Ralf Lesche, Erik Leu, André Lewin, Sabine Maier, Volker Müller, Thomas

Otto, Christian Scholz, Wolfgang A Schulz, Hans-Helge Seifert, Ina Schwope, Heike Ziebarth, Kurt Berlin, Christian Piepenbrock, and Alexander Olek. Tumour class prediction and discovery by microarray-based DNA methylation analysis. *Nucleic acids research*, 30(5):e21, March 2002.

[188] Alexander Meissner, Tarjei S Mikkelsen, Hongcang Gu, Marius Wernig, Jacob Hanna, Andrey Sivachenko, Xiaolan Zhang, Bradley E Bernstein, Chad Nusbaum, David B Jaffe, Andreas Gnirke, Rudolf Jaenisch, and Eric S Lander. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*, 454(7205):766–770, August 2008.

[189] Peter A Jones and Gangning Liang. Rethinking how DNA methylation patterns are maintained. *Nature reviews. Genetics*, 10(11):805–811, November 2009.

[190] Susan J Clark. Action at a distance: epigenetic silencing of large chromosomal regions in carcinogenesis. *Human molecular genetics*, 16 Spec No 1:R88–95, April 2007.

[191] ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, September 2012.

[192] Varley, Katherine E, Gertz, Jason, Bowling, Kevin M, Parker, Stephanie L, Reddy, Timothy E, Pauli-Behn, Florencia, Cross, Marie K, Williams, Brian A, Stamatoyannopoulos, John A, Crawford, Gregory E, Absher, Devin M, Wold, Barbara J, and Myers, Richard M. Dynamic DNA methylation across diverse human cell lines and tissues. *Genome research*, 23(3):555–567, March 2013.

[193] Yukio Yasukochi, Osamu Maruyama, Milind C Mahajan, Carolyn Padden, Ghia M Euskirchen, Vincent Schulz, Hideki Hirakawa, Satoru Kuhara, Xing-Hua Pan, Peter E Newburger, Michael Snyder, and Sherman M Weissman. X chromosome-wide analyses of genomic DNA methylation states and gene expression in male and female neutrophils. *Proceedings of the National Academy of Sciences of the United States of America*, 107(8):3704–3709, February 2010.

[194] Andrew J Sharp, Elisavet Stathaki, Eugenia Migliavacca, Manisha Brahmachary, Stephen B Montgomery, Yann Dupre, and Stylianos E Antonarakis. DNA methylation profiles of human active and inactive X chromosomes. *Genome research*, 21(10):1592–1600, October 2011.

[195] Sanjeev Shukla, Ersen Kavak, Melissa Gregory, Masahiko Imashimizu, Bojan Shutinoski, Mikhail Kashlev, Philipp Oberdoerffer, Rickard Sandberg, and Shalini Oberdoerffer. CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature*, 479(7371):74–79, November 2011.

[196] Mattia Pelizzola, Yasuo Koga, Alexander Eckehart Urban, Michael Krauthammer, Sherman Weissman, Ruth Halaban, and Annette M Molinaro. MEDME: an experimental and analytical methodology for the estimation of DNA methylation levels based on microarray derived MeDIP-enrichment. *Genome research*, 18(10):1652–1659, October 2008.

[197] Daiya Takai and Peter A Jones. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proceedings of the National Academy of Sciences of the United States of America*, 99(6):3740–3745, March 2002.

[198] C E Shannon. Prediction and Entropy of Printed English. *Bell Labs Technical Journal*, 30(1):50–64, January 1951.

[199] T D Schneider. Information Content of Individual Genetic Sequences. *Journal of Theoretical Biology*, 189(4):427–441, December 1997.

[200] R K Saiki, D H Gelfand, S Stoffel, S T Scharf, and R Higuchi. Download Limit Exceeded. *Science (New York, N.Y.)*, 1988.

[201] Crothers, Donald M and Zimm, Bruno H. Theory of the melting transition of synthetic polynucleotides: Evaluation of the stacking free energy. *Journal of molecular biology*, 9(1):1–9, July 1964.

[202] Howard DeVoe and Ignacio Tinoco. The stability of helical polynucleotides: Base contributions. *Journal of molecular biology*, 4(6):500–517, June 1962.

[203] J SantaLucia. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proceedings of the National Academy of Sciences of the United States of America*, 95(4):1460–1465, February 1998.

[204] S Rozen and H Skaletsky. Primer3 on the WWW for general users and for biologist programmers. *Bioinformatics methods and protocols*, 1999.

[205] Linda J W Bosch, Yanxin Luo, Victoria Valinluck Lao, Petur Snaebjornsson, Geert Trooskens, Ilse Vlassenbroeck, Sandra Mongera, Weiliang Tang, Piri Welcsh, James G Herman, Miriam Koopman, Iris Nagtegaal, Cornelis J A Punt, Wim Van Criekinge, Gerrit A Meijer, Raymond J Monnat, Beatriz Carvalho, and William M Grady. WRN promoter CpG island hypermethylation does not predict more favorable outcomes for metastatic colorectal cancer patients treated with irinotecan-based therapy. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 22(18):clincanres.2703.2015–4622, January 2016.

[206] Mario Hermsen, Antoine Snijders, Marta Alonso Guervós, Simone Taenzer, Ulrike Koerner, Jan Baak, Daniel Pinkel, Donna Albertson, Paul van Diest, Gerrit Meijer, and Evelin Schrock. Centromeric chromosomal translocations show tissue-specific differences between squamous cell carcinomas and adenocarcinomas. *Oncogene*, 24(9):1571–1579, February 2005.

[207] D E Grobbee and E B van Veen. *Grobbee: Code of proper secondary use of human tissue... - Google Scholar*. Federation of Biomedical Scientific Societies (FMWV), 2003.

[208] Miriam Koopman, Ninja F Antonini, Joep Douma, Jaap Wals, Aafke H Honkoop, Frans L G Erdkamp, Robert S de Jong, Cees J Rodenburg, Gerard Vreugdenhil, Olaf J L Loosveld, Aart van Bochove, Harm A M Sinnige, Geert-Jan M Creemers, Margot E T Tesselaar, Peter H Th J Slee, Marjon J B P Werter, Linda Mol, Otilia Dalesio, and Cornelis J A Punt. Sequential versus combination chemotherapy with capecitabine, irinotecan, and oxaliplatin in advanced colorectal cancer (CAIRO): a phase III randomised controlled trial. *Lancet (London, England)*, 370(9582):135–142, July 2007.

[209] Josien C Haan, Mariette Labots, Christian Rausch, Miriam Koopman, Jolien Tol, Leonie J M Mekenkamp, Mark A van de Wiel, Danielle Israeli, Hendrik F van Essen, Nicole C T van Grieken, Quirinus J M Voorham, Linda J W Bosch, Xueping Qu, Omar Kabbarah, Henk M W Verheul, Iris D Nagtegaal, Cornelis J A Punt, Bauke Ylstra, and Gerrit A Meijer. Genomic landscape of metastatic colorectal cancer. *Nature Communications*, 5:5457, November 2014.

[210] Ruben Agrelo, Wen-Hsing Cheng, Fernando Setien, Santiago Ropero, Jesus Espada, Mario F Fraga, Michel Herranz, Maria F Paz, Montserrat Sanchez-Cespedes, Maria Jesus Artiga, David Guerrero, Antoni Castells, Cayetano von Kobbe, Vilhelm A Bohr, and Manel Esteller. Epigenetic inactivation of the premature aging Werner syndrome gene in human cancer. *Proceedings of the National Academy of Sciences of the United States of America*, 103(23):8822–8827, June 2006.

[211] Shuji Ogino, Jeffrey A Meyerhardt, Takako Kawasaki, Jeffrey W Clark, David P Ryan, Matthew H Kulke, Peter C Enzinger, Brian M Wolpin, Massimo Loda, and Charles S Fuchs. CpG island methylation, response to combination chemotherapy, and patient survival in advanced microsatellite stable colorectal carcinoma. *Virchows Archiv : an international journal of pathology*, 450(5):529–537, May 2007.

[212] The Cancer Genome Atlas Research Network, John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, 45(10):1113–1120, October 2013.

[213] Ethan Cerami, Jianjiong Gao, Ugur Dogrusoz, Benjamin E Gross, Selcuk Onur Sumer, Bülent Arman Aksoy, Anders Jacobsen, Caitlin J Byrne, Michael L Heuer, Erik Larsson, Yevgeniy Antipin, Boris Reva, Arthur P Goldberg, Chris Sander, and Nikolaus Schultz. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discovery*, 2(5):401–404, May 2012.

[214] Leonie J M Mekenkamp, Karin J Heesterbeek, Miriam Koopman, Jolien Tol, Steven Teerenstra, Sabine Venderbosch, Cornelis J A

Punt, and Iris D Nagtegaal. Mucinous adenocarcinomas: poor prognosis in metastatic colorectal cancer. *European journal of cancer (Oxford, England : 1990)*, 48(4):501–509, March 2012.

[215] Sabine Venderbosch, Iris D Nagtegaal, Tim S Maughan, Christopher G Smith, Jeremy P Cheadle, David Fisher, Richard Kaplan, Philip Quirke, Matthew T Seymour, Susan D Richman, Gerrit A Meijer, Bauke Ylstra, Danielle A M Heideman, Anton F J de Haan, Cornelis J A Punt, and Miriam Koopman. Mismatch repair status and BRAF mutation status in metastatic colorectal cancer patients: a pooled analysis of the CAIRO, CAIRO2, COIN, and FOCUS studies. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 20(20):5322–5330, October 2014.

[216] Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407):330–337, July 2012.

[217] Miriam Koopman, Sabine Venderbosch, Iris D Nagtegaal, Johan H van Krieken, and Cornelis J Punt. A review on the use of molecular markers of cytotoxic therapy for colorectal cancer, what have we learned? *European Journal of Cancer*, 45(11):1935–1949, July 2009.

[218] Anna Hitrik, Ghada Abboud-Jarrous, Natalie Orlovetskie, Raphael Serruya, and Nayef Jarrous. Targeted inhibition of WRN helicase by external guide sequence and RNase P RNA. *Biochimica et biophysica acta*, 1859(4):572–580, April 2016.

[219] Y Yamabe, A Shimamoto, M Goto, J Yokota, M Sugawara, and Y Furuichi. Sp1-mediated transcription of the Werner helicase gene is modulated by Rb and p53. *Molecular and Cellular Biology*, 18(11):6191–6200, November 1998.

[220] Carla Grandori, Kou-Juey Wu, Paula Fernandez, Celine Ngouenet, Jonathan Grim, Bruce E Clurman, Michael J Moser, Junko Oshima, David W Russell, Karen Swisshelm, Scott Frank, Bruno Amati, Riccardo Dalla-Favera, and Raymond J Monnat. Werner syndrome protein limits MYC-induced cellular senescence. *Genes & development*, 17(13):1569–1574, July 2003.

[221] T Kawabe, N Tsuyama, S Kitao, K Nishikawa, A Shimamoto, M Shiratori, T Matsumoto, K Anno, T Sato, Y Mitsui, M Seki, T Enomoto, M Goto, N A Ellis, T Ide, Y Furuichi, and M Sugimoto. Differential regulation of human RecQ family helicases in cell transformation and cell cycle. *Oncogene*, 19(41):4764–4772, September 2000.

[222] Takako Kawasaki, Mutsuko Ohnishi, Yuko Suemoto, Gregory J Kirkner, Zhiqian Liu, Hiroyuki Yamamoto, Massimo Loda, Charles S Fuchs, and Shuji Ogino. WRN promoter methylation possibly connects mucinous differentiation, microsatellite instability and CpG island methylator phenotype in colorectal cancer. *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc*, 21(2):150–158, February 2008.

[223] Shuji Ogino, Katsuhiko Nosho, Gregory J Kirkner, Takako Kawasaki, Jeffrey A Meyerhardt, Massimo Loda, Edward L Giovannucci, and Charles S Fuchs. CpG island methylator phenotype, microsatellite instability, BRAF mutation and clinical outcome in colon cancer. *Gut*, 58(1):90–96, January 2009.

[224] N Knijn, L J M Mekenkamp, M Klomp, M E Vink-Börger, J Tol, S Teerenstra, J W R Meijer, M Tebar, S Riemersma, J H J M van Krieken, C J A Punt, and I D Nagtegaal. KRAS mutation analysis: a comparison between primary tumours and matched liver metastases in 305 colorectal cancer patients. *British journal of cancer*, 104(6):1020–1026, March 2011.

[225] Leonie J M Mekenkamp, Josien C Haan, Danielle Israeli, Hendrik F B van Essen, Jeroen R Dijkstra, Patricia van Cleef, Cornelis J A Punt, Gerrit A Meijer, Iris D Nagtegaal, and Bauke Ylstra. Chromosomal copy number aberrations in colorectal metastases resemble their primary counterparts and differences are typically non-recurrent. *PloS one*, 9(2):e86833, 2014.

[226] Sarah Derks, Cindy Postma, Peter T M Moerkerk, Sandra M van den Bosch, Beatriz Carvalho, Mario A J A Hermsen, Walter Giaretti, James G Herman, Matty P Weijenberg, Adriaan P de Bruïne, Gerrit A Meijer, and Manon van Engeland. Promoter methylation precedes chromosomal alterations in colorectal cancer

development. *Cellular oncology : the official journal of the International Society for Cellular Oncology*, 28(5-6):247–257, 2006.

[227] Hojoon Lee, Patrick Flaherty, and Hanlee P Ji. Systematic genomic identification of colorectal cancer genes delineating advanced from early clinical stage and metastasis. *BMC medical genomics*, 6(1):54, 2013.

[228] Linda J W Bosch, Geert Trooskens, Petur Snaebjornsson, Veerle M H Coupé, Sandra Mongera, Josien C Haan, Susan D Richman, Miriam Koopman, Jolien Tol, Tim De Meyer, Joost Louwagie, Luc Dehaspe, Nicole C T van Grieken, Bauke Ylstra, Henk M W Verheul, Manon van Engeland, Iris D Nagtegaal, James G Herman, Philip Quirke, Matthew T Seymour, Cornelis J A Punt, Wim Van Criekinge, Beatriz Carvalho, and Gerrit A Meijer. Decoy receptor 1 (DCR1) promoter hypermethylation and response to irinotecan in metastatic colorectal cancer. *Oncotarget*, 8(38):63140–63154, September 2017.

[229] Astrid Lièvre, Jean-Baptiste Bachet, Valérie Boige, Anne Cayre, Delphine Le Corre, Emmanuel Buc, Marc Ychou, Olivier Bouché, Bruno Landi, Christophe Louvet, Thierry André, Fréderic Bibeau, Marie-Danièle Diebold, Philippe Rougier, Michel Ducreux, Gorana Tomasic, Jean-François Emile, Frédérique Penault-Llorca, and Pierre Laurent-Puig. KRAS mutations as an independent prognostic factor in patients with advanced colorectal cancer treated with cetuximab. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 26(3):374–379, January 2008.

[230] Miriam Koopman and Cornelis J A Punt. Chemotherapy, which drugs and when. *European journal of cancer (Oxford, England : 1990)*, 45 Suppl 1:50–56, September 2009.

[231] Jolien Tol and Cornelis J A Punt. Monoclonal antibodies in the treatment of metastatic colorectal cancer: A review. *Clinical therapeutics*, 32(3):437–453, March 2010.

[232] Rafael G Amado, Michael Wolf, Marc Peeters, Eric Van Cutsem, Salvatore Siena, Daniel J Freeman, Todd Juan, Robert Sikorski, Sid Suggs, Robert Radinsky, Scott D Patterson, and David D Chang.

Wild-type KRAS is required for panitumumab efficacy in patients with metastatic colorectal cancer. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 26(10):1626–1634, April 2008.

[233] Jolien Tol, Miriam Koopman, Annemieke Cats, Cees J Rodenburg, Geert J M Creemers, Jolanda G Schrama, Frans L G Erdkamp, Allert H Vos, Cees J van Groeningen, Harm A M Sinnige, Dirk J Richel, Emile E Voest, Jeroen R Dijkstra, Marianne E Vink-Börger, Ninja F Antonini, Linda Mol, Johan H J M van Krieken, Otilia Dalesio, and Cornelis J A Punt. Chemotherapy, Bevacizumab, and Cetuximab in Metastatic Colorectal Cancer. *The New England journal of medicine*, 360(6):563–572, February 2009.

[234] Takeshi Nagasaka, Gerald B Sharp, Kenji Notohara, Takeshi Kambara, Hiromi Sasamoto, Hiroshi Isozaki, Donald G MacPhee, Jeremy R Jass, Noriaki Tanaka, and Nagahide Matsubara. Hypermethylation of O6-Methylguanine-DNA Methyltransferase Promoter May Predict Nonrecurrence after Chemotherapy in Colorectal Cancer Cases. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 9(14):5306–5312, November 2003.

[235] S P Chen, S C Chiu, C C Wu, and S Z Lin. The association of methylation in the promoter of APC and MGMT and the prognosis of Taiwanese CRC patients. *Genetic testing and . . .*, 2009.

[236] K Yacqub-Usman, A Richardson, and C V Duong. The pituitary tumour epigenome: aberrations and prospects for targeted therapy. *Nature Reviews . . .*, 2012.

[237] M T Seymour, T S Maughan, J A Ledermann, and C Topham. Different strategies of sequential and combination chemotherapy for patients with poor prognosis advanced colorectal cancer (MRC FOCUS): a randomised controlled trial. *The Lancet*, 2007.

[238] Michael S Braun, Susan D Richman, Philip Quirke, Catherine Daly, Julian W Adlard, Faye Elliott, Jennifer H Barrett, Peter Selby, Angela M Meade, Richard J Stephens, Mahesh K B Parmar, and Matthew T Seymour. Predictive Biomarkers of Chemotherapy Efficacy in Colorectal Cancer: Results From the UK MRC FOCUS Trial. *Journal of Clinical Oncology*, 26(16):2690–2698, June 2008.

[239] Tineke E Buffart, Marianne Tijssen, Thijs Krugers, Beatriz Carvalho, Serge J Smeets, Ruud H Brakenhoff, Heike Grabsch, Gerrit A Meijer, Henry B Sadowski, and Bauke Ylstra. DNA Quality Assessment for Array CGH by Isothermal Whole Genome Amplification. *Cellular oncology : the official journal of the International Society for Cellular Oncology*, 29(4):351–359, 2007.

[240] M M van Noesel, S van Bezouw, and G S Salomons. Tumor-specific down-regulation of the tumor necrosis factor-related apoptosis-inducing ligand decoy receptors DcR1 and DcR2 is associated with dense promoter .... *Cancer research*, 2002.

[241] R C Team. R: A language and environment for statistical computing. 2013.

[242] TERRY M THERNEAU, PATRICIA M GRAMBSCH, and THOMAS R FLEMING. Martingale-based residuals for survival models. *Biometrika*, 77(1):147–160, March 1990.

[243] E Drucker and K Krapfenbauer. Pitfalls and limitations in translation from biomarker discovery to clinical utility in predictive and personalised medicine. *EPMA journal*, 2013.

[244] John P A Ioannidis. Biomarker failures. *Clinical chemistry*, 59(1):202–204, January 2013.

[245] R M Simon, S Paik, and D F Hayes. Use of Archived Specimens in Evaluation of Prognostic and Predictive Biomarkers. *Journal of the National ...*, 2009.

[246] F Medeiros, C T Rigl, and G G Anderson. Tissue handling for genome-wide expression analysis: a review of the issues, evidence, and opportunities. *... of pathology & ...*, 2007.

[247] J Xuan, Y Yu, T Qing, L Guo, and L Shi. Next-generation sequencing in the clinic: promises and challenges. *Cancer letters*, 2013.

[248] R Siegel, J M Ma, and Z H Zou. *ct al. Cancer Statistics, 2014*. Ca Cancer J Clin, 2014.

[249] Hiroko Ohgaki, Pierre Dessen, Benjamin Jourde, Sonja Horstmann, Tomofumi Nishikawa, Pier-Luigi Di Patre, Christoph Burkhard,

Danielle Schüler, Nicole M Probst-Hensch, Paulo César Maiorka, Nathalie Baeza, Paola Pisani, Yasuhiro Yonekawa, M Gazi Yasargil, Urs M Lütolf, and Paul Kleihues. Genetic pathways to glioblastoma: a population-based study. *Cancer Research*, 64(19):6892–6899, October 2004.

[250] R D Wood, M Mitchell, J Sgouros, and T Lindahl. Human DNA repair genes. *Science (New York, N.Y.)*, 2001.

[251] Andreana L Rivera, Christopher E Pelloski, Mark R Gilbert, Howard Colman, Clarissa De La Cruz, Erik P Sulman, B Nebiyou Bekele, and Kenneth D Aldape. MGMT promoter methylation is predictive of response to radiotherapy and prognostic in the absence of adjuvant alkylating chemotherapy for glioblastoma. *Neuro-Oncology*, 12(2):nop020–121, December 2009.

[252] M Esteller and J G Herman. Generating mutations but providing chemosensitivity: the role of O6-methylguanine DNA methyltransferase in human cancer. *Oncogene*, 2004.

[253] Stanton L Gerson. MGMT: its role in cancer aetiology and cancer therapeutics. *Nature reviews. Cancer*, 4(4):296–307, April 2004.

[254] Manel Esteller, Montserrat Sanchez-Cespedes, Rafael Rosell, David Sidransky, Stephen B Baylin, and James G Herman. Detection of Aberrant Promoter Hypermethylation of Tumor Suppressor Genes in Serum DNA from Non-Small Cell Lung Cancer Patients. *Cancer Research*, 59(1):67–70, January 1999.

[255] Mohammadreza Farzanehfar, Hasan Vossoughinia, Raheleh Jabini, Alireza Tavassoli, Hasan Saadatnia, Ahmad Khosravi Khorashad, Mitra Ahadi, Monavvar Afzalaghaee, Ehsan Ghayoor Karimiani, Farzaneh Mirzaei, and Hossein Ayatollahi. Evaluation of Methylation of MGMT (O6-Methylguanine-DNA Methyltransferase) Gene Promoter in Sporadic Colorectal Cancer. *www.liebertpub.com*, 32(7):371–377, July 2013.

[256] Manel Esteller, Jesus Garcia-Foncillas, Esther Andion, Steven N Goodman, Oscar F Hidalgo, Vicente Vanaclocha, Stephen B Baylin, and James G Herman. Inactivation of the DNA-Repair Gene MGMT and the Clinical Response of Gliomas to Alkylating Agents. *dx.doi.org*, 343(19):1350–1354, August 2009.

[257] Roger Stupp, Monika E Hegi, Warren P Mason, Martin J van den Bent, Martin JB Taphoorn, Robert C Janzer, Samuel K Ludwin, Anouk Allgeier, Barbara Fisher, Karl Belanger, Peter Hau, Alba A Brandes, Johanna Gijtenbeek, Christine Marosi, Charles J Vecht, Karima Mokhtari, Pieter Wesseling, Salvador Villa, Elizabeth Eisenhauer, Thierry Gorlia, Michael Weller, Denis Lacombe, J Gregory Cairncross, and René-Olivier Mirimanoff. Effects of radiotherapy with concomitant and adjuvant temozolomide versus radiotherapy alone on survival in glioblastoma in a randomised phase III study: 5-year analysis of the EORTC-NCIC trial. *The lancet oncology*, 10(5):459–466, May 2009.

[258] Paolo Tini, Pierpaolo Pastina, Valerio Nardone, Lucio Sebaste, Marzia Toscano, Clelia Miracco, Alfonso Cerase, and Luigi Pirtoli. The combined EGFR protein expression analysis refines the prognostic value of the MGMT promoter methylation status in glioblastoma. *Clinical neurology and neurosurgery*, 149:15–21, October 2016.

[259] Cameron W Brennan, Roel G W Verhaak, Aaron McKenna, Benito Campos, Houtan Noushmehr, Sofie R Salama, Siyuan Zheng, Debyani Chakravarty, J Zachary Sanborn, Samuel H Berman, Rameen Beroukhim, Brady Bernard, Chang-Jiun Wu, Giannicola Genovese, Ilya Shmulevich, Jill Barnholtz-Sloan, Lihua Zou, Rahulsimham Vegesna, Sachet A Shukla, Giovanni Ciriello, W K Yung, Wei Zhang, Carrie Sougnez, Tom Mikkelsen, Kenneth Aldape, Darell D Bigner, Erwin G Van Meir, Michael Prados, Andrew Sloan, Keith L Black, Jennifer Eschbacher, Gaetano Finocchiaro, William Friedman, David W Andrews, Abhijit Guha, Mary Iacocca, Brian P O'Neill, Greg Foltz, Jerome Myers, Daniel J Weisenberger, Robert Penny, Raju Kucherlapati, Charles M Perou, D Neil Hayes, Richard Gibbs, Marco Marra, Gordon B Mills, Eric Lander, Paul Spellman, Richard Wilson, Chris Sander, John Weinstein, Matthew Meyerson, Stacey Gabriel, Peter W Laird, David Haussler, Gad Getz, and Lynda Chin. The Somatic Genomic Landscape of Glioblastoma. *Cell*, 155(2):462–477, October 2013.

[260] Véronique Quillien, Audrey Lavenu, Lucie Karayan-Tapon, Catherine Carpentier, Marianne Labussière, Thierry Lesimple, Olivier Chinot, Michel Wager, Jérome Honnorat, Stephan Saikali, Frédéric

Fina, Marc Sanson, and Dominique Figarella-Branger. Comparative assessment of 5 methods (methylation-specific polymerase chain reaction, methylight, pyrosequencing, methylation-sensitive high-resolution melting, and immunohistochemistry) to analyze O6-methylguanine-DNA-methyltranferase in a series of 100 g. *Cancer*, 118(17):4201–4211, January 2012.

[261] Annette Bentsen Håvik, Petter Brandal, Hilde Honne, Hanne-Sofie Spenning Dahlback, David Scheie, Merete Hektoen, Torstein Ragnar Meling, Eirik Helseth, Sverre Heim, Ragnhild A Lothe, and Guro Elisabeth Lind. MGMT promoter methylation in gliomas-assessment by pyrosequencing and quantitative methylation-specific PCR. *Journal of translational medicine*, 10(1):36, March 2012.

[262] Manabu Kanemoto, Mitsuaki Shirahata, Akiyo Nakauma, Katsumi Nakanishi, Kazuya Taniguchi, Yoji Kukita, Yoshiki Arakawa, Susumu Miyamoto, and Kikuya Kato. Prognostic prediction of glioblastoma by quantitative assessment of the methylation status of the entire MGMT promoter region. *BMC cancer*, 14(1):641, August 2014.

[263] Paul Muir, Shantao Li, Shaoke Lou, Daifeng Wang, Daniel J Spakowicz, Leonidas Salichos, Jing Zhang, George M Weinstock, Farren Isaacs, Joel Rozowsky, and Mark Gerstein. The real cost of sequencing: scaling computation to keep pace with data generation. *Genome Biology*, 17(1):4731, March 2016.

[264] Ilse Vlassenbroeck, Stéphane Califice, Annie-Claire Diserens, Eugenia Migliavacca, Josef Straub, Ivano Di Stefano, Fabrice Moreau, Marie-France Hamou, Isabelle Renard, Mauro Delorenzi, Bruno Flamion, James DiGuiseppi, Katja Bierau, and Monika E Hegi. Validation of real-time methylation-specific PCR to determine O6-methylguanine-DNA methyltransferase gene promoter methylation in glioma. *The Journal of molecular diagnostics : JMD*, 10(4):332–337, July 2008.

[265] Fabien Subtil and Muriel Rabilloud. Estimating the optimal threshold for a diagnostic biomarker in case of complex biomarker distributions. *BMC Medical Informatics and Decision Making*, 14(1):53, June 2014.

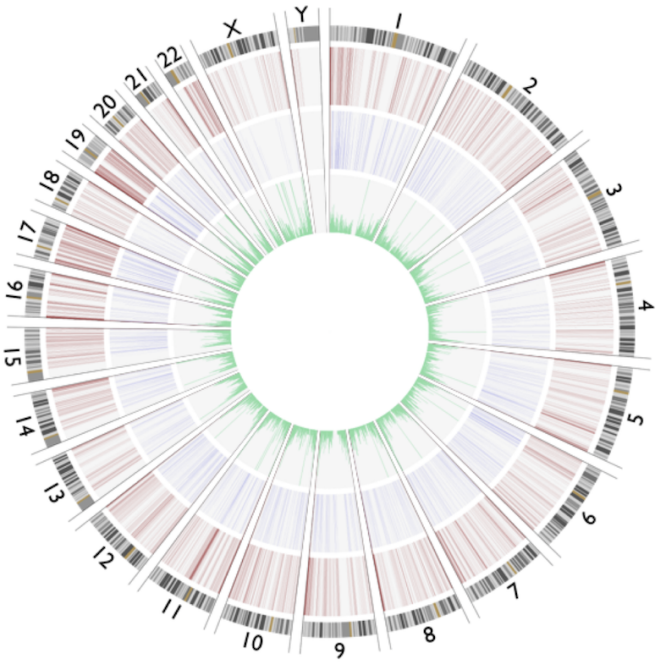
[266] Leander Van Neste, Rianne J Hendriks, Siebren Dijkstra, Geert Trooskens, Erik B Cornel, Sander A Jannink, Hans de Jong, Daphne Hessels, Frank P Smit, Willem J G Melchers, Gisèle H J M Leyten, Theo M de Reijke, Henk Vergunst, Paul Kil, Ben C Knipscheer, Christina A Hulsbergen-van de Kaa, Peter F A Mulders, Inge M van Oort, Wim Van Criekinge, and Jack A Schalken. Detection of High-grade Prostate Cancer Using a Urinary Molecular Biomarker–Based Risk Score. *European urology*, 70(5):740–748, November 2016.

[267] Gillian Z Heller. Survival Analysis: Techniques For Truncated And Censored Data, 2nd Edition. John P. Klein and Melvin L. Moeschberger, Springer-Verlag, New York, 2003. Hardcover. No. of pages: xv +536. ISBN 0-387-95399-X. *Statistics in Medicine*, 23(6):1020–1021, 2004.

[268] Patrick J Heagerty and Yingye Zheng. Survival Model Predictive Accuracy and ROC Curves. *Biometrics*, 61(1):92–105, March 2005.

[269] Philipp Euskirchen, Franck Bielle, Karim Labreche, Wigard P Kloosterman, Shai Rosenberg, Mailys Daniau, Charlotte Schmitt, Julien Masliah-Planchon, Franck Bourdeaut, Caroline Dehais, Yannick Marie, Jean-Yves Delattre, and Ahmed Idbaih. Same-day genomic and epigenomic diagnosis of brain tumors using real-time nanopore sequencing. *Acta . . .*, 134(5):691–703, November 2017.

Curriculum vitae



After graduating as a Bio-Engineer (Master Biotechnology) in 2005, I started my career as a bioinformatician at MDXHealth, a molecular diagnostic company, setting up a pipeline to discover epigenetic markers for early cancer diagnosis. After four years, I returned to the university to obtain at the lab of Bioinformatics and Computational Biology at Ghent University. My PhD revolves around epigenetic profiling and its application in clinical diagnostics. I am also co-founder and lead software developer of Wobblebase Inc., a company located in the bay area that creates personal genomic apps like Genewall (screenshot) and myWobble. I am particularly interested in data visualisation, development and production of bio-analytical software and apps for researchers, specialists and patients.

Personal Data

- **Address** Ooievaarstraat 105, 9000 Ghent - Belgium
- **Phone** +32475712791
- **Email** geert.trooskens@gmail.com

Education

High School Degree

Science - Mathematics
Sint-Pieters College. Jette-Brussels. 1992 - 1999

University Degree

Master of Science and Engineering in Cell and Gene Biotechnology
Ghent University. Ghent. 1999 - 2005

Working Experience

Oncomethylome S.A. , Liege Belgium

Molecular Diagnostics Company

BioInformatician August 2005 - April 2009

- High throughput DNA-Methylation specific PCR analysis focused on early diagnosis of cancer.

Ghent University, Ghent

PhD student / Academic Assistant May 2009 - Present

- Development of a next generation sequencing and visualisation analysis platform.
- Thesis Epigenomic profiling in Cancer
- Teaching practical courses of Bioinformatics I and Bioinformatics II.

MDXHealth, Irvine, CA

Molecular Diagnostics Company

Head of BioInformatics 2015 - Present

- Marker Discovery

- Bioinformatics Software Development for R&D and Clinical Applications.

Wobblebase Inc, Half Moon Bay, CA

Personal Genomics Company

Founder and Lead Software developer 2012 - Present

- Software architect and lead developer of the mobile genome browser Genewall and the personal genomics suite myWobble for iOS

Skills

Languages

- **Dutch** Mother Tongue
- **English** Advanced
- **French** Advanced

Computer Skills

- **Languages** Java, Javascript, HTML5, AngularJS, ActionScript, C, C#, C++, Objective-C, Perl, Python
- **Databases** Mysql, PostgreSQL, SQLite, Oracle
- **Platforms** MacOS, iOS, Windows, Linux, Android

Bioinformatics

- High-throughput Data Analysis
- Next Generation Sequencing Analysis
- Array-based SNP, methylation and expression analysis
- DNA/RNA Thermodynamics
- Variant discovery and genotyping

- DNA Methylation analysis
- Phylogenetics Analysis

Molecular Biology

- (Epi-)Genomics
- DNA Methylation
- Enrichment/Exome/Whole Genome Sequencing
- RNA-seq
- Pharmacogenomics
- Metagenomics
- Molecular phylogenetics and Evolutionary biology

Publications

Peer-Reviewed Publications

Geert Trooskens, David De Beule, Frederik Decouttere and Wim Van Criekinge. Phylogenetic trees: visualizing, customizing and detecting incongruence. Bioinformatics, 21(19):3801-3802, 2005

Dieter Schouppe, Bart Ghesquire, Gerben Menschaert, Winnok H De Vos, Stphane Bourque, **Geert Trooskens**, Paul Proost, Kris Gevaert and Els JM Van Damme. Interaction of the tobacco lectin with histone proteins. Plant physiology, 155(3):1091-1102, 2011

Tina Kyndt, Simon Denil, Annelies Haegeman, **Geert Trooskens**, Tim De Meyer, Wim Van Criekinge and Godelieve Gheysen. Transcriptome analysis of rice mature root tissue and root tips in early development by massive parallel sequencing. Journal of Experimental botany, ():err435, 2012

Mohamad Hamshou, Els JM Van Damme, Gianni Vandenborre, Bart Ghesquire, **Geert Trooskens**, Kris Gevaert and Guy Smagghe. GalNAc/Gal-binding Rhizoctonia solani agglutinin has antiproliferative activity in Drosophila melanogaster S2 cells via MAPK and JAK/STAT signaling. PloS one, 7(4):e33680, 2012

Pierre Dehan, C Canon, **Geert Trooskens**, M Rehli, Carine Munaut, Wim Van Criekinge and Philippe Delvenne. Expression of type 2 orexin receptor in human endometrium and its epigenetic silencing in endometrial cancer. The Journal of Clinical Endocrinology & Metabolism, 98(4):1549-1557, 2013

Tim De Meyer, Evi Mampaey, Michal Vlemmix, Simon Denil, **Geert Trooskens**, Jean-Pierre Renard, Sarah De Keulenaer, Pierre Dehan, Gerben Menschaert and Wim Van Criekinge. Quality evaluation of methyl binding domain based kits for enrichment DNA-methylation sequencing. PloS one, 8(3):e59068, 2013

Renske DM Steenbergen, Mat Ongenaert, Suzanne Snellenberg, **Geert Trooskens**, Wendy F van der Meide, Deeksha Pandey, Noga Bloush-

tainQimron, Kornelia Polyak, Chris JLM Meijer and Peter JF Snijders. Methylationspecific digital karyotyping of HPV16E6E7expressing human keratinocytes identifies novel methylation events in cervical carcinogenesis. The Journal of pathology, 231(1):53-62, 2013

Hongli Ji, Godelieve Gheysen, Simon Denil, Keith Lindsey, Jennifer F Topping, Kamrun Nahar, Annelies Haegeman, Winnok H De Vos, **Geert Trooskens** and Wim Van Criekinge. Transcriptional analysis through RNA sequencing of giant cells induced by Meloidogyne graminicola in rice roots. Journal of experimental botany, 64(12):3885-3898, 2013

Jeroen Crapp, Wim Van Criekinge, **Geert Trooskens**, Eisuke Hayakawa, Walter Luyten, Geert Baggerman and Gerben Menschaert. Combining in silico prediction and ribosome profiling in a genome-wide search for novel putatively coding sORFs. BMC genomics, 14(1):648, 2013

Claudina A Prez-Novo, Yuan Zhang, Simon Denil, **Geert Trooskens**, Tim De Meyer, Wim Van Criekinge, Paul Van Cauwenberge, Luo Zhang and Claus Bachert. Staphylococcal enterotoxin B influences the DNA methylation pattern in nasal polyp tissue: a preliminary study. Allergy, Asthma & Clinical Immunology, 9(1):48, 2013

Sandra Steyaert, Wim Van Criekinge, Ayla De Paepe, Simon Denil, Klaas Mensaert, Katrien Vandepitte, Wim Vanden Berghe, **Geert Trooskens** and Tim De Meyer. SNP-guided identification of monoallelic DNA-methylation events from enrichment-based sequencing data. Nucleic acids research, 42(20):e157-e157, 2014

Malav S Trivedi, Jayni S Shah, Sara Al-Mughairy, Nathaniel W Hodgson, Benjamin Simms, **Geert Trooskens**, Wim Van Criekinge and Richard C Deth. Food-derived opioid peptides inhibit cysteine uptake with redox and epigenetic consequences. The Journal of nutritional biochemistry, 25(10):1011-1018, 2014

Tom Sante, Sarah Vergult, Pieter-Jan Volders, Wigard P Kloosterman, **Geert Trooskens**, Katleen De Preter, Annelies Dheedene, Frank Speleman, Tim De Meyer and Bjrn Menten. ViVar: a comprehensive platform for the analysis and visualization of structural genomic variation. PloS one, 9(12):e113800, 2014

Tim De Meyer, Pierre Bady, **Geert Trooskens**, Sebastian Kurscheid, Jocelyne Bloch, Johan M Kros, Johannes A Hainfellner, Roger Stupp, Mauro Delorenzi and Monika E Hegi. Genome-wide DNA methylation detection by MethylCap-seq and Infinium HumanMethylation450 BeadChips: an independent large-scale comparison.. Scientific reports, 5():15375-15375, 2014

Malav S Trivedi, Nathaniel W Hodgson, Stephen J Walker, **Geert Trooskens**, Vineeth Nair and Richard C Deth. Epigenetic effects of casein-derived opioid peptides in SH-SY5Y human neuroblastoma cells. Nutrition & metabolism, 12(1):54, 2015

Linda JW Bosch, Yanxin Luo, Victoria V Lao, Petur Snaebjornsson, **Geert Trooskens**, Ilse Vlassenbroeck, Sandra Mongera, Weiliang Tang, Piri Welcsh and James G Herman. WRN Promoter CpG Island Hypermethylation Does Not Predict More Favorable Outcomes for Patients with Metastatic Colorectal Cancer Treated with Irinotecan-Based Therapy. Clinical Cancer Research, 22(18):4612-4622, 2016

Grant D Stewart, Thomas Powles, Christophe Van Neste, Alison Meynert, Fiach O'Mahony, Alexander Laird, Dieter Deforce, Filip Van Nieuwerburgh, **Geert Trooskens** and Wim Van Criekinge. Dynamic epigenetic changes to VHL occur with sunitinib in metastatic clear cell renal cancer. Oncotarget, 7(18):25241, 2016

Leander Van Neste, Rianne J Hendriks, Siebren Dijkstra, **Geert Trooskens**, E.B. Cornel, Sander A Jannink, Hans de Jong, Daphne Hessels, Frank P Smit and Willem JG Melchers. Detection of High-grade Prostate Cancer Using a Urinary Molecular BiomarkerBased Risk Score. European urology, 70(5):740-748, 2016

Wina Verlaat, Peter JF Snijders, Putri W Novianti, Saskia M Wilting, Lise MA De Strooper, **Geert Trooskens**, Johan Vandersmissen, Wim Van Criekinge, G Bea A Wisman and Chris JLM Meijer. Genome-wide DNA methylation profiling reveals methylation markers associated with 3q gain for detection of cervical pre-cancer and cancer. Clinical Cancer Research, clincanres. 2641.2016, 2017

Linda Bosch, **Geert Trooskens**, Petur Snaebjornsson, Veerle Coup, Sandra Mongera, Josien Haan, Susan D Richman, Miriam Koopman, Jolien Tol, Tim De Meyer, Joost Louwagie, Luc DeHaspe, Nicole van Grieken, Bauke Ylstra, Henk Verheul, Manon van Engeland, James Herman, Iris Nagtegaal, Philip Quirke, Matthew Seymour, Cornelis Punt, Wim van Crieking, Beatriz Carvalho, and Gerrit Meijer. Decoy receptor 1 (DCR1) promoter hypermethylation and response to irinotecan in metastatic colorectal cancer. Oncotarget. 2017 Accepted with Minor revisions

Geert Trooskens, Annika Malmstrom, Martin Hallbeck, Peter Soderkvist, Greg Jones, Johan Vandersmissen, Hendrik Vandevoorde, Leander Van Neste and Wim Van Criekinge. MGMT Epigenetic Sequencing Assay Predicts Overall Survival in Glioblastoma Patients Receiving Temozolomide. Scientific Reports. Submitted 2017

Conference Proceedings

Wim Van Criekinge, Linda JW Bosch, **Geert Trooskens**, Peter Snaebjornsson, Josien Haan, Lorraine Cheryl Pelosof, Miriam Koopman, Jolien Tol, Joost Louwagie and Luc Dehaspe. Association of DNA promoter hypermethylation of decoy receptor 1 (DCR1) with poor response to irinotecan in metastatic colorectal cancer. Annual meeting of the American Society of Clinical Oncology (ASCO), 31(15, suppl.),@ 2013

L Bosch, **Geert Trooskens**, Petur Snaebjornsson, J Haan, Miriam Koopman, Jolien Tol, T de Meyer, J Louwagie, L Dehaspe and NCT van Grieken. Promoter CpG island hypermethylation of Decoy Receptor 1 (DCR1) is associated with poor response to irinotecan in metastatic colorectal cancer. 26th European congress of Pathology, 465(suppl. 1):S326-S326, 2014

Leander Van Neste, Geert Trooskens, Rianne J Hendriks, Jack Schalken and Wim Van Criekinge. MP53-04 Identification of High-grade Prostate Cancer Using Urine-Based Molecular Biomarkers combined with Clinical Risk Factors. The Journal of Urology, 195(4):e698, 2016

Conference Posters

Geert Trooskens, Tim De Meyer, Simon Denil and Wim Van Criekinge. Charting the methylome. 6th Benelux Bioinformatics Conference:73-73, 2011

Patents

Wim Van Criekinge, Josef Straub, **Geert Trooskens**, Stephen Baylin, James Herman, Kornel Schuebel, Leslie Cope and Leander Van Neste. Detection and prognosis of lung cancer. US Patent App. 12/867,539, 2009

Wim Van Criekinge and **Geert Trooskens** . Method for Personal Genome Data Management. US Patent App. 14/039,860, 2013

Gerrit A Meijer, Beatriz Carvalho, Linda Bosch, **Geert Trooskens** and Wim Van Criekinge. Methylation markers predictive for drug response. US Patent App. 14/438,742, 2013