DIFFERENT APPROACHES TO MEASURING GENE EXPRESSION AND DNA METHYLATION AND THEIR APPLICATION IN CANCER RESEARCH

ALEXANDER KOCH
GHENT UNIVERSITY
2015

PROMOTER

Prof. Wim Van Criekinge
BioBix, Lab of Bioinformatics and Computational Genomics
Ghent University

CO-PROMOTER

Prof. Tim De Meyer
BioBix, Lab of Bioinformatics and Computational Genomics
Ghent University

CO-PROMOTER

dr. Gerben Menschaert
BioBix, Lab of Bioinformatics and Computational Genomics
Ghent University

DEAN

Prof. Guido Van Huylenbroeck

RECTOR

Prof. Anne De Paepe

DIFFERENT APPROACHES TO MEASURING GENE EXPRESSION AND DNA METHYLATION AND THEIR APPLICATION IN CANCER RESEARCH

ALEXANDER
KOCH

Thesis submitted
in fulfilment of the requirements
for the degree of Doctor (PhD)
in Applied Biological
Sciences

DUTCH TRANSLATION OF THE TITLE

Verschillende benaderingen voor het meten van genexpressie en DNA-methylatie en hun toepassing in het kanker onderzoek.

REFERENCE

Koch A. Different approaches to measuring gene expression and DNA methylation and their application in cancer research. *PhD thesis, Ghent University* (2015)

COVER ILLUSTRATION

PRINTING

ISBN

Alexander Koch

Wim Van Criekinge

MEMBERS OF THE EXAMINATION COMMITTEE

Prof. dr. Guy Smagghe – **chairman**
Department of Crop Protection
Ghent University

Prof. dr. Tina Kyndt – **secretary**
Department of Molecular Biotechnology
Ghent University

Prof. dr. Tim De Meyer – **co-promoter**
Department of Mathematical Modelling, Statistics and Bioinformatics
Ghent University

Prof. dr. David Fenyö
Center for Health Informatics & Bioinformatics
New York University

dr. Gerben Menschaert – **co-promoter**
Department of Mathematical Modelling, Statistics and Bioinformatics
Ghent University

Prof. dr. Bart Neyns
Medical Oncology
UZ Brussel

Prof. dr. Wim Van Criekinge – **promoter**
Department of Mathematical Modelling, Statistics and Bioinformatics
Ghent University

Prof. dr. Petra Van Damme
Department of Biochemistry
Ghent University

Prof. dr. Manon Van Engeland
Department of Pathology
Maastricht University

PREFACE

Those of you who read a PhD thesis from time to time might notice that this particular one is not written like most others. From the moment I started writing, it has been my intention to create a book that is not only aimed at the people in my thesis committee. If somebody from my friends or family were to pick this book up (and I hope they do), then I want them to understand what I am writing about.

I have divided my research in three chapters and each chapter starts of with an introduction. The research itself is described as you would expect from a scientific paper, but I have tried to give the introductions a lighter tone, without oversimplifications.

I am happy with the way this thesis turned out and I hope that you, my reader, will find it both interesting and pleasant to read.

WHAT'S IT ALL ABOUT?

RESEARCH GOALS

The main objective of this PhD thesis was to try and better understand the link between epigenetics and gene expression using bioinformatics. Over the course of the past four years, this broad aim was narrowed down to a few specific research questions.

The term epigenetics covers a number of distinct biological processes, such as histone modifications, micro RNA and DNA methylation. The research in this thesis was focused on the latter. It is well established that DNA methylation is a crucial player in the regulation of gene expression, but most studies on its relationship with expression concentrate on expression at the transcript level. We decided to take a different approach and examined expression at the protein level. As demonstrated by several studies, there are substantial differences between transcript and protein expression levels, partly explained by post-transcriptional modifications, which affirms the value of our protein-centric approach. However, high-throughput proteomics has some drawbacks of its own, such as sub-optimal reproducibility and an inadequate sensitivity for low-abundant proteins. To address these drawbacks we turned to a novel technique known as ribosomal sequencing and investigated the potential benefits of integrating this method into our proteomics analyses.

Apart from these more fundamental and technical research questions, we also worked on (epi)genetic analyses in a clinical setting. Strict regulation of DNA methylation and gene expression is vital for the normal functioning of a cell and derangements of this regulation are linked to many diseases. Aberrant DNA methylation for instance has been found in virtually every type of human cancer. There are, however, many ways to take advantage of these aberrations. They can be used to better understand a disease, as targets for therapy or as biomarkers for diagnosis, prognosis or response prediction. The interesting thing about DNA methylation in particular is that the binding of a methyl group to DNA is reversible. The possibility of treating a disease by fixing erroneous DNA methylation has opened the door for a variety of therapeutic opportunities. We studied the use of azacitidine, a demethylating drug, in combination with immunotherapy in the treatment of metastatic lung cancer. To understand how azacitidine-induced demethylation affects lung cancer cells, we analyzed both DNA methylation and gene expression data in several lung cancer cell lines.

Immunotherapy in itself holds a lot of promise for the treatment of cancer, but does not help every patient, comes at a high cost and has many severe side effects.

The ability to select only those patients that will benefit from the treatment is therefore crucial. We tried to tackle this problem by comparing DNA methylation and expression profiles with response to immunotherapy in patients with metastatic melanoma.

OUTLINE

Chapter 1 introduces some of the basic principles of cellular biology and the fundamental elements of gene expression. Special attention is given to the role of epigenetics, and DNA methylation in particular, in the regulation of gene expression. The different techniques that we used in this thesis to study gene expression and DNA methylation are also presented to the reader. The second part of this chapter builds on this introduction and describes the integration of MBD-sequencing with high-throughput proteomics to measure the genome-wide impact of DNA methylation on expression at the protein level.

Chapter 2 introduces a different technique to measure gene expression: ribosomal sequencing, or ribo-seq. The proteomics methods described in chapter 1 have some shortcomings, which we try to resolve with the use of ribo-seq. We explain how the combination of ribo-seq and proteomics improves the overall efficiency of protein identification and enables the discovery of alternative translation start sites.

Chapter 3 focuses on the practical application of gene expression and DNA methylation analysis in cancer research. The first part outlines the biology behind cancer, explains the importance of gene expression and DNA methylation in the development of tumors and introduces immunotherapy as a promising cancer treatment. The next part presents the effects of using azacitidine, a DNA demethylating drug, in the treatment of lung cancer, while the third part demonstrates how expression and DNA methylation analyses could be used to predict response to immunotherapy in melanoma patients. The final part introduces the reader to MEXPRESS, a web tool for the visualization of expression, DNA methylation and clinical data from TCGA, one of the major public cancer databases.

Chapter 4 offers a general conclusion and some future perspectives.

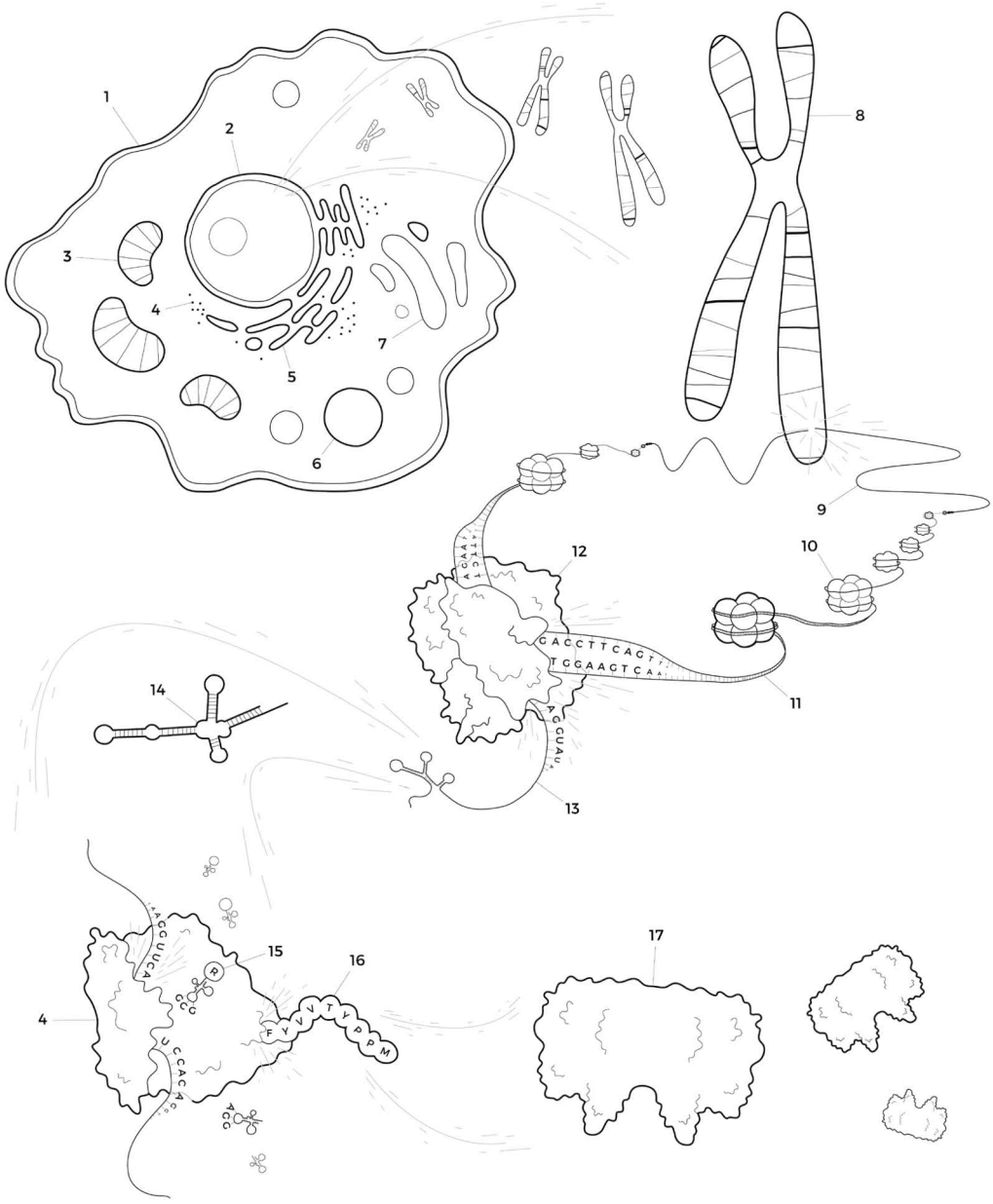
CHAPTER 1

ON GENE EXPRESSION AND DNA METHYLATION

FIGURE 1.1
THE CENTRAL DOGMA

This figure illustrates the central dogma of molecular biology. It shows how genetic information is passed on from our DNA to RNA and to proteins in a process known as gene expression. First, the DNA sequence of gene is transcribed to RNA by RNA polymerase inside a cell's nucleus. Once the transcription is finished, the mature mRNA molecule travels to the endoplasmic reticulum where the RNA sequence is translated to an amino acid sequence by a ribosome. Finally, the resulting string of amino acids folds itself into the correct three-dimensional structure and the protein is ready.

cell membrane	1
nucleus	2
mitochondrium	3
ribosome	4
endoplasmic reticulum	5
lysosome	6
Golgi apparatus	7
chromosome	8
chromatin	9
nucleosome, consists of eight histones	10
DNA	11
RNA polymerase	12
RNA	13
mature mRNA	14
tRNA carrying an amino acid	15
growing protein	16
mature protein	17



THE HUMAN CELL

Think about a human cell. A helper T cell, gall bladder epithelial cell or neuron, any cell. Unless you happened to pick a sperm or egg cell, which have 23 chromosomes instead of the usual 46, your cell of choice contains the same genome as any other type of cell in the human body. All these cells obtain their (sometimes radically) different structures and wide range of functions from the exact same genetic information. This means there has to be a mechanism that controls which part of the genetic information is actively used at any given moment in the life of a cell. Let's talk about some of the basic concepts of molecular biology before we discover what this control mechanism is and how it works.

Everything starts with a molecule known as deoxyribonucleic acid or DNA. Its basic building blocks are nucleotides, which consist of at least one phosphate group, a sugar (deoxyribose in the case of DNA) and one of the following four bases: adenine, thymine, cytosine and guanine or A, T, C and G. These four letters are all that is needed to write our complete genome and all the genetic information it contains. The central dogma of molecular biology describes how this genetic information is converted from DNA to proteins, the main functional elements in a cell (Crick, 1970, Figure 1.1). In short, the dogma states that the information in DNA can be passed on to another DNA molecule (replication) or an RNA molecule (transcription) from which it can then be transferred to a protein (translation). Generally, this is a one-way process, though there are some exceptions. Researchers have for example discovered that certain viruses, such as HIV, and even human enzymes, such as telomerase, can create DNA from an RNA template in a process called reverse transcription. Information transfer from a protein back to RNA or DNA has not been observed. So remember, from DNA to RNA to protein (but not always).

Let's have a closer look now at some of the basic concepts of this central dogma. Our genome contains stretches of DNA that code for proteins and these stretches are what we call genes. The DNA sequence of a gene can be transcribed to a different molecule known as ribonucleic acid or RNA. Just as DNA, this molecule consists of a string of nucleotides. The two main differences between RNA and DNA are that RNA uses ribose as a sugar in its nucleotides (instead of deoxyribose) and that thymine (T) is replaced by uracil (U). After successful transcription, the RNA molecule will contain a copy of the genetic information encoded by the gene's DNA (with every T replaced by a U) and it will carry this information from the nucleus of a cell to the ribosomes. Because of its courier function and to separate it from other types of RNA molecules, this RNA copy of a gene is called messenger RNA or mRNA. The complete set of mRNA molecules or transcripts in a cell at a given time is called the transcriptome. Ribosomes are complex molecular machines that read the mRNA and translate the sequence of A, U, C and Gs to a sequence of amino acids, the building blocks of proteins. A sequence of three bases in the mRNA is known as a codon and every codon can be translated to a specific amino acid. There are 4^3 or 64 possible codons for 20 different amino

acids, meaning that on average each amino acid is specified by about three codons. Despite the fact that a single codon does not code for more than one amino acid, codons can sometimes be linked to an amino acid even if the last base (sometimes referred to as the "wobble" base) does not match.

Before an mRNA molecule is translated into a protein, it undergoes a process we call splicing. The DNA sequence of a gene is not entirely protein coding, it contains stretches of non-protein coding DNA known as introns. The parts that do code for a protein are called exons. Splicing removes the introns from a transcript and joins the remaining protein-coding exons together. Apart from the introns, genes also contain different regulatory elements which are not translated. Sometimes, the protein-coding part of a gene makes up only a tiny fraction of the gene's total length.

Once translation is finished and the resulting protein has folded itself into the correct three-dimensional structure, has (if necessary) formed bonds with other proteins to create a protein complex and has received the appropriate side chains (such as carbohydrates), it is ready to perform its function in the cell or, in the case of membrane or extracellular proteins, outside of it. Similar to the transcriptome, the collection of all the proteins present in a cell at a given moment is called the proteome. When talking about the expression of a gene, we are actually talking about the transcription and translation of this gene. So when we say that a gene is highly expressed, this implies that there is either a lot of mRNA present or a lot of the protein, depending on how the expression was measured. We will discuss the different measurement techniques later on.

A cell is an incredibly intricate and busy system, but behind the apparent chaos of bustling proteins and whirling molecules you will find meticulous control mechanisms. Cells need a tight regulation of gene expression to respond to an ever-changing environment and to manage their life cycle. Once a cell has differentiated into a specific cell type it must maintain its cell-type-specific expression pattern, while repressing all genes that are specific to other cell types, and it has to pass this pattern on to the next generation. The daughter cells of a liver cell for example should also express the necessary liver-cell-specific genes so they maintain the characteristics of a liver cell. Gene expression is regulated on various levels and at different points during the expression process. Transcriptional regulation forms the first checkpoint and as the name suggests it controls the amount of RNA that is produced. A class of proteins known as transcription factors plays an important role in this type of regulation. They recognize special protein binding sites near a gene's coding region: promoters, enhancers, insulators and inhibitors. The promoter of a gene is a cluster of short DNA sequence elements that act as recognition signals for transcription factors and can be found just before the start of a gene's coding region. Once a transcription factor complex is formed at the promoter, an RNA polymerase molecule binds to it and transcription of the gene

can begin. RNA polymerases are enzymes that copy the DNA sequence of a gene into an RNA molecule. Their structure and function are very similar to those of DNA polymerases, the enzymes responsible for DNA replication. As you might have guessed, binding of a transcription factor to an enhancer region will enhance the transcription of the corresponding gene. Binding to an insulator or inhibitor will block or repress transcription. The transcription factors themselves are also regulated, for example by intracellular signaling pathways that change the activity of a transcription factor in response to environmental input.

Epigenetics is another major player in transcriptional regulation. According to the definition of Egger *et al.* (2004), "the term epigenetics defines all meiotically and mitotically heritable changes in gene expression (phenotype) that are not coded in the DNA sequence itself". In this quote, they basically say that epigenetics adds an extra layer of information to the genetic information encoded in the DNA sequence of our genome. A *genetic* change–a change in the sequence of a gene, for example a mutation that replaces a T with a G–can result in a change in the expression of this gene or even in a modified protein with a new function (this is one of the basic genetic mechanisms behind evolution). Genetic changes can have very negative results for a cell. They might result in the cell's death or can cause severe diseases. Processes that affect the expression of a gene and that can pass this change on to future generations without actually changing the DNA sequence fall under *epigenetics*. Three well-known epigenetic mechanisms are DNA methylation, histone modifications and certain classes of non-coding RNA molecules such as micro RNA (miRNA, Bartel, 2004) or long non-coding RNA (lncRNA, Mercer *et al.*, 2009). We will describe these three in more detail a bit further down.

Transcriptional regulation of gene expression manages how much mRNA is created, but it is far from the only available control mechanism. After transcription, but before translation, an mRNA molecule goes through post-transcriptional regulation, a combination of several processing steps that influence the stability and distribution of the transcript and therefore the expression level of the corresponding gene. Splicing is one of the most important processing steps. By including or excluding certain exons, one transcript can result in several different proteins or isoforms. Direct translation control through translational regulation also has an effect on the amount of protein that is produced, but is less common than (post-) transcriptional regulation. Common mechanisms include the control of ribosome recruitment to the first codon and direct adjustments of the protein synthesis process. Post-translational regulation is the final gene expression checkpoint and controls the levels of active protein through small modifications of these proteins. Phosphorylation for example (the addition of a phosphate group to a protein) acts as an "on/off" switch for many proteins, whereas adding ubiquitin to a protein labels it for degradation.

Given that the main objective of this PhD thesis was to try and better understand the impact of epigenetics on gene expression, we must dive a bit deeper into the biology behind it. You could already read how miRNA, histone modifications and DNA methylation are the best-known epigenetic processes and in the following paragraphs we will take a closer look at all three.

Micro RNAs are a type of non-coding RNA molecules, which means that they do not get translated into a protein. The "micro" in their name refers to their short length of approximately 22 nucleotides. MiRNAs bind to specific mRNA targets, thereby marking the targeted mRNA for destruction, which in turn results in a lower expression level for the associated gene (Bartel, 2004). Members of the miR-15/107 group of miRNAs for example regulate gene expression in cell devision, metabolism, stress response and angiogenesis (Finnerty *et al.*, 2010). Another well-known example is miR-21-5p, which is often expressed in colorectal cancer and has been linked to poor disease prognosis (Dong *et al.*, 2014). There are several other classes of non-coding RNAs that control gene expression, such as the lncRNAs, but we will not discuss them further here.

Before we explain how histone modifications regulate gene expression, we need a bit more background information. Our genome is incredibly long, over three billion base pairs long to be exact. If you could glue all of the 46 chromosomes in the nucleus of one of your cells together (head-to-tail) in one long string and stretch it out, you would end up with roughly 2 meters of DNA (Alberts *et al.*, 2002). This has to fit in a nucleus with a diameter of six *micro*meters, or 0.000006 meters. This is where the histones come in. The basic structure of our DNA is that of a double helix, which consists of two intertwined DNA strands held together by hydrogen bonds between adenine-thymine and cytosine-guanine base pairs. The DNA helix itself is wrapped around complexes of eight histones each to form nucleosomes, the basic subunits of chromatin (the macromolecular complex of DNA and proteins our chromosomes are made of). Thanks to this conformation the DNA can be packed much tighter and the total combined length of our chromosomes can be reduced to about 90μm.

The complex three-dimensional structure of our genome is not only important to make it fit in the nucleus, but also to control gene expression. There are two chromatin variations: an open form that enables gene expression, euchromatin, and a tightly packed form that restricts gene expression, heterochromatin. Several functional groups, such as acetyl or methyl, can be added to or removed from histones and depending on the type and exact location of the modifications the chromatin will be open or closed (Table 1.1, Figure 1.2). Acetylation of lysine 9 of histone H3 for example will create a more open structure that allows for transcription, while di or tri-methylation of the same lysine will result in a closed and transcriptionally inactive chromatin structure (Lachner & Jenuwein, 2002).

A third well-known epigenetic mechanism, and the focus of much of the research described in this thesis, is DNA methylation. A methyl group can be bound to a cytosine, most often in a CpG dinucleotide (the "p" refers to the phosphodiester bond between a C and its neighboring G), converting this cytosine to 5-methyl-cytosine, the so-called fifth base. Just like their unmethylated counterparts, these methylated cytosines can still base-pair with guanine in the complementary DNA strand, but there are also some differences. Methylated cytosines can spontaneously deaminate to thymines (Shen *et al.*, 1994), which is why CpG dinucleotides are underrepresented in the human genome (1% instead of the expected 4.41%). Of the CpGs we have, roughly 60% is methylated which means that between 3 to 6% of all cytosines in the genome of a normal human cell are methylated (Jabbari & Bernardi, 2004). The remaining unmethylated CpGs are often found in CpG islands, clusters of (mostly) unmethylated cytosines located in the regulatory regions of many genes (Takai & Jones, 2002). The methylation of CpG-rich regions in the promoter region of a gene for example is generally linked to the repression of that gene's expression (Miranda & Jones, 2007).

Methylated CpGs also attract various proteins, including the histone deacetylase enzymes, which will all collaborate to form a so-called transcription repressor

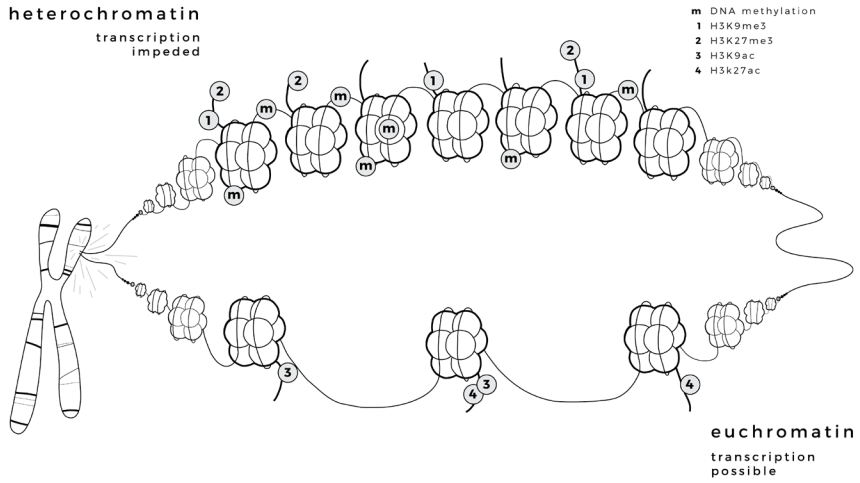


Figure 1.2 DNA methylation, histone modifications and the chromatin structure.

This figure shows the two different chromatin arrangements, heterochromatin (the closed and transcriptionally inactive form) and euchromatin (the open and transcriptionally active form). The structure of chromatin is controlled by DNA methylation and various histone modifications. Typical characteristics of heterochromatin include DNA methylation (**m**), H3K9me3 (**1**) and H3K27me3 (**2**), while euchromatin is characterized by H3K9ac (**3**), H3K27ac (**4**) and a lack of DNA methylation (see Table 1.1 for an overview of some histone modifications).

Table 1.1 Examples of common human histone modifications and their effect on gene expression.

Histone modifications have their own nomenclature. Each modification is named by the affected histone (for example histone 3 or H3), the single-letter abbreviation of the modified amino acid (K for lysine), the position of that amino acid in the protein, the type of modification (me for methylation, ac for acetylation) and the number of modifications. H3K27me3 for example denotes the tri-methylation of lysine 27 in histone 3.

Histone modification	Link with transcription
H3K4me3	high levels have been detected in the promoter regions of active genes (Barski *et al.*, 2007)
H3K27me3	elevated levels correlate with gene repression (Barski *et al.*, 2007)
H3K27ac	associated with active enhancers (Creyghton *et al.*, 2010)
H3K9me3	marks the transcriptionally inactive heterochromatin (Rosenfeld *et al.*, 2009)
H3K9ac	associated with the transcription start sites of genes (Koch *et al.*, 2007)

complex, transforming the surrounding chromatin into heterochromatin (Cedar & Bergman, 2009). As you could read earlier, the acetylation of histones "opens" the chromatin structure. Deacetylation (the removal of these acetyl groups) by the deacetylase enzymes results in the transcriptionally inactive heterochromatin. So besides their individual effects on transcription, there is some important interaction between histone modifications and DNA methylation.

The role of DNA methylation as a controller of gene expression is vital for many biological processes. After an initial "reset" of the DNA methylation profile following fertilization, DNA methylation is used to control cellular differentiation during embryonic development (Jaenisch & Bird, 2003). It is responsible for parental imprinting, whereby one of the two copies of a gene (paternal or maternal) is silenced (Swain *et al.*, 1987), and for the inactivation of one of the two X chromosomes in women (men only have one X chromosome) (Mohandas *et al.*, 1981). DNA methylation also protects our cells from the expression of viral genes (Jahner *et al.*, 1982) and from the potentially disruptive effects of transposable elements (Yoder *et al.*, 1997), stretches of "parasitic", repetitive DNA sequences that can damage our genome by inserting itself more or less randomly in it. The crucial role of DNA methylation in the normal functioning of our cells also implies that if something goes wrong, the results can be disastrous. This is illustrated by the fact that abnormal DNA methylation is found in virtually every type of human cancer (Herman & Baylin, 2003). Chapter three describes the role of DNA methylation in cancer in detail.

TO MEASURE IS TO KNOW

Now that you have been introduced to the concepts of gene expression and DNA methylation, we can discuss some of the techniques that are used to measure them. The overview that follows is by no means exhaustive and focuses on the techniques that were used in this thesis.

As mentioned earlier, gene expression can be measured at either the transcript or the protein level. Both options have their advantages and disadvantages and require very different techniques. Let's start our measurements at the transcript level using microarrays (Figure 1.3). A microarray is basically nothing more than a small (think stamp-sized), flat piece of glass or silicone covered by miniscule DNA spots. These spots contain short (about 60 nucleotides), single-stranded DNA sequences known as probes or oligonucleotides. Researchers can design their own probes to create a fully custom microarray or they can buy one of the many commercially available arrays.

Imagine a simple lab experiment in which you want to compare the gene expression profiles of two samples using a microarray. The basic workflow will start with the extraction of the RNA from your two samples. Using a reverse transcriptase enzyme (responsible for the reverse transcription described earlier) and the right conditions, you can specifically convert only the mRNA to DNA. The resulting DNA molecules are known as complementary DNA or cDNA. Scientists often convert RNA to DNA, because DNA molecules are less fragile, making them easier to work with or store. After the conversion, one sample is fluorescently labeled using a green dye (cyanine 3 or Cy3) and the other with a red dye (cyanine 5 or Cy5). The differently labeled cDNA fragments from both samples are mixed and the mix is added to the microarray.

The probes on a typical gene expression microarray are designed in such a way that they cover a big part of the transcriptome. This means that their sequences must match parts of the sequences of the transcripts we want to detect with the array. Once the labeled cDNA fragments from the two samples have been added to the microarray they will bind to their corresponding probes by base-pairing (the same mechanism by which two complementary DNA strands bind to each other). When enough time has passed to allow for this binding, the array is washed to remove unbound cDNA and then dried. The dry array is now ready to be scanned by a detector. This machine uses a laser to excite the fluorescent dyes and then quantifies the intensity of the resulting fluorescence. Every DNA spot results in two intensities (green for one sample and red for the other) and by

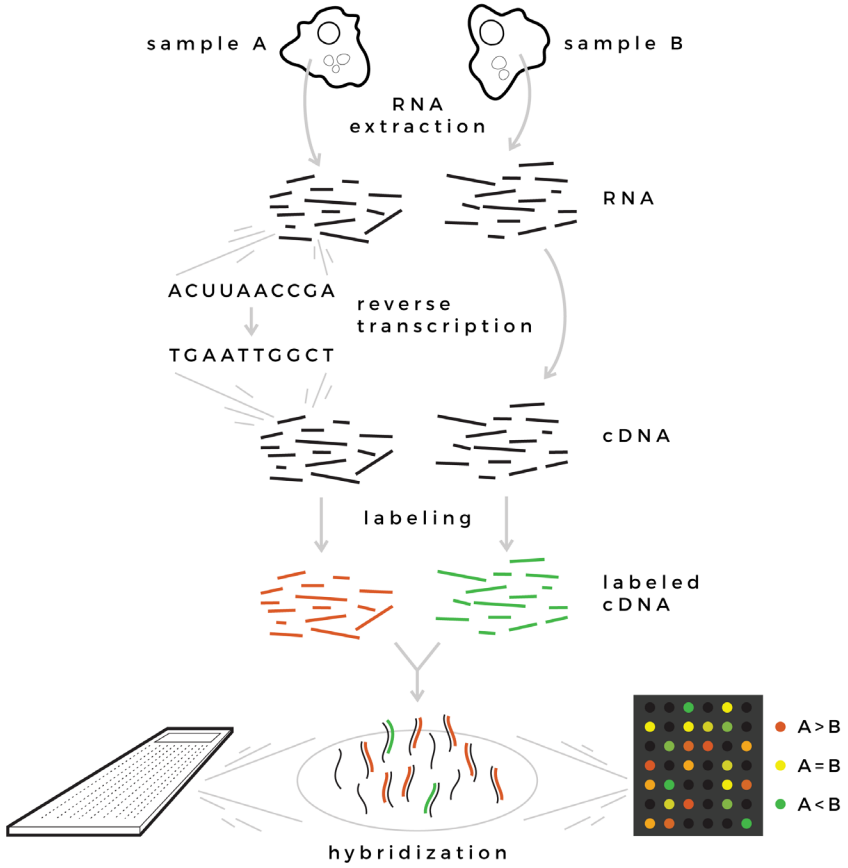


Figure 1.3 A gene expression microarray.

This diagram illustrates the main steps of a gene expression microarray experiment. After the extraction of the RNA from two samples, the RNA fragments are converted to complementary DNA or cDNA through a process we call reverse transcription. The resulting cDNA fragments are then labeled (using different labels for the two samples), mixed and added to the microarray slide. Once on the slide, the cDNA fragments will bind to their matching probes (hybridization). Thanks to the colored labels, we can measure whether a particular cDNA fragment was more present in sample A or in sample B (and thus indirectly whether the corresponding gene was differentially expressed between both samples).

comparing these intensities you can figure out which genes were differentially expressed between the two samples.

Microarrays have been around since the early nineties, so the technology is very well understood. They are relatively cheap and the later versions cover the known transcriptome quite well. However, despite these advantages a different gene expression analysis technique has gained a lot of popularity over the past years: RNA sequencing (RNA-seq). The basic idea behind this technology is that by determining the sequence of the transcripts in a sample and looking up the

location of these sequences in our genome, we can discover which genes were expressed. Because it does not depend on pre-designed probes, an RNA-seq experiment offers the possibility to discover new transcripts as well as genetic variations, such as mutations in the RNA sequence (an indication of a mutation in the corresponding DNA sequence), gene fusions and sample-specific isoforms (Costa *et al.*, 2013). RNA-seq also provides a broad dynamic range, which means that it can be used to accurately measure both very low (only a few transcripts) and very high (thousands of transcripts) gene expression levels. Microarrays on the other hand are limited in their measurements of these extreme expression levels by background noise and signal saturation. When measuring fluorescence there will always be some background signal or "noise" and the signal of genes with very low expression levels might be indistinguishable from this noise. Signal saturation on the other hand can happen for highly expressed genes. At a certain fluorescence intensity, the microarray scanner will no longer be able to accurately tell the difference between two very high expression levels. A disadvantage of RNA-seq used to be the higher cost, but prices have plummeted since the arrival of the so-called next-generation sequencing techniques in the mid 2000s.

The current RNA-seq technology would not have been possible without the initial development of DNA sequencing techniques. By the late forties, the work of Gregor Mendel, Friedrich Miescher, Oswald Avery, Erwin Chargoff and many

Table 1.2 Overview of some of the most commonly used sequencing methods.

This table, which was adapted from Liu *et al.* (2012) and Pareek *et al.* (2011), lists a range of different techniques, from the classic Sanger to the recently developed single molecule sequencing.

Sequencing method	Machine	Number of reads	Time/ run	Advantages	Disadvantages
Sanger sequencing	Sanger 3730xl	—	20 min to 3 h	high quality, long reads	high cost. low throughput
Pyrosequencing	454 GS FLX	1 M	24 h	long reads, fast	high cost, low throughput
Sequencing by synthesis (Illumina)	HiSeq 2000	3 G	~10 days	cheap, high throughput	short reads
Sequencing by ligation (SOLiD sequencing)	SOLiDv4	~1.3 G	7 to 14 days	cheap, accurate	short reads
Single molecule sequencing	Heliscope	< 8 M	~1.5 days	cheap, fast	homopolymer errors
Single molecule sequencing	SMRT	0.5 M per cell	< 1 h	very long reads, can detect DNA methylation	high cost, low throughput

other scientists had made DNA the most likely candidate in the search for the carrier of our genetic information. The discovery of the helical structure of DNA by Francis Crick, James Watson, Maurice Wilkins and Rosalind Franklin, one of the major milestones in science, sealed the deal (Watson & Crick, 1953, Franklin & Gosling, 1953, Wilkins, 1957). Scientists now realized that if they could find a way to read the sequence of A, T, C and Gs that makes up our genome, they would have access to all our genetic information. The sequencing race was on. In 1977 the first complete viral DNA genome was sequenced (Sanger *et al.*, 1977), but for the first human genome we had to wait until 2001 and the completion of the billion-dollar human genome project (Lander *et al.*, 2001, Venter *et al.*, 2001).

Up until the mid 2000s Sanger sequencing was the most popular sequencing technology. Published by Frederick Sanger in 1977, this method is based on the use of labeled dideoxynucleotides (ddNTPs). These molecules are very similar to deoxyribonucleotides (dNTPs), the usual building blocks of DNA, but they lack the hydroxyl group that's needed to string them together. The idea is that when you add a mixture of dNTPs and labeled ddNTPs together with a DNA polymerase to a sample of single-stranded DNA fragments of the genomic region you are interested in, these fragments will be elongated by the polymerase until a ddNTP is randomly incorporated in the growing DNA molecule. This will leave you with a mix of DNA strands of different lengths that all end with a labeled ddNTP (a different label is used for each of the four bases). If you now sort these fragments by length you will be left with a series of fragments that vary in length by just one nucleotide. With the help of the labeling and the right equipment you can now read the DNA sequence of your sample.

The push for faster and cheaper sequencing methods already started in the nineties with the arrival of pyrosequencing, but it wasn't until 454 Life Sciences marketed their "massively parallel" pyrosequencer in 2004 that the reign of the next-generation sequencing techniques started. Today, one of the most popular sequencing methods is the sequencing by synthesis technique developed by Solexa, now part of Illumina (Table 1.2 gives an overview of commonly used sequencing techniques). In this approach, the DNA fragments from a sample are attached to a slide and amplified by a DNA polymerase, resulting in a collection of clusters of identical single-stranded DNA fragments. This amplification will result in increased signal intensities and thus more accurate measurements later on. A mix of the four differently labeled nucleotides is then added to the slide. The labels ensure that only one nucleotide can be added to the fragment at a time (the label blocks DNA elongation) and they allow us to identify the nucleotide that paired with the nucleotide in our fragment. Once a labeled nucleotide has bound to a fragment on the slide and a sensitive camera has registered its signal, the label is removed and washed away together with the remaining labeled nucleotides. By repeating these steps over and over, we can read the DNA sequence of our fragment nucleotide by nucleotide. The process we just described deals with a

single DNA fragment in one fragment cluster on one slide. A current "massively parallel" or next-generation sequencing machine can perform thousands of these processes in parallel over several slides, resulting in millions of reads at the end of the sequencing run.

In this thesis we used this Illumina sequencing method for our RNA-seq analyses. Just as in a microarray experiment the extracted mRNA needs to be converted to cDNA first. Once we have the cDNA, we can sequence it as described above. After sequencing, we are typically left with tens of millions of short reads, but we still don't have any idea which genes were expressed. To figure this out we need specialized bioinformatics tools known as read mappers. One example of such a mapper is Bowtie (Langmead *et al.*, 2009). These tools go through all our sequenced reads and try to map them to the reference human genome. This means that they scan the sequence of every read and then try to find the matching sequence in the reference genome. Given that our genome is about three billion base pairs long, that read sequences don't necessarily exactly match the reference sequence (due to genetic variation or sequencing errors), that reads sometimes map to several locations in the genome and that they can overlap the border between two exons (meaning that we don't have the sequence of the intron between the start and end of the read), read mapping is far from a trivial task. But once we have successfully mapped our reads we can use the number of reads that mapped to a certain gene as a measurement of the expression of this gene in our sample. Using the appropriate normalization and statistical techniques we can also rigidly compare these numbers between different samples.

Microarrays and RNA-seq are popular techniques that are widely used to measure gene expression. In the end though, transcript levels are only a proxy for the number of proteins that are produced in a cell. Several studies have found notable differences in gene expression measured at the transcript and at the protein level (Gry *et al.*, 2009), which can be (partly) attributed to the post-transcriptional regulation we described earlier. So instead of using only the techniques we just described, it might be very useful to try and measure protein instead of transcript levels. Proteomics is the name that is used for the large-scale experimental analysis of protein expression. This term covers various techniques, but is often used to refer to the method we used in this thesis: protein extraction followed by mass spectrometry, also known as shotgun proteomics.

The extraction of proteins from a biological sample results in a complex mixture. To identify the individual proteins in this mix, they are first broken up into smaller parts, known as peptides, by an enzyme such as trypsin. These peptides are then identified and from these identifications we can deduce which proteins were present in the sample. Because we work our way up from identifying the peptide fragments to the identification of the corresponding protein, this is known as a bottom-up technique. The name shotgun proteomics also comes from this

fragmentation step and refers to the firing pattern of an actual shotgun.

The next step in a shotgun proteomics experiment is a first separation of the resulting peptides using reversed-phase high performance liquid chromatography (RP-HPLC). Chromatography is nothing more than a technique to separate a mixture of molecules, in our case peptides. The peptides are dissolved in a liquid solvent (hence liquid chromatography), which travels through a column filled with a solid absorbent material such as silica (known as the stationary phase). The more peptides interact with the column, the slower they will flow through it and the resulting differences in flow rate can be used to separate peptides. The high performance refers to the use of a pump to create a pressure to push the solvent through the column, resulting in a better separation of the peptides. In a normal HPLC experiment the solvent is hydrophobic (non-polar), while the stationary phase is hydrophilic (polar). This is reversed in a reversed-phase HPLC experiment, where a polar solvent and non-polar stationary phase are used, meaning

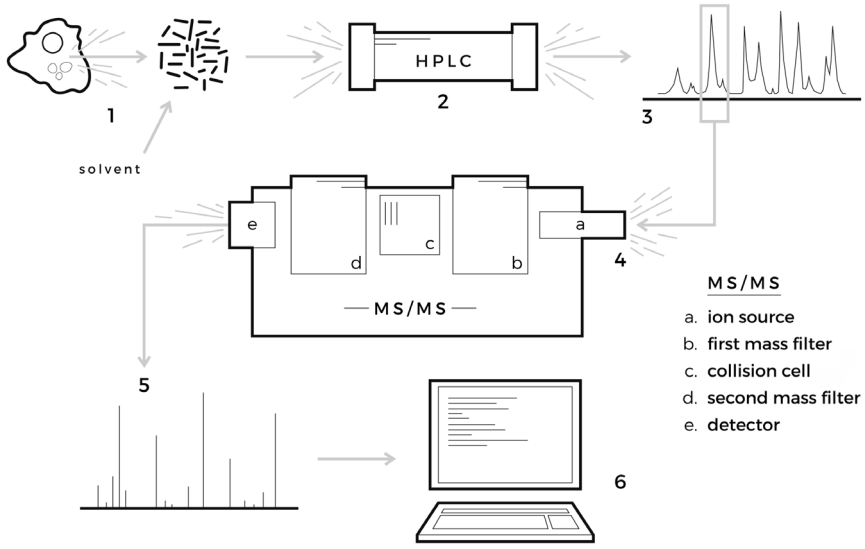


Figure 1.4 The LC-MS/MS workflow.

1 Proteins are extracted from the sample we are interested in, fragmented and then mixed with a solvent. **2** The peptide/solvent mix is pumped through an HPLC column for a first separation. **3** Using the HPLC separation, the complex peptide mixture can be split in batches of peptides with similar size and/or polarity (depending on the type of HPLC column that was used). **4** Each batch is fed into a tandem mass spectrometer, where it is ionized (**a**), filtered on mass and/or charge (**b**), fragmented in even smaller pieces (**c**), filtered again (**d**) and finally, the resulting peptide fragments are picked up by a detector (**e**). **5** The mass spectrometer produces mass spectra that are then compared to existing databases using specialized software (**6**) in order to identify the proteins that were present in our initial sample.

that more polar peptides will travel faster through the column. The solvent that comes out of the column is captured in separate batches and these batches are then used in a liquid chromatography tandem mass spectrometry (LC-MS/MS) experiment to identify the peptides they contain.

An LC-MS/MS analysis can be divided in three steps: separation, detection and identification (Figure 1.4). We start off with a second RP-HPLC. These consecutive separation steps will ultimately improve the identification rate of the proteins in our sample. Next, the separated peptide fractions are fed into a mass spectrometer. This machine vaporizes the solvent together with the peptides it carries, ionizes the peptides using an electron beam and sorts the resulting ions by their mass-to-charge ratio using a varying electromagnetic field. In a tandem mass spectrometry analysis the gas-phase ions from a first MS run are fragmented and then separated once more during a second MS run. The separated ions finally hit an electron multiplier, a small device that generates an electric current each time it is hit by an ion.

The end result of these complicated steps is a mass spectrum, a plot of the intensity of the electrical current generated by ions hitting the electron multiplier as a function of the magnetic field strength (and thus the mass-to-charge ratio of the ionized peptides). These spectra act as a fingerprint for the peptides in our sample. Much like the police looking up the identity of a suspect using his or her fingerprints, we can use the mass spectra to look up the identity of the corresponding peptides in existing protein sequence databases. And once we identified the peptides, we can find out which proteins were present in our sample. This gives you a rather qualitative measurement of the protein content (is protein A present in my sample?). If you are looking for a more quantitative comparison between two samples (how much more or less protein A is there in sample 1 compared to sample 2?), you can label the two samples before the shotgun proteomics analysis. One popular labeling technique uses stable isotopes such as carbon-12 and its heavier isotope carbon-13 (Ong *et al.*, 2003). These isotopes are used to create two versions of an amino acid such as arginine (one with carbon-12 and one with carbon-13), one of which is then fed to the cells in the first sample and the other to the cells in the second sample. The growing cells will incorporate the labeled amino acids in the proteins they create during their normal life cycle. Once labeled, the proteins from both samples can be combined and the mix goes through the whole shotgun proteomics experiment. The labeling will allow you to distinguish the proteins and peptides from both samples at the end of the experiment and special quantification software unveils the expression ratios between the two samples for the identified proteins.

Even though they measure gene expression at the functional protein level instead of the transcript level, these genome-wide proteomics techniques suffer from some shortcomings compared to RNA-seq. As you could read, the technique is

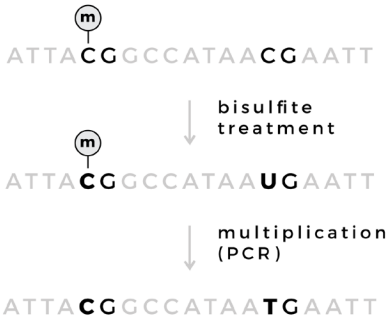


Figure 1.5 Bisulfite conversion.

Treatment of a DNA molecule with bisulfite will convert the unmethylated cytosines to uracil. Multiplication of the bisulfite-treated DNA by using for example polymerase chain reaction (PCR) converts the uracils to thymine. When the resulting DNA fragments are sequenced, the presence of a cytosine indicates that this particular cytosine was methylated (the unmethylated cytosines have been replaced by thymines).

fairly complex and the reproducibility is far from perfect, especially for low-abundant proteins (Fonslow *et al.*, 2011). It can also not rival the massive throughput that can be achieved in a sequencing experiment.

Now that we know how to measure gene expression, we can move on to DNA methylation. The methods that we used in this thesis to measure genome-wide DNA methylation profiles are actually quite similar to the microarray and RNA sequencing techniques we just described. The Illumina Infinium 450k Human Methylation Assay was the specific microarray we used in one of our studies. The underlying principle is similar to the one behind an expression microarray: a glass slide is filled with probes (more than 450,000 probes in our case, as the 450k in the name indicates) whose sequences match the sequences of the genomic regions we are interested in. The binding of DNA fragments from our sample to these probes will give us information about the DNA methylation status of the corresponding genomic regions. We are dealing with DNA molecules, so there is no need for a reverse transcription step. However, we do need to treat the DNA with bisulfite (Figure 1.5). This ion converts the cytosines in a DNA molecule to uracil, but leaves the methylated cytosines largely intact (Frommer *et al.*, 1992). After bisulfite treatment the DNA in our sample is amplified and fragmented. During the amplification step all the uracils that arose from unmethylated cytosines are replaced by thymines. So in the end, the bisulfite treatment reduces the analysis of DNA methylation to a DNA sequence analysis. If we find a C in our sequence, we can assume that the original cytosine was methylated. If we find a T, the original cytosine was most likely not methylated.

The Infinium 450k microarray contains two different kinds of probes. The first type consists of two probes per genomic location, one for a methylated and one for an unmethylated fragment. These probes are designed in such a way that the exact genomic location or locus we are interested in lies at the end of the probe. So probes for the unmethylated loci end with an A (to pair with the T created by the bisulfite treatment) while the probes for the methylated loci end with a G (to pair with the methylated and therefore intact C). Once fragmented, DNA

sequences are allowed to hybridize to the probes after which a mix of labeled ddNTPs is added to the array for a single-base extension of the probe. If the hybridization was perfect (meaning that a fragment with a methylated locus bound to the correct probe and not to the probe for the unmethylated locus, and vice versa), adding a single base to the probe sequence will work; if not, no base can be added. After the staining of the labeled ddNTPs, a scanner is used to examine the intensities of the different probes, which will tell us whether a certain locus was methylated or not.

The second type of probe on the microarray consists of a single probe for both methylated and unmethylated locations. This probe misses its last base (the base that would otherwise pair with the locus of interest), so when we add the labeled ddNTPs after hybridization of our fragment we can tell whether the locus was methylated or not by looking at which base was added to the probe (an A means it was unmethylated, a G means it was methylated). This microarray analysis gives us the methylation status for over 450,000 loci in our genome. Even though this is a lot, we are bound to the loci that are present on the array. If you want to look at the methylation status of other regions, you will have to turn to alternative methods. Several other techniques, such as the expensive but truly whole-genome bisulfite sequencing, are also based on the use of bisulfite treatment (combined with the sequencing of the treated DNA fragments). Reduced representation bisulfite sequencing or RRBS, developed by Meissner *et al.* in 2005, offers a cheaper alternative to whole-genome bisulfite sequencing. The technique is based on the use of a restriction enzyme, which is a protein that recognizes a specific DNA sequence and that cuts the DNA at or near this sequence. By using a methylation-insensitive restriction enzyme that recognizes a sequence with a CpG dinucleotide in it (such as MspI, which targets the CCGG sequence), we can cut the genomic DNA into fragments that contain a CpG at each end. These fragments then go through several processing steps, including size selection and bisulfite conversion, and are finally mapped back to the genome so we can determine which genomic locations were methylated. The use of a restriction enzyme reduces the amount of nucleotides to roughly 1% of the genome, which is why RRBS is cheaper than its whole-genome counterpart. The lower price tag comes at the cost of missing some CpGs, though RRBS still covers more than three quarters of promoter regions as well as most CpG islands (Gu *et al.*, 2011).

A different approach we used in this thesis employs an enrichment step instead of a bisulfite treatment. By using the methylated-cytosine-binding domain (MBD) of a naturally occurring protein we can filter out all methylated DNA fragments from a mixture of methylated and unmethylated fragments (Serre *et al.*, 2010). By sequencing these selected fragments and mapping them back to the reference genome (comparable to an RNA-seq experiment) we can find out which regions of the genome were methylated. This particular method is known as MBD-seq, but several variations exist depending on the type of molecule that is used to enrich

for methylated DNA fragments.

Now that we have introduced you to some of the techniques that can be used to measure gene expression and DNA methylation, it is time to see them in action!

Adapted from this reviewed, but not resubmitted manuscript:

Koch A, Van Damme P, Gevaert K, Trooskens G, Van Criekinge W, De Meyer T & Menschaert G. Measuring the genome-wide impact of DNA methylation at the proteome level in a DNMT knockout human cancer cell model. *J Proteome Res* 2014

After the initial submission to Journal of Proteome Research and subsequent review process, we decided to not resubmit. Instead we plan on integrating these results with RNA and ribo-seq data of the same cell lines. These analyses are currently ongoing and a manuscript is in preparation.

MEASURING THE GENOME-WIDE IMPACT OF DNA METHYLATION AT THE PROTEOME LEVEL IN A DNMT KNOCKOUT HUMAN CANCER CELL MODEL

ABSTRACT

DNA methylation is a crucial epigenetic process involved in embryonic development and cellular differentiation and the impact of specific changes in DNA methylation has been intensely investigated in a large number of diseases. In this study we aimed to investigate the link between DNA methylation and protein expression, both qualitative and quantitative, on a respectively genome and proteome-wide level. Genome-wide DNA methylation profiling was performed using reduced representation bisulfite sequencing (RRBS), while quantitative shotgun and positional proteomics (N-terminal COFRADIC) were used to obtain protein expression data. These methodologies were applied on a wild type HCT116 cell line and a double DNMT knockout HCT116 cell line (DNMT1 -/- and DNMT3b -/-). We found that there were significantly more up-regulated than down-regulated genes in the knockout cells and that these up-regulated genes were characterized by higher levels of promoter methylation in the wild type cells. The experiments also resulted in a list of previously unannotated translation start sites and hinted at the possible use of a methylation-controlled alternative promoter for certain genes. Together, these results confirm the inhibitory effect of promoter methylation on protein expression and suggest a possible role for DNA methylation in alternative promoter control.

INTRODUCTION

DNA methylation is a vital epigenetic process that regulates gene expression. It is involved in embryonic development (Jaenisch & Bird, 2003), parental imprinting (Swain *et al.*, 1987), X chromosome inactivation (Mohandas *et al.*, 1981) and cellular differentiation (Laurent *et al.*, 2010); it provides a template for the chromatin structure (Cedar *et al.*, 2009) and it protects a cell against the expression of viral genes (Jahner *et al.*, 1982). Promoter hypermethylation inhibits gene expression and aberrant DNA methylation is a common feature of virtually every human cancer (Jones & Baylin, 2007, Esteller, 2007). On top of that, at least 90% of the protein-coding genes use alternative transcription and alternative splicing events (Pal *et al.*, 2012), both of which have been linked to DNA methylation. Indeed, Maunakea *et al.* (2010) showed that intragenic DNA methylation plays a role in regulating alternative promoter usage and promoter switching (Maunakea *et al.*, 2010). Increased methylation has also been found at alternatively spliced sites (Anastasiadou *et al.*, 2011, Maunakea *et al.*, 2013); exon recognition is promoted by the recruitment of MeCP2 (methyl CpG binding protein 2) to methylated alternatively spliced exons (Maunakea *et al.*, 2013) and methylation of CTCF binding sites inhibits the binding of the transcription factor CTCF to these sites, which would otherwise lead to the inclusion of weak upstream exons through

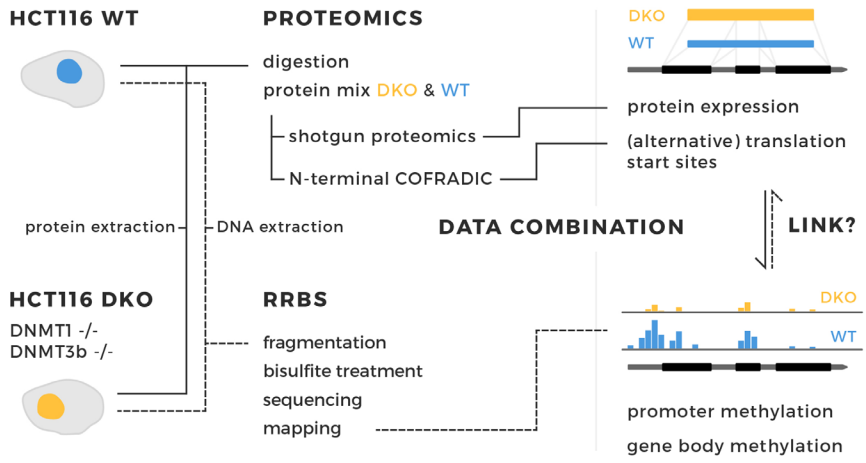


FIGURE 1.3 Experimental overview and data interpretation.

Both the effect of promoter methylation on protein expression and the possible effect of gene-body methylation on translation initiation were investigated using a combination of RRBS, differential shotgun proteomics and N-terminal COFRADIC on two HCT116 colon cancer cell lines: a wild type line (WT) and a double knockout line (DKO). In the DKO line the DNA methyltransferases DNMT1 and DNMT3b are knocked out, removing most DNA methylation.

PolII stalling (Shukla *et al.*, 2011). These findings demonstrate that intragenic DNA methylation might be an important regulator of alternative splicing. Besides the increased molecular understanding of the link between DNA methylation and gene expression, recent research has led to a whole new arsenal of epigenetic biomarkers and potential therapies (Claes *et al.*, 2010).

The effects of DNA methylation on transcription have been described numerous times (Siegfried *et al.*, 1999, Miranda & Jones, 2007, Ball *et al.*, 2009). Evidence of this effect measured on the actual protein level, on the other hand, is scarcer. One advantage of using proteomics techniques for the analysis of gene expression is that compared to RNA sequencing or microarray analysis, protein analysis offers a more direct view on gene expression as most gene-encoded biological functions are controlled and performed by proteins. Translation, and not transcription, has also been identified as the single largest contributor to protein abundance (Schwanhausser *et al.*, 2011). Gry *et al.* (2009) reported differences between expression at the transcript and the protein level, which they attributed to the effect of translational and post-translational modifications. Positional proteomics as used in this study also helps in the discovery of alternative translation initiation events (Menschaert *et al.*, 2013, Van Damme *et al.*, 2014).

Previously, Tang *et al.* (2010) used proteome profiling by 2-D gel electrophoresis (coupled to mass spectrometry) to investigate the effect of 5-aza-2'-deoxycytidine (DAC), a DNMT1 inhibitor, on protein expression in the acute myeloid leukemia HL-60 cell line, thereby identifying 35 candidates. However, the methylation levels of those genes whose protein expression level changed after the DAC treatment were not measured. Another study looked at the association between hepatitis B virus (HBV) modulated DNA methylation and the protein profile of hepatocellular carcinoma cells (Niu *et al.*, 2009). A total of 15 genes silenced by HBV-mediated DNA methylation and reactivated after DAC treatment were identified. Promoter methylation of the identified genes was measured using bisulfite treatment and methylation specific PCR. Xu *et al.* (2013) employed a combination of 2-D gel electrophoresis and MS/MS analysis to find differentially expressed genes in mice testis after exposure to cigarette smoke. They found 31 proteins and for one of these proteins, PEBP1, they examined the methylation status using both bisulfite sequencing and methylation sensitive PCR.
None of these studies offered a genome-wide DNA methylation analysis, unlike the work presented here. In fact, very few papers have been published in which the authors describe the integration of genome-wide protein expression and DNA methylation data. One example is the work by Orozco *et al.* (2015) where bisulfite sequencing was correlated with mass spectrometry data from mouse livers. In our study we aimed to investigate both the effect of promoter methylation at the translational level and the possible effect of gene-body methylation on translation initiation (Figure 1). Reduced representation bisulfite sequencing (RRBS) (Meissner *et al.*, 2005) was used in combination with shotgun and positional

proteomics (N-terminomics) to determine the DNA methylation and protein expression profiles of the HCT116 DNA methyltransferase (DNMT) knockout model (DNMT1 -/- and DNMT3b -/-). Throughout the manuscript we use the terms up and down-regulated when discussing the differences in expression between the DKO and WT cell lines. We use them to refer to the differences in abundance between the two cell lines (up-regulated protein = more abundant in DKO than in WT, down-regulated protein = less abundant in DKO than in WT) and not to refer to any active process.

DNMTs are responsible for de novo DNA methylation (DNMT3a and DNMT3b) (Yoder *et al.*, 1997) and the maintenance of DNA methylation (DNMT1) (Hsieh, 1999). DNMT2 is a fourth, widely conserved DNMT gene. Though it has been linked to tRNA methylation, relatively little is known about its biological role (Goll *et al.*, 2006, Schaefer *et al.*, 2010). Inactivation of DNMT1 and DNMT3b causes the HCT116 cells to lose most of their DNA methylation (Rhee *et al.*, 2002). However, as DNMT3a has not been knocked out in our HCT116 model, these cells still maintain a minimal level of methylation.

RRBS uses a methylation-insensitive restriction enzyme (such as *Bgl*II or *Msp*I) to fragment the genome. These fragments are then treated with bisulfite, which deaminates unmethylated cytosines to uracil. Finally, the fragments are sequenced and mapped back to the reference genome, revealing which cytosines were methylated. The resulting methylation data is expressed as the percentage methylated reads of all the reads that mapped to a certain CpG dinucleotide.

The analysis of a cell's complete set of proteins, or proteome, and the difference between two proteomes requires robust and sensitive techniques such as quantitative shotgun proteomics. Combining shotgun proteomics with metabolic labeling techniques like SILAC (Stable Isotope Labeling by Amino acids in Cell culture, Ong *et al.*, 2002) allows for differential protein expression profiling when comparing cell lines or states using mass spectrometry. Next to the shotgun proteome analysis, the protein content of the HCT116 WT and DKO cell lines was also analyzed by means of positional proteomics. More specifically, N-terminal COFRADIC (COmbined FRActional DIagonal Chromatography, Staes *et al.*, 2011) was used, which, unlike shotgun proteomics, enriches for protein N-terminal peptides by discarding internal peptides (i.e. negative selection of protein N-termini), thus revealing the (alternative) translation initiation landscape. Herein lies, next to a more direct view on the true expression levels as described above, one of the major advantages of the proteomics techniques over commonly used RNA-sequencing methods.

This study presents a novel approach to the genome-wide analysis of the effect of DNA methylation on protein expression by coupling RRBS to shotgun and positional (N-terminal) proteomics (Figure 1.3).

MATERIAL AND METHODS

CELL LINE CULTIVATION

Both HCT116 colorectal carcinoma cell lines, the wild type (WT) and double knockout (DKO, DNMT1 -/-, DNMT3b -/-), were kindly provided by the Johns Hopkins Sidney Kimmel Comprehensive Cancer Center (Baltimore, USA). The DNMT1 and DNMT3b alleles were disrupted using homologous recombination as described by Rhee *et al.* (2002). The HCT116 cells were grown in DMEM medium supplemented with 10% fetal bovine serum (Invitrogen, Carlsbad, CA, USA), 100 units/ml penicillin (Invitrogen) and 100 μg/ml streptomycin (Invitrogen) in a humidified incubator at 37°C and 5% CO2. For the N-terminal COFRADIC analysis (Figure 1.4a), the HCT116 cells were SILAC labeled (Ong *et al.*, 2002) and grown in DMEM medium containing natural (DKO) or $^{13}C_6{}^{15}N_4$ L-arginine (WT)

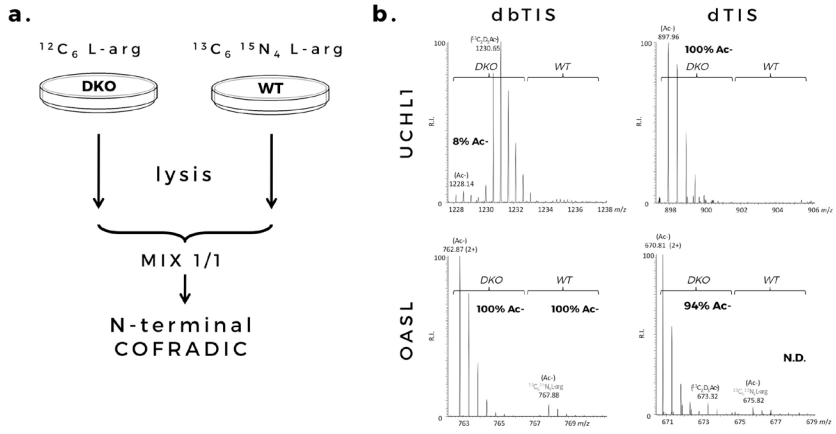


FIGURE 1.4 N-terminal COFRADIC analysis.

a. Experimental setup used for the N-terminome analysis. HCT116 DKO and WT cells were cultivated in $^{12}C_6$ L-arginine and $^{13}C_6{}^{15}N_4$ L-arginine-containing medium respectively. Cells were lysed and their isolated proteomes subjected to N-terminal COFRADIC. **b**. Representative MS spectra of N-termini hinting to unique or up-regulated protein expression in DKO cells. MS spectra of the database annotated (dbTIS) and database unannotated (dTIS) N-terminus of the Ubiquitin carboxyl-terminal hydrolase isozyme L1 protein UCHL1 (^{1}MQLKPMEINPEMLNKVLSR$_{19}$ and ^{6}MEINPEMLNKVLSR$_{19}$; upper two panels) and the 59 kDa 2'-5'-oligoadenylate synthase-like protein OASL (^{2}ALMQELYSTPASR$_{14}$ and ^{4}MQELYSTPASR$_{14}$; lower two panels) are shown. The dbTIS and dTIS-indicative peptides of UCHL1 were respectively partially N-terminally (Nt) acetylated (8%) and fully Nt-acetylated (100%) and hinted at exclusive re-expression in the DKO setup, while the dbTIS and dTIS-indicative peptides of OASL were respectively fully Nt-acetylated and partially Nt-acetylated (94%) and hinted to significant up-regulation in the DKO setup.

(Cambridge Isotope Labs, Andover, MA, USA) at a concentration of 140 μM (i.e. 35% of the normal L-Arg concentration in DMEM) at which arginine to proline conversion was not detected. For the shotgun proteome analyses, HCT116 WT and DKO cells were grown in medium supplemented with natural or $^{13}C_6$ L-arginine (final concentration 140 μM), and natural or $^{13}C_6$ L-lysine (final concentration 800 μM). Media were supplemented with 10% dialyzed fetal bovine serum, 100 units/ml penicillin and 100 μg/ml streptomycin. Cell populations were cultured for at least 6 population doublings.

RRBS

For each cell line, 1 μg genomic DNA was digested overnight using the MspI restriction enzyme (TrueMethyl RRBS protocol) and the resulting fragment mix was purified using the GeneJET PCR purification kit. Next, Illumina adapters were added to the fragments (protocol for use with NEBNext Ultra DNA library prep kit for Illumina), which were then treated with bisulfite using the EZ DNA methylation gold kit (Zymo Research). After PCR amplification and purification, the paired-end (2 × 50 bp) libraries were sequenced on the Illumina HiSeq 2000 platform and finally mapped to the human genome (GRCh37) with Bismarck (Krueger *et al.*, 2011).

SHOTGUN PROTEOME ANALYSES

Cells were lysed in 20 mM NH_4CO_3 (pH 7.9) by three rounds of freeze-thawing. The protein concentration of the cell extracts was measured using Biorad's Protein Assay (Biorad Laboratories, Munich, Germany) and equal amounts of protein material (1 mg each) were mixed. During tryptic digestion, guanidinium hydrochloride (final concentration (f.c.) 0.5 M) and acetonitrile (f.c. 2%) were added. Protein material was digested overnight with sequencing-grade modified trypsin (Promega, Madison, WI, USA; enzyme/substrate, 1/50 (w/w)).

Methionines were uniformly oxidized to sulfoxides prior to RP-HPLC fractionation by adding 20 μl of 3% (w/v) H2O2 to 100 μl sample (f.c. of H_2O_2 was 0.06%) for 30 min at 30°C. 100 μl of this peptide mixture (500 μg peptide material) was subsequently injected onto the RP-column (Zorbax® 300SB-C18 Narrow-bore, 2.1 mm (internal diameter, ID) × 150 mm, 5 μm particles, Agilent). Following 10 min isocratic pumping with solvent A (10 mM ammonium acetate in water/acetonitrile (98:2 v/v), pH 5.5), a gradient was started of 1% solvent B increase per minute (solvent B: 10 mM ammonium acetate in acetonitrile/water (70:30 v/v), pH 5.5). The flow was kept constant at 80 μL/min using Agilent's 1100 series capillary pump with the 100 μL/min flow controller. Fractions of 30 s wide were collected from 20 to 80 min after sample injection. To reduce the LC-MS/MS

analysis time fractions eluting 12 min apart were pooled, resulting in a final set of 24 samples that were vacuum dried and re-dissolved in 20 µl of 20 mM tris(2-carboxyethyl)phosphine (TCEP) in 2% acetonitrile and analyzed by LC-MS/MS.

N-TERMINAL COFRADIC

WT and DKO cells were lysed in 50 mM HEPES pH 7.4, 100 mM NaCl, 0.8% CHAPS and protease inhibitors for 10 min on ice and centrifuged for 15 min at 16,000 g at 4°C. Protein concentrations were measured and equal amounts of protein material (2 mg each) were mixed. This sample was subsequently subjected to N-terminal COFRADIC as described by Staes *et al.* (2011) in order to enrich for N-terminal peptides. For the primary RP-HPLC separation, the equivalent of 1000 µg digested peptide material (before SCX fractionation) was injected onto the RP-column.

LC-MS/MS ANALYSIS

The peptide mixtures obtained from the shotgun proteome samples were introduced into an LC-MS/MS system, the Ultimate 3000 RSLC nano (Dionex, Amsterdam, the Netherlands) in-line connected to an LTQ Orbitrap Velos (Thermo Fisher Scientific, Bremen, Germany), for peptide identification. The sample mixture was loaded on a trapping column (made in-house, 100 µm ID × 20 mm, 5 µm beads C18 Reprosil-HD, Dr. Maisch). After back-flushing from the trapping column, the sample was loaded on a reverse-phase column (made in-house, 75 µm ID × 150 mm, 5 µm beads C18 Reprosil-HD, Dr. Maisch). Peptides were loaded in solvent A' (0.1% trifluoroacetic acid and 2% acetonitrile) and separated with a linear gradient from 2% solvent A'' (0.1% formic acid) to 50% solvent B'' (0.1% formic acid and 80% acetonitrile) at a flow rate of 300 nl/min followed by a wash reaching 100% of solvent B''. The mass spectrometer was operated in data-dependent mode, automatically switching between MS and MS/MS acquisition for the ten most abundant peaks in a given MS spectrum. From the MS/MS data in each LC run Mascot Generic Files were created using Distiller software (version 2.3.2.0).

The generated MS/MS peak lists were searched with Mascot (Perkins *et al.*, 1999) using the Mascot Daemon interface (version 2.3.01, Matrix Science). Searches were performed in the Swiss-Prot database with taxonomy set to human (UniProtKB/SwissProt database versions 2011_05). For the shotgun analysis, methionine oxidation to methionine-sulfoxide was set as fixed modification while pyroglutamate formation of N-terminal glutamine and acetylation of the protein N-terminus were selected as variable modifications. Mass tolerance on precursor ions was set to 10 ppm (with Mascot's C13 option set to 1) and on fragment ions to 0.5 Da.

The peptide charge was set to 1+, 2+, 3+ and the instrument setting was set to ESI-TRAP. Trypsin/P was selected as the enzyme setting, 1 missed cleavage was allowed and cleavage was also allowed when arginine or lysine was followed by proline.

The N-terminal COFRADIC samples were analyzed on the LTQ Orbitrap XL (Thermo Fisher Scientific, Bremen, Germany). The mass spectrometer was operated in data-dependent mode, automatically switching between MS and MS/MS acquisition for the six most abundant peaks in a given MS spectrum. The generated MS/MS peak lists were searched with Mascot using the same parameters as described above but this time $^{13}C_2D_3$-acetylation on lysines, carbamidomethylation of cysteine and methionine oxidation to methionine-sulfoxide were set as fixed modifications. Variable modifications were $^{13}C_2D_3$- acetylation and acetylation of peptide N-termini together with pyroglutamate formation of N-terminal glutamine. The reason we used the heavy $^{13}C_2D_3$-acetylation was to both distinguish between in vitro and in vivo N-terminal acetylation and to allow for the quantification of the degree of N-terminal acetylation. Endoproteinase semi-Arg-C/P (Arg-C specificity with arginine-proline cleavage allowed) was set as enzyme allowing for no missed cleavages. Quantification of the degree of Nt-acetylation was performed as described previously (Van Damme *et al.*, 2011).

All mass spectrometry data were converted using the PRIDE Converter (Barsnes *et al.*, 2009) and have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository (Vizcaino *et al.*, 2013) with the dataset identifier PXD000304 and DOI 10.6019/PXD000304 (http://www.ebi.ac.uk/pride/archive/login, PX reviewer account: username: review48267, password: TTewpyNH).

BIOINFORMATICS AND STATISTICAL ANALYSES

All quantifications ($^{12}C_6$ L-arginine versus $^{13}C_6{}^{15}N_4$ L-arginine) were carried out using the Mascot Distiller Quantitation software (version 2.2.1). Ratios for the peptides were calculated by comparing the XIC peak areas of all matched light versus heavy peptides and all ratios were verified by visual inspection of the MS spectra.

In the shotgun experiment, the methods of robust statistics (Huber, 1981) were applied to the base-2 logarithms of the ratios of the identified peptides to arrive at the protein expression levels. In summary, to identify statistically significant up or down-regulated proteins, a reference set was created using the complete set of true Mascot Distiller peptide ratio values. This resulted in a Huber estimated distribution with a 98% confidence interval (CI) between log2 ratio values 0.47

and 2.14. Only proteins with a ratio outside the 98% CI and that were identified by at least two peptides were initially retained for further analyses. The remaining proteins with only a single peptide were then manually investigated and the best identifications were also retained. For the N-terminal COFRADIC data, we also applied the methods of robust statistics to the log2 ratios of the identified peptides. Both in the shotgun and COFRADIC experiment, peptides or proteins that displayed a ratio reflecting significant up or down-regulation (i.e., outside the 95% CI, $p \leq 0.05$) were considered affected by the double DNMT1 -/-, DNMT3b -/- knock out. Alternative translation start sites were defined as those start sites that do not map to a canonical human Swiss-Prot annotated start site (UniProtKB/SwissProt database version 2011_05).

Custom Perl (version 5.18.2) scripts that use the Biomart API and the Ensembl database (release 75) were created to map the peptides to the human genome (version GRCh37.75). Peptides were mapped by linking their accession and UniProt IDs to Ensembl gene IDs using the Biomart API. Some of the peptides of the N-terminal COFRADIC experiment overlapped and mapped to position 1, 2 and/or 3 (counting from the translation start site) of the same protein. In this case, we aggregated these peptides into a single peptide at position 1 and set its peptide ratio to the mean ratio of the combined peptides. We considered all peptides that mapped to position 1 as having a database-annotated translation initiation site (dbTIS) and the others as having a downstream translation initiation site (dTIS).

A promoter was defined as the region from 1 kb upstream to 200 bp downstream of the transcription start site of the consensus coding sequence. For peptides with an alternative translation start site, the corresponding putative alternative promoter was chosen to range from 1 kb upstream to 200 bp downstream of the start site of the exon that encoded the mapped peptide. Promoter methylation was measured as the mean percentage methylation of all CpGs within the promoter region measured by RRBS.
Data processing and statistical analyses were performed in the R environment (R version 3.1.2, statistical tests used are the proportion test (prop.test) and Kruskal-Wallis test (kruskal.test), both from the stats package).

RESULTS

RRBS

Combining the DKO and WT RRBS data gave us the methylation status of 2,544,642 CpGs. The double knockout of DNMT1 and DNMT3b resulted in a global demethylation as illustrated by the distributions of the RRBS methylation

values in the WT and DKO cell lines (Figure 1.5, Wilcoxon rank sum test (WRT), p value < 2.2e-16, 95% CI = 47.167, 47.368). For the easy comparison with the protein expression data, we created a single promoter methylation value for each protein by calculating the mean methylation of the covered CpGs within the promoter. Using the maximal value was also considered, but resulted in more noise in the data, especially for the DKO cell line (Supplementary Figure 1.1).

SHOTGUN PROTEOMICS

A single shotgun proteome analysis led to the identification of 3,308 unique proteins of which 3,302 were successfully mapped to the human genome. As not all protein identifiers are linked to a gene ID, not all proteins could be mapped on the genome through BioMart. Of the mapped proteins, 244 were significantly up-regulated in the HCT116 DKO cell line and 76 were down-regulated (both $p \leq 0.05$, Table 1.3). The number of up-regulated proteins was significantly higher than the number of down-regulated proteins (proportion test, p value < 2.2e-16, 95% CI = 0.711, 0.807), hinting at the repressive effect of DNA methylation on expression. We used the shotgun protein identifications represented by at least two peptides as well as the manually curated identifications represented by a single peptide for the stringent inspection of up and down-regulated proteins. The promoter regions of the up-regulated proteins were more methylated than the down-regulated proteins in the WT cell line (WRT, p value = 0.0129, 95% CI = -0.948, -4.66e-5) and they were more demethylated in the DKO line (WRT, p value = 0.00429, 95% CI = 5.71e-5, 0.0361). When comparing the promoter methylation in the DKO cells between the up and down-regulated proteins, there was no significant difference (WRT, p value = 0.0529, 95% CI = -0.436, 6.47e-5).

For the 119 up-regulated proteins of which the corresponding genes were demethylated in their promoter region in the DKO cell line (difference between DKO and WT mean promoter methylation > 10), the expression was classified as being under direct methylation control. We defined a demethylated promoter as having a difference in the mean methylation value between DKO and WT > 10 based on the distribution of the methylation differences between DKO and WT (Supplementary Figure 1.2). We kept this cutoff relatively low, because we already reduced the greatest differences between WT and DKO by calculating the mean promoter methylation. Expression of some of these proteins was already known to be under methylation control and has been linked to cancer development. Examples include EML2 (Duong *et al.*, 2012), PLK1 (Ward *et al.*, 2015), ICAM1 (Schuebel *et al.*, 2007, Easwaran *et al.*, 2010) and TRIP10 (Hsiao *et al.*, 2010). Up-regulated proteins whose promoters were not demethylated in the DKO line were considered to be under indirect methylation control.

Interestingly, amongst the up and down-regulated proteins, various interactions

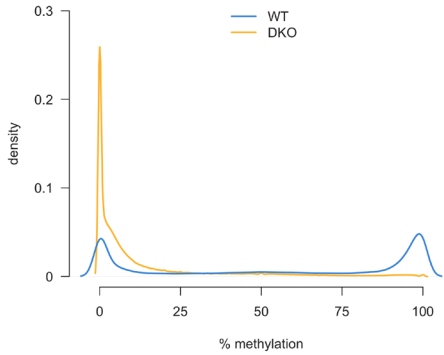


Figure 1.5 Distribution of the methylation data generated by RRBS for both the WT and the DKO cell line.

The difference in distribution of the WT and DKO methylation data illustrates the successful demethylation in the DKO cells.

networks were affected. Supplementary Figure 3 shows a string-db (http://string-db.org) (Franceschini *et al.*, 2013) interaction network of the significantly up and down-regulated proteins and illustrates the complexity of the affected networks. Examples include the immunity-linked TNF/NF-kappaB (NFKB2) and the ubiquitin-mediated degradation pathway (UCHL1) and networks implicated in cell adhesion (ICAM1). Furthermore, the DAVID tool (Huang *et al.*, 2009) was used to analyze the gene ontology enrichment in the up or down-regulated proteins. The best scoring gene ontologies included cytoskeleton organization (UBE2C, CORO1A, TRIP10…) and antigen processing and presentation (ICAM1, CALR, HLA-B…) (Supplementary Table 1.1).

N-TERMINAL COFRADIC

N-terminal COFRADIC was used to map the (differential) translation initiation landscapes of the DKO versus WT cells, resulting in the identification of 1,530 unique N-terminal peptides in total (a combination of peptides with

TABLE 1.3 The number of proteins and peptides up and down-regulated in the DKO cell line identified by shotgun proteomics and N-terminal COFRADIC experiments (DKO vs. WT).

N-terminal peptides identified by means of COFRADIC are divided into a dbTIS (database-annotated TIS) and dTIS (downstream TIS) group.

	shotgun	N-terminal COFRADIC		
		dbTIS	dTIS	total
up-regulated	244	78	18	96
down-regulated	76	31	23	54
total	320	109	41	150

database-annotated translation initiation sites (dbTIS) and peptides with downstream translation initiation sites (dTIS)). Of these, 1,525 N-termini were successfully mapped onto the human genome. Further filtering of the mapped peptides (cf. Materials and Methods) resulted in a final list of 1,283 peptides. Table 1.3 summarizes the numbers of significantly up and down-regulated peptides after DNMT1 and DNMT3b knockout according to their TIS grouping, dbTIS or dTIS. Figure 1.4b shows two representative MS spectra of N-termini that were differentially expressed in the DKO cell line compared to the WT line.

One of the features of the N-terminal COFRADIC technology is its ability to identify N-termini pointing to (alternative) translation initiation sites (Menschaert *et al.*, 2013, Van Damme *et al.*, 2014). Although ribosome leaky scanning and alternative splicing likely cause the majority of these N-terminal protein isoform identifications, the detection of alternative N-termini also enables us to analyze the influence of gene-body methylation on alternative promoter usage. Overall, our N-terminal COFRADIC experiment resulted in a set of 211 dTIS peptides, of which 18 were up and 23 down-regulated (Table 1.3). Interestingly, one of these dTIS peptides ($M\downarrow A_{123}$DANSPPKPLSKPR, found to be 100% Nt-acetylated) mapped to exon 4 of DNMT1 and was uniquely identified in the WT cell line. Exons 3, 4 and 5 of DNMT1 were disrupted by homologous recombination in the DKO cell line, so our data confirmed the successful knockout of DNMT1.

The number of dbTIS peptides up-regulated in the DKO cell line was higher than the number of down-regulated dbTIS peptides (proportion test, p value = 1.053e-5, 95% CI = 0.620, 0.796). For the dTIS peptides no significant difference was found (proportion test, p value = 0.532, 95% CI = 0.288, 0.601). To determine if the expression of dTIS peptides was under methylation control, we made the same comparisons for the dTIS peptides as we did for the proteins from the shotgun experiment. The promoter regions of the up-regulated dTIS peptides were not more demethylated in the DKO line compared to the down-regulated dTIS peptides (WRT, p value = 0.179, 95% CI = -7.19e-6, 34.130). Despite the lack of a statistically significant difference in demethylation between the up and down-regulated dTIS peptides, there still was a notable difference in methylation between WT and DKO for the up-regulated dTIS peptides (paired WRT, p value = 0.00253, 95% CI = 27.112, 63.972) (Figure 1.6). As we were interested in the possible influence of DNA methylation on the expression of alternative translation products, we tried to determine if the expression of any dTIS peptides could have been under methylation control. We aimed to identify those dTIS peptides that were both demethylated (difference between DKO and WT > 10) and up-regulated in the DKO cells. We identified 11 peptides, which mapped to the following genes: PTPN18, PLEKHG3, OASL, EPS8L2, SARG, ZNF511, STAU1, FPGS, DES, CCDC88C and BAIAP3. The alternative promoter regions surrounding the genomic location of these dTIS peptides were analyzed with the FlyBase eukaryotic promoter prediction computational tool (http://www.fruitfly.org/seq_tools/

promoter.html) (Reese *et al.*, 2001) (Table 1.4). Furthermore, using data available through the UCSC genome browser (http://genome-euro.ucsc.edu/) (Kent *et al.*, 2002), the alternative promoter regions were inspected for the presence of CpG islands and the occurrence of trimethylated histone H3 lysine 4 (H3K4me3, Table 1.4) (Santos-Rosa *et al.*, 2002). The genomic regions surrounding the EPS8L2, PTPN18 and ZNF511 dTIS peptide locations showed the most promoter-like characteristics and EPS8L2 was visualized using the UCSC genome browser (Figure 1.7).

DISCUSSION

In this study a combination of genome-wide DNA methylation profiling and proteome analyses using a quantitative shotgun and an N-terminal positional proteomics strategy was used to study the impact of DNA methylation on protein expression and (alternative) translation initiation in a DNMT knockout model (WT and DKO (DNMT1 -/- and DNMT3b -/-) HCT116 colon cancer cell lines). The DNMT1 dTIS N-terminus M↓A_{123}DANSPPKPLSKPR, which maps to isoforms 1 and 2 of DNMT1, was uniquely found in the WT cell line in the N-terminal proteomics data, thereby confirming the success of the knockout.
The shotgun proteomics data were searched for proteins up-regulated in the DKO cell line. Loss of DNA methylation in this cell line resulted in a set of significantly up or down-regulated proteins, including proteins that were uniquely expressed in either the DKO or WT cell line. The number of up-regulated proteins was significantly higher than the number of down-regulated proteins. Together with

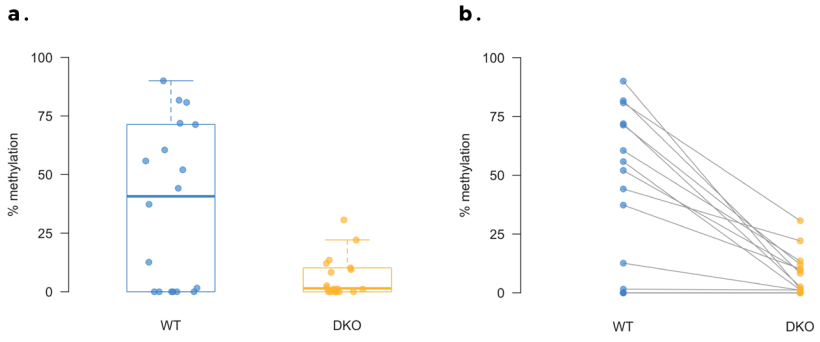


Figure 1.6 Comparison of the promoter methylation for the 18 up-regulated dTIS peptides.

Both the boxplot (**a**) and the scatter-and-line plot (**b**) show the demethylation in the DKO cell line compared to the WT cell line.

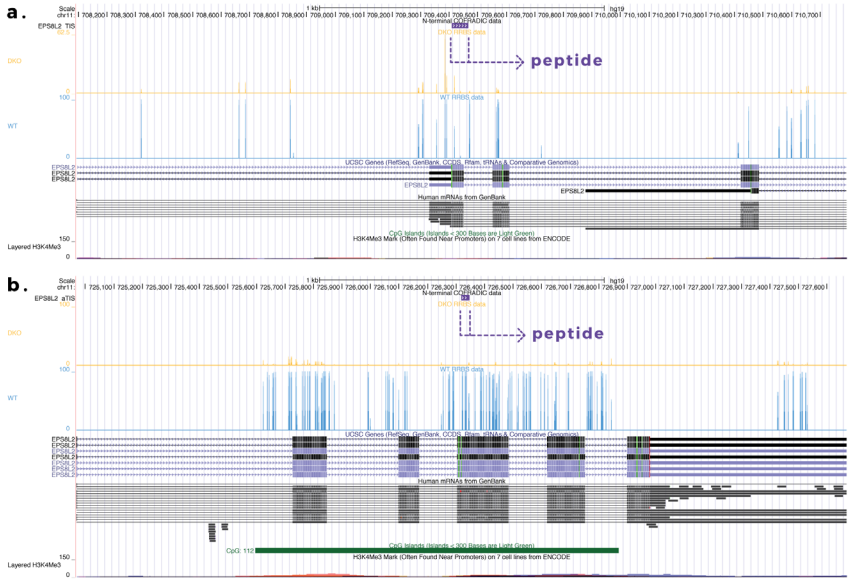


Figure 1.7 Visualization of the RRBS and N-terminal COFRADIC data for EPS8L2.

The different tracks in both **a** (the dbTIS peptide that was identified) and **b** (the dTIS peptide) are from top to bottom: the dTIS/dbTIS peptide, RRBS data for the DKO cell line, RRBS data for the WT cell line, UCSC genes, human mRNA from GenBank, CpG islands and H3K4me3 data. The plots show how the level of methylation in the alternative promoter for the mapped dTIS peptide is reduced in the DKO cell line as compared to the WT cell line, more than for the dbTIS peptide, which was not significantly up-regulated in the DKO cells.

our finding that the promoter regions of these up-regulated proteins were more demethylated than those of the down-regulated or not differentially expressed proteins, this confirmed the inhibitory effect of promoter methylation on gene expression. Some of these genes, such as EML2 (Duong *et al.*, 2012), PLK1 (Ward *et al.*, 2015), ICAM1 (Schuebel *et al.*, 2007, Easwaran *et al.*, 2010) and TRIP10 (Hsiao *et al.*, 2010) have previously been shown to be silenced in different cancer types through promoter methylation, which is reflected in our results, as the expression levels of these genes were higher in the DKO line than in the WT line, while the levels of promoter methylation were lower.

The list of up or down-regulated proteins was also analyzed for gene ontology enrichment and the best scoring gene ontologies included antigen processing and presentation and cytoskeleton organization. DNA hypomethylation has already been linked to the up-regulation of immune related pathways, including antigen processing, in several cancer studies (Wrangle *et al.*, 2013, Li *et al.*, 2014) and to cytoskeletal changes during prostate cancer progression (Schulz *et al.*, 2007).

TABLE 1.4 Analysis of promoter-like properties.

Possible alternative promoter regions were searched for typical promoter characteristics. The demethylation value was calculated by subtracting the mean promoter methylation in WT from the mean promoter methylation value in DKO.

gene	CpG island	H3K4me3	promoter predicted	expression ratio	demethylation
PTPN18	yes	yes	yes	3.49	-11.47
PLEKHG3	no	yes	yes	3.21	-46.94
OASL	no	yes	yes	16.04	-42.51
EPS8L2	yes	yes	yes	41.39	-88.31
SARG	no	no	yes	DKO only	-22.05
ZNF511	yes	yes	yes	7.16	-54.53
STAU1	no	no	yes	2.99	-59.25
FPGS	no	yes	no	2.93	-69.30
DES	no	no	yes	4.69	-50.10
CCDC88C	no	yes	no	DKO only	-27.11
BAIAP3	no	no	no	7.31	-73.41

Reduced methylation also led to an up-regulation of several proteins involved in protein ubiquitination and proteasomal degradation and it has been described previously how DNMT inhibition leads to increased expression of STAT1 and STAT3 (Karpf *et al.*, 1999), two genes found among the up-regulated proteins in our experiment, together with other proteins linked to inflammation.

The relatively high number of up-regulated proteins without promoter methylation in the WT cell line (97 out of 244) could have a biological and a technical explanation. Some proteins will be up-regulated in DKO–even though they are not under promoter methylation control themselves–through the possible compensatory actions of the DKO cells to counteract the hypomethylation, indirect methylation control through transcription factors, signaling pathways or other complex interaction networks (as suggested by Supplementary Figure 1.3), other (post-) transcriptional or (post-) translational events and the possible DNA methylation independent function of DNMTs (Milutinovic *et al.*, 2003, Espada *et al.*, 2011). Some of the promoter methylation might simply not have been picked up by the RRBS experiment.

Another goal of this study was to search the N-terminal COFRADIC and RRBS data for evidence of a correlation between DNA methylation and alternative transcription reflected by (alternative) translation initiation site selection. A comparison of the number of up and down-regulated peptides revealed that there were significantly more up-regulated peptides in the group of dbTIS peptides, which is in accordance with the known inhibitory effect of promoter methylation

on gene expression and thus translation. For the dTIS peptides on the other hand, there was no difference in the number of up and down-regulated peptides. Even though the comparison of the demethylation of alternative promoters between up and down-regulated dTIS peptides failed to produce a statistically significant result, we did observe a difference in the promoter methylation of the up-regulated dTIS peptides between the DKO and WT cell lines (Figure 1.6). We would therefore be careful to dismiss the presence of a link between the expression and methylation of these peptides. Filtering the data for up-regulated dTIS peptides methylated in their alternative promoter returned 11 genes (PTPN18, PLEKHG3, OASL, EPS8L2, SARG, ZNF511, STAU1, FPGS, DES, CCDC88C and BAIAP3), hinting at the possible role of DNA methylation in the regulation of alternative transcription initiation in these cases. Especially the regions surrounding the dTIS site in PTPN18, EPS8L2 and ZNF511 demonstrated several promoter characteristics, such as the presence of a CpG island, trimethylation of lysine 4 of histone 3 and a predicted promoter sequence. Different isoforms of PTPN18 have been described (Gandhi *et al.*, 2005), but no previous mention of an alternative promoter was found. PTPN18 is a member of the protein tyrosine phosphatase family, a group of signaling molecules that regulate cell growth, differentiation, the mitotic cycle and oncogenic transformation (Hunter, 1995). EPS8L2 has been linked to the regulation of actin cytoskeleton remodeling (Offenhauser *et al.*, 2004), while the fusion protein ZNF511-PRAP1 acts as a transcription regulator (Qiu *et al.*, 2011). No mention of alternative promoter usage was found for these two genes either.

Despite certain shortcomings, such as undersampling (Hancock, 2007), proteomics techniques offer an advantage over RNA-sequencing as they measure (differences in) protein abundance. On top of that, N-terminomics is well suited for the identification of N-terminal protein variants or proteoforms (Smith & Kelleher, 2013) and concomitantly enables the detection of alternative translation start sites (Van Damme *et al.*, 2014, Helsens *et al.*, 2011).

CONCLUSION

In this study, we demonstrated how whole-genome methylation profiling and proteomics can be combined to study the influence of DNA methylation on expression at the protein rather than the transcript level. Our findings confirmed the inhibitory effect of promoter methylation on protein expression and hinted at the potential role of DNA methylation in alternative promoter activity.

ACKNOWLEDGEMENTS

We would like to thank Ellen De Meester and Sarah De Keulenaer for their help with the RRBS. K.G. acknowledges support from the Research Foundation - Flanders (FWO-Vlaanderen), project number G.0440.10. P.V.D. is a Postdoctoral Fellow of the Research Foundation - Flanders (FWO-Vlaanderen) and acknowledges support from FWO-Vlaanderen, project number G.0269.13N.

CHAPTER 2

ON RIBOSOMAL SEQUENCING

NEW KID ON THE BLOCK

In the previous chapter we introduced several techniques to measure gene expression, both at the transcript and the protein level, together with some of their advantages and disadvantages. Not only have scientists been working hard on improving these existing techniques, they have also been trying to come up with new ones. Ribosome profiling or ribo-seq is such a new technique (Ingolia, 2010, 2011). Developed at the Weissman lab in San Francisco in 2009, this method aims to combine the throughput, speed and dynamic range of RNA-seq with the biological relevance of measuring expression at the protein level found in proteomics. Ribo-seq uses next-generation sequencing techniques to find out which genes were expressed, just as in a typical RNA-seq experiment. But instead of sequencing the complete transcriptome, only the mRNA molecules that are bound to ribosomes are sequenced. These ribosome-bound mRNA molecules are in the process of being translated to proteins, so ribo-seq actually measures active protein synthesis at the transcript level and not just the amount of transcript that is present. This puts it somewhere halfway between measuring expression at the transcript level and measuring it at the protein level. Even though it is not the same thing as measuring protein levels (remember translational and post-translational regulation) it does come closer than RNA-seq.

Before we can isolate the mRNA molecules bound to ribosomes from the rest of the RNA in our sample, we have to stop translation. This can be achieved by treating the cells in our sample with an antibiotic like cycloheximide. Once we have successfully halted translation, we can add a nuclease to our sample. Nucleases are enzymes that can break the phosphodiester bond between two nucleotides and can therefore be used to dismantle DNA and RNA molecules. The nuclease we added to our sample will happily destroy all the RNA it can find, except for the short (around 30 nucleotides) stretches of mRNA that are protected from the nuclease by a ribosome. If we now isolate the ribosomes and then separate them from the mRNA fragments they sheltered from the nuclease attack, we can convert these mRNA fragments to a cDNA library.

From this point on the experiment is comparable to a traditional RNA-seq experiment. The fragments are sequenced and mapped back to the reference genome, resulting in what are known as ribosome footprints. You could say that every time a ribosome-protected mRNA fragment can be mapped to a genomic location the ribosome leaves a footprint at this location. We can deduce the expression level of a gene by counting the number of footprints we find within it, which is exactly the same as counting the number of RNA-seq reads that mapped to a gene. We can

also use ribo-seq to find the exact translation start site of a gene. Antibiotics such as harringtonine or lactimidomycin specifically halt translation at the initiation site instead of randomly during the translation process as with cycloheximide and can therefore be used to detect for example new translation start sites (Ingolia, 2011, Lee *et al.*, 2012).

In the previous chapter we mentioned that protein sequence databases are used to match the mass spectra from a shotgun proteomics experiment to the corresponding peptides. Ensembl and Swiss-Prot are two examples of such publically available databases. They are priceless and reliable sources of information, but because they only contain protein sequences that have been experimentally verified (or at least predicted), these databases might not always contain all the proteins that are expressed in your specific sample. So to improve the number of identifications in a proteomics experiment scientists often integrate the public databases with their own custom protein sequence database.

This custom database can be created with data from an RNA or ribo-seq analysis of the same sample that was used for the proteomics experiment (Menschaert *et al.*, 2013). Ribo-seq does offer some advantages over RNA-seq though. Because it is not affected by any post-transcriptional regulation, ribo-seq reflects the protein expression levels more closely than RNA-seq. Plus, in an RNA-seq experiment we don't know the precise translation start site, so we need to translate the sequences we find in three or six different reading frames. As you could read in the previous chapter, an RNA molecule can be seen as a string of nucleotide triplets or codons, which can be translated into a string of amino acids. A reading frame is a way to divide the sequence of a transcript in a set of consecutive codons and because a codon is three nucleotides long, there are three possible reading frames in a single transcript. So depending on the exact location of the translation start site, one transcript could code for three possibly radically different proteins. This means that if you do not have any precise information on the translation start site, you have to include all three reading frames and their corresponding protein sequences in your custom database. If you also don't know which DNA strand your transcript came from (remember how the basic structure of our DNA is that of a double helix, consisting of two intertwined DNA strands), you find yourself with six possible reading frames (three for each strand). This issue can be avoided by using ribo-seq, as it allows us to determine the exact translation start site with single-base resolution. Less reading frames to translate means a smaller protein sequence database and thus less time spent searching the database to identify the peptide spectra from the shotgun proteomics experiment.

In the following research paper we describe in detail how we used ribo-seq to create such a custom protein sequence database and how this approach can improve both the number and the quality of protein identifications in a proteomics experiment.

Adapted from:

Koch A*, Gawron D*, Steyaert S, Ndah E, Crappé J, De Keulenaer S, De Meester E, Ming M, Shen B, Gevaert K, Van Criekinge W, Van Damme P and Menschaert G. A proteogenomics approach integrating proteomics and ribosome profiling increases the efficiency of protein identification and enables the discovery of alternative translation start sites. *Proteomics* **14**, 2688–2698 (2014)

* These authors contributed equally

DEEP PROTEOME COVERAGE BASED ON RIBOPROFILING

ABSTRACT

Next-generation transcriptome sequencing is increasingly integrated with mass spectrometry to enhance MS-based protein and peptide identification. Recently, a breakthrough in transcriptome analysis was achieved with the development of ribosome profiling (ribo-seq). This technology is based on the deep sequencing of ribosome-protected mRNA fragments, thereby enabling the direct observation of *in vivo* protein synthesis at the transcript level. In order to explore the impact of a ribo-seq-derived protein sequence search space on MS/MS spectrum identification, we performed a comprehensive proteome study on a human cancer cell line, using both shotgun and N-terminal proteomics, next to ribosome profiling, which was used to delineate (alternative) translational reading-frames. By including protein-level evidence of sample-specific genetic variation and alternative translation, this strategy improved the identification score of 69 proteins and identified 22 new proteins in the shotgun experiment. Furthermore, we discovered 18 new alternative translation start sites in the N-terminal proteomics data and observed a correlation between the quantitative measures of ribo-seq and shotgun proteomics with a Pearson correlation coefficient ranging from 0.483 to 0.664. Overall, this study demonstrated the benefits of ribosome profiling for MS-based protein and peptide identification and we believe this approach could develop into a common practice for next-generation proteomics.

INTRODUCTION

A shotgun proteomics experiment typically involves the fractionation of a complex peptide mixture followed by LC-MS/MS analysis and the identification of peptides using one of several protein or peptide sequence database search tools (Perkins *et al.*, 1999, Craig & Beavis, 2004, Geer *et al.*, 2004). N-terminal proteomics techniques such as N-terminal COFRADIC (combined fractional diagonal chromatography) expand on the results of a typical shotgun experiment by enriching for N-terminal peptides, thus revealing (alternative) translation start sites, while simultaneously measuring co-translational modifications of protein N-termini (Staes *et al.*, 2011). Protein reference databases only contain experimentally verified and/or predicted sequences and are therefore unlikely to contain a comprehensive representation of the actual protein content of a given sample. To resolve this shortcoming, recent efforts have been directed towards the combination of proteomics and next-generation transcriptome sequencing (Nagaraj *et al.*, 2011, Liu *et al.*, 2013, Woo *et al.*, 2014, Pinto *et al.*, 2014). Proteogenomic approaches that delineate translation products based on mRNA sequencing data may improve protein identification in multiple ways. The transcriptome of a sample offers a more representative expression profile than could be obtained with a public database alone while at the same time reducing the search space through the elimination of unexpressed gene products (Wang *et al.*, 2012). The transcript data also contains useful information about sequence variations such as single nucleotide polymorphisms (SNP) or mutations and RNA splice and editing variants (Wang *et al.*, 2012, Ning *et al.*, 2010, 2012), which increases the chances of detecting new proteins or protein forms (Beck *et al.*, 2011, Djebali *et al.*, 2012, Low *et al.*, 2013). Despite the benefits of adding next-generation transcriptome sequencing to an MS-based proteomics experiment, there are still several improvements possible. Because of extensive translation regulation, the presence of a transcript does not necessarily imply the presence of the corresponding protein (Selbach *et al.*, 2008, Sonenberg *et al.*, 2007, Baek *et al.*, 2008). On top of that, several factors, including internal ribosome entry sites, the presence of multiple ORFs per transcript, non-AUG start codons and leaky scanning on top of ribosome frameshifting and stop codon readthrough hamper the prediction of the exact protein sequence(s) from a single transcript sequence (Touriol *et al.*, 2003, Michel *et al.*, 2012, Namy *et al.*, 2012).

Recently, a novel technique has been described that attempts to tackle these limitations: ribosome profiling (Ingolia *et al.*, 2010). Ribosome profiling, or ribo-seq, is based on the deep sequencing of ribosome-associated mRNA fragments, thus enabling the study of *in vivo* protein synthesis at the transcript level. In a ribo-seq experiment, eukaryotic translation is often halted using cycloheximide (CHX). The mRNA that is not protected by ribosomes after the translation halt is digested with nucleases and the monosome-mRNA complexes are isolated. Next, the

protected mRNA sequences are separated from the ribosomes and converted into a DNA library, ready to be sequenced. The sequencing results in a genome-wide snapshot of the mRNA that enters the translation machinery. Additionally, (alternative) translation initiation sites can be studied with sub-codon to single-nucleotide precision through the use of antibiotics such as harringtonine (HARR) or lactimidomycin (LTM), which cause the ribosomes to halt at sites of translation initiation (Ingolia *et al.*, 2011, Lee *et al.*, 2012). When the exact translation start site is known, the ORF can be delineated, thus eliminating the need to translate the transcripts in three or six reading frames. The measurement of mRNA at the translation level, combined with the knowledge of the exact translation start sites, makes ribosome profiling an excellent choice for the creation of a custom protein sequence search space for MS/MS-based peptide identification (Menschaert *et al.*, 2013). It has to be noted that ribo-seq does not generate direct evidence of mature proteins or protein stability and that some non-coding transcripts do not result in a protein product, despite being associated with ribosomes (Guttman *et al.*, 2012, Volders *et al.*, 2013, Bazzini *et al.*, 2014). However, MS-assisted validation may help to resolve both issues. Apart from canonical translation products, ribosome profiling also aids in the identification of unannotated truncated and N-terminally extended protein variants and the validation of these variants can come from matching N-terminal COFRADIC data (Menschaert *et al.*, 2013, Van Damme *et al.*, 2014).

In this study we created a custom protein sequence database based on LTM ORF delineation for the HCT116 cell line, a widely used human colon cancer cell model, to serve as the search space for MS/MS spectra obtained by means of shotgun proteomics and N-terminal COFRADIC (Figure 2.1). Translation products derived from the ribosome profiling data of the HCT116 cells were combined with the public Swiss-Prot protein sequence database (Boeckmann *et al.*, 2003) to build an optimal protein search space for our proteomics data. The addition of ribo-seq data resulted in the identification of 22 new proteins, i.e. proteins that were not contained in the Swiss-Prot database, out of a total of 2,816 protein identifications in our shotgun proteomics experiment. On top of that, the inclusion of ribo-seq data improved the score of 69 proteins as a result of the discovery of proteins with a mutation, new isoforms and homologs and extended protein forms. Out of a total of 1,262 peptides, ribo-seq identified 18 extra N- termini in the COFRADIC experiment compared to Swiss-Prot alone, including 6 N-termini originating from extended protein forms with a near-cognate start site (i.e. the protein does not start with the canonical AUG codon). It needs to be noted that in the shotgun proteomics experiment 312 proteins were uniquely identified using the Swiss-Prot database, emphasizing the importance of proteomics techniques for the validation of next-generation transcriptome sequencing datasets. Finally, the correlation between the ribo-seq and shotgun proteomics data was calculated. Depending on the settings used, the Pearson correlation coefficient between the ribo-seq-derived normalized ribosome-protected fragments (RPF)

counts and the normalized spectral counts of the shotgun experiment (i.e. emPAI (Ishihama *et al.*, 2005) and NSAF (Paoletti *et al.*, 2006) values) ranged from 0.483 to 0.664.

MATERIAL AND METHODS

CELL CULTURE FOR PROTEOMICS

The HCT116 cell line was kindly provided by the Johns Hopkins Sidney Kimmel Comprehensive Cancer Center (Baltimore, USA). Cells were cultivated in DMEM medium supplemented with 10% fetal bovine serum (HyClone, Thermo Fisher Scientific Inc.), 100 units/ml penicillin (Gibco, Life Technologies) and 100 μg/ml streptomycin (Gibco) in a humidified incubator at 37°C and 5% CO2. Prior to the proteomics experiments, the HCT116 cells were subjected to SILAC labeling (Stable Isotope Labeling by Amino acids in Cell culture) (Ong *et al.*, 2002) as part of another experiment that compared the wild type HCT116 cells to a double

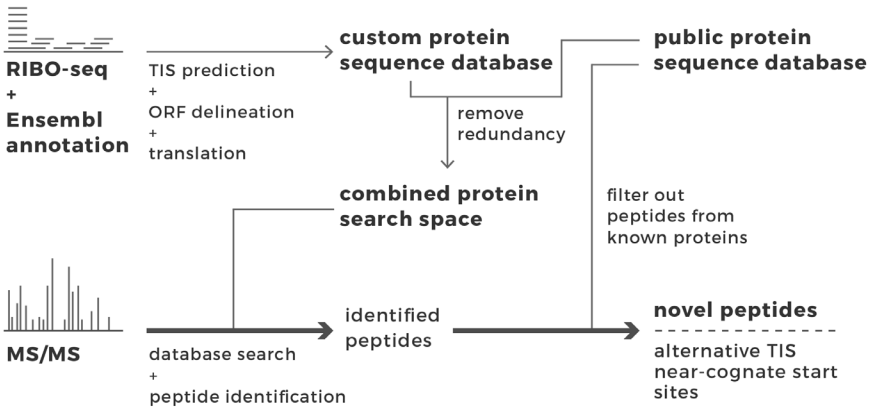


FIGURE 2.1 Proteogenomic strategy for the identification of proteins and peptides using a Swiss-Prot/ribo-seq-derived database.

Ribo-seq was performed twice on the human colon cancer cell line HCT116, once with CHX to halt translation globally and once with LTM to stop translation specifically at translation initiation sites. After translation initiation site (TIS) prediction, the ribo-seq-derived ORFs were translated to create a custom protein sequence database. This database was then combined with the human Swiss-Prot protein sequence database. Proteome samples were prepared from the same HCT116 cells and analyzed using both shotgun proteomics and N-terminal COFRADIC. The proteins and peptides in these samples were then identified using the custom combined protein search space.

knockout line, which was differently labeled (manuscript in preparation, see chapter 1, page 27). For the N-terminal COFRADIC analysis, cells were transferred to media containing 140 μM heavy ($^{13}C_6{}^{15}N_4$) L-arginine (Cambridge Isotope Labs, Andover, MA, USA). For the shotgun proteome analysis, cells were cultured in medium supplemented with 140 μM medium heavy ($^{13}C_6$) L-arginine and 800 μM heavy ($^{13}C_6$) L-lysine. To achieve a complete incorporation of the labeled amino acids, cells were maintained in culture for at least 6 population doublings.

CELL CULTURE AND SAMPLE PREPARATION FOR RIBOSOME PROFILING

The HCT116 cells for the ribosome profiling experiments were cultivated in McCoy's 5A (Modified) Medium (Gibco) supplemented with 10% fetal bovine serum, 2 mM alanyl-L-glutamine dipeptide (GlutaMAX, Gibco), 50 units/ml penicillin and 50 μg/ml steptamycin at 37°C and 5% CO_2. Cultures at 80–90% confluence were treated with 50 μM LTM (Ju *et al.*, 2005, Schneider-Poetsch *et al.*, 2010) or 100 mg/ml CHX (Sigma, USA) at 37°C for 30 min. Subsequently, cells were washed with PBS, harvested by trypsin-EDTA, rinsed again with PBS and recovered by 5 min of centrifugation at 300 × g, all in the presence of CHX to maintain the polysomal state. Cell pellets were resuspended in ice-cold lysis buffer, formulated according to Guo *et al.* (2010) (10 mM Tris-HCl, pH 7.4, 5 mM $MgCl_2$, 100 mM KCl, 1% Triton X-100, 2 mM dithiothreitol (DTT), 100 mg/ml CHX, 1 × complete and EDTA-free protease inhibitor cocktail (Roche)), at a concentration of 40 × 10^6 cells/ml. After 10 min of incubation on ice with periodic agitation, lysed samples were passed across QIAshredder spin columns (Qiagen) to shear the DNA. Subsequently, the flow-throughs were centrifuged for 10 min at 16,000 × g and 4°C. The recovered supernatant was aliquoted, snap-frozen in liquid nitrogen and stored at -80°C for subsequent ribosome footprint recovery and cDNA library generation.

SHOTGUN PROTEOME ANALYSIS

4.2 × 10^6 cells were lysed in 20 mM NH_4HCO_3 pH 7.9 by three rounds of freeze-thawing. Total protein concentration in cell extracts was measured using Biorad's Protein Assay (Biorad Laboratories, Munich, Germany) and 2 mg protein material was used for downstream processing. Digestion was performed overnight using trypsin (Promega, Madison, WI, USA; enzyme/substrate, 1/50) after adding 0.5 M guanidinium hydrochloride and 2% ACN to aid in protein denaturation. Methionines were uniformly oxidized to methionine sulfoxides by adding 20 μl of 3% (w/v) H_2O_2 to 100 μl sample (equivalent to 500 μg proteins) for 30 min at 30°C. For chromatographic separation 100 μl peptide mixture

was then immediately injected onto an RP-HPLC column (Zorbax® 300SB-C18 Narrow- bore, 2.1 mm internal diameter × 150 mm length, 5 μm particles, Agilent). Following 10 min of isocratic pumping with solvent A (10 mM ammonium acetate in water/ACN (98:2 v/v), pH 5.5), a gradient of 1% solvent B increase per minute (solvent B: 10 mM ammonium acetate in ACN/water (70:30 v/v), pH 5.5) was started. The column was then run at 100% solvent B for 5 min, switched to 100% solvent A and re-equilibrated for 20 min. The flow was kept constant at 80 μL/min using Agilent's 1100 series capillary pump with the 100 μL/min flow controller. Fractions of 30 sec wide were collected from 20 to 80 min after sample injection. To reduce LC-MS/MS analysis time, fractions eluting 12 min apart were pooled, vacuum dried and re-dissolved in 20 μl 20 mM tris(2-carboxyethyl) phosphine (TCEP) in 2% acetonitrile.

N-TERMINAL COFRADIC ANALYSIS

HCT116 cells were lysed in 50 mM HEPES pH 7.4, 100 mM NaCl and 0.8% CHAPS containing a cocktail of protease inhibitors (Roche) for 10 min on ice and centrifuged for 15 min at 16,000 g at 4°C. The protein sample was then subjected to N-terminal COFRADIC as described by Staes *et al.* (2011).

LC-MS/MS ANALYSIS

The shotgun proteomics sample was subjected to LC-MS/MS analysis using an Ultimate 3000 RSLC nano HPLC (Dionex, Amsterdam, the Netherlands) in-line connected to an LTQ Orbitrap Velos (Thermo Fisher Scientific, Bremen, Germany). The sample mixture was loaded on a trapping column (made in-house, 100 μm id × 20 mm, 5 μm beads C18 Reprosil-HD, Dr. Maisch). After back flushing from the trapping column, the sample was loaded on a reverse-phase column (made in-house, 75 μm id × 150 mm, 5 μm beads C18 Reprosil-HD, Dr. Maisch). Peptides were loaded in solvent A' (0.1% trifluoroacetic acid, 2% ACN) and separated with a linear gradient from 2% solvent A'' (0.1% formic acid) to 50% solvent B' (0.1% formic acid and 80% ACN) at a flow rate of 300 nl/min followed by a wash reaching 100% solvent B'. The mass spectrometer was operated in data-dependent mode, automatically switching between MS and MS/MS acquisition for the ten most abundant peaks in a given MS spectrum. Mascot Generic Files were created from the MS/MS data in each LC run using the Distiller software (version 2.3.2.0).

The N-terminal COFRADIC sample was analyzed on the LTQ Orbitrap XL mass spectrometer (Thermo Fisher Scientific, Bremen, Germany) which was operated in data-dependent mode, automatically switching between MS and MS/MS acquisition for the six most abundant peaks in a given MS spectrum.

All the MS data were converted using the PRIDE Converter (Barsnes *et al.*, 2009) and are available through the PRIDE database (Martens *et al.*, 2005) with the data-set identifier PXD000304 and DOI 10.6019/ PXD000304 (http://www.ebi.ac.uk/pride/archive/login, PX reviewer account: username: review48267, password: TTewpyNH).

PEPTIDE AND PROTEIN IDENTIFICATION AND INTERPRETATION

The protein and peptide searches were performed against our custom database using X! Tandem Sledgehammer (2013.09.01.1) and OMSSA 2.1.9 in combination with the SearchGui (1.16.4) tool (Vaudel *et al.*, 2011). For the shotgun proteomics experiment, pyroglutamate formation of N-terminal glutamine, acetylation of N-termini (both at peptide level) and methionine oxidation to methionine-sulfoxide were selected as variable modifications. Heavy labelled arginine ($^{13}C_6$) and lysine ($^{13}C_6$) were selected as fixed modifications. Mass tolerance was set to 10 ppm on precursor ions and to 0.5 Da on fragment ions. The peptide charge was set to 2+, 3+, 4+. Trypsin was selected as the enzyme setting, one missed cleavage was allowed and cleavage was also allowed when arginine or lysine was followed by proline.
For the N-terminomics experiment, the generated MS/MS peak lists were searched with Mascot (version 2.3) (Hirosawa *et al.*, 1993). Mass tolerance on precursor ions was set to 10 ppm (with Mascot's C13 option set to 1) and to 0.5 Da on fragment ions. The peptide charge was set to 1+, 2+, 3+ and the instrument setting to ESI-TRAP. Methionine oxidation to methionine- sulfoxide, $^{13}C_2D_3$-acetylation on lysines and carbamidomethylation of cysteine were set as fixed modifications. Variable modifications were $^{13}C_2D_3$ acetylation of N-termini, acetylation of N-termini and pyroglutamate formation of N-terminal glutamine (all at peptide level). $^{13}C_6{}^{15}N_4$ L-arg was set as fixed modification. Endoproteinase semi-Arg-C/P (Arg-C specificity with arginine-proline cleavage allowed) was set as enzyme allowing for no missed cleavages.

Protein and peptide identification and data interpretation were done using the PeptideShaker algorithm (http://code.google.com/p/peptide-shaker, version 0.26.2), setting the FDR to 1% at all levels (peptide-to-spectrum matching, peptide and protein).

RIBOSOME PROFILING

100 µl of the clarified HCT116 cell lysate (equivalent to 4×10^6 cells) was used as input for ribosome footprinting. The A260 absorbance of the lysate was measured with Nanodrop (Thermo Scientific) and for each A260, 5 units of ARTseq

Nuclease (Epicentre) were added to the samples. The nuclease digestion proceeded for 45 min at room temperature and was stopped by adding SUPERase. In Rnase Inhibitor (Life Technologies). Next, the ribosome protected fragments (RPFs) were isolated using Sephacryl S400 spin columns (GE Healthcare) according to the procedure described in 'ARTseq Ribosome Profiling Kit, Mammalian' (Epicentre). The RNA was extracted from the samples using acid 125 phenol : 24 chloroform : 1 isoamyl alcohol and precipitated overnight at -20°C by adding 2 μl glycogen, 1/10th volume of 5 M ammonium acetate and 1.5 volumes of 100% isopropyl alcohol. After centrifugation at 18,840 × g and 4°C for 20 min, the purified RNA pellet was resuspended in 10 μl nuclease free water.

LIBRARY PREPARATION AND SEQUENCING

Libraries were created according to the guidelines described in the ARTseq Ribosome profiling Kit, Mammalian protocol (Epicentre). The RPFs were initially rRNA depleted using the Ribo-Zero Magnetic Kit (Human/Mouse/Rat, Epicentre), omitting the 50°C incubation step. Cleanup of the rRNA depletion reactions was performed through Zymo RNA Clean & Concentrator-5 kit (Zymo Research) using 200 μl binding buffer and 450 μl absolute ethanol. The samples were separated on a 15% urea-polyacrylamide gel and footprints of 26 to 34 nucleotides long were excised. RNA was extracted from the gel and precipitated. The pellet was resuspended in 20 μl nuclease-free water. Next, RPFs were end polished, 3' adaptor ligated, reverse transcribed and PAGE purified. Five μl of circularized template DNA was used in the PCR reaction and amplification proceeded for 11 cycles. The libraries were purified with AMPure XP beads (Beckman Coulter) and their quality was assessed on a High Sensitivity DNA assay chip (Agilent technologies). The concentration of the libraries was measured with qPCR and they were single end sequenced on a Hiseq (Illumina) for 50 cycles. The ribo-seq libraries have been deposited in NCBI's Gene Expression Omnibus (Edgar *et al.*, 2002) and are accessible through the GEO series accession number GSE58207 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE58207).

SWISS-PROT/RIBO-SEQ INTEGRATED DATABASE CONSTRUCTION

The merged database was constructed using all human Swiss-Prot proteins (downloaded from http://www.uniprot.org, version 2014_03) and the translation products obtained from the ribosome profiling experiment (Figure 2.1). The ribo-seq-derived translation products were created from both the predicted (alternative) TIS genomic locations based on the LTM ribosome profiling information (according to Lee *et al.*, 2012) and the corresponding mRNA sequences

obtained from Ensembl (version 70) that displayed overall CHX ribosome protected fragment (RPF) coverage. After reconstructing the amino acid sequences, the Ensembl identifiers were mapped to Swiss-Prot identifiers (to safeguard uniformity) using the pBlast algorithm.

In order to remove redundancy introduced by the combination of the ribo-seq-derived translation products and the Swiss-Prot protein sequences, duplicated sequences were removed, retaining the custom sequence. Moreover, only the longest form of a series of gene translation products (N-terminal extended or canonical) was withheld in the combined database. The custom database contained 68,961 sequences as compared to the 20,264 proteins in UniProtKB-SwissProt version 2014_03. Extra information on the custom DB creation can be found in Menschaert *et al.* (2013).

CORRELATION ANALYSIS

Only the transcripts identified in both Swiss-Prot and the ribo-seq-derived translation products were selected for the correlation analysis. Ribo-seq measurements were expressed as the number of ribosomal footprints per CDS (RPF count), hereby correcting for a possible 3'UTR and 5'UTR bias as suggested by Ingolia *et al.* (2011). Two quantitative measures for protein abundance based on spectral counts (exponentially modified Protein Abundance Index or emPAI (Ishihama *et al.*, 2005) and the Normalized Spectral Abundance Factor or NSAF (Paoletti *et al.*, 2006)) were calculated using the shotgun data. While the first method uses the number of peptides per protein normalized by the theoretical number of peptides, the so-called protein abundance index (PAI), the NSAF method takes both the protein length and the total number of identified MS/MS spectra in an experiment into account. For each dbTIS transcript for which quantitative ribo-seq and shotgun proteomics information was available a Pearson correlation coefficient was calculated between its normalized RPF count and its normalized spectral count. When more than one ribo-seq-derived transcript corresponded with a particular Swiss-Prot protein sequence, the one with the highest normalized RPF count was used. The different normalization and identification approaches were combined with the following additional transcript filtering settings: *i*) no extra cutoffs, *ii*) only dbTIS transcripts with a validated MS/MS-based identification (meaning that the spectral count value was higher than 2), *iii*) only dbTIS transcripts with a total RPF count ≥ 200, and *iv*) only dbTIS transcripts with both a validated MS identification and an RPF count ≥ 200. All correlation coefficients were computed using log-transformed RPF and emPAI/NSAF measures.

RESULTS

A regular shotgun and an N-terminal COFRADIC proteomics experiment were performed on a HCT116 cell line to determine the effect of the addition of ribo-seq-derived translation products to the Swiss-Prot protein sequence database on MS/MS spectrum identification. The shotgun data were used for the overall assessment of protein expression, whereas the N- terminal COFRADIC data were specifically used for the validation of the ribo-seq-predicted translation initiation sites.

SHOTGUN PROTEOMICS

Using the combination of Swiss-Prot and the ribo-seq-derived database, we identified a total of 2,816 proteins in the HCT116 cells (Figure 2.2a). The majority of these proteins (2,482 or 88.1%) were identified in both Swiss-Prot and the custom database. The addition of the ribo- seq data to the protein search space led to 22 extra identifications, which would not have been picked up with just the Swiss-Prot database. Besides 9 previously unannotated protein products, these new identifications included 13 proteins with a mutation and three alternatively spliced isoforms. The inclusion of ribo-seq data also improved protein identification and score significance for 69 proteins since higher peptide coverage was obtained (Supplementary Figure 2.1 shows three examples). The proteins with an improved score coincided with mutation sites (52 proteins), alternatively spliced isoforms (14 proteins) and three N-terminal extensions. The ribo-seq experiment also missed 312 proteins, but these were still picked up thanks to the inclusion of Swiss-Prot in the search space. All the identified proteins and their respective annotations can be found in Supplementary Table 2.1. An approximate analysis of the turnover rate and half-lives of the 312 missed proteins using publically available datasets (Doherty *et al.*, 2009, Sandoval *et al.*, 2013) showed no significant difference between the missed and the other identified proteins (Wilcoxon rank-sum test, $p > 0.05$). A gene ontology enrichment analysis using the DAVID tool (Huang *et al.*, 2009) revealed that several biological process ontologies involving protein transport and localization were significantly enriched in the 312 missed proteins, just as the corresponding cellular localization ontologies linked to the cytoskeleton, cytosol and non-membrane-bounded organelles (Supplementary Table 2.2).

N-TERMINAL COFRADIC

In order to validate the TISs identified by the ribo-seq experiment and thus the

corresponding N-terminal protein isoforms, positional proteomics in the form of N-terminal COFRADIC was applied to the HCT116 cells. After LC-MS/MS analysis and the subsequent combined database search, we identified 1,289 N-terminal peptides (Figure 2.2b). The greater part of these peptides mapped to canonical start sites (1,071 peptides or 83.1%), 208 peptides started downstream of the canonical start site (past protein position 2 in reference to Swiss-Prot), nine peptides mapped to a 5'-extension and one to an uORF. Two examples of proteins with an N-terminal extension or truncation are given in Figure 2.3. Ribo-seq uniquely identified 18 peptides, which would have been missed when only searching Swiss-Prot. Both the N-terminal COFRADIC and ribo-seq experiment provided evidence of translation initiation at near-cognate start sites, which was also reported in previous COFRADIC and ribo-seq studies (Ingolia *et al.*, 2011, Lee *et al.*, 2012). A complete list of all identified N-terminal peptides is provided as Supplementary Table 2.1.

We compared the list of identified protein extensions starting at non-AUG start sites with the previously published list of non-AUG derived N-terminal extensions predicted by Ivanov *et al.* (2011) and found matching evidence for one N-terminally extended protein (Swiss-Prot entry name HDGF_HUMAN; extension of 50 amino acids starting at GTG) out of 9 identified in our proteomics study.

CORRELATION ANALYSIS

We calculated a Pearson correlation coefficient to investigate the relation between the ribo-seq coverage and MS protein abundance measurements. Only transcripts for which quantitative information was available from both the ribo-seq and shotgun proteomics experiments were used in all the plots and calculations. The

TABLE 2.1 Pearson correlation coefficients between MS protein abundance and ribo-seq coverage.

MS protein identifications were performed with an FDR of 1% and protein abundances were calculated as emPAI and NSAF values. The correlation coefficients were computed for each of the following transcript filtering settings: *i*) all dbTIS transcripts without additional thresholds, *ii*) only transcripts with a validated MS identification (i.e. transcripts with a spectral count value > 2), *iii*) only dbTIS transcripts with a total RPF count ≥ 200 and *iv*) only dbTIS transcripts with both a validated MS/MS-based identification and an RPF count ≥ 200.

	i	*ii*	*iii*	*iv*
emPAI	0.488	0.498	0.483	0.518
NSAF	0.608	0.642	0.634	0.664

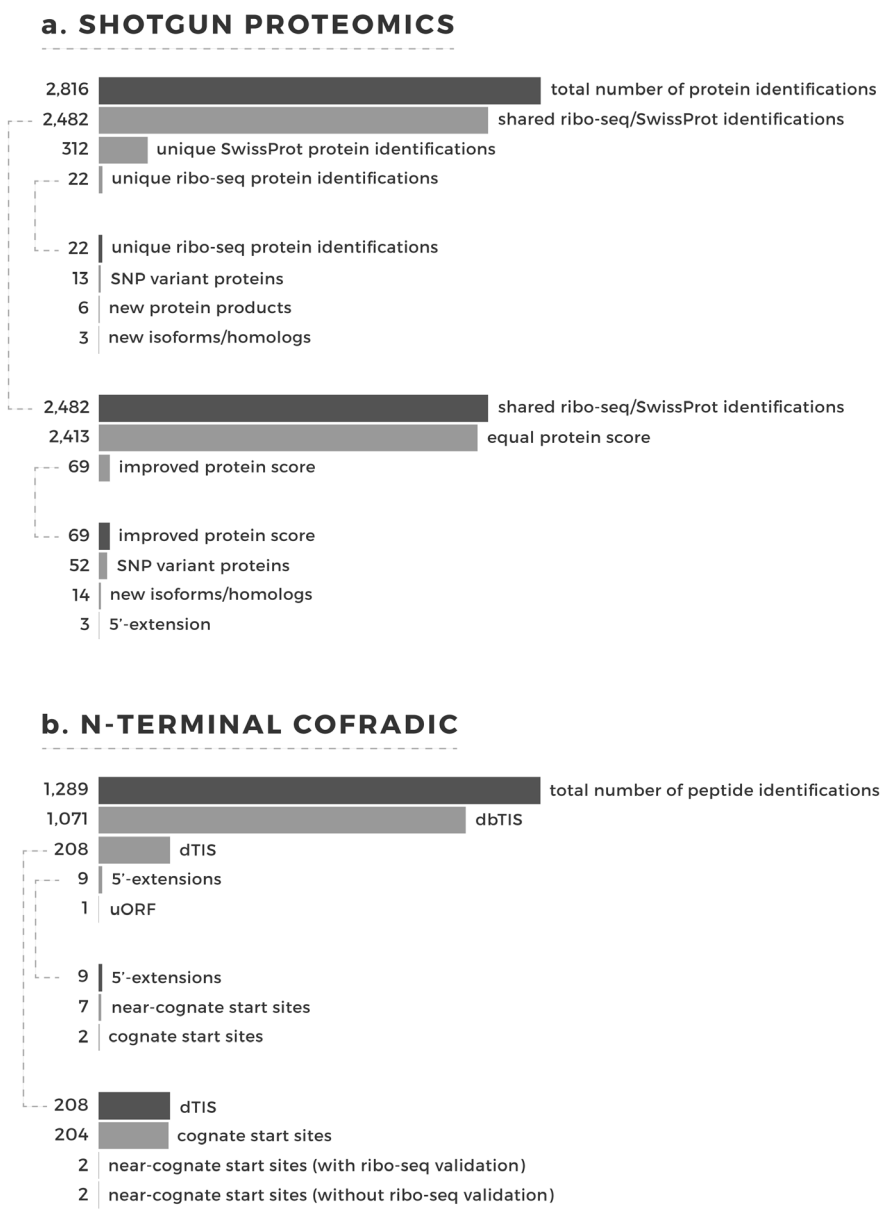


FIGURE 2.2 Bar charts showing the number of protein and peptide identifications obtained from the shotgun proteomics and N-terminal COFRADIC experiments.

a. The custom combined protein sequence database resulted in the identification of 2,816 proteins. Most of these proteins (2,482 or 88.1%) were picked up by both databases independently, while 312 and 22 proteins were uniquely identified in the Swiss-Prot and ribo-seq databases respectively. The 22 unique ribo-seq identifications contained six new

proteins, 13 proteins with a mutation site and three unannotated isoforms. The ribo-seq data also improved the protein identification and score of 69 proteins. **b.** Most of the 1,289 peptides that were found in the custom combined protein sequence database mapped to canonical, annotated N-termini (1,071 dbTIS peptides or 83.1%). Of the remaining N-termini, 208 started downstream of the canonical start site (beyond protein position 2), nine mapped to a 5'-extension and one to an uORF. For both the up- and downstream start sites, we identified several near-cognate start sites.

Pearson correlation values for the different normalization and identification approaches are listed in Table 2.1 and Figure 2.4a shows the correlation plots for the NSAF values, which were better correlated with the ribo-seq coverage than the emPAI values. The highest correlation ($r^2 = 0.664$) was obtained when using only validated dbTIS transcripts with a total RPF count ≥ 200. The correlation coefficients were also calculated for the 312 protein identifications that were present in Swiss-Prot, but not in our ribo-seq-derived search space (Supplementary Figure 2.2). These 312 identifications were missing from the ribo-seq data because no TISs were identified in the LTM-treated cells, but, as there was coverage in the CHX-treated cells, the correlation could still be calculated. The Pearson correlation coefficients ranged from 0.464 to 0.713, depending on the protein selection and normalization procedure, and were similar for the proteins identified in both the Swiss-Prot and ribo-seq database.

We also investigated the link between the correlation and the degree of protein stability. Figure 2.4b shows the correlation plot for validated dbTIS transcripts with an RPF ≥ 200 together with the instability indexes of the proteins. These indexes were obtained with the ExPASy ProtParam tool (Wilkins *et al.*, 1999), where a protein with an instability index < 40 is predicted to be stable and a protein with an index ≥ 40 is considered unstable. The majority of unstable proteins were characterized by lower NSAF and RPF values than the stable proteins. As reported previously, protein stability is among the most significant factors governing the correlation between gene expression and protein abundance (Ning *et al.*, 2010).

DISCUSSION

The successful identification of proteins and peptides from MS/MS spectra depends on a number of factors. A state-of-the-art mass spectrometer that provides high resolution and mass accuracy is a vital element of a proteomics experiment. Solid experimental design and a robust identification pipeline are two other important factors. As even small changes in database search algorithms can lead to different identification results, combining several search engines, such as X!Tandem (Craig *et al.*, 2004) and OMSSA (Geer *et al.*, 2004), helps to increase the

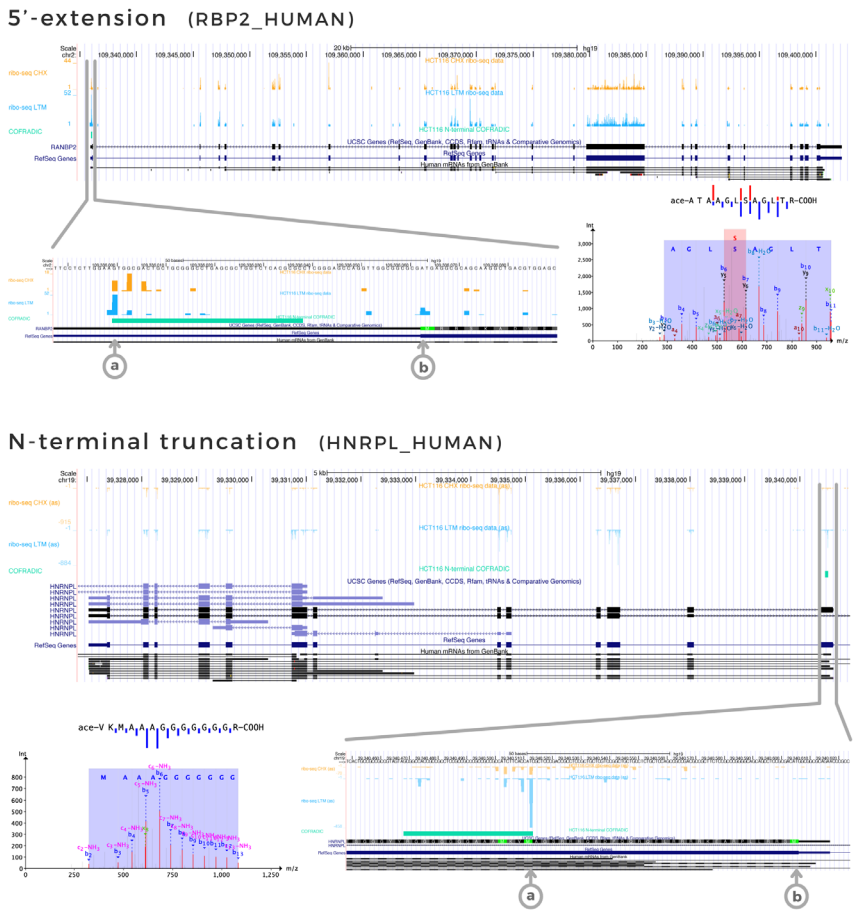


FIGURE 2.3 Depiction of two different N-termini that were predicted by ribo-seq and identified using N-terminal COFRADIC.

This figure shows a 5'-extension (Swiss-Prot entry name RBP2_HUMAN) and an N- terminal truncation (Swiss-Prot entry name HNRPL_HUMAN). The UCSC genome browser (Kent *et al.*, 2002) was used to create the plots of the ribo-seq and N-terminal COFRADIC data and the different browser tracks are from top to bottom: CHX treatment data, LTM treatment data, N-terminal COFRADIC data, UCSC genes, RefSeq genes and human mRNA from GenBank. The different start sites (a: alternative start site, b: canonical start site) are clearly visible in the zoomed genome browser views, just as the three-nucleotide periodicity of the ribo-seq data, especially in the N-terminal truncation image. The MS/MS spectra and sequence fragmentations indicate the confidence and quality of the peptide identifications.

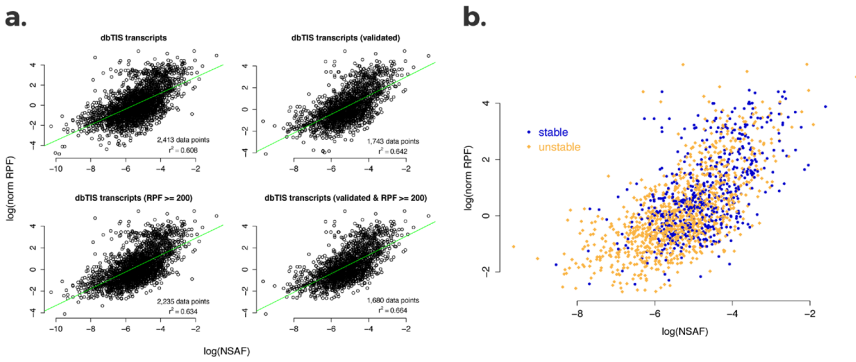


FIGURE 2.4 Correlation plots of protein abundance estimates based on NSAF values and RPF counts.

a. Top left: all dbTIS transcripts; top right: dbTIS transcripts with a validated MS/MS-based identification (i.e. transcripts with a spectral count value > 2); bottom left: dbTIS transcripts with an RPF count ≥ 200; bottom right: dbTIS transcripts with both a validated MS identification and an RPF count ≥ 200. The regression line is shown in green. For each plot, the number of data points used (i.e. the number of dbTIS transcripts) as well as the corresponding Pearson correlation coefficient (r^2) is shown. **b**. Correlation plot with the inclusion of stability data. Only dbTIS transcripts with both a validated MS/MS-based identification and an RPF count ≥ 200 were used (bottom right plot in Figure 2.4a). Instability indexes were determined with the ProtParam tool (Wilkins *et al.*, 1999): proteins with an instability index < 40 were classified as stable and are shown in blue, whereas proteins with an instability index ≥ 40 were classified as unstable and are shown in orange.

number of PSMs (Peptide Spectrum Match) (Searle *et al.*, 2008). A more recent approach to improve the number of PSMs is based on the custom tailoring of the search space through the use of next-generation transcriptome sequencing (Woo *et al.*, 2014, Menschaert *et al.*, 2013). The new and improved protein identifications based on our ribo-seq-derived search space were a first indication of the success of our proteogenomics strategy. Especially the identification of N-terminally extended proteins would not have been possible when using only Swiss-Prot. The positive correlation between protein abundance (measured as NSAF and emPAI values) and the ribo-seq footprint coverage (measured as RPF counts) also justifies the usage of the described proteogenomics approach. It has been described before how NSAF gives a more accurate estimate of protein abundance than emPAI as it uses more information (e.g. fragment ion intensities and protein length) (Colaert *et al.*, 2011, McIlwain *et al.*, 2012). This could explain why the NSAF values correlated better with the ribo-seq data. Interesting to note is that proteins with a lower stability index displayed both lower protein abundances as well as lower RPF counts than their more stable counterparts (Figure 2.4b). Several studies have reported correlation values between mRNA-seq coverage and protein abundance, ranging from 0.41–0.44 (Schwanhausser et al., 2011) to 0.51 (Ning *et al.*, 2012) in mouse and between 0.42 and 0.43 in rat [14]. Nagaraj *et al.* (2011) published

a Spearman's correlation of 0.6 between FPKM-based transcript abundance (Fragments Per Kilobase per Million) and iBAQ-based protein abundance values for the human HeLa cell line. The improved correlation observed in our study (Table 2.1) can be explained by the fact that, because it measures transcripts after they have entered the translation machinery, ribosome profiling is less affected by transcriptional and translational regulation. The ability of ribo-seq to take alternative translation events into account leads to a better delineation of ORFs, which could also improve the correlation. Another advantage of the ribo-seq-derived database was that it allowed us to identify translation initiation from non-AUG start sites at the protein level, for which only limited evidence is available so far (Van Damme *et al.*, 2014, Stern-Ginossar *et al.*, 2012, Slavoff *et al.*, 2013, Branca *et al.*, 2014).

Without the addition of the Swiss-Prot database to our custom search space, a significant amount of proteins would have been missed (unique Swiss-Prot identifications in Figure 2.2). These proteins were missing from the ribo-seq-derived search space because no detectable LTM-signal could be observed. But since the CHX treatment resulted in coverage for these proteins, we could still calculate the correlation between protein abundance and RPF counts (Supplementary Figure 2.2). The abundance values and RPF counts, together with their correlation values, ruled out low abundance or coverage as a reason for the missed identifications. A suboptimal LTM treatment and/or TIS calling could help explain the lack of TIS recognition and the resulting absence of the corresponding proteins from the ribo- seq-derived search space. These results demonstrate the importance of reference databases and MS for the identification and validation of next-generation sequencing-derived translation products.

The combination of N-terminal COFRADIC and ribo-seq data identified a number of alternative TISs. Translation via these start sites produces protein isoforms with a different N-terminus if the new start site maintains the reading frame (e.g. the 5'-UTR extension in Figure 2.3). If the start site is not in the same reading frame, completely different proteins will be generated. The selection of upstream TISs can also lead to the creation of uORFs, which influence the downstream protein synthesis from the main ORF (Wethmar *et al.*, 2010, Medenbach *et al.*, 2011). Roughly half of all mammalian transcripts contain one or more upstream TISs, which are often associated with short ORFs (Lee *et al.*, 2012). In contrast to the previously reported frequent occurrence of uORFs in human and mouse ribosome profiling data (Ingolia *et al.*, 2011, Lee *et al.*, 2012), we were able to identify only one N-terminal peptide of an upstream overlapping ORF in the *PIDD* gene (Supplementary Table 2.1). This limited evidence for uORF protein products could be attributed to several factors, such as a bias towards upstream (near-) cognate start site identification from ribosome profiling data (Michel *et al.*, 2013) or the rapid degradation, small size and possibly low abundance of uORFs.

CONCLUSION

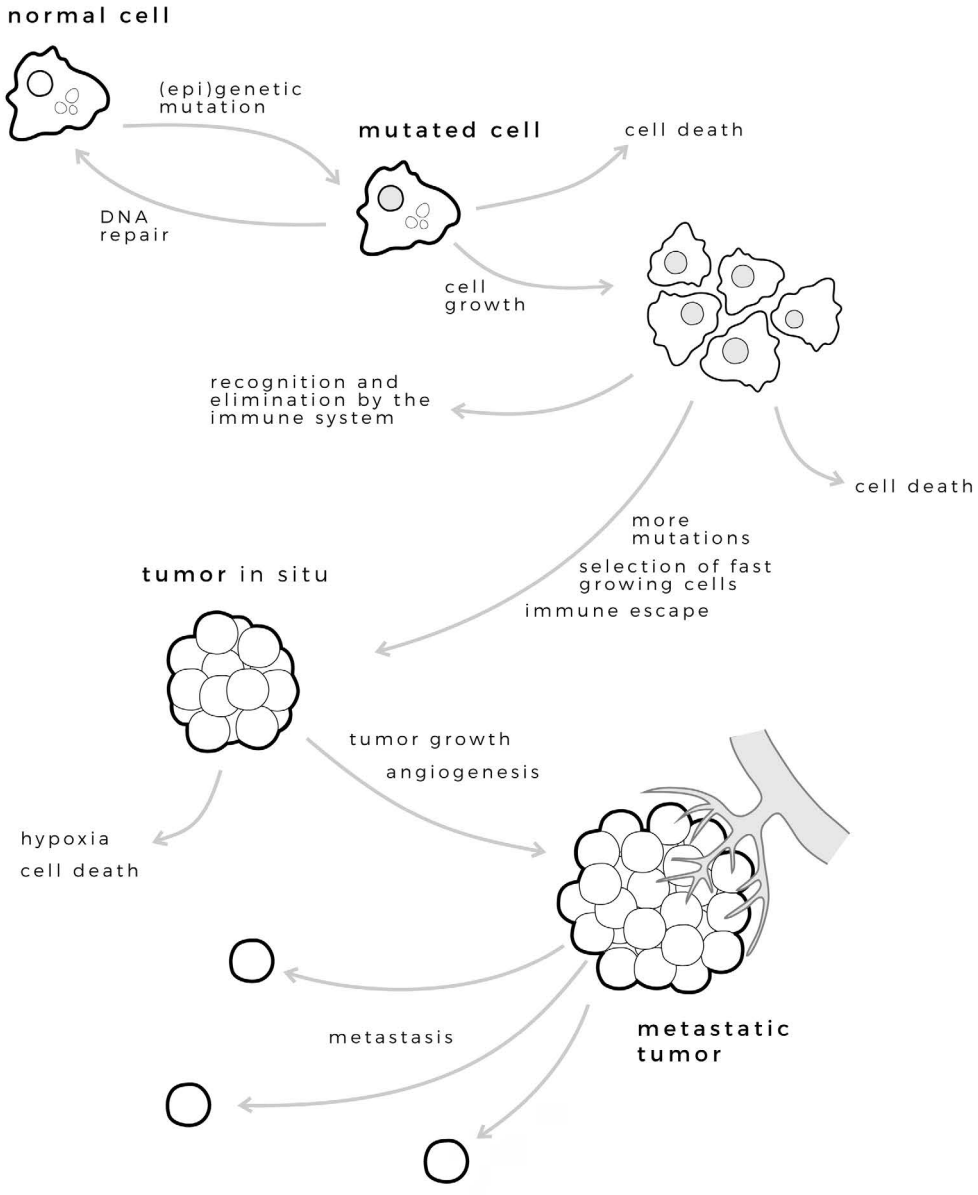
As sequencing techniques become more generally accessible, ribosome profiling has become (Menschaert *et al.*, 2013, Bazzini *et al.*, 2014, Van Damme *et al.*, 2014, Stern-Ginossar *et al.*, 2012, Vasquez *et al.*, 2014) and will continue to be a valuable addition to MS-based protein and peptide identification, possibly taking over the role of mRNA sequencing for ORF delineation. The benefits of ribo-seq include the positive correlation between protein abundance and ribo-seq footprint coverage and the ability to predict TISs with single-nucleotide precision. Despite the advantages of ribo-seq, MS-based validation will remain indispensable, not only for the general identification of proteins (through shotgun proteomics), but also for the validation of ribo-seq-derived (alternative) TIS predictions (by means of N-terminomics techniques such as COFRADIC (Staes *et al.*, 2011)). Furthermore, unlike ribo-seq or any other transcriptome sequencing technique, MS provides true *in vivo* evidence of proteins or peptides, while taking potential co- and post-translational modifications into account. We also found that both reference protein sequence databases and ribo-seq-derived search spaces can miss protein identifications and that the best results were obtained when these databases were combined. Overall, our results show the usefulness of a ribo-seq-based proteogenomics approach. Based on our findings, we constructed an automated pipeline for the easy conversion of ribo-seq data into a custom protein sequence search space that incorporates both sequence variation information and TIS prediction, ready to be searched for protein identifications. This pipeline has meanwhile been published in a follow-up paper (Crappé *et al.*, 2015).

CHAPTER 3

ON CANCER

FIGURE 3.1
FROM NORMAL CELL TO METASTATIC CANCER

This figure gives a general overview of the different steps that separate a normal cell from a tumor cell. It all starts when a a gene involved in cell growth gets hit by a genetic or epigenetic mutation. The cell will often detect these mutations and will then try to repair the damage or will initiate controlled cell death to prevent further mutations. If these control mechanisms fail and the cell acquires more (epi)genetic mutations, the cell might lose all growth inhibitions. After some time the initial mutated cell will have formed a lump of mutated daughter cells, a tumor. This tumor will start to attract blood vessels (angiogenesis) to provide itself with oxygen and nutrients (without it the tumor cells could die of hypoxia, a lack of oxygen) and at a certain point the tumor cells will grow into the surrounding tissues and will spread through the body. The tumor has become metastatic. We have several defence mechanisms to prevent a mutated cell from reaching this point, such as DNA repair and controlled cell death (apoptosis). Our immune system is another example. It is constantly on the lookout for cancer cells and will destroy them if it finds any, though sometimes a tumor cell will manage to escape the immune response.



WHEN EXPRESSION AND METHYLATION GO ASTRAY

Cancer affects us all. The probability that you or someone close to you will at some point be confronted with cancer is unfortunately very high. Researchers from the World Health Organization estimated that in 2012 approximately 14 million people were diagnosed with cancer and they expect this number to rise to 22 million within the next two decades (WHO, 2014). But what is cancer exactly? And how can we use the gene expression and DNA methylation analysis techniques described in the first chapter to study this dreadful disease?

As we mentioned earlier, a cell's life cycle is very tightly regulated. Even death is a carefully orchestrated process, known as apoptosis. Cells can of course also die because of injury or disease, but this type of uncontrolled cell death is called necrosis (a classic example is gangrene). Our cells need the ability to grow and multiply to maintain our bodies. They also need apoptosis and other control mechanisms to keep this growth in check, because a defect in the growth control of a cell can kick-start a frantic proliferation (Figure 3.1). The resulting lump of cells, which usually inherit the growth defect of the original deviant cell, is called a tumor.

We say a tumor is malignant (as opposed to benign) if, on top the uncontrolled growth, it also acquires the ability to spread to other parts of the body (see Hanahan & Weinberg, 2011 for a detailed overview of the different characteristics of a cancer cell). This spreading of malignant tumor cells from one location to the other (often via blood or lymph circulation) is called metastasis. Most of the time it is not the initial tumor, but rather the metastasis that makes cancer so deadly. A primary tumor growing in breast tissue for example might not pose an immediate life-threatening risk, but the moment a metastatic tumor starts ravaging a vital organ like the brain the situation becomes dire. In fact, metastatic tumors are regularly found in crucial organs, such as the lungs, liver, brain or bones. You might also have heard or read about aggressive cancers. A cancer is said to be aggressive when the tumor cells divide and metastasize even faster than in a "normal" or non-aggressive cancer. Aggressive cancers generally have a worse prognosis than their less aggressive counterparts.

So in essence, cancer is a disease characterized by uncontrolled cell growth and the spreading of tumor cells throughout the body. It is actually a collection of diseases, rather than a single one, because different cancers can have very diverse biological backgrounds and often require specific therapies. We name cancers by the type of cell and the organ they originate from. Cancers that arise in epithelial cells, which line the surfaces of our body (for example our skin or the lining of our intestines and lungs), are called carcinomas, whereas sarcomas arise in connective tissue (for example bone or fat cells). Lymphomas and leukemia start in blood-forming cells, while blastomas and germ cell tumors, which occur less often, originate in immature cells or embryonic tissue and in pluripotent cells (often in the testes or ovaries). There are many more subdivisions, but these are

the main ones. So when a doctor diagnoses a patient with a carcinoma of the lung, this means that the primary tumor was found in the lung and that it started from an epithelial cell.

We know that cancer is caused by uncontrolled cell growth, but what exactly makes a cell lose control over its own growth? On a cellular level, the transformation of a healthy cell to a tumor cell is caused by both genetic and epigenetic mutations. Genetic mutations range from single nucleotide polymorphisms or SNPs (when only a single base is changed) to large chromosomal rearrangements (when parts of a chromosome are deleted, duplicated or swapped out with a part from another chromosome) or even changes in the number of chromosomes. An epigenetic mutation, such as a change in the methylation status of the promoter region of a gene, can affect the expression of this gene. If such an epigenetic mutation deactivates a gene that controls cell growth or activates a gene that promotes growth or metastasis, it can encourage the formation of a tumor (Herman & Baylin, 2003). Table 3.1 lists some genes that are known to be mutated and/or hypermethylated in cancer.

A single mutation, genetic or epigenetic, is not enough for a cell to become cancerous, so a typical tumor cell will harbor multiple genetic as well as epigenetic changes. But where do these mutations come from? Between 90 and 95% of all cancers are caused by environmental factors and are known as sporadic cancers (Esteller *et al.*, 2001). The remaining 5 to 10% are non-sporadic or hereditary cancers and have been linked to mutations that are passed on from one generation to next. One of the best-known examples is the mutation of the BRCA1 and BRCA2 genes, which increases a woman's chances of developing breast and ovarian cancer (Robson, 2002). If a woman has several family members with breast cancer, she might decide to get tested for BRCA mutations. In case of a positive test she could opt for more frequent checkups with a doctor or even for preventive surgery (you might remember Angelina Jolie's widely publicized decision to have preventive surgery after she found out she carried the BRCA mutations).

The single most important risk factor for cancer, sporadic or not, is age. It's simple, the older you are, the more likely you are to develop cancer. Lifestyle is another major factor, so even though you cannot change how old you are, there are several ways you can minimize your risk. Smoking for example is to be blamed for almost a fifth of all cancer deaths worldwide (Kuper *et al.*, 2002). Next in the list of things to avoid if you want to reduce your risk are a lack of physical exercise, obesity and an unhealthy diet (overeating, lots of alcohol and a lack of fruit and vegetables) (Kushi *et al.*, 2006). Infectious diseases (human papilloma virus for example has been linked to cervical cancer) (zur Hausen, 1996) and radiation (don't forget to wear sunscreen!) (Elwood & Jopson, 1997) are also significant cancer-causing environmental factors. In the end, all these different causes, whether it's exposure to asbestos (Straif *et al.*, 2009) or a hepatitis C virus infection (Waghray *et al.*, 2015),

produce genetic and epigenetic changes on a cellular level. These changes result (directly or indirectly) in the up or down regulation of certain genes or even in the production of mutated proteins with a novel function, transforming a normal cell into a tumor cell.

There is an abundance of cancer treatments available, some very specific (only work for certain patients with a certain type of cancer), some more general (can be used in several types of cancer). We should also note that, because there are so many different cancer types, it is unlikely that we will ever have a single, one-size-fits-all therapy for cancer. Before they decide which therapy or combination of therapies to use, doctors will try to collect as much information as they can about the patient and the cancer. Among many other things, they evaluate the type of cancer, tumor location and size, whether or not the tumor has metastasized and the presence of certain mutations. Correct staging of the cancer is

Table 3.1. A list of genes with well-known mutations and/or abberant DNA methylation in cancer.

The genes listed in this table have well-described mutations or are hypermethylated (and therefore inactivated) in one or more cancer types (we give an example for each gene). Most of the genes that are listed here act as tumor suppressors and are inactivated in cancer cells. One exception is the oncogene BRAF. There is a particular mutation that increases the activity of BRAF, which leads to oncogenesis via unregulated MAPK signaling (Wan *et al.*, 2004). For some of these genes, such as BRAF or BRCA1, clinical tests are commercially available.

Gene	(epi)mutation information
BRCA1	mutations have been linked to increased risk of developing breast and ovarian cancer (King *et al.*, 2003) and several clinical tests are available inactivation of BRCA1 expression through promoter hypermethylation has also been detected (Tapia *et al.*, 2008)
APC	inactivating mutations can be found in most sporadic colorectal cancers (Markowitz & Bertagnolli, 2009)
CDKN2A (p16)	frequently mutated in pancreatic cancer (Rozenblum *et al.*, 1997)
BRAF	frequently mutated in metastatic melanomas (Kainthla *et al.*, 2013) vemurafenib is a commercially available BRAF inhibitor (Bollag *et al.*, 2012)
KRAS	mutations in this gene have been found in colorectal cancer and can be used to predict response to treatment with cetuximab (Lièvre *et al.*, 2006)
GSTP1	promoter hypermethylation has been observed in more than 80% of prostate cancers (Bastian *et al.*, 2004)
MGMT	epigenetic silencing of this DNA repair gene makes glioblastoma patients more susceptible to temozolomide treatment (Hegi *et al.*, 2005)
PTEN	together with TP53 one of the most commonly mutated genes in prostate cancer (Chen *et al.*, 2005)
TP53 (p53)	common mutations in a wide range of cancer types, including colon. lung, breast and liver cancer (Hollstein *et al.*, 1991)
RASSF1	promoter methylation correlates with advanced tumor stage and poor prognosis in bladder cancer (Lee *et al.*, 2001)

also vital in choosing the correct treatment. During its development, a cancer goes through several phases or stages. These stages range from stage 0, an early, localized tumor, to stage 4, a metastasized tumor that has spread to other organs. While a tumor might be surgically removed without any further therapy if it is in an early stage, late stage tumors are much more difficult to treat and generally have a worse outcome.

Once the doctors have the information they need, they can decide how to treat their patient. The three predominant types of treatment are surgery, radiation therapy and chemotherapy. Especially in the early stages, a surgeon can try to cure a cancer by completely removing the tumor from a patient's body. However, the moment a tumor has metastasized, it becomes very difficult to surgically remove all tumor cells. In blood cancers, such as leukemia, surgery is of little use even at the early stages, as these tumor cells circulate around the body through the bloodstream instead of forming a solid lump of cells.

Apart from surgery, doctors can also choose to use radiotherapy to kill malignant cells and shrink tumors. Irradiating tumor cells with X-rays causes DNA damage, which stops them from growing and kills them. To minimize any unwanted damage, these X-rays are aimed at the tumor from different angles at once so that the radiation dose is higher in the tumor than in the surrounding healthy tissue. In addition to radiotherapy, doctors can also use chemotherapy to kill cancer cells. This type of treatment relies on the use of various drugs that interfere with cell division. Cancer cells divide faster than normal cells, so these drugs will mostly affect the fast-growing cancer cells. Chemotherapy can have severe side effects, because cancer cells are not the only fast-growing cells in our bodies. The epithelial cells that line the stomach for example, or hair follicle cells, are naturally fast-dividing cells that can be killed by chemotherapy, resulting in the characteristic gastrointestinal problems, such as nausea and vomiting, and hair loss. Patients generally receive a combination of the three predominant treatments. Chemo and radiotherapy are for example often used together to reduce the size of a tumor before a surgeon will try to remove it.

The therapies we just described come with severe side effects, so severe sometimes that the patient succumbs to the treatment instead of the cancer. It is actually not difficult to kill tumor cells, not at all. The big problem however, is to only kill tumor cells and not the healthy cells. In their search for a more targeted therapy that does not harm healthy cells, some scientists have set their sights on the immune system. Our immune system is a very complex and intricate system that involves several organs and specialized cells. A detailed overview of these organs and cells along with their functions is beyond the scope of this introduction, so we will focus on the role of the immune system in cancer and why scientists have been trying to harvest its power in the battle against this disease.
Most people will immediately associate the immune system with infectious

diseases, for example caused by a bacterial infection, but it is also absolutely vital in the detection and destruction of tumor cells. Certain immune cells (natural killer and killer T cells) are constantly patrolling our bodies, looking for so-called non-self or foreign antigens (antigens are molecules that can bind to the receptors of certain immune cells). A bacterial or viral protein for example that is normally not present in our cells will be recognized as a foreign antigen. Our cells express a protein complex on their membranes known as the major histocompatibility complex (MHC). The function of this complex is to present a sample of the proteins that are present within the cell to the extracellular environment (Janeway *et al.*, 2001).

The proteins in a cell, whether they belong to the cell itself or to a pathogen that has invaded the cell, are constantly synthesized and dismantled. Some of the protein fragments or peptides that result from the destruction process are not further degraded to the individual amino acids the protein was made of, but are instead transported to the cell membrane where they bind to the MHC. When a natural killer or killer T cell comes by it checks the antigens presented by the MHC to see if any of these antigens are foreign. If it does not find a non-self antigen, the killer cell will move on to the next cell. If it does find a non-self antigen, it will destroy the cell it was inspecting.

If the immune cells only react to non-self antigens and tumor cells originate from our own "self" cells, how can our immune system recognize these tumor cells? Well, there are several possibilities. Tumor cells carry many genetic mutations and some of these mutations will change the corresponding proteins in such a way that they are recognized as foreign. We also mentioned earlier that some viruses are known to cause cancer, so in this case the tumor cells (which were infected by the virus) may express viral antigens that can be recognized by the immune system. Finally, the gene expression profile of a tumor cell is very different from a normal cell and sometimes genes that are normally expressed at a low level are highly expressed in a tumor cell. Our immune system can then use the antigens derived from these genes to separate tumor from normal cells. Frank Macfarlane Burnet, a Nobel Prize-winning giant of immunology, was the first scientist to propose the concept of immune surveillance, in which the immune system recognizes and destroys tumor cells, in the 1950s (Burnet, 1970).

Despite the constant immune surveillance, there are several ways for tumor cells to escape an immune response. Sometimes tumor cells express less MHC proteins and can therefore avoid recognition by the killer cells. Tumor cells also divide quickly and their genome is rather unstable, so there is a good chance that new mutations will pop up during the growth of a tumor. If these mutations hit the antigens that the killer cells were able to recognize, the tumor cells could become unrecognizable to the immune system. Tumors can also create a so-called immunotolerant microenvironment. By suppressing genes and signals that would

otherwise stimulate T cell proliferation and activation or by expressing certain genes that downregulate the immune system, a tumor can turn its close environment into a safe haven where it can't be harmed by the immune system.

One example of a gene that downregulates the immune response is PD-1. This gene is expressed on the surface of T cells where it acts as an "off switch". If a specific protein (encoded by the PD-L1 gene) binds to the PD-1 receptor, the T cell can't be activated (Okazaki & Wang, 2005). The ability to prevent T cell activation is very useful, for example to reduce the risk of autoimmunity (when immune cells attack their own host body). The downside is that tumors can abuse this mechanism. Researchers have found that many tumors express PD-L1 and are therefore able to suppress T cell activation (Iway *et al.*, 2002). CTLA-4 is another example of such an "off switch" on T cells that is used by some tumors to prevent T cell activation (Hodi *et al.*, 2003).

The idea behind immunotherapy is that if we could somehow overcome the immune escape of tumor cells, if we could give the immune system of a cancer patient a nudge in the right direction, we would have a highly specific treatment at our disposal, much more specific than radio or chemotherapy. Researchers have already come up with a lot of different immunotherapy techniques, some more successful than others. The different types of immunotherapy can be split up in two main groups: active and passive immunity. Examples of passive immunotherapy are the administration of tumor-specific antibodies or killer cells. Antibodies are specialized proteins that bind to specific antigens on pathogens, infected cells or tumor cells. By binding to a tumor cell, an antibody tags the cell for destruction by the immune system. The other passive immunity technique involves the isolation of inactive killer cells from a patient and the activation of these cells in the lab, followed by their reinjection into the patient. Despite rapid initial responses to these treatments, they often fail to generate long-term immunity. They are called passive because there is no stimulation of any existing (but insufficient) immune responses, like in active immunity. Instead, the patient receives a sort of "ready-made" solution.

One way to actively stimulate a patient's existing, but failing, immune response against a tumor is to administer non-specific immune system stimulants. Injecting certain bacteria or general antibodies into a patient will evoke an inflammatory response, which in turn might result in a general activation of T cells and the elimination of tumor cells. Two more common and more successful approaches, which we used in this thesis, are the vaccination of a patient with tumor cells or tumor antigens and the use of cytokines and costimulators. It is easy to understand how the vaccination against certain viruses that are known to cause cancer (such as the hepatitis B virus or the human papilloma virus) will reduce the incidence of these cancers, but this is not the only option. Vaccination against a tumor-inducing virus is often preventive, rather than therapeutic, which means

it will reduce the chance that a tumor will develop, but it might not help fight established tumors.

One example of a therapeutic vaccine involves an antigen-presenting immune cell we have not discussed yet: the dendritic cell (Timmerman & Levy, 1999). Immature dendritic cells patrol the tissues that are in contact with the outside world, like our skin and the lining of our lungs and stomach, where they constantly sample the cells they meet, looking for foreign antigens (Banchereau & Steinman, 1998). When an immature dendritic cell stumbles across such an antigen, it ingests the antigen, processes it and displays it on its surface using the MHC. The dendritic cell is now a mature dendritic cell and it travels to a lymph node where it shows the antigen it found to all the T cells in this node. Once they have been exposed to the antigen, the T cells become activated and they are ready to go after any cell that carries this antigen.

Before we can use a patient's dendritic cells, we need to separate them from the patient's blood. Next, the dendritic cells are incubated in the lab together with certain tumor antigens. One group of antigens that is often used is the MAGE gene family, a subgroup of the cancer-testis antigens (Wilgenhof *et al.*, 2011). These genes are known as cancer-testis antigens, because they are normally only expressed in the germ cells of the testes. However, research has shown that they are also expressed in some cancers, for example in melanoma, hence the "cancer" in their name (Scanlan *et al.*, 2004). Once the dendritic cells have processed the antigens, they can be reinjected in the patient where they will present the antigens to the T cells, which might initiate an immune response against the tumor. In the third part of this chapter you can read how we used dendritic cell vaccines to treat metastatic melanoma patients.

Another way to actively stimulate the immune system, besides the vaccination we just described, is based on the role of costimulators in the immune response. The recognition of an antigen by the T cell receptor is not the only thing that controls the activity of a T cell. Activation requires several other signals (from the so-called costimulators) and T cells also have different receptors that are responsible for deactivation, such as PD-1 and CTLA-4. Earlier we explained how some tumors avoid destruction by using these "off-switches" to shut down T cells. In addition to the dendritic cell therapy, we also used an antibody to block CTLA-4 and thus prevent tumor-induced T cell deactivation in our melanoma study.

In the second part of this chapter we describe a therapy that targets the PD-1/PD-L1 interaction in lung cancer (Wrangle *et al.*, 2013). Just as with the CTLA-4 blocking therapy, the idea is that by interfering with the PD-1/PD-L1 interaction, we can stop a tumor from deactivating T cells. In our lung cancer project we actually focused on the combination of two treatments: the anti-PD-1/PD-L1 immunotherapy and the demethylating drug azacytidine. Not all tumors express

PD-L1 and if they don't, it doesn't make sense to use a therapy that targets PD-L1, unless we could somehow increase the expression of PD-L1 in these tumors. As it turns out, we might actually be able to do just that.

Azacytidine is a molecule whose structure resembles that of cytosine, the nucleotide in our DNA that can be methylated, and it can be used to remove methylation in living cells (Jones & Taylor, 1980). In low concentrations azacytidine inhibits the enzymes that are responsible for DNA methylation (the DNA methyltransferases or DNMTs) and in high concentrations it replaces cytosine in both DNA and RNA (Stresemann & Lyko, 2008). It is similar enough to cytosine to take its place, but different enough that it can't be methylated like cytosine, so wherever a methylated cytosine is replaced by azacytidine, methylation will be removed. In our paper we describe how lung cancer patients might benefit from a treatment with azacytidine before they receive immunotherapy. We found that this initial demethylation treatment increased the expression of PD-L1 in some tumors, which made them more susceptible for the anti-PD-1/PD-L1 immunotherapy that followed.

The gene expression and DNA methylation analysis techniques we described in the previous chapter are a crucial part of cancer research. Among many other things, they are used to investigate the differences between normal and tumor cells, to study what effect treatments have or to figure out why some patients respond to therapy and others don't. One particularly active field of cancer research is focused on the development of expression and DNA methylation biomarkers. Biomarkers are certain biological characteristics that can be used to say something about a person's health. The BRCA1 and BRCA2 mutations we mentioned earlier are examples of genetic biomarkers that can be used to predict a woman's chances of developing breast cancer (Robson, 2002). For these two genes there is a well-known causal link between their mutation and the development of breast cancer. This is not always the case though. For a lot of biomarkers we do not know if and how they cause a certain disease, we only know they are associated with that disease.

If researchers find a gene that is expressed in tumor cells, but not in normal cells, or a certain region of a gene that is methylated in patients that respond to a treatment while patients without methylation don't respond, they can use this information to develop new therapies or new methods to check whether a patient should get a certain therapy or not. The tailoring of treatments to individual patients is part of the relatively recent push towards personalized medicine. The idea behind personalized medicine is that instead of offering the same therapy to everyone who suffers from a certain disease, each patient should receive a treatment that is fully customized to her or his biologic background. In the early days of personalized medicine, the focus was mainly on the patient's genetic background, but nowadays it encompasses many other features (such as gene expression and DNA

methylation). The ultimate goal is to improve diagnosis and therapy efficiency.

In our melanoma project, we used RNA-seq to find expression differences between patients that benefit from the immunotherapy and those that don't. It would be extremely valuable to have an expression signature that separates responders from non-responders, because the current immunotherapies cost tens of thousands of euros and can have severe side effects. Such a signature could save some patients a costly and potentially dangerous therapy that wouldn't even help them.

We will round off this introduction with a relatively recent development in cancer research. Over the last few years a lot of resources have been poured into large-scale cancer genomics projects such as The Cancer Genome Atlas (TCGA). Using the high-throughput technologies from chapter one these projects analyze hundreds of patient samples for dozens of different cancer types and bundle the results in publically available databases. The TCGA database for example contains gene expression (at the transcript and the protein level), DNA methylation, mutation and clinical data. The goal of these projects is to enable researchers to test their hypotheses in much larger patient groups than they would otherwise have access to or even to come up with completely new ideas. Both in our lung cancer and melanoma project we used the TCGA database to check the findings from our small-scale studies. One downside of these databases is that they are often not easily accessible, especially to researchers without an informatics background. We decided to try and tackle this issue by developing an easy-to-use web tool for the visualization of the TCGA data on a single-gene level. MEXPRESS was born. The final part of this chapter describes this tool in detail.

Adapted from:

Wrangle J*, Wang W*, Koch A*, Easwaran H, Helai PM, Xiaoyu P, Vendetti F, Van Criekinge W, De Meyer T, Du Z, Parsana P, Rodgers K, Yen R, Zahnow CA, Taube JM, Brahmer JR, Tykodi SS, Easton K, Carvajal RD, Jones PA, Laird PW, Weisenberger DJ, Tsai S, Juergens RA, Topalian SL, Rudin CM, Brock MV, Pardoll D and Baylin SB. Alterations of immune response of non-small cell lung cancer with azacytidine. *Oncotarget* **4**, 2067–2079 (2013)

* These authors contributed equally

ALTERATIONS OF IMMUNE RESPONSE OF NON-SMALL CELL LUNG CANCER WITH AZACYTIDINE

ABSTRACT

Innovative therapies are needed for advanced Non-Small Cell Lung Cancer (NSCLC). We have undertaken a genomics-based, hypothesis-driving approach to query an emerging potential that epigenetic therapy may sensitize to immune checkpoint therapy targeting PD-L1/PD-1 interaction. NSCLC cell lines were treated with the DNA hypomethylating agent azacytidine (AZA - Vidaza) and genes and pathways altered were mapped by genome-wide expression and DNA methylation analyses. AZA-induced pathways were analyzed in The Cancer Genome Atlas (TCGA) project by mapping the derived gene signatures in hundreds of lung adeno (LUAD) and squamous cell carcinoma (LUSC) samples. AZA up-regulates genes and pathways related to both innate and adaptive immunity and genes related to immune evasion in several NSCLC lines. DNA hypermethylation and low expression of IRF7, an interferon transcription factor, tracks with this signature particularly in LUSC. In concert with these events, AZA up-regulates PD-L1 transcripts and protein, a key ligand-mediator of immune tolerance. Analysis of TCGA samples demonstrates that a significant proportion of primary NSCLC have low expression of AZA-induced immune genes, including PD-L1. We hypothesize that epigenetic therapy combined with blockade of immune checkpoints–in particular the PD-1/PD-L1 pathway–may augment response of NSCLC by shifting the balance between immune activation and immune inhibition, particularly in a subset of NSCLC with low expression of these pathways. Our studies define a biomarker strategy for response in a recently initiated trial to examine the potential of epigenetic therapy to sensitize patients with NSCLC to PD-1 immune checkpoint blockade.

INTRODUCTION

Innovative strategies are needed to treat the world's most common cause of cancer death, non-small cell lung cancer (NSCLC) (Siegel *et al.*, 2013, Youlden *et al.*, 2008). Less than a quarter of lung adenocarcinomas (LUAD) harbor genetic abnormalities for which targeted therapies have been derived. Early responses are often robust for these but are generally followed by acquired resistance (Vadakara *et al.*, 2012, Shepherd *et al.*, 2005). Lung squamous cell carcinoma (LUSC) has no approved targeted therapies and few effective chemotherapeutic options beyond the first line of therapy. In the current study, we offer a genomics-based, hypothesis-driving analysis to suggest a rationale for a novel combinatorial therapeutic approach to efficacious treatments for advanced NSCLC. The backdrop for the present study comes from our initial clinical trials in our Stand up to Cancer project (SU2C) in which patients with advanced, heavily-pretreated NSCLC received a form of "epigenetic therapy" combining low doses of the DNA hypomethylating agent azacytidine (AZA - Vidaza) and the HDAC inhibitor entinostat (Juergens *et al.*, 2011). Only two of now 65 patients treated to date have had RECIST (Response Evaluation Criteria In Solid Tumors) criteria responses to this therapy alone, but these were very robust and durable. A group of patients followed for 8 to 26 months responded to multiple different therapeutic regimens given subsequently, suggesting a "priming" effect of epigenetic therapy. Twenty-five percent of these patients with both LUAD and LUSC experienced RECIST criteria responses to their subsequent regimens. These subsequent therapies included not only standard chemotherapies but also immunotherapy targeting the PD-1 immune-checkpoint which when given alone has yielded responses in 16 to 17% of patients with advanced NSCLC (Brahmer *et al.*, 2012, 2013, Topalian *et al.*, 2012) (Supplementary Figure 3.1). While the number of patients who have received epigenetic therapy followed by immune checkpoint blockade is small, a clinical trial to evaluate potential sensitization to PD-1 immune checkpoint blockade with epigenetic therapy in patients with NSCLC has now begun.

This trial will be biopsy driven and offer the opportunity to examine hypotheses generated in the present pre-clinical work in order to develop biomarker strategies. In this regard, one of the key therapy agents being employed in the trial is AZA, a nucleotide analog DNA demethylating agent which blocks the activity of all three biologically active DNA methyltransferases (DNMT's) and also triggers degradation of these proteins in the nucleus (Gabbara & Bhagwat, 1995, Stresemann & Lyko, 2008). With respect to sensitization potential of this drug for immune responses, such targeting of DNMT's is known to induce increased expression of promoter DNA hypermethylated cancer testes antigens and also is reported to up-regulate other individual facets of the tumor immune stimulating profile, including major histocompatibility antigens, and transcription factors *IRF7* and *IRF5* (Li & Tainsky, 2011, Kulaeva *et al.*, 2003, Simova *et al.*, 2011, Fonsatti *et al.*,

2007, Claus *et al.*, 2005, Karpf *et al.*, 1999). In this regard, we previously reported that elements of such immune pathway activation were produced by low doses of DNA demethylating agents in a genomics based, pre-clinical approach (Tsai *et al.*, 2012). These studies demonstrated how low doses of AZA, which avoid early, cytotoxic and off-target effects, can provide a memory for a "reprogramming"-like effect on hematopoietic and selected examples of solid tumor cells (Tsai *et al.*, 2012). We hypothesize in this work that these effects may underlie the fact that significantly lowering doses of DNMT inhibitors in the clinic may account for the markedly decreased toxicity, and significant clinical efficacy, which has led to FDA approval of AZA for myelodysplasia (MDS) (Silverman *et al.*, 2002).

Initially, we focused our pre-clinical studies for low dose AZA on NSCLC. By first deriving genomic signatures of gene expression responses and DNA methylation for treated NSCLC lines, we observed in most cell lines a complex, multi-faceted up-regulation, involving hundreds of genes of the immune profile of these cells which includes the target of immune checkpoint therapy, the tumor ligand PD-L1. Moreover, using this extensive genomic signature, we have been able to specifically query hundreds of primary NSCLC samples in the Cancer Genome Atlas project (TCGA) for how basal expression of these immune genes and related DNA methylation events group lung cancers. We define a stark clustering of subsets of primary LUAD and LUSC for an "immune evasion" signature, which relates highly to events for low interferon pathway signaling and includes low levels of PD-L1 (Khong & Restifo, 2002, Tomasi *et al.*, 2006, Pardoll, 2012). Low expression of these genes closely matches those up-regulated by AZA treatment of the NSCLC cell lines. We hypothesize that these may be cancers which would benefit from AZA priming together with immune checkpoint therapy and outline a signature that may identify predictive biomarkers from biopsies forthcoming in the current trial.

MATERIALS AND METHODS

CLINICAL DATA

Institutional review board approved informed consent signed by each patient allowed the collection of clinical data following treatment on trial with epigenetic therapy. Relevant data were obtained by chart review. Representative images demonstrating responses to therapy were obtained from computed tomography series employed in the assessment of patient responses to anti- PD1 or anti-PD-L1 directed immune-checkpoint therapy. Assessment of response to treatment was performed by a single reference radiologist who employed (RECIST 1.0) to generate measurements for target lesions to be followed over the course of therapy.

Change in target lesions from baseline (%) is calculated by summing the diameter of all target lesions at each radiographic tumor evaluation and calculating percentage change at a given time point ([(Target Lesion SumTimepoint X/ Target Lesion SumBaseline)-1]*100).

TCGA SAMPLES

Level 3 RNA-Seq data (Illumina HiSeq RNA-Seq platform, Illumina, Inc., San Diego, CA, USA) were downloaded for 353 NSCLC samples (129 LUAD / 224 LUSC) and 54 adjacent non-tumor lung tissue samples from the TCGA Data Portal (https://tcga-data.nci.nih.gov/tcga/). Similarly, level 1 DNA methylation data (Illumina Infinium HumanMethylation450 BeadChip, Illumina, Inc., San Diego, CA, USA) were downloaded for 353 NSCLC samples (222 LUAD / 149 LUSC) and 74 adjacent non- tumor lung tissue samples. Among these, data for 174 NSCLC samples (80 LUAD / 94 LUSC) and 21 adjacent non-tumor lung tissue samples were available on both of the above platforms.

RNA-SEQ DATA ANALYSIS

We used TCGA level 3 RNA-Seq data (already normalized and quantified at gene levels), and presented as RPKM values (Reads Per Kilobase per Million mapped reads). To construct heatmaps: 1) Values of 0 (indicating no reads observed for a gene) in the RPKM data were set to NA; 2) the remaining RPKM values were log2 transformed; 3) genes from X and Y chromosomes were removed; and 4) heatmaps were made using the "heatmap.2" function in "gplots" package from CRAN being centered and scaled in the row direction, and using the default functions for computing distance and hierarchical clustering (or being specifically ordered in column according to the order of other heatmaps). Expression spectrums for individual genes were displayed in five quartile intervals following the order of associated heatmaps of the RNA-Seq data.

INFINIUM DNA METHYLATION DATA ANALYSIS

TCGA level 1 DNA methylation data contain raw binary intensity data files. Raw data files were imported into R (http://www.r-project.org) to calculate beta values (beta value Infinium = M / [U + M], M: mean intensities of the Methylated bead type, U: mean intensities of the Unmethylated bead types), M values (M value Infinium = log2 [M / U]) and detection p values (calculated by comparing probes to negative control probes to determine if signals are significantly different from the background) using the "methylumi" package from Bioconductor

(Gentleman *et al.*, 2004). Beta values and M values for probes with detection p value > 0.05 were considered not significantly different from background and were masked as NA. TCGA methylation data were first assessed for batch effects by principle component analysis (PCA) on the M values. To accomplish this, data points from X chromosome and Y chromosome as well as data points that are associated with SNPs (Single Nucleotide Polymorphisms) were removed, and the first two principle components were used for plotting. Spearman's correlation coefficients between methylation (beta value of probe, Illumina Infinium HumanMethylation450 BeadChip) and gene expression (RPKM value of gene, Illumina HiSeq RNA-Seq platform) were calculated using TCGA samples with available data on both platforms. For a particular gene, only methylation probes that have a negative Spearman's correlation coefficient and an adjusted p value (FDR) for the coefficient < 0.01 were considered informative and their relative distances to the corresponding transcriptional start site (TSS) of the genes were calculated from genomic coordinates obtained from the UCSC genome browser (http://genome.ucsc.edu). Heatmaps of the M values of informative probes were made using the "heatmap.2" function in "gplots" package from CRAN being centered and scaled in the row direction, and ordered according to the associated heatmaps of the RNA-Seq data in column and to the relative distances to TSS in row.

For in vitro DNA methylation values, DNA was extracted from cell lines that were either untreated or treated with AZA at day 3, at the end of treatment, and day 10 (7 days post end of treatment) and analyzed by the Illumina Infinium HumanMethylation450 BeadChips (Illumina, Inc., San Diego, CA, USA). Raw data were imported into R using the "methylumi" package from Bioconductor. Data points for probes with detection p value > 0.05 were masked as NA. Δ beta values (Δ beta value = beta value AZA – beta value Mock) were calculated and used to make boxplots. Heatmaps were made similarly like those for the TCGA data using informative probes defined by the TCGA data.

EXPRESSION MICROARRAY DATA

For in-vitro RNA extracted from cell lines treated with AZA, analyses were done at exactly the same time points as for DNA methylation above. Analyses from wild type colon cancer, HCT116 cells, and genetic knockout counterparts for DNA methyltransferases (DKO cells) were also performed. Expression microarrays were carried out using Agilent Human 4 × 44k expression arrays (Agilent Technologies, Santa Clara, CA, USA, Cat#: G4112F). Within-array and between-array normalization was performed using Loess and Aquantile normalization, respectively (Smyth & Speed, 2003). Median of the M values (M value Expression = log2 [AZA / Mock] OR log2 [DKO / HCT116]) was determined for multiple probes associated with the same gene.

GENE SET ENRICHMENT ANALYSIS (GSEA)

For each of the eight lung cancer cell lines (H838, H1299, H358, H1270, A549, H460, HCC4006, HCC827) a ranked gene list was created (genes were sorted by decreasing M value). These eight ranked gene lists were entered in the GSEA tool (Subramanian *et al.*, 2005, 2007) and the enrichment of both Kegg (Nakaya *et al.*, 2013) and Reactome (Joshi-Tope *et al.*, 2005) pathways in these lists was calculated (default parameters). A gene set was selected when it was enriched in any of the eight cell lines (p value < 0.05 and false discovery rate < 0.25). The normalized enrichment scores (NES) for the gene sets in each cell line were used to create the heatmaps. When a certain gene set was not significant in a cell line, it was assigned a NES of 0.

TRANSCRIPTION FACTOR ANALYSIS

Expression and methylation data were analyzed to find genes whose re-expression was linked to demethylation after AZA treatment. Genes were selected based on a set of cut-offs, both for the methylation and expression values: A gene was considered to be re-expressed when at day 3 or day 10 the median M value of all the probes linked to that gene was higher than 0.5. Infinium probes were analyzed separately at their distances from the transcription start site for each gene examined. For a probe to be called demethylated, it had to have a beta value higher than 0.5 in the mock treatment and a difference in beta value between mock and AZA treatment had to be at least 0.25. Only probes that were associated with a CpG island and that were located within 1000 bp upstream and 1000 bp downstream of the transcription start site were used in the analyses. The probes that passed these filters were validated using the TCGA methylation and expression data (see the definition of informative probes in the "Infinium DNA Methylation Data" section of Methods). Only genes that had an expression-methylation correlation value < -0.25 and a false discovery rate < 0.05 were retained. To better understand the biological implications of the re-expressed genes, the gene lists were searched for transcription factors. Two human transcription factor lists obtained from Ravasi *et al.* (2010) and Vaquerizas *et al.* (2011, 2012) were combined and the resulting list was matched to the lists of demethylated and re-expressed genes. The targets of IRF7 from the list of genes that are 4-fold or more up-regulated in H2170 by AZA were similarly identified using the TranscriptomeBrowser database (Lopez *et al.*, 2008).

FLOW CYTOMETRY METHODS (FACS)

Frozen cells were thawed in 37°C and washed once with flow-washing buffer.

Aliquots of single-cell suspension were then stained with fluorescent-labeled antibodies for 15 minutes at room temperature. Each sample was washed twice and re-suspended in flow-washing buffer and analyzed by FACSCalibur. The following antibodies were used: CD274 (12-5983- 42 Ebiosciences), HLA abc (12-9983-42 Ebiosciences), CD276(331606 Biolegend), CD119(558934 BD), B2 microblogumin(551337BD), CD58(555921BD). Changes between AZA treated and mock cells are calculated using mean fluorescence intensities (MFI) and the formula
$\log2([(MFI_{antibody,\ treated})-(MFI_{isotype,\ treated})]/\ [(MFI_{antibody,\ mock})-\ (MFI_{isotype,\ mock})])$.

PSCAN

PSCAN (http://159.149.160.51/pscan/, Zambelli *et al.*, 2009) is an online software tool that predicts the association of user defined gene-lists with transcription factors by scanning promoter sequences of co-regulated or co-expressed genes looking for over- or under-represented motifs. RefSeq IDs of the gene lists were obtained from BioMart (http://www. biomart.org/) and analyzed in PSCAN. Scanned promoter region was -450 to +50 base pairs around the transcription start site and employing TRANSFAC as the database for co-regulated or co-expressed genes.

RESULTS

CLINICAL DATA

Six patients who received treatment on a clinical trial of epigenetic therapy for advanced treatment-refractory NSCLC were placed on trials for immunotherapy targeting the PD-1/PD-L1 immune tolerance checkpoint. Of these six patients three have experienced durable partial responses to immunotherapy now ongoing for 14 to 26 months, and the other two had stable disease lasting 8.25 and 8.5 months. (Supplementary Figure 3.1, Supplementary Table 3.1) For comparison, 41-46% of NSCLC patients on these two trials of immunotherapy alone, one for anti-PD-1 and the other for anti-PD-L1 therapy, passed 24 weeks without progression and 16-17% had durable partial response rates (Brahmer *et al.*, 2012, 2013, Topalian *et al.*, 2012).

AZA INDUCED IMMUNE RESPONSE IN NON-SMALL CELL LUNG CANCER CELL LINES

We used our previously validated pre-clinical model to examine how AZA alters expression of key pathways in NSCLC cell lines (Tsai *et al.*, 2012). Cells were treated in vitro with 500 nM AZA for 72 hours then harvested immediately after withdrawal of drug and again one week later for genome wide methylation and expression studies. To the point of the clinical suggestion that epigenetic therapy may provide sensitization to subsequent immune-checkpoint blockade, we agnostically noted that one or more of the top ten pathways emerging for each cell line were immune related. The genes involved are important to the interaction of both innate and adaptive anti-tumor immunity. As earlier mentioned, other groups have described the ability of AZA to up-regulate individual immune pathway steps relative to assembly of major histocompatibility antigens (HLA Class I), interferon pathway genes, and cancer-testis antigens (Li & Tainsky, 2011, Kulaeva *et al.*, 2003, Simova *et al.*, 2011, Fonsatti *et al.*, 2007, Claus *et al.*, 2005, Karpf *et al.*, 1999). However, our current analysis reveals a more complex, concordant, broad immune gene signature. Gene Set Enrichment Analysis showed AZA induced up-regulation of multiple immune-related pathways in a manner roughly correlating to the degree of demethylation in response to AZA treatment (Figure 3.2a, Supplementary Table 3.2). Each of these components has a demonstrated role in immune tolerance pathways associated with immune checkpoints and immune evasion. Some of these genes have low expression associated with cancer-specific promoter region DNA hyper-methylation, and increased expression after treatment with DNA demethylating drugs (Li & Tainsky, 2011, Kulaeva *et al.*, 2003). In this regard, it is noteworthy that when compared to normal bronchial epithelial cells, NSCLC is known to exhibit diminished innate immune responses to viral-like stimuli involving intertwined pathways of cell- intrinsic responses to infection and inflammation (Li & Tainsky, 2011).

ANTIGEN PRESENTATION

A key step in tumor recognition and killing by cytotoxic T-cells involves recognition of peptides derived from tumor-specific antigens or up-regulated shared antigens bound to HLA Class I antigens expressed by the tumor cells (Raghavan *et al.*, 2008). As recognized by others, AZA increases expression of multiple cancer testes antigens including multiple MAGE family genes, whose expression has been shown to be suppressed by promoter hypermethylation (Fonsatti *et al.*, 2007, Claus *et al.*, 2005) (Figure 3.2g). AZA up-regulates not only transcripts of HLA Class I antigens but also a series of genes including, beta-2-microglobulin (*B2M*), *CD58*, *TAP1*, and the immuno-proteasome subunits *PMSB9* and *PSMB8* which encode proteins required for endoplasmic reticulum processing of, transport to,

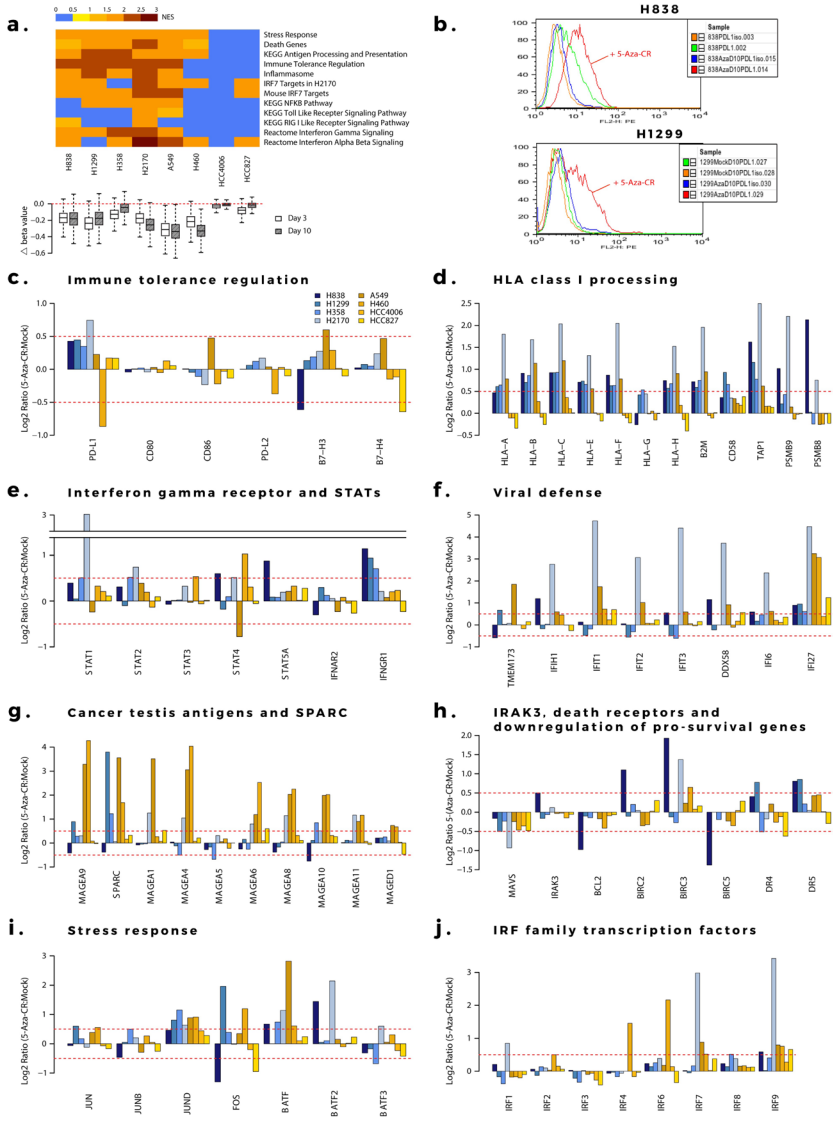


FIGURE 3.2 Azacytidine alters gene expression in NSCLC cell lines for multiple immune related pathways.

a. Top panel: Gene Set Enrichment Analysis (GSEA) for pathways up-regulated by azacytidine. Normalized enrichment scores are plotted as a heat map. Bottom panel: boxplot showing degree of demethylation in each cell line, as measured by the difference in beta values between the AZA and mock-treated cells immediately after drug withdrawal and 7 days later. **b**. FACS analysis shows increased level of cell surface PD-L1 after AZA treatment by day 10 in NSCLC lines H838 and H1299. **c**. to **j**. AZA-mediated expression changes at day 10 in key genes from pathways outlined in **a**. Y axis = Ratio of expression values (log2) of AZA -treated vs. mock-treated cells; X-axis = gene names.

and anchoring to the cell surface , and recognition of surface HLA class I subunits (Raghavan *et al.*, 2008, Procko & Gaudet, 2009, Challa-Malladi *et al.*, 2011) (Figure 3.2d). We find generally good correlation between HLA Class I, B2M, CD58, and B7-H3 transcripts and protein on the cell surface by flow cytometry (Supplementary Figure 3.2). Importantly, mutations potentially contributing to immune evasion have been described in *HLA-A* in a small percentage of LUSC and of *B2M* and *CD58* in other tumor types (Challa-Malladi *et al.*, 2011, Hammerman *et al.*, 2012).

TYPE I AND II INTERFERON SIGNALING

A second key issue for immune cell interaction with tumor cells is that, *in vivo*, AZA administration to tumor-bearing mice has been shown to induce antigen processing and presentation genes, particularly when administered with CpG TLR9 agonists, and this is largely attributed to interferon-γ production by lymphocytes (Simova *et al.*, 2011). While the lymphocyte-specific γ-interferon is not induced in NSCLC lines with AZA treatment, there is up-regulation of the interferon-γ receptor (*IFNGR1*) as well as of multiple STAT genes, including *STAT1*, the major *IFNGR1* signal transducer (Figure 3.2e).

PROGRAMMED CELL DEATH AND VIRAL DEFENSE

The re-expressed genes in the above mentioned pathways are downstream targets of interferon response pathways in a fashion closely linked to pro-inflammatory and viral defense responses (Strowig *et al.*, 2012, Ishikawa *et al.*, 2009, Sharma & Fitzgerald, 2010, Hsu *et al.*, 2012). In turn, triggering of these responses can have both tumor repressing activities, such as apoptosis, or tumor promoting events and this paradox has been termed "the dual face" of inflammation (Ishikawa *et al.*, 2009, Sharma & Fitzgerald, 2010, Dunn *et al.*, 2002). In this regard, we see key subsets of immune related genes that are up-regulated by AZA with potential for inhibiting tumor growth including *IFI27*, which encodes a protein triggering apoptosis in late stages of chronic viral infection (Cheriyath *et al.*, 2011) (Figure 3.2f). Simultaneously, there is down-regulation of the anti-apoptotic gene, *MAVS*, a change which accompanies activation of the RIG I signaling pathway in response to viral challenge (Sharma and Fitzgerald, 2010, Hsu *et al.*, 2012, Xu *et al.*, 2010) (Figure 3.2h). Downstream events in viral response include, especially in line H838, simultaneous increases for expression of BIRC family autophagy genes and simultaneous decreases in the anti-apoptotic genes *BCL2* and *BIRC5* (*SURVIVIN*) (Yang & Klionsky, 2010) (Figure 3.2h). Indeed, suppression of *SURVIVIN* is known to be triggered by the viral induction of *IRAK3*, which encodes an IL-1 receptor associated kinase (De Carvalho *et al.*, 2012). *IRAK3* is, again

in H838 cells, up-regulated by AZA concordantly with the death related genes mentioned just above (Figure 3.2h). These dynamics are similar to those for colon cancer cells where *IRAK3* is silenced in association with promoter-region DNA hypermethylation and when reactivated by induced demethylation, is associated with *SURVIVIN* down-regulation (De Carvalho *et al.*, 2012).

PD-L1 EXPRESSION

The key to immune checkpoint therapy is antibody targeting of either the receptor PD-1 on immune cells and or the ligand PD-L1 on tumor cells. In the clinical trials for immune check point blockade to date involving NSCLC patients, a subset showed no responses when their tumors did not express cell surface PD-L1 (Brahmer *et al.*, 2012, Topalian *et al.*, 2012, Pardoll, 2012). In this regard, when treated with AZA, several NSCLC cell lines up-regulate PD-L1, not only at the transcript level but also at the cell surface protein level (Figure 3.2b, c). Notably, this AZA increase of *PD-L1* in cell lines is far more consistent than for *PD-L2*, a second dendritic cell/macrophage ligand for the CTL PD-1 receptor, or other checkpoint ligands such as *B7- H3* and *B7-H4* (Figure 3.2c). Similarly, *CD80* and *CD86*, the ligands for CTLA4, another therapeutically targeted immune checkpoint receptor, are not altered (Figure 3.2c). *PD-L1* expression in tumor cells can either be driven by cell-intrinsic mechanisms or by a process termed adaptive resistance, through interferon-γ signaling and subsequent activation of STAT transcription factors, which we also see induced by AZA (Figure 3.2e).

AZA ALTERS THE IMMUNO-PHENOTYPE OF NSCLC THROUGH ITS EFFECT ON DNA METHYLTRANSFERASES

A key issue for all of the above responses is whether these represent attributes of AZA as a targeted therapy. In this regard, this drug, particularly at less toxic doses, specifically targets the three biologically active DNMT's, acting to directly inhibit their catalytic sites and triggering degradation of these proteins in the nucleus (Gabbara & Bhagwat, 1995, Santi *et al.*, 1984). We thus queried how our complex, immune-related, pharmacologic responses compare to simultaneous genetic depletion of two of the three DNMT's. We compared HCT116 colon cancer cells and HCT116 double knock out (DKO) cells that have been genetically disrupted to give severe haplo-insufficiency of DNMT1, and complete absence of DNMT3B, enzymes for DNA methylation maintenance and *de novo* DNA methylation, respectively (Rhee *et al.*, 2002). These cells have lost the majority of their genome-wide DNA methylation and have de-methylation of many cancer specific, promoter region, DNA hypermethylated CpG islands with corresponding re-expression of genes silenced in the wild type HCT 116 cells (Rhee *et al.*,

2002). From the standpoint of the present studies, the immune-related expression alterations in DKO versus wild type HCT116 are remarkably similar to the AZA induced changes in NSCLC cells (Figure 3.3). We conclude that previously described off target effects of high dose AZA including incorporation into RNA and DNA as an abnormal nucleotide (Stresemann & Lyko, 2008) do not appear to be required for the drug's effect that we have defined.

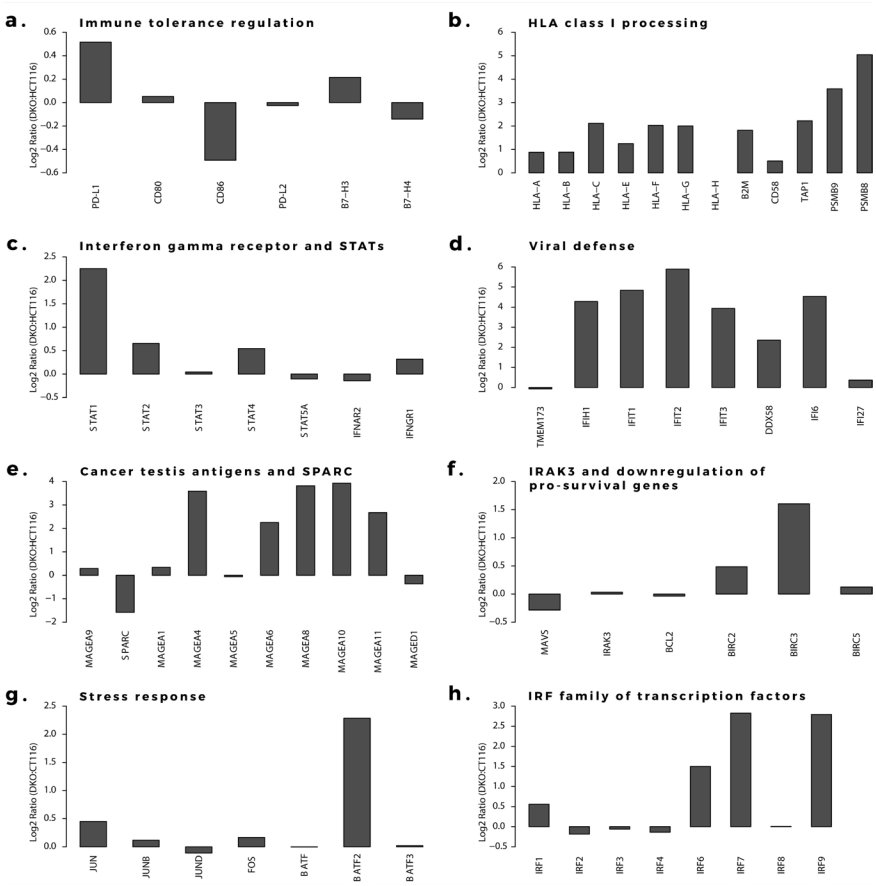


FIGURE 3.3 Genetic knock out of DNA Methyltransferases mimics the effects of azacytidine mediated immune pathway up-regulation.

Gene expression alterations when comparing wild-type HCT116 colon cancer cells to their isogenic DNMT1 and 3B knockout counterpart (DKO). The gene expression differences are given as the log2 ratio of expression in DKO over wild-type HCT116 (Y-axis) and the gene panels, A-H correspond to panels **c** to **j** in Figure 3.2 for the NSCLC cell lines treated with AZA.

UP-REGULATION OF IMMUNE RELATED TRANSCRIPTION FACTORS BY AZACYTIDINE

In order to find specific genes re-expressed in response to AZA which may be driving immune-related changes we extensively filtered our genome-wide expression and methylation data from cell line experiments to identify transcription factors meeting criteria of epigenetically re-expressed genes. We found approximately 300 genes with high baseline promoter region CpG island methylation, promoter demethylation of 25% or more after treatment and an increase in expression of log2(0.5) (1.4- fold) or greater after treatment (Figure 3.4a, Supplementary Table 3.3). Nearly 17% are in an interferome database (Samarajiwa *et al.*, 2009) (http://www.interferome.org), and 19% are transcription factors. The transcription factor IRF7 has been reported by others to be hypermethylated in cancer, as it is in our NSCLC line with the lowest basal expression (Li & Tainsky, 2011, Bidwell *et al.*, 2012, Jee *et al.*, 2009, Lu *et al.*, 2000). It is up-regulated in response to AZA in several cell lines, most prominently in the LUSC cell line H2170, showing a 9-fold increase (Figure 3.2j). IRF7 is an upstream activator of functions in cellular pathways recognizing the virus response element VRE-A to increase transcription of genes involved in type 1 IFN signaling (Li & Tainsky, 2011). There is a significant association of *IRF7* transcription targets with genes driving several of our GSEA enrichment scores for the immune pathway alterations observed in response to AZA (Figure 3.4b).

IMMUNE-PHENOTYPES WITHIN HISTOLOGIES IN THE CANCER GENOME ATLAS

From our analysis suggesting *IRF7* to be a potentially important cancer-specific hypermethylation induced down-regulation event, we sought to create a list of functionally derived genes closely associated with its re-expression. Examining H2170, the LUSC cell line with the greatest up-regulation of *IRF7* we hypothesized that other genes highly up-regulated in this cell line might be targets of this transcription factor (Figure 3.2j). Filtering expression array data, 114 genes where found to be 4-fold or more up-regulated in response to AZA in the H2170 (Supplementary Table 3.4). The association of this functionally derived gene list with *IRF7* is confirmed by PScan analysis ($p = 7.6e\text{-}18$) (Figure 3.4b). These data suggest that IRF7 silencing by DNA methylation in tumors could result in suppression of immune-regulatory genes important for the surveillance of tumors by cytotoxic immune mechanisms. Other studies have reported an immune-evasion signature dependent on IRF7 in breast and melanoma (Bidwell *et al.*, 2012, Carretero *et al.*, 2012). To test if such relation between IRF7 and

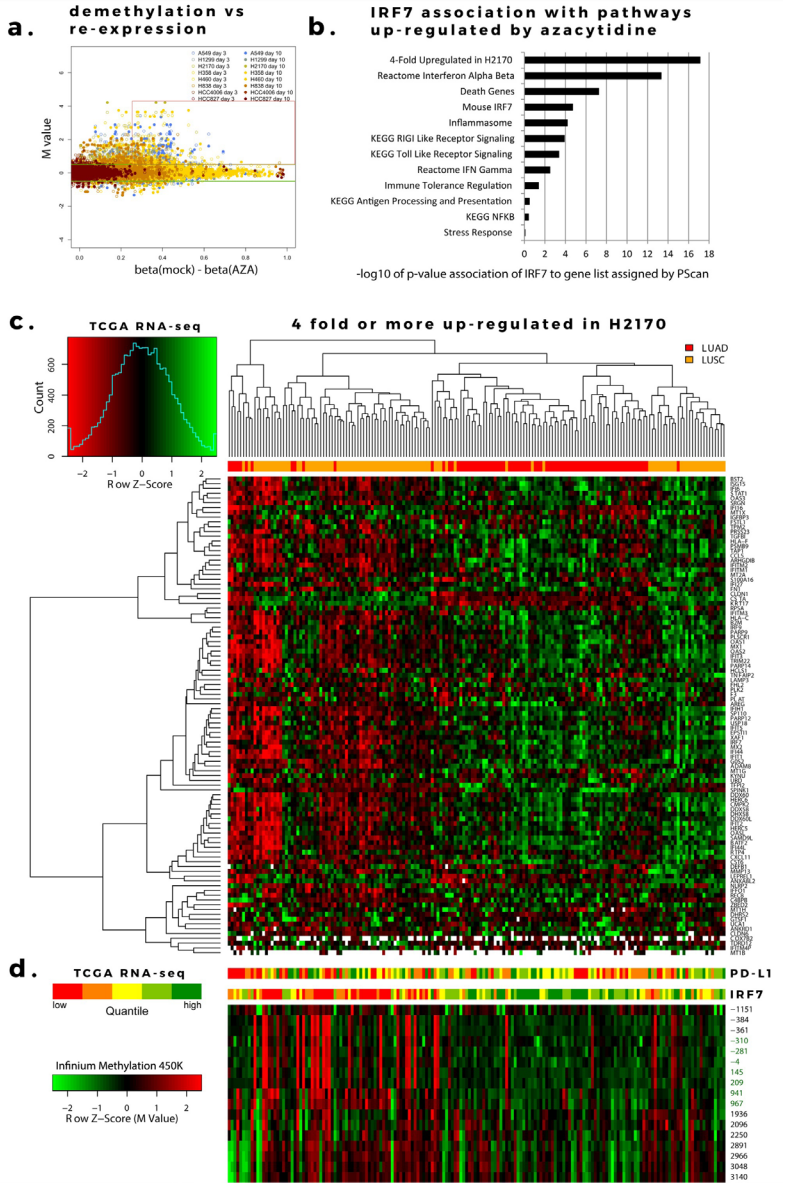


FIGURE 3.4 Identification of azacytidine up-regulated transcription factors and interferon signaling related genes, and their clustering of primary Non-Small Cell Lung Cancer in TCGA.

a. Identification of genes in Non-Small Cell Lung Cancer cell lines with low basal expression and high basal promoter region DNA methylation which are demethylated and re-expressed after AZA treatment. The red box encompasses genes meeting these criteria which are described specifically in methods. Among these, IRF7, a key immune-related transcription factor, was up-regulated in multiple cell lines. **b**. Pathways up-regulated in NSCLC cell lines in response to AZA are enriched for IRF7 targets as determined by PScan analysis (-log10 of p-values) and gene set enrichment analysis. **c**. Heatmap of RNA-Seq expression

levels in primary lung cancers from TCGA for genes 4-fold or more induced by AZA in the LUSC cell line H2170, the cell line with the greatest degree of IRF7 up-regulation. Top bar: red indicates LUAD and orange indicates LUSC samples. Genes used in the heatmap are listed in supplemental table 4. **d.** Bar panels show expression of PD-L1 and IRF7 in five quantile intervals (red for lower and green for higher expression). Heatmap immediately below IRF7 expression bar shows corresponding Infinium platform DNA-methylation levels (Z-scores, red for more and green for less methylated) across the promoter region. Positions relative to transcription start site are shown to the right. CpG-island probes are labeled in green. Sample order in bar plots and methylation heatmap is maintained from the main heatmap.

immune-regulatory genes exist in primary LUAD and LUSC tumors, we analyzed the expression of these genes as a function of IRF7 expression, and its promoter methylation status. We found that low expression of these genes describes a subgroup, particularly among LUSC, in TCGA samples which clusters tightly with high promoter region DNA methylation and low expression of *IRF7* (Figures 3.4c, 3.4d and 3.5). Finally, expression levels of *PD-L1*, the key tumor ligand targeted in the anti-checkpoint immunotherapy trials, tracks quite well with the above immune evasion signature in subgroups of not only LUSC, but also LUAD, as especially well visualized in heatmaps for individual immune related pathways, which each track closely with an immune evasion signature in the LUSC and LUAD (Figure 3.5).

DISCUSSION

In the present work, we have used an in-vitro model to derive a pre-clinical understanding of the immunomodulatory effects of clinically relevant doses of AZA in NSCLC that may underpin its potential to "prime" for subsequent response to PD-1 pathway blockade. We characterize an AZA induced expression signature of immune genes and pathways in NSCLC known to play a role in the down-regulation of immune surveillance of cancer. However, concomitant with induction of the immune genes comprising both innate and adaptive immunity is the up-regulation of a primary immune inhibitory ligand, PD-L1. Our data therefore suggest a mechanism by which epigenetic therapy might improve the outcome of treatment of patients with NSCLC with PD-1/ PD-L1immune checkpoint blockade. By matching these basal gene expression and DNA methylation patterns, including that of a core interferon pathway transcription factor, IRF7 in the TCGA project, we extrapolate our in vitro AZA-induced gene signature to hundreds of primary NSCLC cancers. These results suggest that a major effect of AZA treatment is the alteration of tumor immune-inducing pathways that could lead to susceptibility

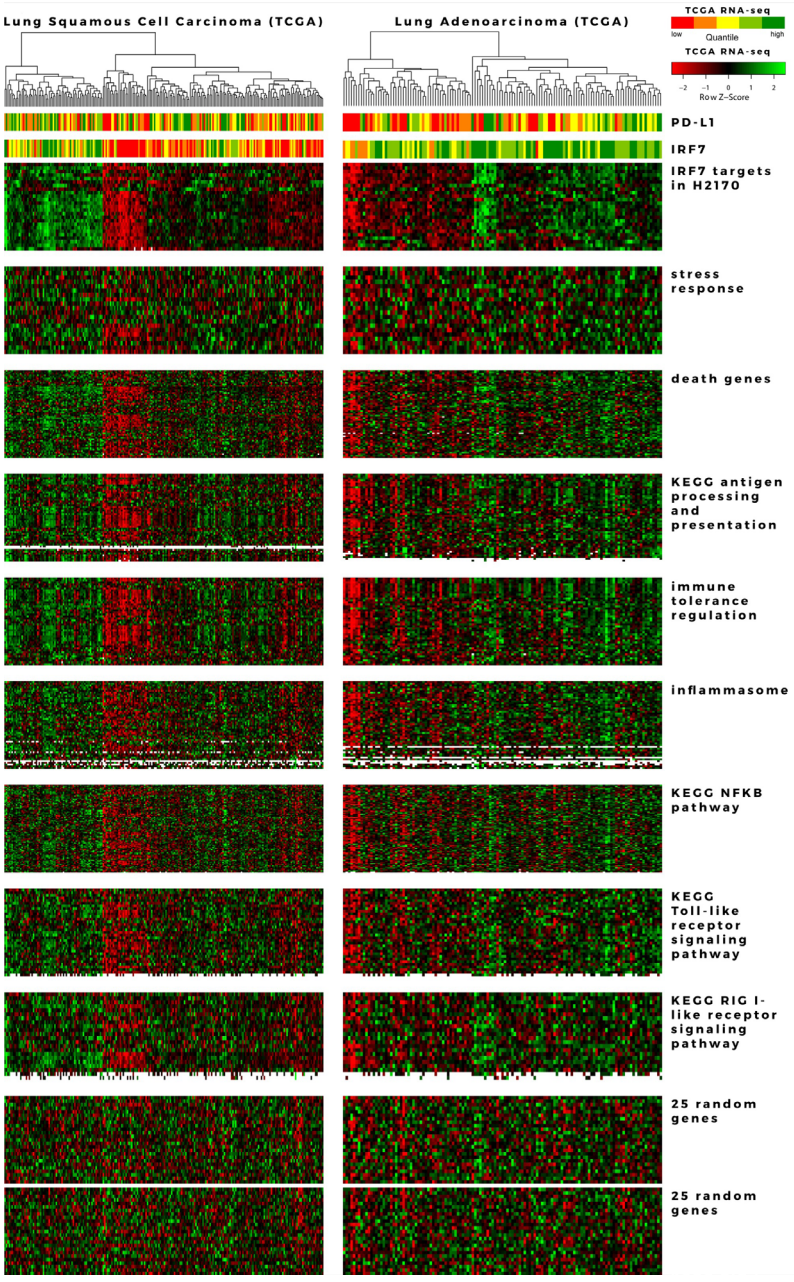


FIGURE 3.5 Relationship of azacytidine-induced, immune-related pathways to primary lung tumors grouped by expression of IRF7-associated genes.

TCGA samples are ordered by unsupervised clustering based on genes highly up-regulated in H2170, which are enriched for IRF7-targets, represented in the topmost heat map. Order of samples is maintained in all lower heat maps. PD-L1 and IRF7 expression are depicted in the top bar panels as in Figure 3.4d. Supplemental Table 3.5 table shows the overlaps

of genes from each pathway represented in the heat maps. That the observed clustering pattern is not due to chance or batch effect is demonstrated using random sets of 25 genes shown in the bottom two panels.

of tumor cells themselves to immune attack by T cells. In particular, because the inhibitory ligand PD-L1 is up-regulated by AZA in our cell lines, and subsets of primary tumors have concordant low-expression of AZA induced immune genes and PD-L1, we suggest that combination of epigenetic therapy and PD-1 pathway blockade might produce a synergistic anti-tumor response.

Our findings provide a basis for biomarker approaches that we will test in a just initiated trial for patients with advanced LUAD and LUSC, aimed at validating the promise for sensitization by epigenetic therapy to immune checkpoint therapy. If we continue to see robust patient efficacy, our data may prove key to determining which individuals are likely to benefit from the epigenetic therapy approaches we are testing in clinical trials by evaluating gene panels for expression and DNA methylation in pre and post-drug administration biopsies.

ACKNOWLEDGMENTS

This paper was supported by grants from CA058184, National Cancer Institute (NCI), Stand Up To Cancer Epigenetics Dream Team (SU2C) - the Samuel Waxman Cancer Research Foundation and the Hodson Trust. We gratefully acknowledge the TCGA consortium for creation of the public database from which we queried RNA-Seq gene expression data and DNA methylation status for selected genes. Drs. Laird (Principal Investigator - USC) and Baylin (co-Principal Investigator - JHU) lead the epigenetic analyses in TCGA, while Dr. Weisenberger leads the efforts to perform the DNA methylation analyses. We thank Kathy Bender for the preparation of the manuscript.

Adapted from:

Seremet T*, Koch A*, Wilgenhof S, Schreuer M, Jansen Y, Del Marmol V, Lièvard D, Thielemans K, Schats K, Kockx M, Van Criekinge W, Coulie PG, De Meyer T, Van Baren N & Neyns B. Identification of a predictive signature for response to ipilimumab-based immunotherapy in metastatic melanoma based on immunohistochemical, RNA-sequencing and epigenetic profiling – *manuscript in preparation for submission*

* These authors contributed equally

A PREDICTIVE SIGNATURE FOR RESPONSE TO IMMUNOTHERAPY IN MELANOMA METASTASES

ABSTRACT

Ipilimumab (ipi) improves the survival of patients with advanced melanoma and combination of ipi with an autologous monocyte-derived DC therapy (TriMixDC-MEL) may further improve patient outcome. A predictive melanoma tissue signature for the clinical efficacy of ipi and TriMixDC is needed to optimize individualized treatment strategies. We analyzed the expression and DNA methylation profiles of metastatic tumors from melanoma patients who were treated with ipi, TriMixDC-MEL or a combination of both, using immunohistochemistry, RNA sequencing and MBD sequencing. Patients were classified in three groups (high, intermediate and no clinical benefit, or HCB, ICB and NCB) based on their response to therapy. We found a higher number of CD8+, PD-L1+ and CD20+ cells in the HCB group compared to the NCB group. The RNA sequencing experiment resulted in a list of 195 genes that were differentially expressed between HCB and NCB samples (false discovery rate < 0.05). This gene list was enriched for immune-related ontologies and included many genes that reflect a humoral (*IGHM*, *IGHA1*, *IGHV3-23*, *BANK1*, *SELL*, *IGLV3-1*, *IGHG1*, *IGSF6*) and cellular (*CD69*, *FYB*, *CARD11*, *CD244*, *TIGIT*) immune response. Differential expression did not appear to be driven by differences in DNA methylation. Together with the immunohistochemistry results, the RNA sequencing analysis revealed a distinct immune system-related expression profile in metastatic melanoma patients that responded to immunotherapy. These results are a first step towards the development of gene signature for the prediction of therapy response, but further validation is needed.

INTRODUCTION

Melanomas are antigenic tumors against which most melanoma patients spontaneously mount T cell responses. Immunotherapy aims at increasing these spontaneous responses or at stimulating new ones. The first drug that improved the survival of metastatic melanoma patients was ipilimumab (ipi), a CTLA-4 targeting monoclonal antibody and T cell activation checkpoint inhibitor (Hodi *et al.*, 2010). Ipi improves overall survival (OS) of melanoma patients and induces a long-term survival benefit with a plateau at 3 years in 20% of the patients (Maio *et al.*, 2015, Schadendorf *et al.*, 2015). Different strategies to improve the efficacy of ipi are currently being investigated, including combinatorial schemes with vaccination approaches as well as its use together with antibodies that block PD-1, another T cell activation checkpoint inhibitor (Postow *et al.*, 2015). In our center we established an autologous monocyte-derived dendritic cells (DC) vaccination approach combining intradermal and intravenous administration. The DC were electroporated with synthetic messenger RNA (mRNA) that encodes a CD40 ligand, a constitutively active Toll-like receptor 4 and CD70, together with mRNA encoding fusion proteins of a human leukocyte antigen (HLA)-class II targeting signal (DC-LAMP) and a melanoma-associated antigen, either MAGE-A3, MAGE-C2, tyrosinase or gp100 (TriMixDC) (Van Lint *et al.*, 2014). This vaccination therapy already showed anti-tumor activity in patients with advanced melanoma (Wilgenhof *et al.*, 2013).

We conducted a phase II clinical trial to investigate the activity of the TriMixDC in combination with ipi. This combination showed superior activity compared to ipi alone with a 38% best overall response rate (BORR) (Neyns *et al.*, 2014), indicating that combining DC vaccination with immunomodulatory agents translates in a better clinical outcome. One difficulty in clinical practice is to identify the patients that will have a good outcome on immunotherapy. Recent findings showed that a specific neoantigen landscape is present in tumors that respond to ipi (Snyder *et al.*, 2014). The identification of this neoantigen landscape involves a complex analysis that requires both whole-exome sequencing and patient-specific HLA typing to identify candidate tumor neoantigens for each patient. We need predictive biomarkers that are easier to test in the clinical setting in order to further optimize individualized treatment strategies. In the present study we used immunohistochemical (IHC), RNA-sequencing (RNA-seq), and whole genome DNA methylation analyses in order to characterize the profile that identifies long-term responders to ipi and TriMixDC-based immunotherapy.

MATERIAL AND METHODS

PATIENTS AND TISSUE SAMPLES

Melanoma metastases samples were collected between January 2011 and May 2013 from patients who received immunotherapy in academic trials carried out at the VUB university hospital (http://clinicaltrials.gov NCT01676779 and NCT01302496) or ipi outside of a clinical trial after its approval. Informed consent has been obtained from all patients. The metastatic tumors were divided in three or two parts depending on their size. In general one part was processed for formalin-fixed, paraffin-embedded (FFPE) preservation, the second part was frozen immediately and the third part was preserved in RNAlater stabilization reagent. Samples were collected before or after therapy onset when progression occurred. The FFPE samples were used for diagnosis confirmation in the Pathology Department of our hospital and further automated quantification of immune cells with Definiens platform in HistoGeneX Laboratories, while freshly frozen and RNAlater samples were used for translational research. The first clinical trial was a 2-arm 1-stage randomized and controlled phase II study for disease-free patients without any prior systemic therapy. The second clinical trial was a 2-stage phase II single-arm trial for patients with AJCC stage III (unresectable) or stage IV melanoma of the skin, or unknown primary site.

In the first trial patients received combined intradermal and intravenous administration of autologous monocyte-derived DCs electroporated with synthetic messenger RNA (mRNA) encoding a CD40 ligand, a constitutively active Toll-like receptor 4 and CD70, together with mRNA encoding fusion proteins of a human leukocyte antigen (HLA)-class II targeting signal (DC-LAMP) and a melanoma-associated antigen, either MAGE-A3, MAGE-C2, tyrosinase or gp100 (TriMixDC). In the second trial (TriMixIpi) patients received TriMixDC vaccination (4×10^6 cells id and 20×10^6 iv, q3wks × 4) combined with ipi (10 mg/kg q3wks × 4), followed by ipi maintenance therapy (10 mg/kg q12w, in patients who were progression-free at week 24) (Supplementary Figure 3.3). For the purpose of biomarker analyses, clinical activity was defined as a three-level clinical benefit derived from investigator assessment of BORR: high clinical benefit (complete responders and long-term partial responders), intermediate clinical benefit (stable disease or partial response < 24 weeks) and no clinical benefit (progressive disease).

PATIENT AND SAMPLE CHARACTERISTICS

Table 3.2a illustrates the clinical characteristics of the 25 patients included in the analysis. The median age of this patient population almost equally distributed by gender was 44 years and ranged from 25 to 67. The majority of the patients (56%) received concomitant administration of ipi and TriMixDC in the TriMixIpi trial (https://clinicaltrials.gov NCT01302496) (Supplementary Figure 3.3), while 36% received ipi alone. Two patients received TriMixDC vaccination as first immunotherapy in the randomized controlled phase II trial available at the VUB university hospital (http://clinicaltrials.gov NCT01676779). Five patients (20%) presented durable partial or complete response and these were considered to have high clinical benefit (HCB). Among them four patients were included in the TriMixIpi trial, and one patient received ipi alone (Supplementary Table 3.1). 14 patients (56%) showed no clinical benefit (NCB) from immunotherapy, while six patients (24%) showed a partial response or stable disease as BORR, but no longer than 24 weeks. These six patients were grouped in an intermediate clinical benefit (ICB) group. The overall survival in the HCB was 34 months with four responses still ongoing compared with 10 months in the NCB group (Kruskal-Wallis test, p value = 0.001), while the overall survival for ICB was 26,5 months with only two still ongoing (Kruskal-Wallis test, p value = 0.03) (Figure 3.4). The majority of the samples (73,1%) were skin, subcutaneous or lymph node metastases that are accessible and can be easily removed by surgery. The remaining samples were from lung, liver, small intestine, brain, and adrenal gland metastatic tumors (Table 3.2b).

IMMUNOHISTOCHEMISTRY

Seven μm-thick cryosections were obtained from frozen OCT-embedded tissue samples, air dried, and stored at -80°C until use. The cryosections were thawed and fixed in 4% paraformaldehyde before staining. Consecutive sections from the same tumor sample were stained on the automated Dako Autostainer system using Dako/Thermo reagents. Sections were incubated with unlabeled primary antibody, washed, and incubated with a secondary polyclonal goat anti-mouse antibody coupled to horseradish peroxidase. Staining was performed for HE and 12 markers: PanMel, MCSP, CD3, CD8, CD20, CD163, DC-LAMP, Casp-3, Ki-67, PHH3, HLA class I and vWF.

Stained slides were digitized by automated whole-slide image capture, using a Mirax Midi scanner (Carl Zeiss MicroImaging), equipped with a Zeiss Plan-Apochromat 20_ NA 0.80 objective lens and a Hitachi HV-F22 acquisition camera, providing an object pixel size of 0.23 μm. Image acquisition was controlled with the Mirax Scan software (Zeiss). Image files were analyzed with the Mirax Viewer

Table 3.2 Clinical characteristics of the melanoma patients.

a. This table lists the number of patients in function of their gender, the therapy they received and their response to this treatment. b. Here we list the number of samples together with their tissue origin. There are more samples than patients, because we sometimes obtained more than one sample from the same patient.

a.

variable	# patients
gender	
male	12
female	13
type of immunotherapy	
TriMixIpi	13
Ipi	10
TriMixDC	2
response to immunotherapy	
HCB	5
ICB	6
NCB	14

b.

variables	# samples
biopsy time point	
before treatment	15
after treatment	11
tissue type	
lymph node	8
skin	4
nodule sc	7
adrenal gland	1
liver	1
small intestine	2
brain	1
lung	2

software (Zeiss) and Pannoramic Viewer (http://www.3dhistech.com/pannoramic_viewer).

DNA AND RNA EXTRACTION FROM FROZEN TISSUE SAMPLES

RNAlater samples or, when RNAlater samples were not available, a series of 10 cryosections of 20 μm thickness from frozen metastases, were processed for the simultaneous extraction of DNA and RNA by ultracentrifugation through a caesium chloride (CsCl) gradient, followed by extraction and purification with the NucleoSpin RNA II kit. A first assessment of the RNA and DNA concentration was performed by spectrophotometry with the NanoDrop tool.

MBD-SEQ

MBD-sequencing was performed using the MethylCap kit (Diagenode, Belgium) as described by De Meyer *et al.* (2013) with some minor modifications. After quality control (QC) of the extracted genomic DNA with the picogreen dsDNA assay (Life Technologies, USA) the DNA was sheared using the Covaris S2 ultrasonicator to obtain 200 bp fragments (intensity 5, duty cycle 10%, 200 cycles/burst,

190 s). A second QC step was performed using the high sensitivity DNA chip (Diagenode) and 300 ng of the sheared DNA of each sample was subjected to MBD capture with the MethylCap kit. This kit uses the methyl-binding domain of the MeCP2 protein to enrich for methylated DNA fragments. Next, the fragment library for the sequencing experiment was prepared using the NEBNext Ultra DNA library prep kit (New England Biolabs, USA) and the NEBNext Multiplex Oligos for Illumina protocol (index primers set 1). Adapters were diluted 1/10 before ligation and fragments were size selected on an E-Gel EX agarose 2% gel (Invitrogen, Life Technologies) by cutting out 300 ± 50 bp fragments and purifying them with the Zymoclean Gel DNA recovery kit (Zymo Research, USA). The library was amplified with 13 PCR cycles and purified a second time using AMPureXP beads. A High Sensitivity DNA Chip (Agilent Technologies, USA) was used for a final QC of the libraries and the concentrations were assessed by qPCR according to the Illumina protocol "qPCR quantification protocol guide". The samples were sequenced on an Illumina Genome Analyzer IIx and the resulting paired-end reads (2 × 50 bp) were mapped to the human genome (GRCh37) using Bowtie (Langmead *et al.*, 2009). We have previously created a map of the human methylome (available online at http://www.biobix.be/sample-page/) based on the DNA methylation profiling of 80 different cell line and tissue samples. This map was used to delineate a list of a finite number of genomic locations where methylation can be observed (referred to as methylation cores), thereby reducing data complexity. The mapped reads of the different melanoma samples were converted to methylation cores and for each core the peak height was calculated as the maximum coverage within this core. The full list of methylation cores that were used can be found at http://www.biobix.be/map-of-the-human-methylome/.

RNA-SEQ

The QC of the extracted total RNA was performed using the RNA 6000 pico chip (Agilent Technologies) and the Ribogreen assay (Life Technologies). Isolation of mRNA, synthesis of cDNA and library preparation were carried out using the NEBNext Ultra Directional RNA Library Prep kit and the NEBNext Multiplex Oligos for Illumina protocol (index primers set 1, New England Biolabs). 14 PCR cycles were used to amplify the libraries and the DNA 1000 chip (Agilent Technologies) was used for the QC. The final mRNA concentrations were assessed by qPCR (as described for MBD-seq). The libraries were sequenced on the HiSeq 2000 platform (2 × 50bp) and the resulting paired-end reads were mapped to the human genome (GRCh37) using TopHat 2 (Kim *et al.*, 2013). Per-gene count values were calculated with the HTSeq-count software (Anders *et al.*, 2015). Two technical repeats were available for each sample and both read counts were summed to combine the repeats. Finally, the read counts were aggregated per gene by considering the maximal value.

DATA ANALYSES AND STATISTICAL METHODOLOGY

The immunohistochemical evaluation of the samples provided categorical data as well as continuous data when quantification of the different cells types was made. Fisher's exact test was used for the analyses of the contingency tables for the categorical variables. The continuous variables were characterized by median, percentile 25 and percentile 75, and the Kruskal-Wallis test was used for statistical analyses of the IHC results. The analyses were performed using R and SPSS software. Before the differential analysis both the MBD and RNA-seq datasets were filtered to remove low coverage data points (average coverage > 1 and 25% of the samples must have a coverage > 1). The differential analysis of the MBD and RNA-seq data was performed in R (version 3.1.2) with the edgeR package (version 3.8.5, Robinson *et al.*, 2010), which was developed for differential expression analysis of RNA-seq data, and which also works well for other types of genome-scale count data such as MBD-seq data. Before the differential analysis, the MBD-seq data was normalized using quantile normalization (towards average profile, rounding was used to maintain count character of the data), while the RNA-seq data was normalized using the trimmed mean of M values method available in the edgeR package. The complete-linkage clustering method was used to create heatmaps of the MBD and RNA-seq data. For the gene ontology enrichment analysis, single lists of genes ranked by their false discovery rates were entered into the GOrilla tool (Eden *et al.*, 2009).

RESULTS

AN IMMUNE GENE SIGNATURE THAT DIFFERENTIATES BETWEEN HCB AND NCB PATIENTS

The analysis of the differences in expression between the HCB patients and the NCB patients resulted in a list of 195 genes with an FDR < 0.05 (Figure 3.6a). In order to understand the biological significance of these genes, we looked at their biological function and found 42 immune system-related genes (Figure 3.6a). The majority of the differentially expressed genes reflected a humoral (*IGHM*, *IGHA1*, *IGHV3-23*, *BANK1*, *SELL*, *IGLV3-1*, *IGHG1*, *IGSF6*) and cellular immune response (*CD69*, *FYB*, *CARD11*, *CD244*, *TIGIT*). Likewise, the gene ontology enrichment analysis using the full list of 195 genes ranked by their FDR produced a list of 65 significantly enriched (FDR < 0.05) biological process ontologies (Supplementary Table 3.2). Many of these ontologies were linked to immune system processes

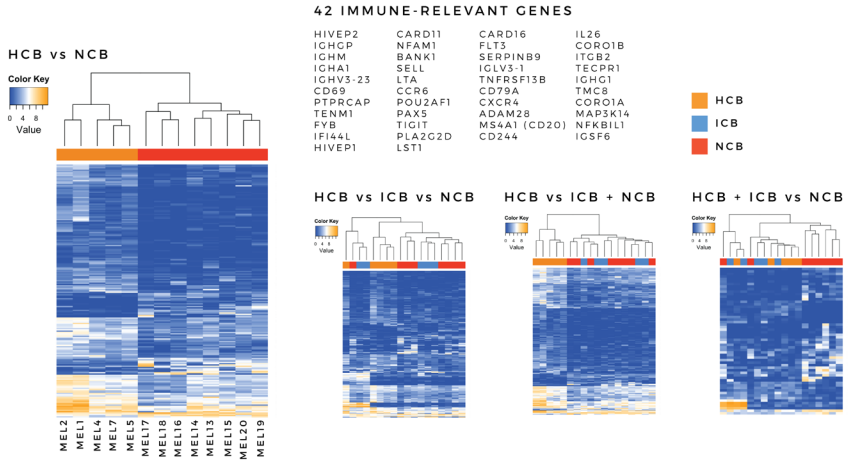


FIGURE 3.6 Gene expression analysis of HCB, ICB and NCB patients.

a. Supervised hierarchical clustering of patients with high clinical benefit (HCB) and those with no clinical benefit (NCB) using the list of 195 genes that were differentially expressed between these two groups (FDR < 0.05). We manually investigated this list and found 42 immune-related genes. **b**. These three heatmaps show the supervised clustering of the three response groups (HCB, ICB and NCB) using the results from the following comparisons: HCB vs. ICB vs. NCB, HCB & ICB vs. NCB and HCB vs. ICB & NCB. We found that when HCB and ICB samples were combined in a single group or when we compared the groups against eachother, the HCB and NCB samples were no longer perfectly separated. When the ICB and NCB groups were combined, the HCB group clustered separately again. Together, these results indicate that the expression profile of the ICB samples resembled the NCB profile more than the HCB.

and immune cell migration. Three further comparisons were made between the three groups of patients: HCB vs. ICB vs. NCB, HCB+ICB vs. NCB, and HCB vs. ICB+NCB (239, 58, and 253 differentially expressed genes respectively). These comparisons showed that the expression profile of the HCB group was different from both the NCB and the ICB groups (Figure 3.6b).

EXPLORATORY ANALYSIS OF IMMUNE INFILTRATE BY IMMUNOHISTOCHEMISTRY ON FROZEN SECTIONS

In order to see whether the RNA-seq-based immune signature reflected the presence of immune cells in the tumor samples, IHC analysis was performed on frozen sections available from the same block that was used for RNA-seq. 26 sam-

ples from 24 patients were evaluated by immunohistochemistry for tumor and immune markers (Supplementary Table 3.1). These included 17 samples from the 18 samples analyzed by RNA-seq. Sample MEL27 was only used for RNA/DNA extraction, and not for IHC due to its small tissue size. First, the infiltration of T cells in the tumors was assessed by CD3 and CD8 immunostainings. We observed five patterns of T cell infiltration, including a heterogeneous one with two different patterns present within the same tumor sample (Figure 3.7). The infiltration patterns were annotated with capital letters from A to E based on the following characteristics: the number of T cells present at the invasive margin (the border between tumor nets and immune infiltrate), the level of T cell confinement at the invasive margin and the number of T cells present inside of the tumor nests. The majority of the samples presented infiltration pattern B or C (86% in the HCB group, 80% in the ICB group, and 50% in the NCB group respectively). We did not observe statistically significant differences in terms of infiltration patterns between the three patient groups (Supplementary Table 3.3). For further analyses, the patterns A and B were combined in pattern 1 and the patterns C, D and E in pattern 2 (Figure 3.7). In the NCB group, we observed that infiltration pattern 2 (high number of T cells, not confined to the invasive margin) was only present in samples that were removed after immunotherapy onset (Fisher's exact test, $p = 0.01$). Only pattern 1 and the heterogeneous infiltration were observed in the samples collected before therapy onset in the NCB group (Supplementary Table 3.4). Additionally, CD20+ cells were present in 71,4% of the samples from the HCB group, whereas in the NCB group only 25% of the samples contained CD20+ cells (Supplementary Table 3.4).

AUTOMATED QUANTIFICATION OF CD8+ AND PD-L1+ CELLS ON FFPE SECTIONS

FFPE sections obtained from the same tumor samples were stained for CD8 and PD-L1 markers, and computerized image analysis was performed. An overall percentage of infiltration that takes the number of CD8+ cells at the invasive margin and the number of PD-L1+ cells at the center of the tumor into consideration was attributed to every sample. The median CD8+ infiltration values in the three groups were 4.3 for HCB, 3.6 for ICB and 2.4 for NCB (Supplementary Table 3.5). A tendency for higher infiltration of CD8+ cells was observed in the HCB compared to NCB samples. However, the differences across the three clinical benefit groups were not statistically significant when a Kruskal-Wallis test was performed, mainly due to the small sample size (data not shown). Furthermore, a higher overall percentage of PD-L1+ cells was observed in the HCB compared to NCB samples (median of 11.1 vs. 6.8) without reaching statistical significance. Two independent pathologists evaluated the percentage of PD-L1+ cells in the tumor component as well as in the immune component (Supplementary Table 3.5). Again the same trend was observed: a higher number of PD-L1+ cells in

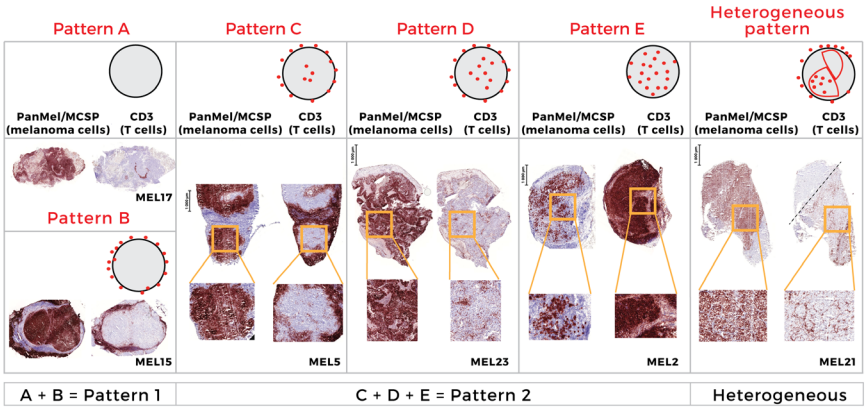


FIGURE 3.7 Infiltration patterns of immune cells in melanoma metastases.

Representative sections are shown for patterns A through E, as well as for the heterogeneous infiltration pattern observed in 2 samples from the NCB group. A schematic drawing and two IHC stainings illustrate each pattern type: PanMel or MCSP for tumor cells and CD3 for T cells. Both tumor and T cells are shown in red in IHC sections. Additional zoomed-in images are shown for patterns C, D, E and the heterogeneous patterns. Infiltration patterns A and B were combined in pattern 1 and patterns C, D and E were combined in pattern 2.

both tumor and immune components in the HCB group as compared to the NCB group (a median of 17.5 vs. 0 for the tumor component and 4.5 vs. 1 for the immune component).

DIFFERENTIAL METHYLATION FOR HIGH CLINICAL BENEFIT VS. NO CLINICAL BENEFIT PATIENTS

The differential analysis of all annotated methylation cores (intron, exon and promoter regions) between HCB and NCB resulted in a list of 107 cores, associated with 92 genes (FDR < 0.05, Supplementary Figure 3.5). Of these cores, 70 were located in introns, 19 in exons and 18 in promoter regions. The gene ontology enrichment analysis on the genes linked to the differentially methylated cores resulted in a list of 56 ontologies with an FDR < 0.001 (Supplementary Table 3.6). Many of these ontologies were associated with the development and function of neurons. No overlap was found between the lists of differentially expressed and differentially methylated genes. The list of differentially expressed genes was used in the unsupervised clustering of the methylation data for these genes (and vice versa), but no clustering of the HCB and NCB in separate groups was observed (data not shown).

DISCUSSION

Using RNA-seq we identified differences in the gene expression profile of tumors of metastatic melanoma patients who experienced high clinical benefit from ipi and/or TriMixDC therapy compared to the tumors of patients with no clinical benefit. These differences were linked to immune system genes, and reflected a complex humoral and cellular immune response. The gene expression differences could also be partly attributed to the infiltration of immune cells in the tumors. However, the mere presence of immune cells within a tumor was in itself not enough to elicit a response, as evidenced by the presence of immune cells in the NCB samples.

The subcutaneous nodule MEL2 that was removed from a long-term responder during the lesion-regression period showed a very high infiltration by T and B cells. This lesion illustrated the "perfect" immune response that takes place when a tumor is efficiently attacked and eliminated by the immune system. It also suggested that a humoral immune response is needed for tumor eradication and that this humoral response contributes to an appropriate adaptive immune response at the tumor site, while at the same time counteracting immune tolerance towards the tumor cells. Additionally, CD20+ B cells were found in close proximity to CD8+ T cells similar to the observations of Nielsen *et al.* (2012) in ovarian cancer, where the presence of both CD20+ and CD8+ lymphocytes was associated with prolonged survival. Moreover, large numbers of peritumoral B cells in metastatic lymph nodes were associated with favorable outcome in oro and hypopharyngeal carcinoma (Pretscher *et al.*, 2009).

The role of B cells in anti-tumor immunity is still a matter of debate (Germain *et al.*, 2015) although, over the last three years, several clinical studies have shown a positive association between better clinical outcome and high B cell tumor densities in hepatocellular carcinoma (Shi *et al.*, 2013), metastatic colorectal cancer (Meshcheryakova *et al.*, 2014), lung cancer (Germain *et al.*, 2014), and oral squamous cell carcinoma (Wirsing *et al.*, 2014). Furthermore, Yuan *et al.* (2011) showed that patients with pre-existing serological immunity (in this case to NY-ESO1 antigen) and detectable specific CD8+ T cells were twice as likely to experience clinical benefit after ipilimumab treatment (Yuan *et al.*, 2011). One additional question is whether immune responses are originally primed locally at the tumor site or if they are generated in lymphoid organs and migrate to the disease site. Our results suggested that a complex humoral immune response in the invaded lymph nodes might indicate a better response to ipi combined with a DC vaccination. Nevertheless, the specificity of the humoral immune response within primary or metastatic tumor sites still needs to be proven.

We did not find a link between the gene expression and DNA methylation anal-

yses. The differentially expressed genes did not show differences in methylation between the HCB and NCB groups, and vice versa, suggesting that the expression of differentially expressed genes was not under DNA methylation control. Interestingly, a gene ontology enrichment analysis revealed that the differentially methylated genes were enriched for ontologies linked to the nervous system. This result might reflect the fact that melanocytes are derived from the neural crest, just like peripheral and enteric neurons (Huang & Saint-Jeannet, 2004, Cichorek et al., 2013). Given that there are elaborate interactions between the immune and nervous system (Steinman, 2004) and that several genes are involved in both these systems (Lepelletier et al., 2007, Guo et al., 2013), the differential methylation might also hint at the role of the immune system in patient response, even though there was no clear overlap with the RNA-seq results.

CONCLUSION

The differences in gene expression between HCB group (long-term responders to ipi-based immunotherapy) and NCB translated in a list of immune genes reflecting both a humoral and cellular immune response. The differential analysis of methylation between HCB and NCB resulted in a list of 107 cores, associated with 92 genes. No link could be established between the RNA-seq and MBD-seq analyses due to the fact that the expression of differentially expressed genes was not under DNA methylation control. A trend for higher number of CD8+ cells, PD-L1+ cells (both in the tumor and immune component) as well as CD20+ cells was observed in the samples from the HCB group compared to NCB group. Furthermore the melanoma metastases from the NCB that presented with a higher immune infiltrate (pattern 2) were all removed after therapy onset. The results presented in this manuscript offer a first hint at the biological differences between HCB and NCB metastatic melanoma patients treated with immunotherapy as well as a starting point for further experiments to find an actual biomarker for the prediction of this response.

Adapted from:

Koch A, De Meyer T, Jeschke J, Van Criekinge W. MEXPRESS: Visualizing Expression, DNA Methylation and clinical TCGA Data. *BMC Genomics* **16**, 636 (2015)

MEXPRESS

ABSTRACT

BACKGROUND

In recent years, increasing amounts of genomic and clinical cancer data have become publicly available through large-scale collaborative projects such as The Cancer Genome Atlas (TCGA). However, as long as these datasets are difficult to access and interpret, they are essentially useless for a major part of the research community and their scientific potential will not be fully realized. To address these issues we developed MEXPRESS, a straightforward and easy-to-use web tool for the integration and visualization of the expression, DNA methylation and clinical TCGA data on a single-gene level (http://mexpress.be).

RESULTS

In comparison to existing tools, MEXPRESS allows researcher to quickly visualize and interpret the different TCGA datasets and their relationships for a single gene, as demonstrated for GSTP1 in prostate adenocarcinoma. We also used MEXPRESS to reveal the differences in the DNA methylation status of the PAM50 marker gene MLPH between the breast cancer subtypes and how these differences were linked to the expression of MPLH.

CONCLUSIONS

We have created a user-friendly tool for the visualization and interpretation of TCGA data, offering clinical researchers a simple way to evaluate the TCGA data for their genes or candidate biomarkers of interest.

BACKGROUND

Over the last few years, large-scale cancer genomics projects have had a significant impact on cancer research. The goal of these projects is to create extensive, publicly available and multidimensional oncogenomic datasets using high-throughput technologies. These datasets allow researchers to compare the genomic sequences, epigenetic profiles and transcriptomes of cancer cells to those of normal cells or cells of different cancer (sub)types. The Cancer Genome Atlas (TCGA), a joint effort of the National Cancer Institute and the National Human Genome Research Institute, is an example of such a project (http://cancergenome.nih.gov/).

New findings derived from the statistical and data mining analysis of TCGA data are published regularly and have already proven to be a valuable addition to cancer research (The Cancer Genome Atlas Network, 2008, 2011, 2013, 2014). Large-scale datasets like TCGA also provide a validation platform for newly identified biomarkers and they are becoming a standard tool for current biomarker research. Another powerful aspect of the TCGA data is the possibility to correlate different types of data. Promoter DNA methylation for example influences gene expression, and aberrant methylation is found in almost every human cancer (Herman & Baylin, 2003). The ability to compare these data in a large number of cancer patients is therefore extremely valuable, especially for the identification of DNA methylation biomarkers. Given the growing importance of large-scale datasets for cancer research, intuitive data visualization tools are increasingly crucial to help researchers understand the data, especially when multiple samples and datasets have to be compared.

A number of visualization tools, each focused on one or more specific research questions, are available for TCGA data and offer a wide range of visualization methods (Zhang *et al.*, 2007, Cerami *et al.*, 2012, Goldman *et al.*, 2013, Thorvaldsdottir *et al.*, 2013). There is however no tool available that offers fast and straightforward visualization and interpretation of the expression, methylation and clinical data in TCGA, as well as the relation between these different data types. Such a tool could be of particular use to the large community of clinical researchers without bioinformatics expertise who are looking for a way to explore genes of interest or candidate biomarkers in the TCGA data.

Here we introduce MEXPRESS, an intuitive web tool for the fast and straightforward querying and visualization of the clinical, expression and methylation data in TCGA and the relationship between these datasets on a single-gene level. MEXPRESS was designed after the principles of graphical excellence as described by Edward Tufte (Tufte, 1983) to ensure that the complex and multidimensional TCGA data would be presented in a clear, precise and efficient way to the user. It is generally accepted that analysis and visualization tools intended for a broad

research audience should be easy to use and should not require computational or bioinformatics expertise (Cerami *et al.*, 2012, Thorvaldsdottir *et al.*, 2013, Perez-Llamas & Lopez-Bigas, 2011, Schroeder *et al.*, 2013). MEXPRESS was therefore developed to have virtually no learning curve, allowing especially clinical researchers to get their results fast without having to invest time in learning yet another tool.

IMPLEMENTATION

Ease of use is a key feature of MEXPRESS. Just three simple steps are needed to create a plot: a user has to enter a gene name, select one of the available cancer types and click the plot button. The resulting figure (Figures 3.8 and 3.9) shows the selected gene together with its transcripts and any CpG islands. Next to the gene, blue line plots illustrate the methylation data for each probe location (Infinium HumanMethylation450 microarray data). A yellow line plot displays the RNA-seq-derived expression data and grey bar plots represent the values of the clinical parameters. The numbers on the far right indicate the significance of the relation (correlation coefficient or p value, depending on the data types compared) between each row of data (clinical, expression or methylation) and the selected "sorter". By default, expression is the selected "sorter", which means that the samples are ordered by their expression value. Clicking on one of the clinical parameters will reorder the samples based on the selected variable and the relationships will be recalculated. The resulting images can be downloaded in PNG or SVG file format.

TCGA DATA

We downloaded the following TCGA data from the TCGA ftp site: level 3 per-gene RNA-seq v2 expression data (UNC IlluminaHiSeq_RNASeqV2), level 3 DNA methylation data (JHU_USC HumanMethylation450) and clinical data in Biotab format (both clinical patient and tumor sample data). Bash scripts running on the back-end Linux server check the TCGA ftp site monthly for any data updates, which are then automatically uploaded to the database. Whenever TCGA publishes data for new cancer types, these will also be included in MEXPRESS. Before the upload, R scripts (R version 3.0.2) process the data to address missing values, to combine separate files into one where necessary, to reformat the data and to generate SQL scripts for the data upload. The RNA-seq data is log-transformed before being used to draw the plots and only a selection of the most relevant clinical parameters (for which data is available) is shown in the MEXPRESS plots in order to reduce data clutter.

OTHER DATA SOURCES

For the breast invasive carcinoma samples, we downloaded a table with the expression subtype (normal, basal, luminal A, luminal B and Her2) for each sample from the UCSC cancer genome browser (Goldman *et al.*, 2013). The CpG island data was downloaded from the UCSC genome browser (Kent *et al.*, 2002) using the table browser with the following settings: clade: Mammal, genome: Human, assembly: Feb. 2009 (GRCh37/hg19), group: Regulation, track: CpG Islands, table: cpgIslandsExt. The exon and transcript annotation was obtained from Ensembl using the BioMart tool (Ensembl Genes 75, *Homo sapiens* genes GRCh37.p13). We designed MEXPRESS in such a way that it will be easy in the future to include new types of data, such as mutation or proteomics data.

STATISTICAL ANALYSES

We recreated all the statistical functions used in MEXPRESS in Javascript, with the Pearson correlation and the non-parametric Wilcoxon's rank-sum test being the two main functions. The former is used to compare two types of data that both have more than 2 levels (e.g. expression and methylation data), whereas the latter is used to calculate the difference of a variable between two groups (e.g. the difference in expression between male and female). To correct for multiple comparisons, we included a false discovery rate correction step (Benjamini & Hochberg, 1995).

MEXPRESS WEBSITE

The MEXPRESS site runs on an Apache server and uses PHP to interact with the back-end database. It employs Javascript, the jQuery Javascript library (version 1.11.0), Ajax autocomplete for jQuery (version 1.2.10, https://github.com/devbridge/jQuery-Autocomplete) and the d3.js Javascript library (version 3.0.6, http://d3js.org/) to create the interactive plots and to perform the calculations for the statistical analyses. When a user downloads a figure, the SVG image is converted into a PNG image using Inkscape, an open source vector graphics editor (http://www.inkscape.org/). The backbone of MEXPRESS is a MySQL database that contains the TCGA data needed for the visualizations. PHP scripts handle the database queries, package the results in JSON and send them back to the user. All the MEXPRESS code (back-end, front-end and data processing) can be cloned or downloaded from this GitHub repository: https://github.com/akoch8/mexpress.

RESULTS AND DISCUSSION

One of the best-studied examples of epigenetic aberrations in human cancer is the hypermethylation of the *GSTP1* promoter region in prostate cancer, leading to the transcriptional silencing of *GSTP1* (Brooks *et al.*, 1998, Millar *et al.*, 1999, Henrique & Jeronimo, 2004). Using MEXPRESS, this effect can be observed in the TCGA data. Figure 3.8a shows the default MEXPRESS plot for *GSTP1* in prostate adenocarcinoma with the samples sorted by their *GSTP1* expression value. It is immediately clear that the normal samples cluster towards higher *GSTP1* expression and that there is a negative correlation between expression and methylation around the promoter region. The p value for the comparison of expression between normal and tumor samples (Wilcoxon rank-sum test, P = 0.001) and the Pearson correlation coefficients (ranging from -0.675 to -0.792 around the promoter region) confirm the visual interpretation of the data. When the samples are rearranged based on the sample type (normal vs. tumor), this difference in methylation and expression between normal and tumor samples stands out even more (Figure 3.8b). It is not possible to create a similar figure that allows a comparable interpretation using one of the existing tools, as they lack the necessary data implementation and/or features, making them less suitable for clinical researchers (Table 3.3, Supplementary Figures 3.6, 3.7, 3.8 and 3.9).

Breast cancer is a heterogeneous disease that covers a myriad of subtypes. Each subtype has distinct biological features, leading to differences in clinical outcome and response to treatment. Perou *et al.* (2000) were the first to describe breast cancer subtypes based on gene expression patterns and it was found that these subtypes (luminal-like, basal-like, Her2-enriched and normal-like) have significantly different survival times (Sorlie *et al.*, 2001). The classification of breast cancer samples into these subtypes (based on the PAM50 gene signature (Parker *et al.*, 2009)) is available in MEXPRESS, allowing users to compare expression, methylation and clinical data between the different subtypes. One member of the PAM50 signature is the gene *MLPH*. Using MEXPRESS, it becomes clear that *MLPH* expression is negatively correlated with DNA methylation in the promoter region (a so far unpublished result) and that expression and methylation, as well as *HER2*, estrogen and progesterone receptor status, differ between the breast cancer subtypes (Figure 3.9).

Traditional genome browsers, such as the UCSC genome browser (Kent *et al.*, 2002), present data as horizontally stacked genomic tracks, which is very useful to display different types of location-bound genomic data. This allows users to observe differences within a track or between a limited number of tracks from different samples. MEXPRESS rotates this more traditional "genome browser view" and organizes samples vertically and the different data types horizontally. This simple transformation offers a very different view of the data, resulting

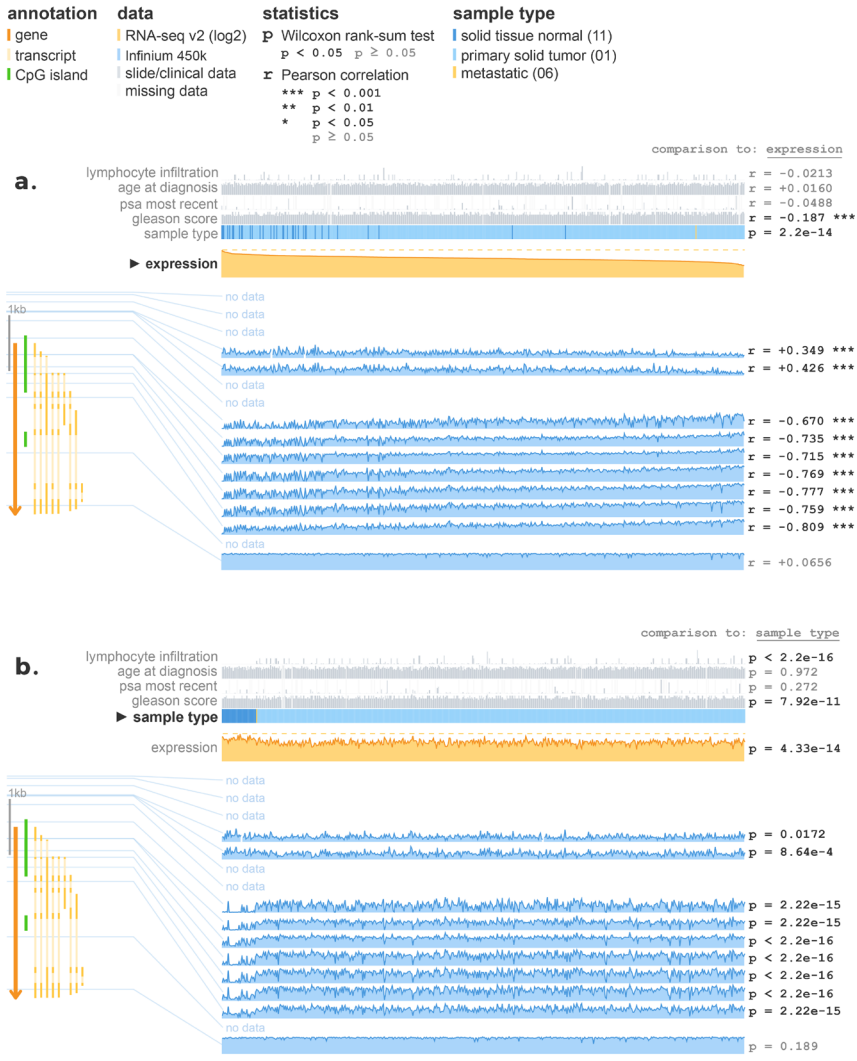


FIGURE 3.8 Visualization of the TCGA data for GSTP1 in prostate adenocarcinoma using MEXPRESS.

a. In the default MEXPRESS plot, the samples are ordered by their expression value. This view shows how GSTP1 expression and promoter methylation are negatively correlated, which is confirmed by the Pearson correlation coefficients on the right. It also indicates that normal samples tend to have higher GSTP1 expression than tumor samples. **b**. When reordered by sample type, the differences in expression and methylation between normal and tumor samples become even more apparent.

in an easier interpretation of the differences between samples than could be achieved through a conventional genome browser, especially when comparing hundreds of samples at the same time. It also allows for the easy comparison of location-bound genomic features, such as DNA methylation, to expression data or clinical information. The combination of this visualization approach with a simple user interface and the strengths listed in Table 1 sets MEXPRESS apart from existing tools when it comes to visualizing and integrating the expression, DNA methylation and clinical TGCA data.

CONCLUSION

Along with their expanding size, the value and significance of large-scale oncogenomics datasets will continue to rise in the coming years. This growth creates a need for intuitive and straightforward tools that enable researchers to quickly analyze and visualize the data of interest. The tool presented here offers a unique set of features, including its ease of use and the integrated visualization of different data types over hundreds of samples. It may therefore help to quickly test hypotheses that concern the discovery of DNA methylation or expression-based biomarkers.

TABLE 3.3 A comparison of different tools for the visualization of TCGA data.

As illustrated by the supplementary figures (Supplementary Figures 1, 2, 3 and 4), there are obvious differences between existing tools and MEXPRESS in both the data and the features these tools offer. This table lists the most relevant of these differences, thereby highlighting some of the strengths and weaknesses of each tool. (CGW = Cancer Genomics Workbench, IGV = Integrative Genomics Viewer)

	UCSC genome browser	cBioPortal	CGW	IGV	MEXPRESS
All TCGA cancer and data types available	yes	yes	no	no	no
Integration of expression, DNA methylation and clinical data	no	no	no	no	yes
Statistical interpretation of the relationships	no	yes	no	no	yes
Registration and download required	no	no	no	yes	no

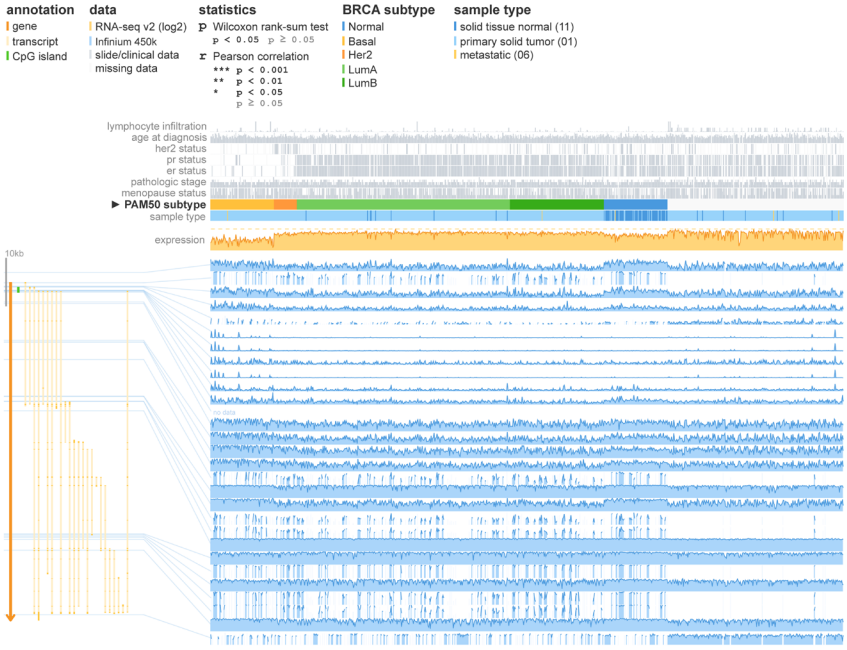


FIGURE 3.9 MEXPRESS view of the TCGA data for MLPH in breast invasive carcinoma.

The samples are ordered by breast cancer subtype, revealing clear differences in expression and methylation, as well as HER2, estrogen and progesterone receptor status, between the different subtypes.

ACKNOWLEDGEMENTS

We would like to thank Gerben Menschaert for his help in revising this manuscript.

CHAPTER 4

GENERAL CONCLUSIONS AND FUTURE PERSPEC-TIVES

Throughout this thesis we introduced several different methods that are commonly used to measure gene expression and DNA methylation and we described how these methods are applied in clinical cancer research. Every chapter had its own focus. The first one presented a combination of genome-wide protein (shotgun proteomics and N-terminal COFRADIC) and DNA methylation (reduced representation bisulfite sequencing or RRBS) measurement techniques, which we used to study the effect of DNA methylation on expression at the protein level. Not only did this experiment confirm the inhibitory effect of promoter DNA methylation on gene expression (thereby validating our approach), it also offered some insight into the possibility of DNA methylation-controlled alternative transcription. This integration was only the first step in a larger project. The final goal is to combine the protein and DNA methylation data with RNA and ribo-seq data in order to create a truly comprehensive overview of gene expression in our model cell lines. These analyses will include some possibly very interesting correlation studies as well as the merging of all the datasets with pathway information and their visualization in Cytoscape (Shannon *et al.*, 2003).

In the second chapter we examined how we can improve protein identifications in a proteomics experiment by using ribosomal sequencing. Looking at the results of our analyses, we believe that a proteogenomics approach that combines proteomics and transcriptomics (particularly ribosome profiling) offers some distinct advantages over the separate use of these methods and that this type of approach will be more widely used in the future. Our lab has already acted on this belief by developing PROTEOFORMER (Crappé *et al.*, 2015), an automated pipeline for the integration of proteomics and ribosomal profiling data that we made available to other researchers through the Galaxy platform (Goecks *et al.*, 2010). Proteogenomics is a relatively new field of research and is still evolving rapidly. The arrival of ribosome profiling for example has had and will continue to have an important impact on the field. We will keep following up on these evolutions and have for example already planned a new release of our PROTEOFORMER tool (support of other species, option to select transcripts…).

The combination of the work presented in the first two chapters could lead to some interesting biological and technical insights. Every technique we used has its inherent shortcomings. By integrating them, we can use each technique's strengths to overcome another technique's weaknesses. One example is the combination of transcriptomics and proteomics. Compared to proteomics, RNA-seq offers a higher throughput and better reproducibility, whereas proteomics provides gene expression measurements at the functional protein level. This is quite relevant, especially knowing that the correlation between transcript and protein levels is not always perfect (Ning *et al.*, 2012). The combination of the two techniques not only gives us a more complete view of the expression profile of a sample, it can also tell us something about the control of gene expression. It will be interesting to see what the result of adding ribo-seq data to our RNA-seq, proteomics and

RRBS data will be. We have to keep in mind that we are working with a single cell line model (wild type and double knockout), but despite this limited setup we should still be able to demonstrate the feasibility of the integration of the different techniques (RRBS, proteomics, RNA-seq and ribo-seq). We are not saying that the experiment could not benefit from additional cell lines, but the complete analysis of just our model cell line could already result in some interesting hypotheses.

The third chapter described how high-throughput transcriptomics and DNA methylation profiling techniques can be used to better understand cancer biology and to develop new treatment strategies. The two cancer research projects described in chapter three also touched on two other important subjects, immunotherapy and personalized medicine. Cancer cells can sometimes evade or alter the immune response that would normally kill them and a lot of research is focused on stimulating and enhancing the immune response of cancer patients to their tumors. In the lung cancer study for example, we described how the treatment of lung cancer patients with a demethylating drug might sensitize them to anti-PD-1/PD-L1 immunotherapy. Just as with the more traditional treatments such as chemo or radiotherapy, some patients will benefit from immunotherapy while others will not. In our melanoma study, patients were treated with a dendritic cell-based therapeutic vaccine and/or ipilimumab, a CTLA-4 inhibitor. By analyzing the gene expression differences between the patients that responded positively to the treatment and those who did not, we took a first step towards the development of an expression signature that could be used to predict a patient's response. This concept of using a biomarker to select the most appropriate treatment for a patient is one of the main ideas behind personalized medicine. Genetic mutations, gene expression, DNA methylation and many other types of markers are already used in hospitals to tailor treatment strategies to individual patients, saving some of them from expensive, ineffective and potentially harmful treatments (Tian *et al.*, 2012).

The results from both the lung cancer and the melanoma study are first steps in the development of new treatment strategies. Ultimately, the goal is to improve patient care, but it is obvious that despite their interesting results, these two studies are preliminary and they should be interpreted as such. The main goal of the lung cancer project was to better understand the biology behind the improved response of patients to anti-PD-1/PD-L1 treatment after they received azacytidine, not the development of a gene expression or DNA methylation biomarker. The clinical part of the study would have been much too small to achieve that. However, our pre-clinical findings (immune stimulation after azacitidine treatment) have already been confirmed in other studies and cancer types, for example by Li *et al.* (2014). If these results are to improve patient care, they will first have to be validated in independent, large-scale clinical studies.

Just as for the lung cancer project, the final goal of the melanoma project is to im-

prove patient care, but through the development of a biomarker for response to immunotherapy rather than a new treatment strategy. The main take-away from our study is that the difference between responders and non-responders seems to be linked to the immune response and the presence of CD8+, PD-L1+ and CD20+ cells within the tumor. One of the next steps should be repeating the experiment in a larger patient cohort, followed by the selection of potential markers and their validation in a separate group of patients. Once a promising gene expression signature has been found (for example through the analysis of RNA-seq data), the best gene expression biomarkers will have to be further validated using a different technique such as immunohistochemistry, PCR, proteomics or western blot. A publically available dataset like the one from TCGA could also be used, but the problem with our melanoma study is that we are specifically interested in the response to immunotherapy and there is almost no such response data available in TCGA. Some more experiments are already planned for the melanoma project, including the immunohistochemical analysis of about 40 FFPE (formalin fixed and paraffin embedded) patient samples. This analysis will give us more information on the different (immune) cells present in and around the tumors. There will also be another RNA-seq analysis of about 50 patients that were treated with anti-PD-1/PD-L1 therapy.

Given our experience with proteogenomics, it might also be a good idea to integrate this technique in the melanoma project. The proteogenomic analysis of tumors is a still largely unexploited research domain and could give some very interesting results. Boja & Rodriguez (2014) explain how proteogenomic approaches could be a valuable addition to cancer research and how they can be used to better understand the disease. Polyakova *et al.* (2015) and Shukla *et al.* (2015) look at proteogenomics from a more practical and applied point of view and describe how proteogenomics techniques can be used to identify tumor neoantigens, which could be used in immunotherapy, or to develop sensitive diagnostic and prognostic biomarkers. Snyder *et al.* (2015) for example describe how melanoma tumors with a higher mutational load (*i.e.* a higher number of neoantigens) are more susceptible to the immune response, a finding that–when confirmed–could have important implications for our understanding of the complex interplay between a tumor and our immune system. These types of analyses will also greatly benefit from appropriate data processing tools, such as PROTEOFORMER, and visualization tools, like MEXPRESS.

It needs to be noted that biomarker research is very difficult, with roughly 1% of all proposed and investigated biomarkers making it to the clinical practice. As we described above, if we want the results from our lung cancer and melanoma studies to have an impact on cancer therapy and the well-being of cancer patients, further analyses and validation, ideally in several studies using patients from different hospitals or even countries, are vital. A biomarker that is validated in one study will not necessarily validate in another one. Yi *et al.* (2011) describe how one

of their colorectal cancer DNA methylation biomarkers could not be validated in a patient cohort from a different country. They cite possible biological differences between the patient groups (there are epigenetic differences between different populations and races) and differences in screening programs as possible explanations. This finding perfectly illustrates the complexity of biomarker research as well as the importance of independent validation studies

Both the search for personalized treatments and the rise of immunotherapy are partly fueled by the ever-improving high-throughput "omics" techniques and show no signs of slowing down. Large-scale projects, such as the Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov), offer researchers unprecedented opportunities, but as the size of these cancer databases keeps growing, the development of specialized analysis methods will become increasingly important. In an attempt to respond to this need, we created MEXPRESS (http://mexpress.be), an online visualization tool for the TCGA data. MEXPRESS offers researchers a very fast and straightforward way to check the expression and DNA methylation levels of the gene they are interested in, as well as the correlation with the available clinical data. The examples presented in this thesis clearly show how MEXPRESS could be used to generate hypotheses. The first one illustrates the known correlation between gene expression and DNA methylation for the gene GSTP1 in prostate cancer. This example acts as a "proof-of-concept". We chose the gene MLPH for the second example, because it allows us to demonstrate the potential power of MEXPRESS. MPLH is a member of the PAM50 gene expression signature that distinguishes the different breast cancer subtypes (Parker et al., 2009), so we know that the expression of MLPH should vary between the subtypes. The MEXPRESS plot confirms this gene expression variation, but it also clearly displays the link between gene expression and DNA methylation for MLPH. No papers have been published on the DNA methylation control of MLPH expression or on how this gene is differentially methylated between the breast cancer subtypes. Therefore, this plot demonstrates how MEXPRESS could be used to generate a new hypothesis. MEXPRESS makes it very easy for researchers to check the results of a small-scale study in the hundreds of samples offered by TCGA. Even if they are only interested in for example the expression of a certain gene, any potential link between their gene's expression and its DNA methylation status or even some clinical characteristics will immediately stand out. Of course, we intend to keep MEXPRESS up-to-date with latest TCGA data releases and plan to add new types of data whenever possible.

Even though the different chapters of this thesis deal with quite diverse fields of research, they do share a common theme, namely the integration of various "omics" datasets. I believe that the integration of such -omics data presents the next big step forward in cancer research (see Cui *et al.* (2015) for a recent review on the subject). Over the last few years, several factors, such as ever-improving high throughput techniques, increasing computer power and an impressive drop

in sequencing costs, have both enabled and stimulated this data integration trend, creating exciting opportunities as well as significant bottlenecks along the way, especially in the medical field. As we have tried to demonstrate in this thesis, this evolution touches on many fields of research, from basic biology and methodology to applied clinical studies.

SUMMARY

SUMMARY

A photoreceptor cell in your retina and one of your heart muscle cells carry the same genome with virtually the same genetic information. How can our cells achieve such a spectrum of looks and functions from a single genetic blueprint? Or on a more basic level, how can a cell control which genes to express and which ones not? This question can be (partly) answered by epigenetics, the collection of heritable changes in gene expression that are not encoded in the DNA sequence itself. DNA methylation is one of the main epigenetic processes and its role in the control of gene expression is well established. For example, when a stretch of DNA around the start of a gene (known as that gene's promoter) is methylated, the gene will often not be expressed. Gene expression can be measured in multiple ways and at different points in the expression process. A gene's DNA sequence is first transcribed to an RNA molecule, which is subsequently translated by a ribosome to a stretch of amino acids that will compose the final protein. Most of the studies that investigate the link between DNA methylation and gene expression measure the expression at the RNA or transcript level. Because of numerous control mechanisms that stand between a transcript and the protein it encodes, this measurement is just an approximation of the expression at the functional protein level. We performed a study in which we integrated genome-wide DNA methylation data with the protein-level gene expression data from a proteomics experiment. The results confirmed the known inhibitory effect of promoter methylation on gene expression, thereby validating our approach.

Despite its advantages, such as the biological relevance of measuring the functional end product of gene expression instead of intermediary transcripts, high-throughput proteomics comes with some notable drawbacks. Poor reproducibility and inadequate sensitivity for low-abundant proteins are the two main ones. We tried to improve the number of identified proteins in a proteomics experiment using ribosomal profiling (ribo-seq). This technique is similar to the more common RNA-sequencing approaches in that it measures gene expression at the transcript level. The novelty of ribo-seq is that only the small stretches of RNA that are bound to ribosomes are sequenced. The result is that ribo-seq does not just measure gene expression at the transcript level, it measures active protein synthesis. In a typical high-throughput proteomics experiment, a protein sequence database is used to identify the proteins in a sample. We combined existing public protein databases with a custom database we created using our own ribo-seq data. This approach helped improve the identification of proteins in our samples and it allowed us to identify proteins that we would have missed without the ribo-seq analysis.

In addition to this more basic and methodology-focused research, we also studied the role of gene expression and DNA methylation in cancer. Cancer cells arise when the expression of genes that control cell growth is disrupted. This disruption can have many different causes, from small mutations over large chromosomal rearrangements to changes in the DNA methylation status of a gene. Researchers have in fact found abnormal DNA methylation profiles in almost every known type of cancer. We used several techniques to analyze gene expression and DNA methylation in two different projects. In the first one, we tried to understand the effects of a treatment with azacytidine, a demethylating chemical, on lung cancer cells. In a small-scale clinical trial we noticed that patients who received azacytidine before they received immunotherapy (anti-PD-1/PD-L1) responded better to the latter than those who did not. PD-1 is a receptor found on T cells and when a PD-L1 molecule binds to this receptor, the T cell is inactivated. Some tumor cells use this system to their advantage by expressing PD-L1, thereby deactivating the T cells that might otherwise kill them. The anti-PD-1/PD-L1 therapy blocks this interaction and prevents tumor cells from evading the immune response. Our analyses revealed that a significant number of lung cancer tumor cells have a low expression of PD-L1 and will therefore not respond to the anti-PD-1/PD-L1 therapy. However, we also noticed that the azacytidine treatment increased the expression PD-L1, making the previously non-responsive tumor cells sensitive to the immunotherapy. Together, these results suggest that epigenetic therapy (such as the removal of DNA methylation using azacytidine) could be used to sensitize a patient's tumors to immunotherapy.

Our second cancer project involved two different types of immunotherapy, this time for the treatment of metastatic melanoma. The first therapy blocks the T cell receptor CTLA-4, which acts as an "off switch" for the T cell, similar to the PD-1 receptor. The second therapy is based on the use of dendritic cells, which help T cells recognize tumor cells. Sometimes, dendritic cells fail to recognize the tumor cells in a cancer patient. To resolve this problem, dendritic cells can be isolated from the patient's blood, sensitized in the lab to a gene that is specifically expressed by the tumor cells and then injected back into the patient. In some patients these two approaches work very well, while in others we see no response. Because these treatments are very expensive and potentially harmful it would be very useful to know which patients will benefit from the therapy before they receive it. This is why we compared gene expression and DNA methylation profiles between a group of patients that responded to the therapies and a group that did not. Our analyses resulted in a list of immune system-related genes whose expression varied between the responders and the non-responders. We also found several T cell markers among the differentially expressed genes, which reflected the higher number of infiltrating immune cells in the tumors that responded. These results are a first step towards the development of a gene signature that could be used to predict a patient's response. Furthermore, they indicate that the immune system likely plays an important role in the treatment of melanoma.

In recent years, a lot of effort in cancer research has been focused on the creation of large-scale cancer databases that contain the data of hundreds or even thousands of patients. The Cancer Genome Atlas (TCGA) is an example of such an initiative and its publically available database contains a wide range of genomic, clinical, gene expression and DNA methylation datasets. Databases like TCGA are a valuable resource for cancer researchers and have already generated a lot of discoveries. Their usability and accessibility could still be improved though, especially for clinical researchers without a bioinformatics background. That is why we developed MEXPRESS, a web tool for the quick and simple visual integration of the clinical, gene expression and DNA methylation data in the TCGA database.

SAMENVATTING

Een fotoreceptorcel in jouw netvlies en een van de cellen in jouw hartspier bevatten hetzelfde genoom met nagenoeg identiek dezelfde genetische informatie. Hoe kunnen onze cellen zo een brede waaier aan vormen en functies uit één enkel genetisch bouwplan halen? Of op een meer fundamenteel niveau, hoe kan een cel controleren welke genen tot expressie komen en welke niet? Epigenetica, de overerfbare veranderingen in genexpressie die niet door het DNA gecodeerd worden, vormt (voor een deel) het antwoord op deze vraag. DNA-methylatie is een van de voornaamste epigenetische processen en de rol er van in de controle van genexpressie is welgekend. Wanneer bijvoorbeeld een stukje DNA rond de start van een gen (gekend als de promoter regio) gemethyleerd is, dan komt dit gen vaak niet tot expressie. Genexpressie kan op verschillende manieren en op verschillende momenten tijdens het expressieproces gemeten worden. De DNA-sequentie van een gen wordt eerst gekopieerd naar een RNA molecule. Deze RNA molecule wordt op zijn beurt door een ribosoom vertaald naar een sliert aminozuren, die dan uiteindelijk het eiwit vormt. Studies die het verband tussen DNA-methylatie en genexpressie onderzoeken meten de genexpressie meestal op het RNA of transcriptniveau. Aangezien er verschillende controlemechanismen actief zijn voor, tijdens en na de vertaling van een transcript naar een eiwit, is deze meting slechts een benadering van de hoeveelheid eiwit die er uiteindelijk in een cel aanwezig is. Daarom hebben we een studie uitgevoerd waarin we genoom-wijde DNA-methylatie data gelinkt hebben aan expressiedata op proteïneniveau uit een proteomics experiment. Onze resultaten bevestigden het gekende remmende effect van promoter methylatie op genexpressie en valideerden bijgevolg onze aanpak.

Ondanks de voordelen, zoals bijvoorbeeld de biologische relevantie van expressiemetingen op het niveau van het functionele eiwit in plaats van de tussenliggende transcripten, hebben de zogenoemde high-throughput proteomics technieken ook enkele nadelen. De gebrekkige reproduceerbaarheid en de lage gevoeligheid voor eiwitten die slechts in geringe mate voorkomen zijn de belangrijkste. We hebben geprobeerd het aantal identificaties in een proteomics experiment te verhogen door gebruik te maken van ribosoomprofilering (ribo-seq). Net zoals de vaak gebruikte RNA sequencing technieken meet ribo-seq genexpressie op het transcriptniveau. Het grote verschil is dat bij ribo-seq enkel de stukjes RNA die gebonden zijn aan ribosomen gesequenced worden. Hierdoor meet ribo-seq niet zomaar gen expressie op het transcript niveau, maar meet het in feite de eigenlijke eiwitsynthese. In een typisch high-throughput proteomics experiment wordt een eiwitsequentie databank gebruikt om de proteïnen in een

staal te identificeren. In onze studie hebben we bestaande databanken gecombineerd met een aangepaste databank gebaseerd op onze eigen ribo-seq data. Deze aanpak verbeterde de identificatie van eiwitten in onze stalen en het zorgde er voor dat we eiwitten konden identificeren die we zonder de ribo-seq data niet zouden gevonden hebben.

Naast dit meer fundamenteel en op methodologie gefocust onderzoek hebben we ook de rol van gen expressie en DNA-methylatie in kanker onderzocht. Kankercellen ontstaan wanneer de expressie van genen die de celgroei controleren verstoord wordt. Deze storing kan verschillende oorzaken hebben, van kleine mutaties in de DNA-sequentie over grootschalige chromosomale herschikkingen tot veranderingen in de DNA-methylatie status van een gen. Zo hebben onderzoekers afwijkende DNA-methylatieprofielen gevonden in nagenoeg elk type kanker. We hebben verschillende technieken gebruikt om genexpressie en DNA-methylatie te analyseren in twee projecten. In het eerste project hebben we geprobeerd de gevolgen van een behandeling met azacytidine, een demethylerende molecule, op longkankercellen beter te begrijpen. In een kleinschalig klinisch onderzoek stelden we vast dat patiënten die azacytidine toegediend hadden gekregen voor ze met immunotherapie (anti-PD-1/PD-L1) behandeld waren beter reageerden op deze immunotherapie. PD-1 is een receptor die zich op het celmembraan van T cellen bevindt. Wanneer een PD-L1 molecule met deze receptor bindt, dan wordt de T cel gedeactiveerd. Sommige tumorcellen slagen er in dit systeem uit te buiten door zelf PD-L1 te produceren en zo de T cellen die hen anders zouden kunnen vernietigen uit te schakelen. De anti-PD-1/PD-L1 behandeling blokkeert de interactie tussen de PD-1 receptor en PD-L1 en voorkomt zo dat tumor cellen kunnen ontsnappen aan het immuunsysteem. Onze analyses toonden aan dat een beduidend deel van de longkankertumorcellen weinig of geen PD-L1 produceren en bijgevolg niet vatbaar zijn voor de anti-PD-1/PD-L1 therapie. We stelden echter ook vast dat de behandeling met azacytidine de expressie van PD-L1 kan verhogen en de tumor cellen dus toch gevoelig zou kunnen maken voor de immunotherapie. Deze resultaten geven aan dat epigenetische therapieën (zoals het verwijderen van DNA methylatie met azacytidine) zouden kunnen gebruikt worden om de tumoren van een patiënt gevoeliger te maken voor immunotherapie.

In ons tweede project hebben we twee verschillende soorten immunotherapie voor de behandeling van uitgezaaide huidkanker onderzocht. De eerste therapie blokkeert de T cel receptor CTLA-4 die, net zoals PD-1, als een soort "uit-knop" voor de T cel functioneert. De tweede therapie is gebaseerd op het gebruik van dendritische cellen. Dit zijn een type immuuncellen die T-cellen helpen bij het herkennen van tumorcellen. Soms lukt het de dendritische cellen echter niet om de kankercellen te herkennen. Om dit probleem op te lossen worden de dendritische cellen uit het bloed van een huidkankerpatiënt gefilterd en in het lab in contact gebracht met een gen dat specifiek op de tumorcellen tot expressie

komt, waarna ze terug in de patiënt geïnjecteerd worden. In sommige patiënten werken beide behandelingen goed, maar in veel patiënten zien we geen effect. Omdat het hier over bijzonder dure en potentieel gevaarlijke therapieën gaat, zou het zeer waardevol zijn om te weten welke patiënten voordeel kunnen halen uit de behandeling alvorens er mee te starten. Daarom hebben we de genexpressie en DNA-methylatieprofielen vergeleken tussen een groep patiënten die reageerde op de behandelingen en een groep waarbij dit niet het geval was. Deze analyse resulteerde in een lijst van genen gerelateerd aan het immuunsysteem waarvan de expressie verschilde tussen beide groepen. We vonden ook verschillende T-cel-specifieke genen in deze lijst, wat overeenstemde met het hogere aantal immuuncellen in de tumoren die reageerden op de immunotherapie. Deze resultaten zijn een eerste stap in de ontwikkeling van een set genen waarvan de expressie kan gebruikt worden om de patiënten te identificeren die zullen reageren op de behandeling. Verder geven ze ook aan dat er waarschijnlijk een belangrijke rol is weggelegd voor het immuunsysteem van een huidkankerpatiënt in zijn of haar behandeling.

De laatste jaren zijn er veel middelen naar de ontwikkeling van grootschalige kankerdatabanken gegaan. Deze databanken bevatten informatie over honderden tot duizenden patiënten. The Cancer Genome Atlas (TCGA) is een voorbeeld van een dergelijk initiatief en de publiek beschikbare TCGA databank bevat een breed gamma aan genomische, klinische, genexpressie en DNA-methylatiedatasets. Dergelijke databanken vormen een uiterst waardevol instrument voor onderzoekers en hebben reeds tot verscheidene ontdekkingen geleid. Hun gebruiksvriendelijkheid kan wel nog een stuk beter, zeker voor klinische onderzoekers zonder bio-informatica-achtergrond. Daarom hebben we MEXPRESS ontwikkeld. MEXPRESS is een online applicatie voor de snelle en eenvoudige visuele integratie van de klinische, genexpressie en DNA-methylatie data uit de TCGA databank.

REFERENCES

Alberts B, Johnson A, Lewis J, Raff M, Roberts K *et al.* Molecular biology of the cell, 4th edition. *Garland Science* (2002)

Anastasiadou C, Malousi A, Maglaveras N & Kouidou S. Human epigenome data reveal increased CpG methylation in alternatively spliced sites and putative exonic splicing enhancers. *DNA Cell Biol* **30**, 267–275 (2011)

Anders S, Pyl PT & Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics **31**, 166–169 (2015)

Baek D, Villen J, Shin C, Camargo FD, Gygi SP *et al.* The impact of microRNAs on protein output. *Nature* **455**, 64–71 (2008)

Ball MP, Li JB, Gao Y, Lee JH, LeProust EM *et al.* Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nat Biotechnol* **27**, 361–368 (2009)

Banchereau J & Steinman RM. Dendritic cells and the control of immunity. *Nature* **392**, 245–252 (1998)

Barski A, Cuddapah S, Cui K, Roh TY, Schones DE *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–837 (2007)

Barsnes H, Vizcaino JA, Eidhammer I & Martens L. PRIDE Converter: making proteomics data- sharing easy. *Nat Biotechnol* **27**, 598–599 (2009)

Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116**, 281–297 (2004)

Bastian PJ, Yegnasubramanian S, Palapattu GS, Rogers CG, Lin X *et al.* Molecular biomarker in prostate cancer: the role of CpG island hypermethylation. *Eur Urol* **46**, 698–708 (2004)

Bazzini AA, Johnstone TG, Christiano R, Mackowiak SD, Obermayer B *et al.* Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J* **33**, 981–993 (2014)

Beck M, Schmidt A, Malmstroem J, Claassen M, Ori A *et al.* The quantitative proteome of a human cell line. *Mol Syst Biol* **7**, 549 (2011)

Benjamini Y & Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Methodol* **57**, 289–300 (1995)

Bidwell BN, Slaney CY, Withana NP, Forster S, Cao Y *et al.* Silencing of Irf7 pathways in breast cancer cells promotes bone metastasis through immune escape. *Nat Med* **18**, 1224–1231(2012)

Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A *et al.* The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* **31**, 365–370 (2003)

Boja ES & Rodriguez H. Proteogenomic convergence for understanding cancer pathways and networks. *Clin Proteomics* **11**, 22 (2014)

Bollag G, Tsai J, Zhang J, Zhang C, Ibrahim P *et al.* Vemurafenib: the first drug approved for BRAF-mutant cancer. *Nat Rev Drug Discov* **11**, 873–886 (2012)

Brahmer JR, Tykodi SS, Chow LQ, Hwu WJ, Topalian SL *et al.* Safety and activity of anti-PD-L1 antibody in patients with advanced cancer. *N Engl J Med* **366**, 2455–2465 (2012)

Brahmer JR, Horn L, Antonia SJ, Spigel DR, Gandhi L *et al.* Survival and long-term follow-up of the phase I trial of nivolumab (Anti-PD-1; BMS-936558; ONO-4538) in patients (pts) with previously treated advanced non-small cell lung cancer (NSCLC). *ASCO Meeting Abstracts* **31**, 8030 (2013)

Branca RM, Orre LM, Johansson HJ, Granholm V, Huss M *et al.* HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics. *Nat Methods* **11**, 59–62 (2014)

Brooks JD, Weinstein M, Lin X, Sun Y, Pin SS *et al.* CG island methylation changes near the GSTP1 gene in prostatic intraepithelial neoplasia. *Cancer Epidemiol Biomarkers Prev* **7**, 531–536 (1998)

Burnet FM. The concept of immunological surveillance. *Prog Exp Tumor Res* **13**, 1–27 (1970)

Carretero R, Wang E, Rodriguez AI, Reinboth J, Ascierto ML *et al.* Regression of melanoma metastases after immunotherapy is associated with activation of antigen presentation and interferon-mediated rejection genes. *Int J Cancer* **131**, 387-395 (2012)

Cedar H & Bergman Y. Linking DNA methylation and histone modification: patterns and paradigms. *Nat Rev Genet* **10**, 295–304 (2009)

Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO *et al.* The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov* **2**, 401–404 (2012)

Challa-Malladi M, Lieu YK, Califano O, Holmes AB, Bhagat G *et al.* Combined genetic inactivation of beta2-Microglobulin and CD58 reveals frequent escape from immune recognition in diffuse large B cell lymphoma. *Cancer Cell* **20**, 728–740 (2011)

Chen Z, Trotman LC, Shaffer D, Lin HK, Dotan ZA *et al.* Crucial role of p53-dependent cellular senescence in suppression of Pten-deficient tumorigenesis. *Nature* **436**, 725–730 (2005)

Cheriyath V, Leaman DW & Borden EC. Emerging roles of FAM14 family members (G1P3/ISG 6-16 and ISG12/ IFI27) in innate immunity and cancer. *J Interferon Cytokine Res* **31**, 173–181 (2011)

Cichorek M, Wachulska M, Stasiewicz A & Tyminska A. Skin melanocytes: biology and development. *Postepy Dermatol Alergol* **30**, 30–41

Claes B, Buysschaert I & Lambrechts D. Pharmaco-epigenomics: discovering therapeutic approaches and biomarkers for cancer therapy. *Heredity (Edinb)* **105**, 152–160 (2010)

Claus R, Almstedt M & Lubbert M. Epigenetic treatment of hematopoietic malignancies: in vivo targets of demethylating agents. *Semin Oncol* **32**, 511–520 (2005)

Colaert N, Vandekerckhove J, Gevaert K & Martens L. A comparison of MS2-based label-free quantitative proteomic techniques with regards to accuracy and precision. *Proteomics* **11**, 1110–1113 (2011)

Costa V, Aprile M, Esposito R & Ciccodicola A. RNA-Seq and human complex diseases: recent accomplishments and future perspectives. *Eur J Hum Genet* **21**, 134–142 (2013)

Craig R & Beavis RC. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **20**, 1466–1467 (2004)

Crappé J, Ndah E, Koch A, Steyaert S, Gawron D *et al.* PROTEOFORMER: deep proteome coverage through ribosome profiling and MS integration. *Nucleid Acids Res* **43**, e29 (2015)

Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci USA* **107**, 21931–21936 (2010)

Crick F. Central dogma of molecular biology. *Nature* **227**, 561–563 (1970)

Cui H, Dhroso A, Johnson N & Korkin D. The variation game: cracking complex genetic disorders with NGS and omics data. *Methods* **79**, 18–31 (2015)

De Carvalho DD, Sharma S, You JS, Su SF, Taberlay PC *et al.* DNA methylation screening identifies driver epigenetic events of cancer cell survival. *Cancer Cell* **21**, 655–667 (2012)

De Meyer T, Mampaey E, Vlemmix M, Denil S, Trooskens G *et al.* Quality evaluation of methyl binding domain based kits for enrichment DNA-methylation sequencing. *PloS One* **8**, e59068 (2013)

Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T *et al.* Landscape of transcription in human cells. *Nature* **489**, 101–108 (2012)

Doherty MK, Hammond DE, Clague MJ, Gaskell SJ & Beynon RJ. Turnover of the human proteome: determination of protein intracellular stability by dynamic SILAC. *J Proteome Res* **8**, 104–112 (2009)

Dong Y, Yu J & Ng SS. MicroRNA dysregulation as a prognostic biomarker in colorectal cancer. *Cancer Manag Res* **6**, 405–422 (2014)

Dunn GP, Bruce AT, Ikeda H, Old LJ & Schreiber RD. Cancer immunoediting: from immunosurveillance to tumor escape. *Nat Immunol* **3**, 991–998 (2002)

Duong CV, Emes RD, Wessely F, Yacqub-Usman K, Clayton RN *et al.* Quantitative, genome-wide analysis of the DNA methylome in sporadic pituitary adenomas. *Endocr Relat Cancer* **19**, 805–816 (2012)

Easwaran HP, Van Neste L, Cope L, Sen S, Mohammad HP *et al.* Aberrant silencing of cancer-related genes by CpG hypermethylation occurs independently of their spatial organization in the nucleus. *Cancer Res* **70**, 8015–8024 (2010)

Eden E, Navon R, Steinfeld I, Lipson D & Yakhini Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* **10**, 48 (2009)

Edgar R, Domrachev M & Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* **30**, 207–210 (2002)

Egger G, Liang G, Aparicio A & Jones PA. Epigenetics in human disease and prospects for epigenetic therapy. *Nature* **429**, 457–463 (2004)

Elwood JM & Jopson J. Melanoma and sun exposure: an overview of published studies. *Int J Cancer* **73**, 198–203 (1997)

Espada J, Peinado H, Lopez-Serra L, Setien F, Lopez-Serra P *et al.* Regulation of SNAIL1 and E-cadherin function by DNMT1 in a DNA methylation-independent context. *Nucleic Acids Res* **39**, 9194–9205 (2011)

Esteller M. Epigenetic gene silencing in cancer: the DNA hypermethylome. *Hum Mol Genet* **16**, R50–59 (2007)

Esteller M, Fraga MF, Guo M, Garcia-Foncillas J, Hedenfalk I *et al.* DNA methylation patterns in hereditary human cancers mimic sporadic tumorigenesis. *Hum Mol Genet* **10**, 3001–3007 (2001)

Finnerty JR, Wang WX, Hébert SS, Wilfred BR, Mao G *et al.* The miR-15/107 group of microRNA genes: evolutionary biology, cellular functions, and roles in human diseases. *J Mol Biol* **402**, 491–509 (2010)

Fonsatti E, Nicolay HJ, Sigalotti L, Calabro L, Pezzani L *et al.* Functional up-regulation of human

leukocyte antigen class I antigens expression by 5-aza-2'-deoxycytidine in cutaneous melanoma: immunotherapeutic implications. *Clin Cancer Res* **13**, 3333–3338 (2007)

Fonslow BR, Carvalho PC, Academia K, Freeby S, Xu T *et al.* Improvements in proteomic metrics of low abundance proteins through proteome equalization using ProteoMiner prior to MudPIT. *J Proteome Res* **10**, 3690–3700 (2011)

Franklin RE & Gosling RG. Molecular configuration in sodium thymonucleate. *Nature* **421**, 400–401 (1953)

Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M *et al.* STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res* **41**, D808–815 (2013)

Frommer M, McDonald LE, Millar DS, Collis CM, Watt F *et al.* A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc Natl Acad Sci USA* **89**, 1827–1831 (1992)

Gabbara S & Bhagwat AS. The mechanism of inhibition of DNA (cytosine-5-)-methyltransferases by 5-azacytosine is likely to involve methyl transfer to the inhibitor. *Biochem J* **307**, 87–92 (1995)

Gandhi TK, Chandran S, Peri S, Saravana R, Amanchy R *et al.* A bioinformatics analysis of protein tyrosine phosphatases in humans. *DNA Res* **12**, 79–89 (2005)

Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M *et al.* Open mass spectrometry search algorithm. *J Proteome Res* **3**, 958–964 (2004)

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5**, R80 (2004)

Germain C, Gnjatic S & Dieu-Nosjean MC. Tertiary lymphoid structure-associated B cells are key players in anti-tumor immunity. *Front Immunol* 6, 67 (2015)

Germain C, Gnjatic S, Tamzalit F, Knockaert S, Remark R *et al.* Presence of B cells in tertiary lymphoid structures is associated with a protective immunity in patients with lung cancer. *Am J Respir Crit Care Med* **189**, 832–844 (2014)

Goecks J, Nekrutenko A, Taylor J & the Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol* **11**, R86 (2010)

Goldman M, Craft B, Swatloski T, Ellrott K, Cline M *et al.* The UCSC Cancer Genomics Browser: update 2013. *Nucleic Acids Res* **41**, D949–954 (2013)

Goll MG, Kirpekar F, Maggert KA, Yoder JA, Hsieh CL *et al.* Methylation of tRNAAsp by the DNA methyltransferase homolog Dnmt2. *Science* **311**, 395–398 (2006)

Gry M, Rimini R, Stromberg S, Asplund A, Ponten F *et al.* Correlations between RNA and protein expression profiles in 23 human cell lines. *BMC Genomics* **10**, 365 (2009)

Gu H, Smith ZD, Bock C, Boyle P, Gnirke A *et al.* Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. *Nat Protoc* **6**, 468–481 (2011)

Guo H, Ingolia NT, Weissman JS & Bartel DP. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature* **466**, 835–840 (2010)

Guo XK, Liu YF, Zhou Y, Sun XY, Qian XP *et al.* The Expression of Netrin-1 in the Thymus and Its Effects on Thymocyte Adhesion and Migration. *Clin Dev Immunol* 462152 (2013)

Guttman M & Rinn JL. Modular regulatory principles of large non-coding RNAs. *Nature* **482**, 339–346 (2012)

Hammerman PS, Hayes DN, Wilkerson MD, Schultz N, Bose R *et al.* Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519–525 (2012)

Hanahan D & Weinberg RA. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011)

Hancock WS. An analytical chemist's perspective. *J Proteome Res* **6**, 1633 (2007)

Hegi ME, Diserens AC, Gorlia T, Hamou MF, de Tribolet N *et al.* MGMT gene silencing and benefit from temozolomide in glioblastoma. *N Engl J Med* **352**, 997–1003 (2005)

Helsens K, Van Damme P, Degroeve S, Martens L, Arnesen T *et al.* Bioinformatics analysis of a Saccharomyces cerevisiae N-terminal proteome provides evidence of alternative translation initiation and post-translational N-terminal acetylation. *J Proteome Res* **10**, 3578–3589 (2011)

Henrique R & Jeronimo C: Molecular detection of prostate cancer: a role for GSTP1 hypermethylation. *Eur Urol* **46**, 660–669 (2004)

Herman JG & Baylin SB. Gene silencing in cancer in association with promoter hypermethylation. *N Engl J Med* **349**, 2042–2054 (2003)

Hirosawa M, Hoshida M, Ishikawa M & Toya T. MASCOT: multiple alignment system for protein sequences based on three-way dynamic programming. *Comput Appl Biosci* **9**, 161–167 (1993)

Hodi FS, Mihm MC, Soiffer RJ, Haluska FG, Butler M *et al.* Biologic activity of cytotoxic T lymphocyte-associated antigen 4 antibody blockade in previously vaccinated metastatic melanoma and ovarian carcinoma patients. *Proc Natl Acad Sci USA* **100**, 4712–4717 (2003)

Hodi FS, O'Day SJ, McDermott DF, Weber RW, Sosman JA *et al.* Improved survival with ipilimumab in patients with metastatic melanoma. *N Engl J Med* **363**, 711–723 (2010)

Hollstein M, Sidransky D, Vogelstein B & Harris CC. p53 mutations in human cancers. *Science* **253**, 49–53

Hsiao SH, Lee KD, Hsu CC, Tseng MJ, Jin VX *et al.* DNA methylation of the Trip10 promoter accelerates mesenchymal stem cell lineage determination. *Biochem Biophys Res Commun* **400**, 305–312 (2010)

Hsieh CL. In vivo activity of murine de novo methyltransferases, Dnmt3a and Dnmt3b. *Mol Cell Biol* **19**, 8211–8218 (1999)

Hsu TH, Chu CC, Jiang SY, Hung MW, Ni WC *et al.* Expression of the class II tumor suppressor gene RIG1 is directly regulated by p53 tumor suppressor in cancer cell lines. *FEBS Lett* **586**, 1287–1293 (2012)

Huang da W, Sherman BT & Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* **37**, 1–13 (2009)

Huang da W, Sherman BT & Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44–57 (2009)

Huang X & Saint-Jeannet JP. Induction of the neural crest and the opportunities of life on the edge. *Dev Biol* **275**, 1–11 (2004)

Huber PJ. Robust statistics. *Wiley* (1981)

Hunter T. Protein kinases and phosphatases: the yin and yang of protein phosphorylation and signaling. *Cell* **80**, 225–236 (1995)

Ingolia NT. Genome-wide translational profiling by ribosome footprinting. *Methods Enzymol* **470**, 119–142 (2010)

Ingolia NT, Lareau LF & Weissman JS. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147**, 789–802 (2011)

Ishihama Y, Oda Y, Tabata T, Sato T, Nagasu T *et al.* Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Mol Cell Proteomics* **4**, 1265–1272 (2005)

Ishikawa H, Ma Z & Barber GN. STING regulates intracellular DNA-mediated, type I interferon-dependent innate immunity. *Nature* **461**, 788–792 (2009)

Ivanov IP, Firth AE, Michel AM, Atkins JF & Baranov PV. Identification of evolutionarily conserved non-AUG-initiated N-terminal extensions in human coding sequences. *Nucleic Acids Res* **39**, 4220–4234 (2011)

Iwai Y, Ishida M, Tanaka Y, Okazaki T, Honjo T *et al.* Involvement of PD-L1 on tumor cells in the escape from host immune system and tumor immunotherapy by PD-L1 blockade. *Proc Natl Acad Sci USA* **99**, 12293–12297 (2002)

Jabbari K & Bernardi G. Cytosine methylation and CpG, TpG (CpA) and TpA frequencies. *Gene* **333**, 143–149 (2004)

Jaenisch R & Bird A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat Genet* **33**, 245–254 (2003)

Jahner D, Stuhlmann H, Stewart CL, Harbers K, Lohler J *et al.* De novo methylation and expression of retroviral genomes during mouse embryogenesis. *Nature* **298**, 623–628 (1982)

Janeway CA, Travers P, Walport M & Shlomchik MJ. Immunology, 5th edition: the immune system in health and disease. *Garland Science* (2001)

Jee CD, Kim MA, Jung EJ, Kim J & Kim WH. Identification of genes epigenetically silenced by CpG methylation in human gastric carcinoma. *Eur J Cancer* **45**, 1282–1293 (2009)

Jones PA & Baylin SB. The epigenomics of cancer. *Cell* **128**, 683–692 (2007)

Jones PA & Taylor SM. Cellular differentiation, cytidine analogs and DNA methylation. *Cell* **20**, 85–93 (1980)

Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E *et al.* Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res* **33**, D428–432 (2005)

Ju J, Lim SK, Jiang H, Seo JW & Shen B. Iso-migrastatin congeners from Streptomyces platensis and generation of a glutarimide polyketide library featuring the dorrigocin, lactimidomycin, migrastatin, and

NK30424 scaffolds. *J Am Chem Soc* **127**, 11930–11931 (2005)

Juergens RA, Wrangle J, Vendetti FP, Murphy SC, Zhao M *et al.* Combination epigenetic therapy has efficacy in patients with refractory advanced non-small cell lung cancer. *Cancer Discov* **1**, 598–607 (2011)

Kainthla R, Kim KB & Falchook GS. Dabrafenib for treatment of BRAF-mutant melanoma. *Pharmgenomics Pers Med* **7**, 21–29

Karpf AR, Peterson PW, Rawlins JT, Dalley BK, Yang Q *et al.* Inhibition of DNA methyltransferase stimulates the expression of signal transducer and activator of transcription 1, 2, and 3 genes in colon tumor cells. *Proc Natl Acad Sci USA* **96**, 14007–14012 (1999)

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH *et al.* The human genome browser at UCSC. *Genome Res* **12**, 996–1006 (2002)

Khong HT & Restifo NP. Natural selection of tumor variants in the generation of "tumor escape" phenotypes. *Nat Immunol* **3**, 999–1005 (2002)

Kim D, Pertea G, Trapnell C, Pimentel H, Kelly R *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**, R36 (2013)

King MC, Marks JH, Mandell JB & New York Breast Cancer Study Group. Breast and ovarian cancer risks due to inherited mutations in BRCA1 and BRCA2. *Science* **302**, 643–646 (2003)

Koch CM, Andrews RM, Flicek P, Dillon SC, Karaöz U *et al.* The landscape of histone modifications across 1% of the human genome in five human cell lines. *Genome Res* **17**, 691–707 (2007)

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

Kulaeva OI, Draghici S, Tang L, Kraniak JM, Land SJ *et al.* Epigenetic silencing of multiple interferon pathway genes after cellular immortalization. *Oncogene* **22**, 4118–4127 (2003)

Kuper H, Adami HO & Boffetta P. Tobacco use, cancer causation and public health impact. *J Intern Med* **251**, 455–466 (2002)

Kushi LH, Byers T, Doyle C, Bandera EV, McCullough M *et al.* American Cancer Society Guidelines on Nutrition and Physical Activity for cancer prevention: reducing the risk of cancer with healthy food choices and physical activity. *CA Cancer J Clin* **56**, 254–281 (2006)

Lachner M & Jenuwein T. The many faces of histone lysine methylation. *Curr Opin Cell Biol* **14**, 286–298 (2002)

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001)

Langmead B, Trapnell C, Pop M & Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25 (2009)

Laurent L, Wong E, Li G, Huynh T, Tsirigos A *et al.* Dynamic changes in the human methylome during differentiation. *Genome Res* **20**, 320–331 (2010)

Lee MG, Kim HY, Byun DS, Lee SJ, Lee CH *et al.* Frequent epigenetic inactivation of RASSF1A in human bladder carcinoma. *Cancer Res* **61**, 6688–6692 (2001)

Lee S, Liu B, Lee S, Huang SX, Shen B *et al.* Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proc Natl Acad Sci USA* **109**, E2424–2432 (2012)

Lepelletier Y, Smaniotto S, Hadj-Slimabe R, Villa-Verde DM, Nogueira AC *et al.* Control of human thymocyte migration by Neuropilin-1/Semaphorin-3A-mediated interactions. *Proc Natl Acad Sci USA* **104**, 5545–5550 (2007)

Li H, Chiappinelli KB, Guzzetta AA, Easwaran H, Yen RW *et al.* Immune regulation by low doses of the DNA methyltransferase inhibitor 5-azacitidine in common human epithelial cancers. *Oncotarget* **5**, 587–598 (2014)

Li Q & Tainsky MA. Epigenetic silencing of IRF7 and/or IRF5 in lung cancer cells leads to increased sensitivity to oncolytic viruses. *PLoS One* **6**, e28683 (2011)

Lièvre A, Bachet JB, Le Corre B, Boige V, Landi B *et al.* KRAS mutation status is predictive of response to cetuximab therapy in colorectal cancer. *Cancer Res* **66**, 3992–3995 (2006)

Liu S, Im H, Bairoch A, Cristofanilli M, Chen R *et al.* A chromosome-centric human proteome project (C- HPP) to characterize the sets of proteins encoded in chromosome 17. *J Proteome Res* **12**, 45–57 (2013)

Lopez F, Textoris J, Bergon A, Didier G, Remy E *et al.* Transcriptome-Browser: a powerful and flexible toolbox to explore productively the transcriptional landscape of the Gene Expression Omnibus database. *PLoS One* **3**, e4001 (2008)

Low TY, van Heesch S, van den Toorn H, Giansanti P, Cristobal A *et al.* Quantitative and qualitative

proteome characteristics extracted from in-depth integrated genomics and proteomics analysis. *Cell Rep* **5**, 1469–1478 (2013)

Lu R, Au WC, Yeow WS, Hageman N & Pitha PM. Regulation of the promoter activity of interferon regulatory factor-7 gene. Activation by interferon snd silencing by hypermethylation. *J Biol Chem* **275**, 31805–31812 (2000)

Maio M, Grob JJ, Aamdal S, Bondarenko I, Robert C *et al.* Five-Year Survival Rates for Treatment-Naive Patients With Advanced Melanoma Who Received Ipilimumab Plus Dacarbazine in a Phase III Trial. *J Clin Oncol* **33**, 1191–1196 (2015)

Markowitz SD & Bertagnolli MM. Molecular origins of cancer: Molecular basis of colorectal cancer. *N Engl J Med* **361**, 2449–2460 (2009)

Martens L, Hermjakob H, Jones P, Adamski M, Taylor C *et al.* PRIDE: the proteomics identifications database. *Proteomics* **5**, 3537–3545 (2005)

Maunakea AK, Chepelev I, Cui K & Zhao K. Intragenic DNA methylation modulates alternative splicing by recruiting MeCP2 to promote exon recognition. *Cell Res* **23**, 1256–1269 (2013)

Maunakea AK, Nagarajan RP, Bilenky M, Ballinger TJ, D'Souza C *et al.* Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature* **466**, 253–257 (2010)

McIlwain S, Mathews M, Bereman MS, Rubel EW, MacCoss MJ *et al.* Estimating relative abundances of proteins from shotgun proteomics data. *BMC Bioinformatics* **13**, 308 (2012)

Medenbach J, Seiler M & Hentze MW. Translational control via protein-regulated upstream open reading frames. *Cell* **145**, 902–913 (2011)

Meissner A, Gnirke A, Bell GW, Ramsahoye B, Lander ES *et al.* Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res* **33**, 5868–5877 (2005)

Menschaert G, Van Criekinge W, Notelaers T, Koch A, Crappé J *et al.* Deep proteome coverage based on ribosome profiling aids mass spectrometry-based protein and peptide discovery and provides evidence of alternative translation products and near-cognate translation initiation events. *Mol Cell Proteomics* **12**, 1780–1790 (2013)

Mercer TR, Dinger ME & Mattick JS. Long non-coding RNAs: insights into functions. *Nat Rev Genet* **10**, 155–159 (2009)

Meshcheryakova A, Tamandl D, Bajna E, Stift J, Mittlboeck M *et al.* B cells and ectopic follicular structures: novel players in anti-tumor programming with prognostic power for patients with metastatic colorectal cancer. *PloS One* **9**, e99008 (2014)

Michel A, O'Connor P, Choudhury RK, Firth A Li GW *et al.* Elucidating mechanisms of translation with computational analysis of ribo-seq data. *EMBO Conference Series: Protein Synthesis and Translational Control. Heidelberg, Germany* (2013)

Michel AM, Choudhury KR, Firth AE, Ingolia NT, Atkins JF *et al.* Observation of dually decoded regions of the human genome using ribosome profiling data. *Genome Res* **22**, 2219–2229 (2012)

Millar DS, Ow KK, Paul CL, Russell PJ, Molloy PL *et al.* Detailed methylation analysis of the glutathione S-transferase pi (GSTP1) gene in prostate cancer. *Oncogene* **18**, 1313–1324 (1999)

Milutinovic S, Zhuang Q, Niveleau A & Szyf M. Epigenomic stress response. Knockdown of DNA methyltransferase 1 triggers an intra-S-phase arrest of DNA replication and induction of stress response genes. *J Biol Chem* **278**, 14985–14995 (2003)

Miranda TB & Jones PA. DNA methylation: the nuts and bolts of repression. *J Cell Physiol* **213**, 384–390 (2007)

Mohandas T, Sparkes RS & Shapiro LJ. Reactivation of an inactive human X chromosome: evidence for X inactivation by DNA methylation. *Science* **211**, 393–396 (1981)

Nagaraj N, Wisniewski JR, Geiger T, Cox J *et al.* Deep proteome and transcriptome mapping of a human cancer cell line. *Mol Syst Biol* **7**, 548 (2011)

Nakaya A, Katayama T, Itoh M, Hiranuka K, Kawashima S *et al.* KEGG OC: a large-scale automatic construction of taxonomy-based ortholog clusters. *Nucleic Acids Res* **41**, D353–357 (2013)

Namy O, Rousset JP, Napthine S & Brierley I. Reprogrammed genetic decoding in cellular gene expression. *Mol Cell* **13**, 157–168 (2004)

Neyns B, Wilgenhof S, Corthals J, Heirman C & Thielemans K. Phase II study of autologous mRNA electroporated dendritic cells (TriMixDC-MEL) in combination with ipilimumab in patients with pretreated advanced melanoma. *ASCO Annual Meeting* (2014)

Nielsen JS, Sahota RA, Milne K, Kost SE, Nesslinger NJ *et al.* CD20+ tumor-infiltrating lymphocytes have an atypical CD27- memory phenotype and together with CD8+ T cells promote favorable prognosis in

ovarian cancer. *Clin Cancer Res* **18**, 3281–3292 (2012)

Ning K & Nesvizhskii AI. The utility of mass spectrometry-based proteomic data for validation of novel alternative splice forms reconstructed from RNA-Seq data: a preliminary assessment. *BMC bioinformatics* **11**, S14 (2010)

Ning K, Fermin D & Nesvizhskii AI. Comparative analysis of different label-free mass spectrometry based protein abundance estimates and their correlation with RNA-Seq gene expression data. *J Proteome Res* **11**, 2261–2271 (2012)

Niu D, Sui J, Zhang J, Feng H & Chen WN. iTRAQ-coupled 2-D LC-MS/MS analysis of protein profile associated with HBV-modulated DNA methylation. *Proteomics* **9**, 3856–3868 (2009)

Offenhauser N, Borgonovo A, Disanza A, Romano P, Ponzanelli I *et al.* The eps8 family of proteins links growth factor stimulation to actin reorganization generating functional redundancy in the Ras/Rac pathway. *Mol Biol Cell* **15**, 91–98 (2004)

Okazaki T & Wang J. PD-1/PD-L pathway and autoimmunity. *Autoimmunity* **38**, 353–357 (2005)

Ong SE, Blagoev B, Kratchmarova I, Kristensen DB, Steen H *et al.* Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics* **1**, 376–386 (2002)

Ong SE, Kratchmarova I & Mann M. Properties of 13C-substituted arginine in stable isotope labeling by amino acids in cell culture (SILAC). *J Proteome Res* **2**, 173–181 (2003)

Orozco LD, Morselli M, Rubbi L, Guo W, Go J *et al.* Epigenome-wide association of liver methylation patterns and complex metabolic traits in mice. *Cell Metab* **21**, 905–917 (2015)

Pal S, Gupta R & Davuluri RV. Alternative transcription and alternative splicing in cancer. *Pharmacol Ther* **136**, 283–294 (2012)

Paoletti AC, Parmely TJ, Tomomori-Sato C, Sato S, Zhu D *et al.* Quantitative proteomic analysis of distinct mammalian Mediator complexes using normalized spectral abundance factors. *Proc Natl Acad Sci USA* **103**, 18928–18933 (2006)

Pardoll DM. The blockade of immune checkpoints in cancer immunotherapy. *Nat Rev Cancer* **12**, 252–264 (2012)

Parker JS, Mullins M, Cheang MC, Leung S, Voduc D *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol* **27**, 1160–1167 (2009)

Perez-Llamas C & Lopez-Bigas N. Gitools: analysis and visualisation of genomic data using interactive heat-maps. *PloS One* **6**, e19541 (2011)

Perkins DN, Pappin DJ, Creasy DM & Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551–3567 (1999)

Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS *et al.* Molecular portraits of human breast tumours. *Nature* **406**, 747–752 (2000)

Pinto SM, Manda SS, Kim MS, Taylor K, Selvan LD *et al.* Functional annotation of proteome encoded by human chromosome 22. *J Proteome Res* **13**, 2749–2760 (2014)

Polyakova A, Kuznetsova K & Moshkovskii S. Proteogenomics meets cancer immunology: mass spectrometric discovery and analysis of neoantigens. *Expert Rev Proteomics* **15**, 1–9 (2015)

Pretscher D, Distel LV, Grabenbauer GG, Wittlinger M, Buettner M *et al.* Distribution of immune cells in head and neck cancer: CD8+ T-cells and CD20+ B-cells in metastatic lymph nodes are associated with favourable outcome in patients with oro- and hypopharyngeal carcinoma. *BMC Cancer* **9**, 292 (2009)

Procko E & Gaudet R. Antigen processing and presentation: TAPping into ABC transporters. *Curr Opin Immunol* **21**, 84–91 (2009)

Qiu GH, Leung CH, Yun T, Xie X, Laban M *et al.* Recognition and suppression of transfected plasmids by protein ZNF511-PRAP1, a potential molecular barrier to transgene expression. *Mol Ther* **19**, 1478–1486 (2011)

Raghavan M, Del Cid N, Rizvi SM & Peters LR. MHC class I assembly: out and about. *Trends Immunol* **29**, 436–443 (2008)

Ravasi T, Suzuki H, Cannistraci CV, Katayama S, Bajic VB *et al.* An atlas of combinatorial transcriptional regulation in mouse and man. *Cell* **140**, 744–752 (2010)

Reese MG. Application of a time-delay neural network to promoter annotation in the Drosophila melanogaster genome. *Comput Chem* **26**, 51–56 (2001)

Rhee I, Bachman KE, Park BH, Jair KW, Yen RW *et al.* DNMT1 and DNMT3b cooperate to silence genes in human cancer cells. *Nature* **416**, 552–556 (2002)

Robinson MD, McCarthy DJ & Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010)

Robson ME. Clinical considerations in the management of individuals at risk for hereditary breast and ovarian cancer. *Cancer Control* **9**, 457–465 (2002)

Rosenfeld JA, Wang Z, Schones DE, Zhao K, DeSalle R *et al.* Determination of enriched histone modifications in non-genic portions of the human genome. *BMC Genomics* **10**, 143

Rozenblum E, Schutte M, Goggins M, Hahn SA, Panzer S *et al.* Tumor-suppressive pathways in pancreatic carcinoma. *Cancer Res* **57**, 1731–1734 (1997)

Samarajiwa SA, Forster S, Auchettl K & Hertzog PJ. INTERFEROME: the database of interferon regulated genes. *Nucleic Acids Res* **37**, D852–857 (2009)

Sandoval PC, Slentz DH, Pisitkun T, Saeed F, Hoffert JD *et al.* Proteome-wide measurement of protein half- lives and translation rates in vasopressin-sensitive collecting duct cells. *J Am Soc Nephrol* **24**, 1793–1805 (2013)

Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR *et al.* Nucleotide sequence of bacteriophage ϕX174 DNA. *Nature* **265**, 687–695 (1977)

Sanger F, Nicklen S & Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA* **74**, 5463–5467 (1977)

Santi DV, Norment A & Garrett CE. Covalent bond formation between a DNA-cytosine methyltransferase and DNA containing 5-azacytosine. *Proc Natl Acad Sci USA* **81**, 6993–6997 (1984)

Santos-Rosa H, Schneider R, Bannister AJ, Sherriff J, Bernstein BE *et al.* Active genes are tri-methylated at K4 of histone H3. *Nature* **419**, 407–411 (2002)

Scanlan MJ, Simpson AJ & Old LJ. The cancer/testis genes: review, standardization, and commentary. *Cancer Immun* **4**, 1 (2004)

Schadendorf D, Hodi FS, Robert C, Weber JS, Margolin K *et al.* Pooled Analysis of Long-Term Survival Data From Phase II and Phase III Trials of Ipilimumab in Unresectable or Metastatic Melanoma. *J Clin Oncol* 33, 1889–1894 (2015)

Schaefer M & Lyko F. Solving the Dnmt2 enigma. *Chromosoma* **119**, 35–40 (2010)

Schneider-Poetsch T, Ju J, Eyler DE, Dang Y, Bhat S *et al.* Inhibition of eukaryotic translation elongation by cycloheximide and lactimidomycin. *Nat Chem Biol* **6**, 209–217 (2010)

Schroeder MP, Gonzalez-Perez A & Lopez-Bigas N. Visualizing multidimensional cancer genomics data. *Genome Med* **5**, 9 (2013)

Schuebel KE, Chen W, Cope L, Glockner SC, Suzuki H *et al.* Comparing the DNA hypermethylome with gene mutations in human colorectal cancer. *PLoS Genet* **3**, 1709–1723 (2007)

Schulz WA, Alexa A, Jung V, Hader C, Hoffmann MJ *et al.* Factor interaction analysis for chromosome 8 and DNA methylation alterations highlights innate immune response suppression and cytoskeletal changes in prostate cancer. *Mol Cancer* **6**, 14 (2007)

Schwanhäusser B, Busse D, Li N, Dittmar G, Schuchhardt J *et al.* Global quantification of mammalian gene expression control. *Nature* **473**, 337–342 (2011)

Searle BC, Turner M & Nesvizhskii AI. Improving sensitivity by probabilistically combining results from multiple MS/MS search methodologies. *J Proteome Res* 7, 245–253 (2008)

Selbach M, Schwanhausser B, Thierfelder N, Fang Z, Khanin R *et al.* Widespread changes in protein synthesis induced by microRNAs. *Nature* **455**, 58–63 (2008)

Serre D, Lee BH & Ting AH. MBD-isolated Genome Sequencing provides a high-throughput and comprehensive survey of DNA methylation in the human genome. *Nucleic Acids Res* **38**, 391–9 (2010)

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498–2504 (2003)

Sharma S & Fitzgerald KA. Viral defense: it takes two MAVS to Tango. *Cell* **141**, 570–572 (2010)

Shen JC, Rideout WM & Jones PA. The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA. *Nucleic Acids Res* **22**, 972–976 (1994)

Shepherd FA, Rodrigues Pereira J, Ciuleanu T, Tan EH, Hirsh V *et al.* Erlotinib in previously treated non-small-cell lung cancer. *N Engl J Med* **353**, 123–132 (2005)

Shi JY, Gao Q, Wang ZC, Zhou J, Wang XY *et al.* Margin-infiltrating CD20(+) B cells display an atypical memory phenotype and correlate with favorable prognosis in hepatocellular carcinoma. *Clin Cancer Res* **19**, 5994–6005 (2013)

Shukla HD, Mahmood J & Vujaskovic Z. Integrated proteo-genomic approach for early diagnosis and

prognosis of cancer. *Cancer Lett* (2015)

Shukla S, Kavak E, Gregory M, Imashimizu M, Shutinoski B *et al.* CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature* **479**, 74–79 (2011)

Siegel R, Naishadham D & Jemal A. Cancer statistics. *CA Cancer J Clin* **63**, 11–30 (2013)

Siegfried Z, Eden S, Mendelsohn M, Feng X, Tsuberi BZ *et al.* DNA methylation represses transcription in vivo. *Nat Genet* **22**, 203–206 (1999)

Silverman LR, Demakos EP, Peterson BL, Kornblith AB, Holland JC *et al.* Randomized controlled trial of azacitidine in patients with the myelodysplastic syndrome: a study of the cancer and leukemia group B. *J Clin Oncol* **20**, 2429–2440 (2002)

Simova J, Pollakova V, Indrova M, Mikyskova R, Bieblova J *et al.* Immunotherapy augments the effect of 5-azacytidine on HPV16-associated tumours with different MHC class I-expression status. *Br J Cancer* **105**, 1533–1541 (2011)

Slavoff SA, Mitchell AJ, Schwaid AG, Cabili MN, Ma J *et al.* Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nat Chem Biol* **9**, 59–64 (2013)

Smith LM, Kelleher NL & Consortium for Top Down Proteomics. Proteoform: a single term describing protein complexity. *Nat Methods* **10**, 186–187 (2013)

Smyth GK & Speed T. Normalization of cDNA microarray data. *Methods* **31**, 265–273 (2003)

Snyder A, Makarov V, Merghoub T, Yuan J, Zaretsky JM *et al.* Genetic basis for clinical response to CTLA-4 blockade in melanoma. *N Engl J Med* **371**, 2189–2199 (2014)

Sonenberg N & Hinnebusch AG. New modes of translational control in development, behavior, and disease. *Mol Cell* **28**, 721–729 (2007)

Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci USA* **98**, 10869–10874 (2001)

Staes A, Impens F, Van Damme P, Ruttens B, Goethals M *et al.* Selecting protein N-terminal peptides by combined fractional diagonal chromatography. *Nat Protoc* **6**, 1130–1141 (2011)

Steinman L. Elaborate interactions between the immune and nervous systems. *Nat Immunol* **5**, 575–581 (2004)

Stern-Ginossar N, Weisburd B, Michalski A, Le VT, Hein MY *et al.* Decoding human cytomegalovirus. *Science* **338**, 1088–1093 (2012)

Straif K, Benbrahim-Tallaa L, Baan R, Grosse Y, Secretan B *et al.* A review of human carcinogens—Part C: metals, arsenic, dusts, and fibres. *Lancet Oncol* **10**, 453–454 (2009)

Stresemann C & Lyko F. Modes of action of the DNA methyltransferase inhibitors azacytidine and decitabine. *Int J Cancer* **123**, 8–13 (2008)

Strowig T, Henao-Mejia J, Elinav E & Flavell R. Inflammasomes in health and disease. *Nature* **481**, 278–286 (2012)

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* **102**, 15545–15550 (2005)

Subramanian A, Kuehn H, Gould J, Tamayo P & Mesirov JP. GSEA-P: a desktop application for Gene Set Enrichment Analysis. *Bioinformatics* **23**, 3251–3253 (2007)

Swain JL, Stewart TA & Leder P. Parental legacy determines methylation and expression of an autosomal transgene: a molecular mechanism for parental imprinting. *Cell* **50**, 719–727 (1987)

Takai D & Jones PA. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc Natl Acad Sci USA* **99**, 3740–3745 (2002)

Tang C, Tan T, Xiao Y, Ruan L, Li C *et al.* Screening for methylation-silenced genes in acute myeloid leukemia HL-60 cell line by a quantitative proteomic approach. *Zhong Nan Da Xue Xue Bao Yi Xue Ban* **35**, 641–648 (2010)

Tapia T, Smalley SV, Kohen P, Muñoz A, Solis LM *et al.* Promoter hypermethylation of BRCA1 correlates with absence of expression in hereditary breast cancer tumors. *Epigenetics* **3**, 157–163 (2008)

The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068 (2008)

The Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609–615 (2011)

The Cancer Genome Atlas Research Network. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med* **368**, 2059–2074 (2013)

The Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543–550 (2014)

Thorvaldsdottir H, Robinson JT & Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* **14**, 178–192 (2013)

Tian Q, Price ND & Hood L. Systems cancer medicine: towards realization of predictive, preventive, personalized and participatory (P4) medicine. *J Intern Med* **271**, 111–121 (2012)

Timmerman JM & Levy R. Dendritic cell vaccines for cancer immunotherapy. *Annu Rev Med* **50**, 507–529 (1999)

Tomasi TB, Magner WJ & Khan AN. Epigenetic regulation of immune escape genes in cancer. *Cancer Immunol Immunother* **55**, 1159–1184 (2006)

Topalian SL, Hodi FS, Brahmer JR, Gettinger SN, Smith DC *et al.* Safety, activity, and immune correlates of anti-PD-1 antibody in cancer. *N Engl J Med* **366**, 2443–2454 (2012)

Touriol C, Bornes S, Bonnal S, Audigier S, Prats H *et al.* Generation of protein isoform diversity by alternative initiation of translation at non-AUG codons. *Biol Cell* **95**, 169–178 (2003)

Tsai HC, Li H, Van Neste L, Cai Y, Robert C *et al.* Transient Low Doses of DNA-Demethylating Agents Exert Durable Antitumor Effects on Hematological and Epithelial Tumor Cells. *Cancer Cell* **21**, 430–446 (2012)

Tufte ER. The visual display of quantitative information. *Graphics Press* (1983)

Ulloa-Montoya F, louahed J, Dizier B, Gruselle O, Spiessens B *et al.* Predictive gene signature in MAGE-A3 antigen-specific cancer immunotherapy. *J Clin Oncol* **31**, 2388–2395 (2013)

Vadakara J & Borghaei H. Personalized medicine and treatment approaches in non-small-cell lung carcinoma. *Pharmgenomics Pers Med* **5**, 113–123 (2012)

Van Damme P, Gawron D, Van Criekinge W & Menschaert G. N-terminal proteomics and ribosome profiling provide a comprehensive view of the alternative translation initiation landscape in mice and men. *Mol Cell Proteomics* **13**, 1245–1261 (2014)

Van Damme P, Hole K, Pimenta-Marques A, Helsens K, Vandekerckhove J *et al.* NatF contributes to an evolutionary shift in protein N-terminal acetylation and is important for normal chromosome segregation. *PLoS Genet* 7, e1002169 (2011)

Van Lint S, Wilgenhof S, Heirman C, Corthals J, Breckpot K *et al.* Optimized dendritic cell-based immunotherapy for melanoma: the TriMix-formula. *Cancer Immunol Immunother* **63**, 959–967 (2014)

Vaquerizas JM, Akhtar A & Luscombe NM. Large-scale nuclear architecture and transcriptional control. *Sub-cellular biochemistry* **52**, 279–295 (2011)

Vaquerizas JM, Teichmann SA & Luscombe NM. How do you find transcription factors? Computational approaches to compile and annotate repertoires of regulators for any genome. *Methods Mol Biol* **786**, 3–19 (2012)

Vasquez JJ, Hon CC, Vanselow JT, Schlosser A & Siegel TN. Comparative ribosome profiling reveals extensive translational complexity in different Trypanosoma brucei life cycle stages. *Nucleic Acids Res* **42**, 3623–3637 (2014)

Vaudel M, Barsnes H, Berven FS, Sickmann A & Martens L. SearchGUI: An open-source graphical user interface for simultaneous OMSSA and X!Tandem searches. *Proteomics* **11**, 996–999 (2011)

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001)

Vizcaino JA, Cote RG, Csordas A, Dianes JA, Fabregat A *et al.* The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Res* **41**, D1063–1069 (2013)

Volders PJ, Helsens K, Wang X, Menten B, Martens L *et al.* LNCipedia: a database for annotated human lncRNA transcript sequences and structures. *Nucleic Acids Res* **41**, D246–251 (2013)

Waghray A, Murali AR & Menon KN. Hepatocellular carcinoma: From diagnosis to treatment. *World J Hepatol* 7, 1020–1029 (2015)

Wan PT, Garnett MJ, Roe SM, Lee S, Niculescu-Duvaz D *et al.* Mechanism of activation of the RAF-ERK signaling pathway by oncogenic mutations of B-RAF. *Cell* **116**, 855–867 (2004)

Wang X, Slebos RJ, Wang D, Halvey PJ, Tabb DL *et al.* Protein identification using customized protein sequence databases derived from RNA-Seq data. *J Proteome Res* **11**, 1009–1017 (2012)

Ward A, Sivakumar G, Kanjeekal S, Hamm C, Labute BC *et al.* The deregulated promoter methylation of the Polo-like kinases as a potential biomarker in hematological malignancies. *Leuk Lymphoma*, 1–11 (2015)

Watson JD & Crick FH. The structure of DNA. *Cold Spring Harb Symp Quant Biol* **18**, 123–131 (1953)

Wethmar K, Begay V, Smink JJ, Zaragoza K, Wiesenthal V *et al.* C/EBPbetaDeltauORF mice–a genetic model for uORF-mediated translational control in mammals. *Genes Dev* **24**, 15–20 (2010)

Wilgenhof S, Van Nuffel AM, Corthals J, Heirman C, Tuyaerts S *et al.* Therapeutic vaccination with an autologous mRNA electroporated dendritic cell vaccine in patients with advanced melanoma. *J Immunother* **34**, 448–456 (2011)

Wilgenhof S, Van Nuffel AM, Benteyn D, Corthals J, Aerts C *et al.* A phase IB study on intravenous synthetic mRNA electroporated dendritic cell immunotherapy in pretreated advanced melanoma patients. *Ann Oncol* **24**, 2686–2693 (2013)

Wilkins MH. Physical studies of the molecular structure of deoxyribose nucleic acid and nucleoprotein. *Cold Spring Harb Symp Quant Biol* **21**, 75–90 (1957)

Wilkins MR, Gasteiger E, Bairoch A, Sanchez JC, Williams KL *et al.* Protein identification and analysis tools in the ExPASy server. *Methods Mol Biol* **112**, 531–552 (1999)

Wirsing AM, Rikardsen OG, Steigen SE, Uhlin-Hansen L & Hadler-Olsen E. Characterisation and prognostic value of tertiary lymphoid structures in oral squamous cell carcinoma. *BMC Clin Pathol* **14**, 38 (2014)

Woo S, Cha SW, Merrihew G, He Y, Castellana N *et al.* Proteogenomic database construction driven from large scale RNA-seq data. *J Proteome Res* **13**, 21–28 (2014)

World Health Organization. World Cancer Report 2014. *IARC* (2014)

Wrangle J, Wang W, Koch A, Easwaran H, Mohammad HP *et al.* Alterations of immune response of Non-Small Cell Lung Cancer with Azacytidine. *Oncotarget* **4**, 2067–2079 (2013)

Xu W, Fang P, Zhu Z, Dai J, Nie D *et al.* Cigarette smoking exposure alters pebp1 DNA methylation and protein profile involved in MAPK signaling pathway in mice testis. *Biol Reprod* **89**, 142 (2013)

Xu Y, Zhong H & Shi W. MAVS protects cells from apoptosis by negatively regulating VDAC1. *Mol Cell Biochem* **375**, 219 (2010)

Yang Z & Klionsky DJ. Mammalian autophagy: core molecular machinery and signaling regulation. *Curr Opin Cell Biol* **22**, 124–131 (2010)

Yi JM, Dhir M, Van Neste L, Downing SR, Jeschke J *et al.* Genomic and epigenomic integration identifies a prognostic signature in colon cancer. *Clin Cancer Res* **17**, 1535–1545 (2011)

Yoder JA, Soman NS, Verdine GL & Bestor TH. DNA (cytosine-5)-methyltransferases in mouse cells and tissues. Studies with a mechanism-based probe. *J Mol Biol* 270, 385–395 (1997)

Yoder JA, Walsh CP & Bestor TH. Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet* **13**, 335-340 (1997)

Youlden DR, Cramb SM & Baade PD. The International Epidemiology of Lung Cancer: geographical distribution and secular trends. *J Thorac Oncol* **3**, 819–831 (2008)

Yuan J, Adamow M, Ginsberg BA, Rasalan TS, Ritter E *et al.* Integrated NY-ESO-1 antibody and CD8+ T-cell responses correlate with clinical benefit in advanced melanoma patients treated with ipilimumab. *Proc Natl Acad Sci USA* **108**, 16723–16728 (2011)

Zambelli F, Pesole G & Pavesi G. Pscan: finding over- represented transcription factor binding site motifs in sequences from co-regulated or co-expressed genes. *Nucleic Acids Res* **37**, W247–252 (2009)

Zhang J, Finney RP, Rowe W, Edmonson M, Yang SH *et al.* Systematic analysis of genetic alterations in tumors using Cancer Genome WorkBench (CGWB). *Genome Res* **17**, 1111–1117 (2007)

A BIG THANK YOU

Writing these acknowledgements was more difficult than I thought it would be. Not because I have a hard time thanking people for their friendship and support (far from it), but because I couldn't stop thinking back on everything that happened these past few years and how I ended up where I am today (sitting here at my desk struggling to get these words on paper and to keep my mind from drifting of).

My hopes of becoming a world-class rower crashed and burned in the fall of 2010. Ready to revive the aspiring scientist in me, I started looking for interesting PhD positions. I had only two demands, the PhD had to be bioinformatics related (computers are the future!) and I wanted to go abroad. Finding a position abroad proved to be more difficult than I thought, but when I talked to Wim about the possibility of starting a PhD in his lab he told me he was looking for someone to keep his collaboration with Johns Hopkins in Baltimore going. Wim, I want to thank you for the opportunities you gave me. I had a great time at Hopkins and met many wonderful people there. Hari, Subodh, Tina, Ben and Stacy, thank you for all the good times in Baltimore! And who in the world would open up his house for a lonely Belgian kid without a roof over his head? Calvin, that's who. You're the best.

There are many exciting moments throughout a PhD, just as there are mind-numbingly boring days. It was a pleasure to have a nice group of colleagues on my side to enjoy the good days with and to get me through the bad ones. Thank you Klaas, Sandra, Jeroen & Jeroen, Joachim, Geert, Vladimir, Elvis, Simon and Daisy. Tim and Gerben, I want to thank you for your guidance and support. This thesis would not have existed without it. And Gerben, thank you for your infectious enthusiasm for science. I'm not the first and will not be the last to be inspired by it. Without Gerben, I would not have participated in the 2013 ASMS conference in Minneapolis. I would not have met David, Kelly and Jennifer and I would not have spent two months in David's lab at NYU. David, thank you for taking me in. I did not have to cross the Atlantic Ocean to find more excellent collaborators though. Petra, Bart and Teo, it was a pleasure to work with you.

There is of course more to life than just academia (luckily!) and what better way to take your mind off research than to enjoy the company of your best friends? Karen, Karel, Thomas, Evi and Christophe, it was great to work my way through university by your side. Despite a shattered dream or two, rowing has continued to be an important part of my life, not in the least for the friends it gave me. Vera, Jolien, Wouter, Jo and Renne, nothing beats a night out at Ozman's or some time on the water together to push an irritating software bug or a rejected paper to the back of my mind. And then there's the Zebrastraat boys. Kristof, Pieter, Lander and Henryk, I guess I have to thank you for not calling me a nerd every single day? Above all, I consider myself very lucky to have such an international group of amazing people around me.

Last but not least, from friends to family. Mama, papa, my favorite brother and Eve, thank you for supporting me, no matter what I choose to do, through success and failure. I could have introduced you as part of the Baltimore gang, but, Jana, you're part of the family now. Who would have thought I'd find my girl in the greatest city in America? Jana, I admire your intelligence and strength. You are an inspiration. And you know what they say, don't you? A couple that publishes together stays together!

Four years ago I could have never predicted where I would be today. I can't wait to see what the next years will bring! Thank you all for sharing this beautiful ride with me.

CURRICULUM VITAE

PERSONAL INFORMATION

Alexander Koch
º 22/07/1986 Gent, Belgium
Patijntjestraat 150
9000 Gent
Belgium
+32 (0)485/90 42 70
alexander_koch86@hotmail.com

EDUCATION & EXPERIENCE

FEBRUARY 2011 – NOW

PhD student at the lab of bioinformatics and computational genomics, department of mathematical modeling, statistics and bioinformatics, Ghent University (Ghent, Belgium).
Visiting trainee at the Baylin lab, Johns Hopkins University School of Medicine (Baltimore, USA, Aug – Oct 2011 & June – Sep 2012).
Visiting trainee at the Fenyö lab, New York University (New York, USA, Oct – Dec 2014).

JULY 2009 – SEPTEMBER 2010

Member of the Belgian national rowing team. Raced at three world cup races and the European championships.

SEPTEMBER 2004 – JUNE 2009

Bachelor and Master's degree in bioscience engineering, major in biotechnology, at Ghent University (Ghent, Belgium).
International student at the University of Natural Resources and Applied Life Sciences (Vienna, Austria, fall semester of 2008).

SEPTEMBER 1998 – JUNE 2004

Royal Athenaeum Voskenslaan (Ghent, Belgium), studying science and mathematics.
Finalist at the Flemish mathematics Olympiad (2001).

PUBLICATIONS

Seremet T*, Koch A*, Jansen Y, Schreuer M, Wilgenhof S, Del Marmol V, Lièvard D, Thielemans K, De Meyer T, Van Criekinge W, Coulie P, Van Baren N, Neyns B. Identification of a predictive signature based on immunohistochemical, RNA-seq and epigenetic profiling of melanoma metastases for response to ipilimumab-based immunotherapy. (2015) – *in preparation*

Koch A, De Meyer T, Jeschke J, Van Criekinge W. MEXPRESS: Visualizing Expression, DNA Methylation and clinical TCGA Data. *BMC Genomics* **16**, 636 (2015)

Crappé J, Ndah E, Koch A, Steyaert S, Gawron D, De Keulenaer S, De Meester E, De Meyer T, Van Criekinge W, Van Damme P, Menschaert G. PROTEOFORMER: deep proteome coverage through ribosome profiling and MS integration. *Nucleic Acids Res* **43**, e29 (2015)

Koch A*, Gawron D*, Steyaert S, Ndah E, Crappé J, De Keulenaer S, De Meester E, Ma M, Shen B, Gevaert K, Van Criekinge W, Van Damme P, Menschaert G. A proteogenomics approach integrating proteomics and ribosome profiling increases the efficiency of protein identification and enables the discovery of alternative translation start sites. *Proteomics* **14**, 2688–2698 (2014)

Menschaert G, Van Criekinge W, Notelaers T, Koch A, Crappé J, Gevaert K, Van Damme P. Deep proteome coverage based on ribosome profiling aids mass spectrometry-based protein and peptide discovery and provides evidence of alternative translation products and near-cognate translation initiation events. *Mol Cell Proteomics* **12**, 1780–1790 (2013)

Wrangle J*, Wang W*, Koch A*, Easwaran H, Helai PM, Xiaoyu P, Vendetti F, Van Criekinge W, De Meyer T, Du Z, Parsana P, Rodgers K, Yen R, Zahnow CA, Taube JM, Brahmer JR, Tykodi SS, Easton K, Carvajal RD, Jones PA, Laird PW, Weisenberger DJ, Tsai S, Juergens RA, Topalian SL, Rudin CM, Brock MV, Pardoll D and Baylin SB. Alterations of immune response of non-small cell lung cancer with azacytidine. *Oncotarget* **4**, 2067–2079 (2013)

Jeschke J, Van Neste L, Glöckner SC, Dhir M, Calmon MF, Deregowski V, Van Criekinge W, Vlassenbroeck I, Koch A, Chan TA, Cope L, Hooker CM, Schuebel KE, Gabrielson E, Winterpacht A, Baylin SB, Herman JG, Ahuja N. Biomarkers for detection and prognosis of breast cancer identified by a functional hypermethylome screen. *Epigenetics* **7**, 701–709 (2012)

* these authors contributed equally

CONFERENCES

American Society of Clinical Oncology (ASCO): Annual Meeting. Chicago, USA (May 29 – June 2, 2015) – *P*

Big Data Science Symposium. Gent, Belgium (May 11, 2015) – *A*

The Society for Melanoma Research (SMR): 11th International Congress. Zurich, Switzerland (November 13 – 17, 2014) – *P*

Cancer Plan (Support Committee Action 29): Support to Translational Research. Brussels, Belgium (July 3, 2014) – *P*

European Melanoma Experience Exchange (euMEET): Focus on Therapy. Brussels, Belgium (December 11 – 12, 2013) – *T*

Mini Symposium: "Bridging the gap between two omics worlds: transcriptomics and proteomics". Gent, Belgium (November 29, 2013) – *A*

MRP Bioinformatics. From Nucleotides to Networks (N2N): 2013 Event. Zwijnaarde, Belgium (September 13, 2013) – *A*

61st ASMS Conference on Mass Spectrometry and Allied Topics. Minneapolis, USA (June 9 – 13, 2013) – *P*

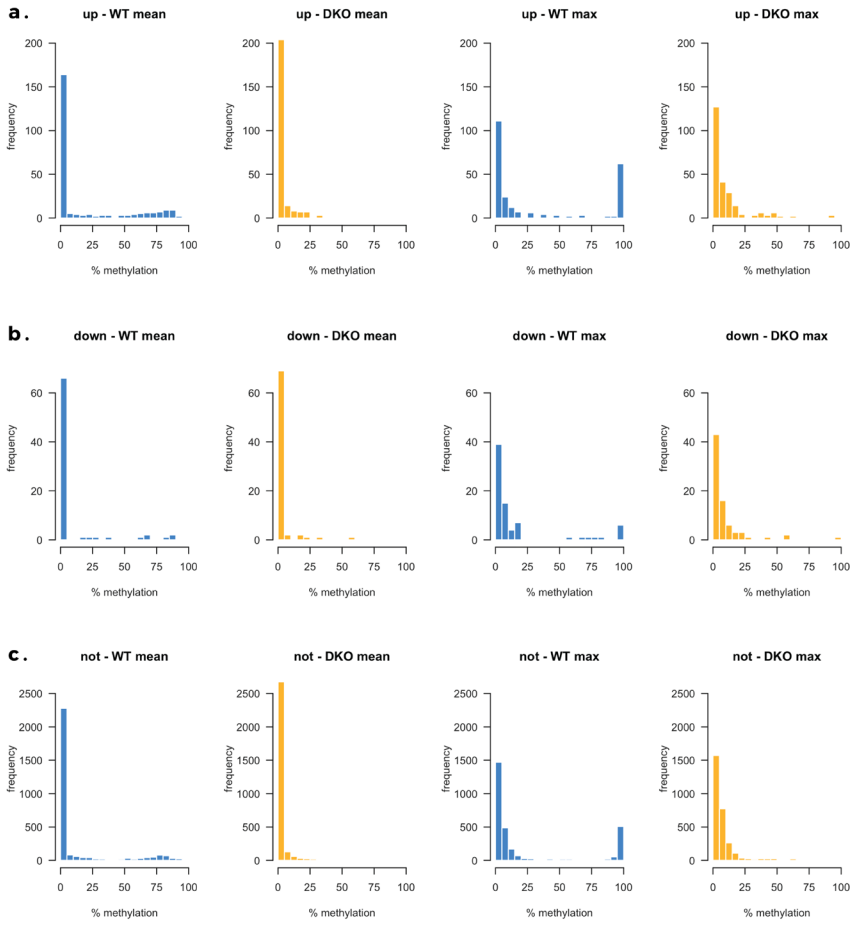
Belgium Proteomics Association Conference (BePAc). Gent, Belgium. (November 29 – 30, 2012) – *A*

MRP Bioinformatics. From Nucleotides to Networks (N2N): Kick-off Event. Zwijnaarde, Belgium (May 4, 2011) – *A*

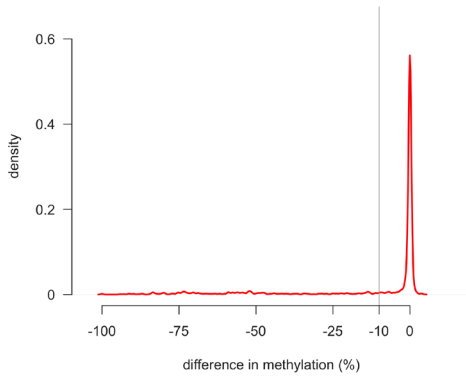
A – attending
P – poster presentation
T – talk

APPENDIX

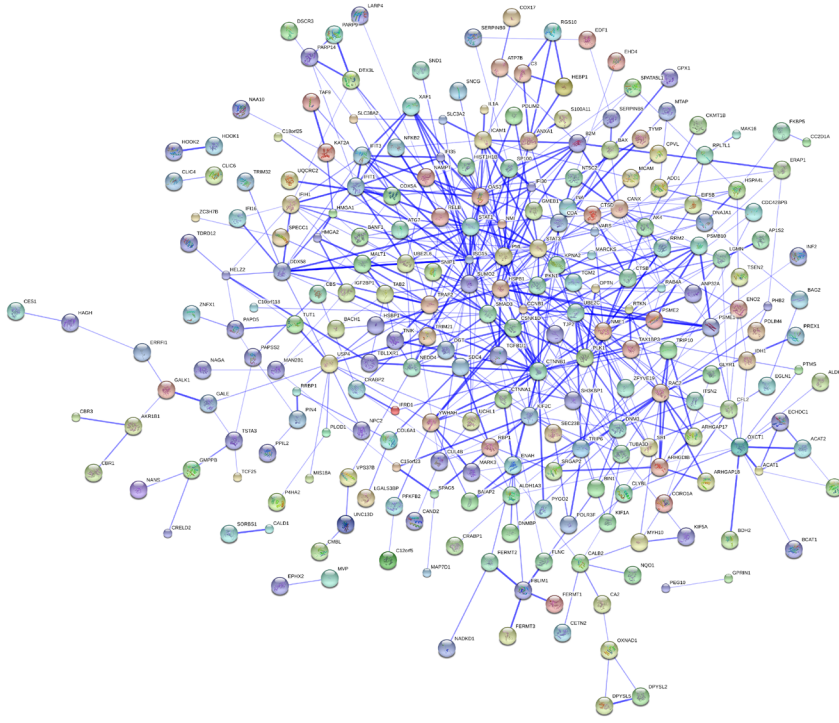
MEASURING THE GENOME-WIDE IMPACT OF DNA METHYLATION AT THE PROTEOME LEVEL IN A DNMT KNOCKOUT HUMAN CANCER CELL MODEL



Supplementary Figure 1.1 histograms of the WT (blue) and DKO (yellow) promoter methylation data for the up-regulated proteins (a), the down-regulated proteins (b) and the proteins that were not differentially expressed between DKO and WT (c). The two left columns show the histograms of the mean promoter methylation data, whereas the two columns on the right show the maximal promoter methylation value. When comparing the mean to the maximal values, we noticed a higher amount of noise for the latter and particularly for the up and down-regulated proteins. Based on these histograms, we decided to use the mean promoter methylation values and not the maximal values in the subsequent analyses.



Supplementary Figure 1.2 Distribution plot of the differences in promoter methylation between DKO and WT. We calculated the difference in promoter methylation for all the proteins we identified in the shotgun experiment by subtracting the WT from the DKO promoter methylation. The peak around zero corresponds to promoters that were not demethylated and we decided to put our cutoff to call a promoter demethylated in DKO to the left of the base of this peak. Given that calculating the mean promoter methylation already reduced the greatest differences between WT and DKO, we kept the cutoff close to the peak.



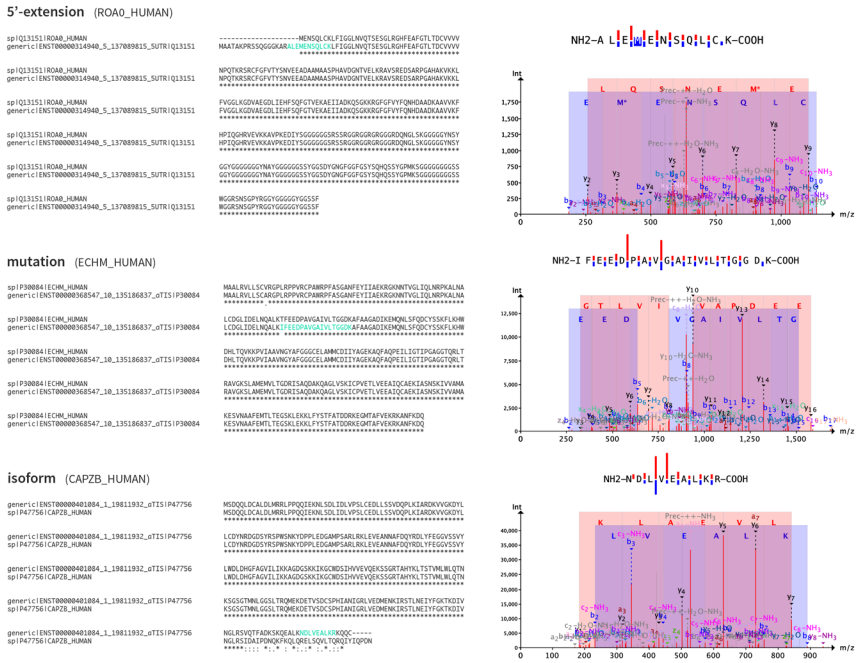
Supplementary Figure 1.3 STRING protein interaction network of the up and down-regulated proteins as deduced from the shotgun proteome analysis (http://string-db.org) (Franceschini *et al.*, 2013). The proteins without any connections are not shown in this plot.

Supplementary Table 1.1
Result of the gene ontology enrichment analysis using the DAVID tool on the list of genes that were found to be significantly up or down-regulated in the shotgun experiment.

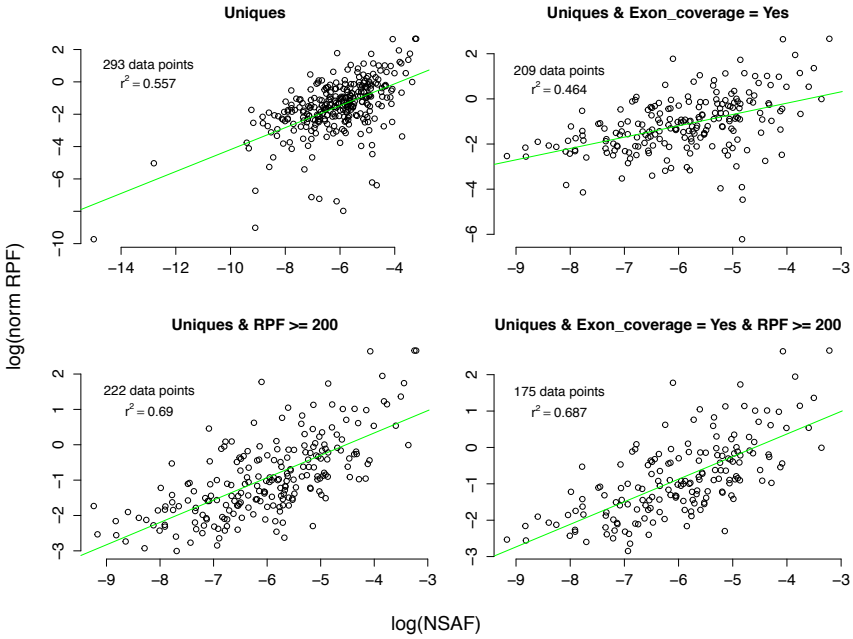
Category	Term	PValue	Genes	Fold Enrichment	FDR
GOTERM_BP_FAT	GO:0007010~cytoskeleton organization	1.86E-5	P31146, Q9UPN3, Q9NYT0, P35611, Q9BX66, Q15642, P52566, Q96ED9, Q96R06, Q5TZA2, P35222, P41208, Q16352, Q99661, P35580, Q8TCU6, Q05682, P15153, Q27J81, O00762, Q9UJC3, Q9Y5S2, Q96AC1	2.855	0.032
GOTERM_BP_FAT	GO:0006511~ubiquitin-dependent protein catabolic process	4.25E-5	P09936, Q9UL46, P40306, Q13107, Q06323, P46934, O95260, P14635, P28062, Q13620, O00762, Q9UK22, O43294, Q9BZK7, P28065, Q13049	3.578	0.073
GOTERM_BP_FAT	GO:0007017~microtubule-based process	7.03E-5	Q9UPN3, P09936, Q12756, Q96ED9, Q96R06, Q12840, O60282, Q5TZA2, P35222, P41208, Q99661, Q13748, O00762, Q9UJC3, P52292, Q9BW19	3.422	0.121
GOTERM_BP_FAT	GO:0044093~positive regulation of molecular function	8.17E-5	P42224, P40306, Q9NX02, P84022, Q9UDY8, P53350, P28062, P23497, P28065, Q13586, Q9UL46, P35611, P05362, Q06323, Q12933, O60825, P15531, P14635, P21980, Q05682, Q07812, O00762, O60869, P31431, Q16512, P29590	2.401	0.140
GOTERM_BP_FAT	GO:0019882~antigen processing and presentation	1.41E-4	P61769, Q01201, P28062, P01892, P05362, Q06323, Q9NZ08, P28065, P13284	5.868	0.242
GOTERM_BP_FAT	GO:0051437~positive regulation of ubiquitin-protein ligase activity during mitotic cell cycle	2.45E-4	P14635, Q9UL46, P53350, P28062, P40306, O00762, Q06323, P28065	6.366	0.420
GOTERM_BP_FAT	GO:0051443~positive regulation of ubiquitin-protein ligase activity	2.94E-4	P14635, Q9UL46, P53350, P28062, P40306, O00762, Q06323, P28065	6.184	0.504
GOTERM_BP_FAT	GO:0051439~regulation of ubiquitin-protein ligase activity during mitotic cell cycle	3.21E-4	P14635, Q9UL46, P53350, P28062, P40306, O00762, Q06323, P28065	6.097	0.550
GOTERM_BP_FAT	GO:0051351~positive regulation of ligase activity	3.81E-4	P14635, Q9UL46, P53350, P28062, P40306, O00762, Q06323, P28065	5.930	0.653
GOTERM_BP_FAT	GO:0031396~regulation of protein ubiquitination	5.10E-4	P14635, Q9UL46, P53350, P28062, P40306, Q9UK22, O00762, Q06323, P28065	4.870	0.872
GOTERM_BP_FAT	GO:0051098~regulation of binding	5.35E-4	Q05682, Q9GZT9, P35611, Q07812, O60869, P05362, P30533, P23497, P84022, Q9UDY8, P15531	3.890	0.915
GOTERM_BP_FAT	GO:0051438~regulation of ubiquitin-protein ligase activity	5.72E-4	P14635, Q9UL46, P53350, P28062, P40306, O00762, Q06323, P28065	5.550	0.978
GOTERM_BP_FAT	GO:0043161~proteasomal ubiquitin-dependent protein catabolic process	5.82E-4	P14635, Q9UL46, P28062, P40306, Q9UK22, O00762, Q06323, Q9BZK7, P28065	4.775	0.996
GOTERM_BP_FAT	GO:0010498~proteasomal protein catabolic process	5.82E-4	P14635, Q9UL46, P28062, P40306, Q9UK22, O00762, Q06323, Q9BZK7, P28065	4.775	0.996
GOTERM_BP_FAT	GO:0051340~regulation of ligase activity	7.18E-4	P14635, Q9UL46, P53350, P28062, P40306, O00762, Q06323, P28065	5.344	1.227
GOTERM_BP_FAT	GO:0031398~positive regulation of protein ubiquitination	8.93E-4	P14635, Q9UL46, P53350, P28062, P40306, O00762, Q06323, P28065	5.154	1.524
GOTERM_BP_FAT	GO:0048871~multicellular organismal homeostasis	9.58E-4	P31146, P35611, P15153, Q07812, P01583, P07203, P35222, P40763	5.093	1.634
GOTERM_BP_FAT	GO:0030036~actin cytoskeleton organization	9.95E-4	P35580, P31146, Q9NYT0, Q8TCU6, Q05682, P35611, Q27J81, P15153, Q9BX66, Q15642, P52566, Q9Y5S2, Q96AC1	3.113	1.697

DEEP PROTEOME COVERAGE BASED ON RIBOPROFILING

The supplementary tables for this paper can be found online: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4391000/



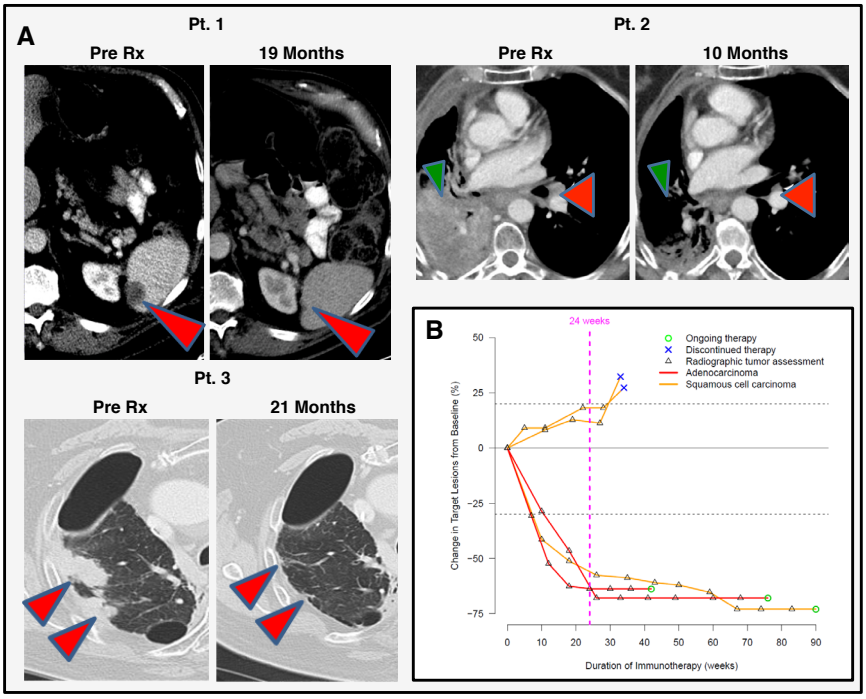
Supplementary Figure 2.1 Examples of improved identifications in the shotgun proteomics experiment. The addition of ribo-seq data to the proteomics experiment improved the identification and score significance for 69 proteins and three representative examples are depicted here. The left column shows the Clustal Omega alignment of the ribo-seq-derived amino acid sequences to the Swiss-Prot sequences with the relevant peptide identifications highlighted in cyan. The column on the right shows the corresponding fragmentation spectra and peptide sequence fragmentations.



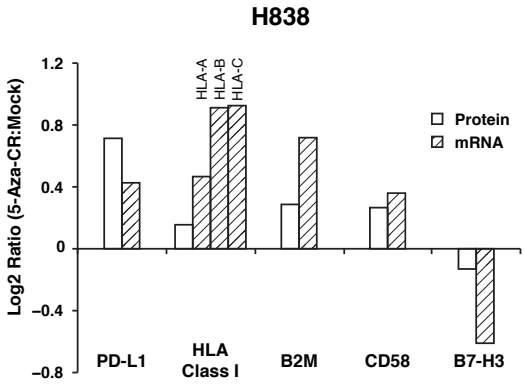
Supplementary Figure 2.2 Correlation plots of protein abundance based on NSAF values and RPF counts for the proteins uniquely identified in Swiss-Prot. Some transcripts were not contained in our custom database because the LTM treatment and/or TIS calling failed to identify these TISs. Correlations could still be calculated as the CHX treatment did result in detectable coverage for these transcripts. The number of data points used in every plot was lower than the total number of unique Swiss-Prot identifications (312), because whenever a Swiss-Prot protein corresponded to multiple transcripts only the transcript with the highest normalized RPF value was used. Top left: all transcripts; top right: transcripts with ribo-seq coverage in all exons; bottom left: all transcripts with an RPF count ≥ 200; bottom right: transcripts with both coverage in all exons and an RPF count ≥ 200. The regression line is shown in green. For each plot, the number of data points used (i.e. the number of dbTIS transcripts) as well as the corresponding Pearson correlation coefficient (r^2) is shown.

ALTERATIONS OF IMMUNE RESPONSE OF NON-SMALL CELL LUNG CANCER WITH AZACYTIDINE

The supplementary tables for this paper can be found online: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3875770/



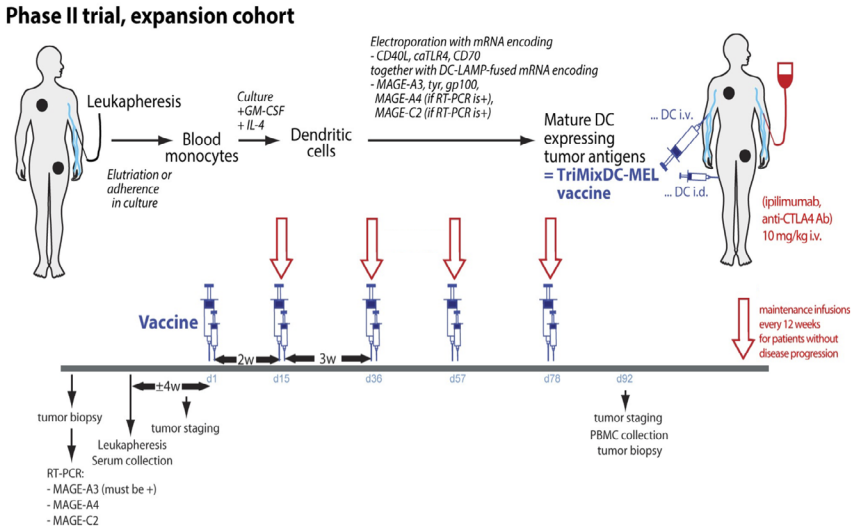
Supplementary Figure 3.1 Outcomes for five patients treated with immune checkpoint immunotherapy after epigenetic therapy. (A) Scans for 3 patients (Pt.) with RECIST criteria responses to either PD-1 or PD-L1 therapy. All scan interpretations were performed by a single radiologist and lesions used to measure tumor shrinkage between pre- and during immunotherapy at specified times are shown by red arrows (metastasis in the spleen- Pt.1; lung tumor lesions- Pt.2; lymph node in right central chest with metastases –Pt. 2. Green arrow denotes large area of the right lung collapsed behind airway obstruction by tumor and resolving by the 10 month period after immunotherapy. (B) Spider plot of sequential scan measurements (Y-axis) of lesions relative to time of treatment initiation with anti-PD-1 or anti-PD-L1 shown in panel (A) by weeks (X-axis) with a decrease of 30% qualifying as RECIST criteria response (green circles). Blue crosses indicate tumor increase of > 20% qualifying as disease progression. The 24 weeks point denoted by the dashed vertical line represents a duration of treatment after which disease stabilization is conventionally considered to represent clinical benefit.



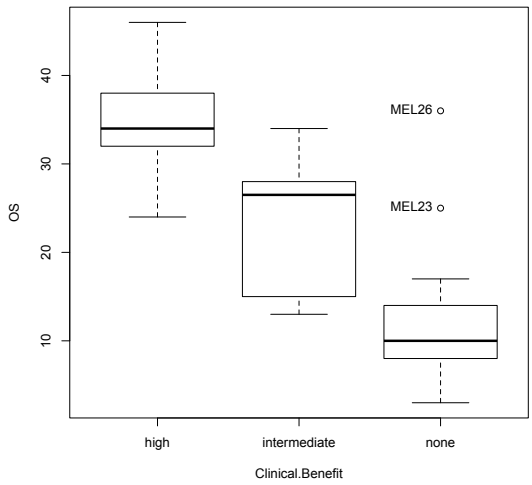
Supplementaray Figure 3.2 Comparison of expression array data to flow cytometry for select cell surface Protein proteins in H838. Clear bars represent the log2 ratio of mRNA mean fluorescence intensity of AZA over mock treated cells. Hashed bars represent the M-values of expression array (log2[AZA:Mock]). For HLA Class I, the antibody used in flow cytometry does not discriminate subtypes of class I molecules. Individual class I molecule subtype transcript data are available from the Agilent array platform and is presented. Changes between AZA treated and mock cells are calculated using mean fluorescence intensities (MFI) and the formula

$\log2([(MFI_{antibody,\ treated})-(MFI_{isotype,\ treated})]/\ (MFI_{antibody,\ mock})-(MFI_{isotype,\ mock})]$

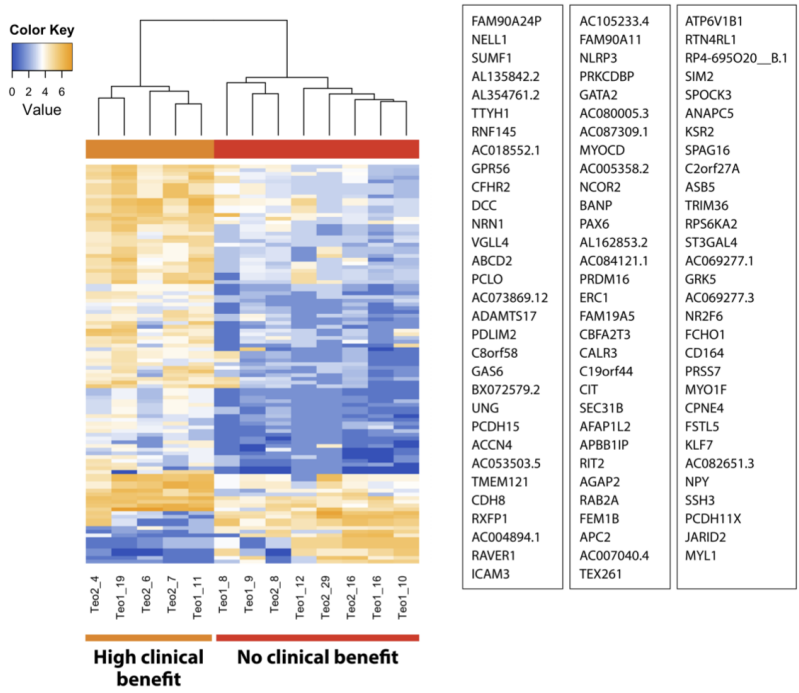
A PREDICTIVE SIGNATURE FOR RESPONSE TO IMMUNOTHERAPY IN MELANOMA METASTASES



Supplementary Figure 3.3 Study design for the TriMixIpi trial (http://clinicaltrials.gov NCT01302496)



Supplementary Figure 3.4 Comparison of the overal survival for the three response groups.



Supplementary Figure 3.5 Result of the differential DNA methylation analysis. 107 methylation cores were found to be differentially methylated between the HCB and NCB groups (edgeR analysis, FDR < 0.05). This corresponded to the list of 92 genes shown on the image.

Supplementary Table 3.1 Patient clinical benefit and characterization of melanoma metastases samples. (LN = lymph node)

Sample ID	Patient ID	Response BOR	OS mo	Obs	Median OS by groups mo	Biopsy time point	Type of immunotherapy	Pattern of T cell infiltration	CD20+ cells B cells	CD163+ cells Mo	Ki67+ cells Proliferation	Tissue type	Processed for RNA-Seq and MBD-Seq
High Clinical Benefit					34								
MEL1	PAT1	PR	46			Before	Ipi	B	yes	yes	no	LN	Yes
MEL2	PAT2	PR	32+	This sample is a regression lesion		After	TriMixIpi	E	yes	yes	yes	Sc nodule	Yes
MEL3	PAT3	CR	34+			Before	TriMixIpi	C	yes	yes	yes	LN	Yes
MEL4								C	yes	yes	yes	LN	No
MEL5	PAT4	CR	38+			Before	TriMixIpi	C	yes	yes	yes	LN	Yes
MEL6	PAT5	CR	24+			Before	TriMixIpi	B	no	yes	yes	Skin MTS	Yes
MEL7								C	no	yes	yes	LN	No
Intermediate Clinical Benefit					26,5								
MEL8	PAT6	SD	27+	PD at w36		Before	TriMixIpi	E	yes	yes	yes	Lung	Yes
MEL9	PAT7	PR	26+	New lesions treated by surgery and radiotherapy before w24, reached CR		Before	Ipi	C	no	yes	yes	Skin MTS	Yes
MEL10	PAT8	SD	13	PD at w20		Before	Ipi	C	yes	yes	no	Sc nodule	Yes
MEL11	PAT9	PR	15	PD at w24		After	TriMixDC	B	yes	yes	no	Skin MTS	Yes
MEL12	PAT10	SD	34	PD at w24		After	TriMixIpi	C	no	yes	yes	Sc nodule	No
MEL27	PAT25	SD	28	PD at w24		After	Ipi	NA	NA	NA	NA	Liver	Yes
No Clinical Benefit					10								
MEL13	PAT11	PD	17	First line was TriMixDC but MTS was removed after receiving also Ipi		After	TriMixDC	E	no	yes	yes	Skin MTS	Yes
MEL14	PAT12	PD	13			After	TriMixIpi	C	no	yes	yes	Small intestine	Yes
MEL15	PAT13	PD	3			Before	TriMixIpi	B	yes	yes	yes	LN	Yes
MEL16	PAT14	PD	13			After	TriMixIpi	C	yes	yes	yes	LN	Yes
MEL17	PAT15	PD	14			After	TriMixIpi	A/B	no	yes	yes	Brain	Yes
MEL18	PAT16	PD	8	First line was Ipi but MTS was removed after receiving also TriMixIpi		After	Ipi	C	no	yes	yes	Liver	Yes
MEL19	PAT17	PD	6			Before	Ipi	NE	NE	yes	NE	Adrenal gland	Yes
MEL20	PAT18	PD	9			Before	TriMixIpi	B	no	yes	no	Sc nodule	Yes
MEL21	PAT19	PD	5			Before	Ipi	Heterogenous A and C	no	yes	yes	Small intestine	No
MEL22	PAT20	PD	10			Before	Ipi	Heterogenous E and C	no	yes	yes	Lung	No
MEL23	PAT21	PD	25	First line was Ipi but MTS was removed after receiving Ipi, TriMixIpi and Ipi reinduction		After	Ipi	D	no	yes	yes	Sc nodule	No
MEL24	PAT22	PD	10			After	TriMixIpi	E but weak infiltration	no	yes	yes	Sc nodule	No
MEL25	PAT23	PD	10			Before	TriMixIpi	NE	NE	NE	NE	LN	No
MEL26	PAT24	PD	36	Experienced DC vaccination, Ipi, Ipi reinduction, DC vaccination combined with Ipi, and TIL therapy		After	Ipi	C	yes	yes	yes	Sc nodule	No

Supplementary Table 3.2 Result of the gene ontology enrichment analysis of the genes that were differentially expressed between HCB and NCB patients. The analysis was performed using the online GOrilla tool. Gene ontologies that are linked to the immune system are marked in blue.

Biological Processes

Description	FDR q-value
immune system process	7.08E-7
humoral immune response	5.01E-7
immune response	4.5E-6
regulation of immune system process	2.38E-5
signal transduction	3.48E-4
defense response	3.01E-3
positive regulation of immune system process	2.75E-3
cell surface receptor signaling pathway	4.35E-3
leukocyte migration	4.01E-3
cytokine-mediated signaling pathway	4.45E-3
ventricular septum morphogenesis	4.54E-3
positive regulation of lamellipodium organization	6.17E-3
leukocyte chemotaxis	6.02E-3
response to other organism	6.03E-3
positive regulation of homotypic cell-cell adhesion	6.37E-3
regulation of lymphocyte activation	8.26E-3
negative regulation of actin nucleation	9.87E-3
cell chemotaxis	1.11E-2
antibacterial humoral response	1.19E-2
lymphocyte activation	1.15E-2
regulation of leukocyte activation	1.12E-2
positive regulation of cell adhesion	1.25E-2
positive regulation of cell-cell adhesion	1.41E-2
regulation of lamellipodium organization	1.46E-2
leukocyte activation	1.48E-2
positive regulation of leukocyte cell-cell adhesion	1.47E-2
antimicrobial humoral response	1.45E-2
dendritic cell chemotaxis	1.62E-2
cell activation	1.71E-2
positive regulation of response to stimulus	1.74E-2
response to stimulus	1.71E-2
leukocyte cell-cell adhesion	1.69E-2
cardiac septum morphogenesis	1.72E-2
positive regulation of immune response	1.73E-2
defense response to other organism	1.75E-2
cell motility	1.94E-2
cell migration	1.92E-2
regulation of cell activation	2.04E-2
leukocyte aggregation	2.21E-2
regulation of immune response	2.67E-2
positive regulation of T cell activation	2.84E-2
lymphocyte aggregation	3.19E-2
immune response-activating cell surface receptor signaling pathway	3.26E-2
dendritic cell migration	3.36E-2
canonical Wnt signaling pathway	3.98E-2
B cell receptor signaling pathway	3.95E-2
midbrain-hindbrain boundary maturation during brain development	3.89E-2
central nervous system morphogenesis	3.81E-2
cerebellum formation	3.73E-2
cellular calcium ion homeostasis	4.1E-2
cellular divalent inorganic cation homeostasis	4.32E-2
regulation of antigen processing and presentation	4.25E-2
locomotion	4.27E-2
tissue morphogenesis	4.22E-2
immunological synapse formation	4.19E-2
regulation of biological process	4.24E-2
negative regulation of cellular component movement	4.3E-2
regulation of homotypic cell-cell adhesion	4.44E-2
positive regulation of cytosolic calcium ion concentration	4.46E-2
biological regulation	4.53E-2
immune response-activating signal transduction	4.76E-2
T cell aggregation	4.72E-2
T cell activation	4.65E-2
regulation of response to stimulus	4.75E-2
response to external biotic stimulus	4.7E-2

Cellular Component

Description	FDR q-value
external side of plasma membrane	6.42E-6
side of membrane	8.23E-5
immunoglobulin complex	5.13E-4
intrinsic component of plasma membrane	7.13E-4
immunoglobulin complex, circulating	1.67E-3
integral component of plasma membrane	1.79E-3
T cell receptor complex	2.09E-3
plasma membrane part	1.91E-3
IgM immunoglobulin complex, circulating	4.38E-3
IgM immunoglobulin complex	3.94E-3
pentameric IgM immunoglobulin complex	3.58E-3
secretory IgA immunoglobulin complex	1.9E-2
IgA immunoglobulin complex	1.75E-2
IgA immunoglobulin complex, circulating	1.63E-2
monomeric IgA immunoglobulin complex	1.52E-2
polymeric IgA immunoglobulin complex	1.42E-2
receptor complex	2.06E-2
blood microparticle	2.78E-2
extracellular space	3.24E-2
intrinsic component of membrane	3.27E-2
alpha-beta T cell receptor complex	3.64E-2
plasma membrane	4.06E-2
membrane part	4.5E-2

Molecular Function

Description	FDR q-value
transmembrane signaling receptor activity	2.39E-4
signaling receptor activity	1.27E-3
receptor activity	1.26E-3
signal transducer activity	1.9E-3
molecular transducer activity	1.52E-3
antigen binding	1.78E-3
G-protein coupled receptor activity	1.79E-3
immunoglobulin receptor binding	2.35E-2
frizzled binding	3.82E-2
receptor binding	4.38E-2
cytokine receptor activity	4.44E-2

Supplementary Table 3.3 Melanoma metastases immune infiltration analyses in the three clinical benefit groups.

A

		Clinical Benefit							
		high		intermediate		none		Total	
		Count	Column Valid N %	Count	Column Valid N %	Count	Column Valid N %	Count	Column Valid N %
Pattern of T cell infiltration	A	0	0,0%	0	0,0%	1	8,3%	1	4,2%
	B	2	28,6%	1	20,0%	2	16,7%	5	20,8%
	C	4	57,1%	3	60,0%	4	33,3%	11	45,8%
	D	0	0,0%	0	0,0%	1	8,3%	1	4,2%
	E	1	14,3%	1	20,0%	2	16,7%	4	16,7%
	Heterogenous	0	0,0%	0	0,0%	2	16,7%	2	8,3%
Pattern of T cell infiltration grouped	1	2	28,6%	1	20,0%	3	25,0%	6	25,0%
	2	5	71,4%	4	80,0%	7	58,3%	16	66,7%
	Heterogenous	0	0,0%	0	0,0%	2	**16,7%**	2	8,3%
CD20+ cells	no	2	**28,6%**	2	40,0%	9	**75,0%**	13	54,2%
	yes	5	**71,4%**	3	60,0%	3	**25,0%**	11	45,8%
CD163+ cells	yes	7	100,0%	5	100,0%	13	100,0%	25	100,0%
Ki67+ cells	no	1	14,3%	2	40,0%	1	8,3%	4	16,7%
	yes	6	85,7%	3	60,0%	11	91,7%	20	83,3%

B Crosstab for the 3 clinical benefit groups

	Fisher's Exact Test	Value	p-value {Exact Sig. (2-sided)}
		1,896	,950
Pattern of T cell infiltration grouped			
CD20+ cells		4,228	,128
Ki67+ cells		2,486	,285

Supplementary Table 3.4 Pattern of T cell infiltration in the NCB group in samples removed before or after therapy onset.

NO CLINICAL BENEFIT			Pattern of T cell infiltration			
			1	2	Heterogen.	Total
Biopsy time point	After	Count	1	7	0	8
		% within Pattern of T cell infiltration grouped	33,3%	100,0%	0,0%	66,7%
	Before	Count	2	0	2	4
		% within Pattern of T cell infiltration grouped	66,7%	0,0%	100,0%	33,3%
Total		Count	3	7	2	12
		% within Pattern of T cell infiltration grouped	100,0%	100,0%	100,0%	100,0%
Fisher's Exact Test		p-value 0,01				

Supplementary Table 3.5 Automated image analysis and quantification of CD8+ and PD-L1+ cells. The median with percentile 25 and percentile 75 for overall CD8+ cells and PD-L1+ cells infiltration are shown for the three clinical benefit group. The same parameters are shown for the pathologist assessment of the PD-L1 staining in the tumor compartment as well as in the immune compartment.

		Clinical Benefit			
		high	intermediate	none	Total
DEFINIENS CD8 % marker area Overall	Count	7	6	14	27
	Valid N	6	3	11	20
	Median	**4,3**	**3,6**	**2,4**	3,3
	Percentile 25	2,7	,2	,3	,4
	Percentile 75	9,5	11,0	4,9	5,1
DEFINIENS PDL1 % marker area Overall	Count	7	6	14	27
	Valid N	6	3	7	16
	Median	**11,1**	**1,3**	**6,8**	5,1
	Percentile 25	1,7	,2	,1	,4
	Percentile 75	20,1	4,4	9,8	11,9
PATHOLOGIST % PDL1 Tumor Component	Count	7	6	14	27
	Valid N	6	4	13	23
	Median	**17,5**	**1,0**	**0,0**	2,0
	Percentile 25	0,0	0,0	0,0	0,0
	Percentile 75	20,0	11,0	5,0	20,0
PATHOLOGIST % PDL1 Immune Component	Count	7	6	14	27
	Valid N	6	4	13	23
	Median	**4,5**	**2,5**	**1,0**	2,0
	Percentile 25	1,0	1,0	0,0	0,0
	Percentile 75	5,0	6,5	5,0	5,0

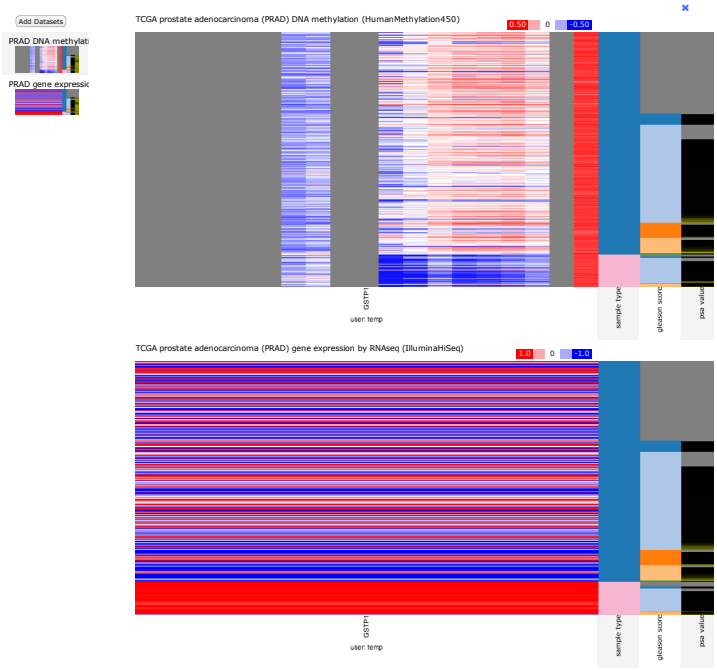
Supplementary Table 3.6 – *part 1*
Result of the gene ontology enrichment analysis of the genes that were differentially methylated between HCB and NCB patients. The analysis was performed using the online GOrilla tool. Note the numerous neuron-related ontologies

GO PROCESS

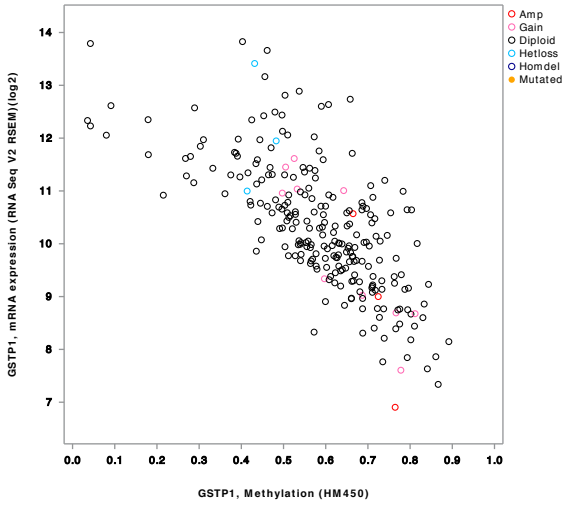
Description	*FDR q-value*	*Description*	*FDR q-value*
developmental process	2.49E-9	regulation of cell projection organization	3.09E-5
single-organism developmental process	4.7E-9	positive regulation of multicellular organismal process	3.2E-5
neuron projection guidance	3.95E-9	positive regulation of developmental process	3.24E-5
axon guidance	2.96E-9	regulation of cell morphogenesis involved in differentiation	4.4E-5
anatomical structure development	4.24E-9	cellular developmental process	6.35E-5
multicellular organismal process	1.02E-8	cell-cell signaling	8.16E-5
regulation of multicellular organismal process	8.78E-9	biological regulation	8.55E-5
single-multicellular organism process	1.04E-8	positive regulation of molecular function	8.46E-5
regulation of neurogenesis	2.16E-8	single organism signaling	9.05E-5
regulation of developmental process	3.69E-8	system development	9.23E-5
regulation of neuron differentiation	5.15E-8	signaling	9.27E-5
regulation of nervous system development	9.21E-8	positive regulation of cell differentiation	9.92E-5
synaptic transmission	1.47E-7	positive regulation of biological process	9.91E-5
single-organism process	2.06E-7	positive regulation of cell development	1.2E-4
regulation of multicellular organismal development	2.25E-7	cell differentiation	1.31E-4
regulation of cell differentiation	2.52E-7	pattern specification process	1.46E-4
movement of cell or subcellular component	3.37E-7	cell development	1.45E-4
regulation of neuron projection development	4.32E-7	positive regulation of cellular process	1.56E-4
regulation of cell development	5.16E-7	regulation of biological process	1.97E-4
single-organism cellular process	1.06E-6	positive regulation of neuron projection development	2.15E-4
neuron differentiation	3.06E-6	neurological system process	3.17E-4
positive regulation of neurogenesis	5.69E-6	anatomical structure morphogenesis	3.6E-4
organ development	5.56E-6	regulation of cellular process	4.26E-4
positive regulation of neuron differentiation	1.21E-5	regulation of dendrite development	5.03E-4
regulation of anatomical structure morphogenesis	1.19E-5	positive regulation of metabolic process	6.41E-4
positive regulation of nervous system development	1.2E-5	cell communication	7.93E-4
cell adhesion	1.87E-5	regulation of small GTPase mediated signal transduction	8.32E-4
biological adhesion	2.5E-5	inorganic cation transmembrane transport	8.48E-4

Supplementary Table 3.6 – *part 2*
Result of the gene ontology enrichment analysis of the genes that were differentially methylated between HCB and NCB patients. The analysis was performed using the online GOrilla tool. Note the numerous neuron-related ontologies

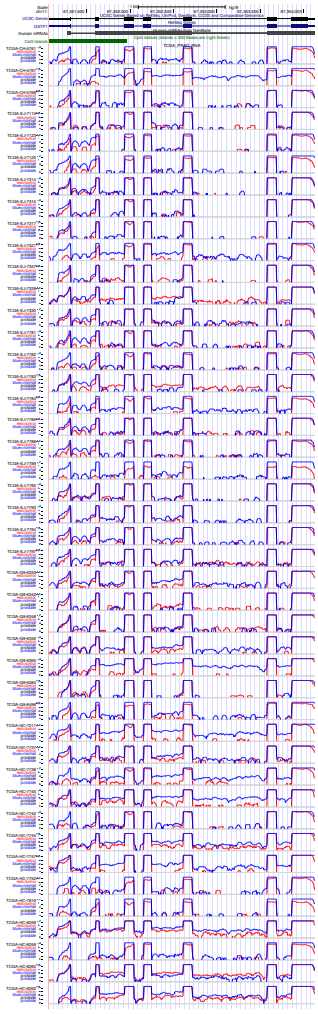
GO FUNCTION		GO COMPONENT	
Description	***FDR q-value***	***Description***	***FDR q-value***
cytoskeletal protein binding	2.55E-4	neuron part	4.59E-11
cation channel activity	2.7E-4	synaptic membrane	2.18E-9
sequence-specific DNA binding	2.18E-4	postsynaptic membrane	3.68E-9
actin binding	2.79E-4	neuron projection	1.1E-8
carbohydrate derivative binding	3.63E-4	synapse part	3.92E-8
gated channel activity	3.47E-4	cell junction	3.29E-8
motor activity	5.97E-4	postsynaptic density	1.63E-7
ion channel activity	5.56E-4	cell projection	1.91E-7
substrate-specific channel activity	6.95E-4	axon part	1.46E-5
sequence-specific DNA binding RNA polymerase II transcription factor activity	6.63E-4	plasma membrane part	3.17E-5
ion binding	6.31E-4	synapse	1.87E-4
calcium ion transmembrane transporter activity	7.26E-4	plasma membrane	2.09E-4
metal ion transmembrane transporter activity	8.24E-4	ion channel complex	2.35E-4
		cation channel complex	3.22E-4
		transmembrane transporter complex	7.8E-4
		receptor complex	8.45E-4



Supplementary Figure 3.6 UCSC Cancer Genome Browser (CGB) visualization of the GSTP1 methylation, expression and clinical TCGA data in prostate adenocarcinoma as shown in Figure 3.8. Like MEXPRESS, the CGB allows the samples to be ordered by a clinical parameter, showing differences in methylation and expression between for example the normal and tumor samples. Unlike MEXPRESS however, it is not possible to rank the samples by their expression values or to integrate the expression and methylation data.



Supplementary Figure 3.7 cBioPortal visualization of the correlation between the TCGA expression and methylation data for GSTP1 in prostate adenocarcinoma. Using the cBioPortal tool the correlation between the expression and methylation data for a gene can be visualized, though only for one probe. It is not possible to integrate the expression and methylation data with clinical parameters or to compare the methylation data to the genomic location of the probes as shown in Figure 3.8.



Supplementary Figure 3.8 A Cancer Genome Workbench (CGWB) view of the TCGA expression data for GSTP1 in prostate adenocarcinoma. The CGWB is based on the UCSC genome browser and allows a user to plot the expression data for a (limited) number of samples. It offers a more detailed profile of the expression data as compared to the per-gene aggregated expression value shown in MEXPRESS, but cannot integrate the expression profiles with methylation and clinical data.

Supplementary Figure 3.9 Integrative Genomics Viewer (IGV) visualization of the GSTP1 expression and methylation TCGA data in glioblastoma multiforme. The IGV only offers TCGA expression and methylation data for glioblastoma multiforme and ovarian serous cystadenocarcinoma, so no direct comparison could be made to the visualization of the GSTP1 data in prostate adenocarcinoma as shown in Figure 3.8. Instead, the GSTP1 (a) expression and (b) methylation data is plotted for glioblastoma. The IGV does not offer a direct comparison of the expression and methylation data and does not integrate these datasets with the clinical parameters available in TCGA.