Differential DNA methylation profiles modulating phenotypes. Regions of tissue-specific DNA methylation and their relation to gene expression, their evolutionary conservation and their application as molecular biomarkers

Dissertation

zur

Erlangung des Doktorgrades (Dr. rer. nat)

der

Mathematisch-Naturwissenchaftlichen Fakultät

der

Rheinischen Friedrich-Wilheims-Universität Bonn

vorgelegt von

Rene Gabriel Cortese

aus

Mercedes (BA) Argentinien

Bonn 2007

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Rheinischen Friedrich-Wilhelms-Universität Bonn

Wissenchaftliche Betreeung

Dr. Florian Eckhardt

Brahms AG
Neuendorfs. 25
D-16761 Hennigsdorf

1. Referent:

Prof. Dr. Klaus Olek

Labor für Abstammungsbegutachtungen
Marie-Curie-Straße 1
D-53359 Rheinbach

2. Referent:

Prof. Dr. Karl-Heinz Scheidtmann

Institut für Genetik
Universität Bonn
Abt. Molekulargenetik
Römerstr. 164
D-53117 Bonn

Tag der mündlichen Prüfung: 17.03.2008

Diese Dissertation ist auf dem Hochschulsschriftenserver der ULB Bonn http://hss.ulb.uni-bonn.de/diss_online elektronish publiziert.

Erscheinungsjahr: 2008

Acknowledgments

This work was performed under the direction of Prof. Dr. Klaus Olek (University of Bonn) and the supervision of Dr. Florian Eckhardt at Epigenomics AG, Berlin.
The practical work was carried out at Epigenomics AG, which also founded the vast majority of the experimental work.
Some sections of this thesis were partially founded by the European Union: Conservation of DNA methylation patterns in human/mouse orthologues (HEROIC program) and Conservation of DNA methylation upon gene duplications (MolPAGE program).

After so many years and so much interaction with so many people, it is quite difficult to thank all without being unfair and forgetting someone. However, I would like to acknowledge very special people, who contributed to make this possible.
In first place, I would like to thank Prof. Dr. Klaus Olek for accepting and directing the work and his constant support. Furthermore, I am very grateful to my supervisor at Epigenomics, Dr. Florian Eckhardt, for supporting, guiding and encouraging me in all my projects.
Moreover, I would like to thank the whole staff of Epigenomics AG (actual and former directors, managers and colleagues) for sharing this time with me, for their support and valuable discussions. Very especially, I would like to thank Dr. Alexander Olek for giving me the chance to work in the Company and for introducing me in the fascinating world of Epigenetics. In addition, I would like to thank Dr. Ralf Lesche for guiding my first steps.
I would like to express as well my gratitude to Dr. Thomas Brune (University of Magdeburg) and to the Human Epigenome Project team, especially to Dr. Stephan Beck (Cambridge University, UK), for the very exciting collaborations in the Laminopathies and HEP projects, respectively.

Becoming more emotional, I would like to thank very especially my family (Mamá, Papá, Karina, Rodrigo, Ramiro, Gabriel, Cecilia, Malena and Tobias) and friends in

Argentina (I should mention everyone, but it is hard due to the number and quality of friends I have been blessed, I hope you understand) for being so close, no matter the distances. Without them and their unconditional support, nothing of this dream had been possible.
Finally, I would like to thank my very good friends in Germany, Hong, Marco and Manuel, and my pretty nice "Family in Europe" (Carolina, Joaquin and Vicky) for their tireless support and being always at my side all along the road.

I dedicate this work to my family and friends with all my heart.
Thanks a lot, vielen Dank, muchas gracias, to all of them...

Contents

Abstract

Tissue- specific DNA methylation plays a major regulatory role. The association of increased tissue-specific DNA methylation and gene silencing is widely accepted despite the lack of experimental evidence. I have approached this topic in two complementary directions. First, using direct sequencing of bisulfite-converted DNA and semiquantitative RT-PCR, I show that increased methylation correlates with decreased expression of the cognate transcript in healthy adult tissues and primary cells in 16 out of 43 (39%) tissue-specific differentially methylated regions (T-DMRs) located at the 5'-UTR of annotated genes. In the second approach, I have first determined the genome-wide expression profiles of fetal and healthy adult human lung and then studied the methylation status of 43 differentially expressed genes. Among them, highly methylated 5'-UTR regions correlate with decreased expression levels in 19% of the cases. Furthermore, this approach allowed the discovery of four differentially expressed genes in fetal lung (*MEOX2*, *MDK*, *LAPTM5* and *FGFR3*), whose differential methylation arise as potential biomarker for lung cancer.

Next, I have studied the conservation of T-DMRs throughout the evolution. First, using direct sequencing of bisulfite-converted DNA, I show that the majority (69.4%) out of 61 orthologous regions in human and mouse differ by less than 20% in their methylation, indicating significant conservation. Additionally, by comparative DNA methylation analysis of three different gene duplication events leading to functional gene families, unprocessed and processed pseudogenes, I show that for genes that evolved recently and for some unprocessed pseudogenes, tissue-specific DNA methylation and RNA expression are conserved upon duplication.

Finally, I have studied the interaction of genetic and epigenetic alterations in modulating phenotypes in non-oncogenic diseases. The *RARB* gene appears highly methylated in patients with familial partial lipodystrophy (FPLD) compared to progeria patients and healthy controls. This differential methylation might explain the different phenotypes observed despite similar genetic backgrounds.

Altogether, this work presents vast experimental evidence supporting the hypothesis of differential DNA methylation patterns influencing, and in may cases determining, phenotypes in healthy and diseased tissues.

1. Introduction

1.1. Generalities about epigenetic and DNA methylation

1.1.1. The concept of epigenetics

Epigenetics refers to the study of heritable changes in genome function that occur without a change in DNA sequence. From a functional point of view, research on this field approaches issues such how patterns of gene expression are passed from one cell to its descendants, how gene expression changes during cell differentiation, and how environmental factors can change the way genes are expressed.
Epigenetics studies focus on three related molecular mechanisms: DNA methylation, histone code changes and RNA interference. Among them, DNA methylation represents the most stable epigenetic modification. DNA methylation marks are copied after DNA synthesis resulting in heritable changes in chromatin structure [1].

1.1.2. DNA methylation. Principles and definitions

DNA methylation is the biochemical addition of a methyl group (CH_3) to a nucleotide molecule. In mammalian DNA, this addition occurs predominantly to cytosines, especially in the context of a cytosine-guanosine (CpG) dinucleotides. In healthy cells, the modified methyl cytosine (mC) is present at a 2%-5% level of all cytosines [2]. CpG sites are present very significantly less than expected (5-10 fold) from the overall base composition of the DNA and unevenly distributed throughout the genome. While the vast majority of the genome is CpG poor, about 1% consists of CpG rich areas, typically related to the transcription start sites (TSS) of the genes. These CpG regions are referred as "CpG islands" and are mainly unmethylated when located nearby the TSS of expressed genes, in clear contrast to the mainly, but not exclusively, hypermethylated rest of the genome [3].
As the annotation of the human genome was completed, a clear definition of the criteria for naming a CpG dense region as "CpG island" became indispensable. The generally accepted definition of what constitutes a CpG island was proposed in 1987

by Gardiner-Garden and Frommer as being a 200-bp stretch of DNA with a C+G content of 50% and an observed/expected CpG ratio of 0.6 [4]. However, different authors and annotation databases may add their own small variations to this definition. For example, current version of the Ensembl database (release 42, as February 2007) CpG island definition requires a minimum length of 1000 bp, a minimal observed/expected ratio of 0.6 and a minimal GC content of 50%. In the present work, I took this definition and only CpG islands annotated in Ensembl database are named as such. All other CpG dense regions not fulfilling these requirements are called "CpG rich regions" accordingly.

1.1.3. DNA methyltransferases

Various methyltransferase enzymes (DNMT) catalyze cytosine methylation. In eukaryotes, there are three DNMT families (1, 2 and 3). DNMT1 and DNMT3 proteins comprise two domains: the N-terminal "regulatory" domain and C-terminal "catalytic" domain; in contrast DNMT2 has only the "catalytic" domain. Four DNMTs have been identified so far: DNMT1, DNMT2, DNMT3A and DNMT3B. In addition, DNMT3L has been found as a stimulator of the DNMT3A and DNMT3B enzymes, reviewed in [5].

DNMT1 enzyme showed a 5-10 fold preference for hemimethylated DNA [6]. This property suggests that DNMT1 reproduces parental DNA methylation pattern into the daughter DNA strand, therefore this enzyme is referred as responsible for "maintenance methylation". Functional experiments in mammalian cells also assigned a role for DNMT1 in the overall genome methylation, X- chromosome inactivation, allelic silencing of imprinted genes and silencing of incorporated transposons, reviewed in [7]. Function and roles for DNMT2 are not completely elucidated. Okano and colleagues showed that Dnmt2 is not essential for global *de novo* or maintenance methylation of DNA in embryonic stem (ES) cells [8]. In addition, purified mouse Dnmt2 protein has weak DNA methyltransferase activity and probably related to a loose ttnCGga(g/a) consensus sequence, suggesting that DNMT2 has a more specialized role than other mammalian DNA methyltransferases [9]. DNMT3a and DNMT3b are generally regarded as *de novo* DNA

methyltransferases. During early development (when most of the *de novo* methylation occurs), both enzymes are highly expressed and essential for the establishment of new methylation patterns and subsequent proper development [10]. All methyltransferases share the catalytic mechanism. The AdoMet cofactor (S-adenosylmethionine) contributes the methyl groups to the reaction [11]. The reaction starts when cytosine is flipped out of DNA and inserted into the binding pocket of a methyltransferase enzyme (Figure 1.1.1). In the active site, the catalytic cysteine thiolate forms a transition state intermediate with the carbon-6 of the cytosine's ring. This creates a reactive 4–5 enamine, which attacks the methyl group provided by the cofactor. After the methyl group transfer to the C-5 position of the cytosine ring, proton abstraction from this position causes reformation of the 5,6 double bond and the release of the enzyme by β-elimination [12], [13].

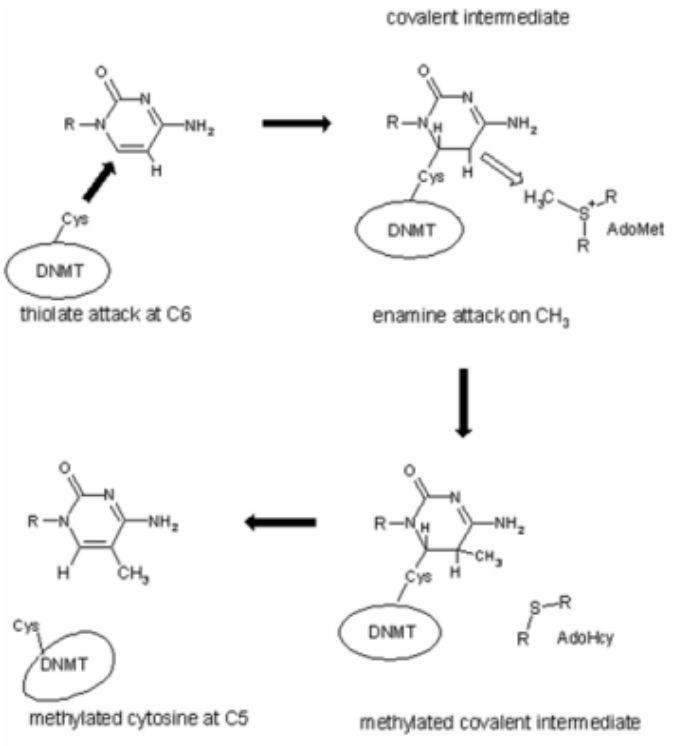


Figure 1.1.1: **Reaction catalyzed by DNA methyltransferases.** The reaction starts with a thiolate attack at position C6 and subsequent formation of a covalent intermediate. This intermediate will gain the CH_3 group at the C5 position from the donor, AdoMet, by an enamine attack. Finally enzyme is released by ß-elimination. Reproduced from Siedlecki and Zielenkiewicz [7]

1.2. Epigenetic models of gene expression regulation

Several models describe the way DNA methylation directs gene silencing. The simplest model states that methylation at particular CpGs within the promoter region of a gene avoids the binding of transcription factor and further assembly of the transcription machinery [14] (Figure 1.2.1).

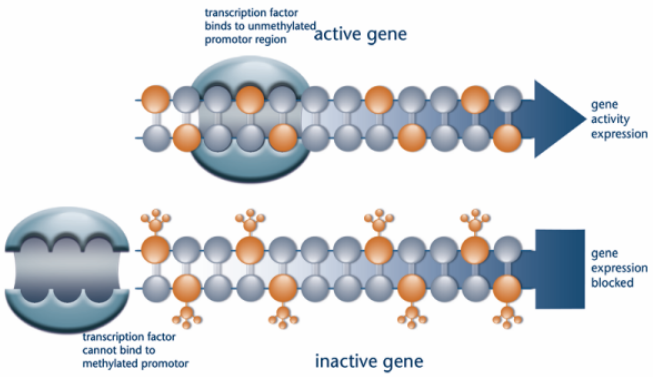


Figure 1.2.1: DNA methylation blocks transcription by avoiding the binding of transcription factors. Schematic representation showing the simplest model of transcriptional regulation by DNA methylation. Cytosines are depicted as orange circles, while all other nucleotides are blue. Transcription factors can bind to unmethylated DNA at the promoter region promoting the expression of the cognate gene (Upper graphic). Methyl groups (represented as lollipops at the corresponding cytosine position) avoid the binding of transcription factors and the gene expression is blocked (Lower graphic).

Although this model may explain gene silencing in several genes where target sequences contain single CpG position that become methylated [15], [16], [17], [18], more complex molecular interactions drive this process. In the vast majority, proteins containing domains that bind methylated CpGs appear to recognize methyl cytosines enabling interactions with protein complexes that drive transcription and chromatin packing. Currently, five methyl-CpG binding proteins are recognized in mammals, reviewed in [19]. Four of these proteins, MeCP2, MBD1, MBD2 and MBD4 bind methylated CpGs through a conserved protein motif called the methyl-CpG binding domain (MBD). Unlike, Kaiso binds methylated DNA through a zinc finger motif. Four of the methyl-CpG binding proteins, MBD1, MBD2, MeCP2 and Kaiso, displayed transcription repression activity *in vitro* by interacting with histone

deacetylase complexes [20], [21]. In contrast, MBD4 is implicated in DNA repair [22], [23].

Knockout mice for *Mecp2* and *Mbd2* did not show increased expression of endogenous genes or parasitic DNAs [24], [25]. However, the redundant functions among the methyl-CpG binding family may compensate the deficiency of one its members.

Methyl-CpG binding proteins interact with histone deacetylases enzymes (HDACs) [26], [27], [28], [29]. Histones become acetylated in several residues leading to an open chromatin conformation [30]. HDACs remove the acetyl groups from these residues resulting on close chromatin packing, which makes DNA inaccessible to transcription factors and other elements enabling transcription. Futscher and colleagues [31] proposed a model of interaction between methylated DNA, methyl-CpGs binding proteins, HDACs and other protein factors to explain the relation between DNA methylation status of the promoter of the *SERPINB5* gene and its complete silencing in skin fibroblasts and expression in epithelial skin cells. In this model (Figure 1.2.2); epithelial cells in skin have an unmethylated promoter that is occupied by the transcriptional regulators Ap1 and p53. Surrounding histones are acetylated limiting histone-histone interactions and providing an open chromatin structure that enables transcription. In contrast, the promoter in skin fibroblasts is fully methylated. Methylated DNA is associated to methyl-CpG binding proteins, which attract HDACs and chromatin remodeling complexes. Deacetylated histones result in local close chromatin packing and subsequent gene silencing.

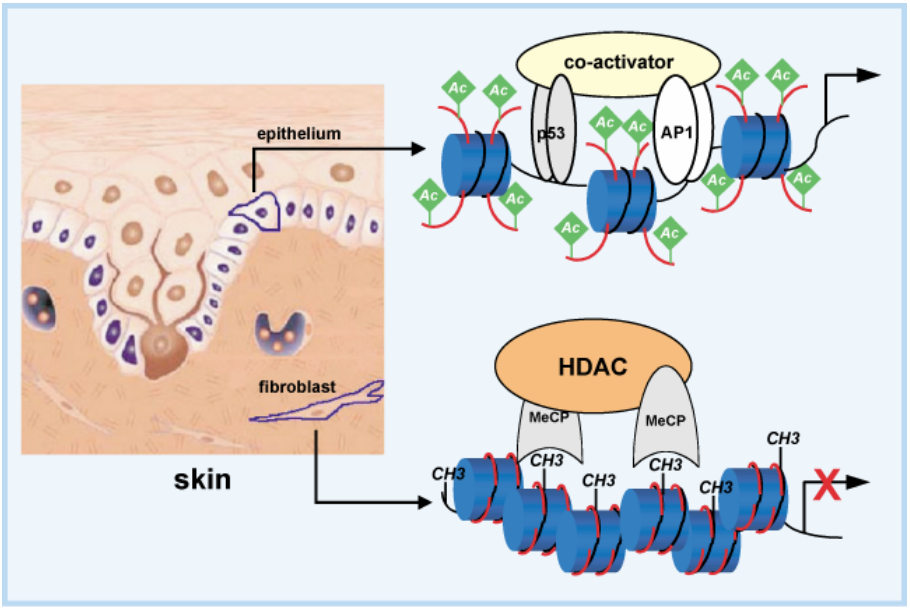


Figure 1.2.2: Model of cell-type specific control of *SERPINB5* expression by DNA methylation. The epithelial cells in skin have an unmethylated *SERPINB5* that is occupied by transcriptional regulators. The histones (blue) are acetylated (Ac) and chromatin conformation is open, allowing *SERPINB5* expression. In contrast, promoter in fibroblasts is methylated. MeCP2 binds to methylated DNA and attracts HDAC, leading to a local change in chromatin conformation and further gene silencing. Reproduced from Costello and Vertino [32]

In agreement with this model, Harikrishnan and colleagues [33] demonstrated that MeCP2 interaction with methylated DNA results in the recruitment of the SWI/SNF complex. This complex would be responsible for chromatin remodeling in an ATP-dependent manner.

Methylation of lysine residues in histones represents another epigenetic mechanism for gene silencing. Similar to hypoacetylation, methylation at lysine position leads to transcriptional repression. Vire and colleagues demonstrated that a histone methyltransferase, EZH2, interacts -within the context of the Polycomb repressive complexes 2 and 3- with DNA methyltransferases and associates to further gene silencing [34].

Csankovszki and colleagues [35] added a new layer to the model, the role of an antisense transcript interacting with DNA methylation and histone modifications. They demonstrated *in vitro* a synergic effect of Xist RNA, DNA methylation and histone hypoacetylation in maintaining X chromosome inactivation.

In addition, methylation in silencer or insulator elements might block the binding of the cognate binding proteins, potentially abolishing their repressive activities in gene expression [36], [37], [38]. Figure 1.2.3 displays a model to explain how DNA methylation of these elements might influence the binding of proteins to DNAs, thus altering the expression of a gene.

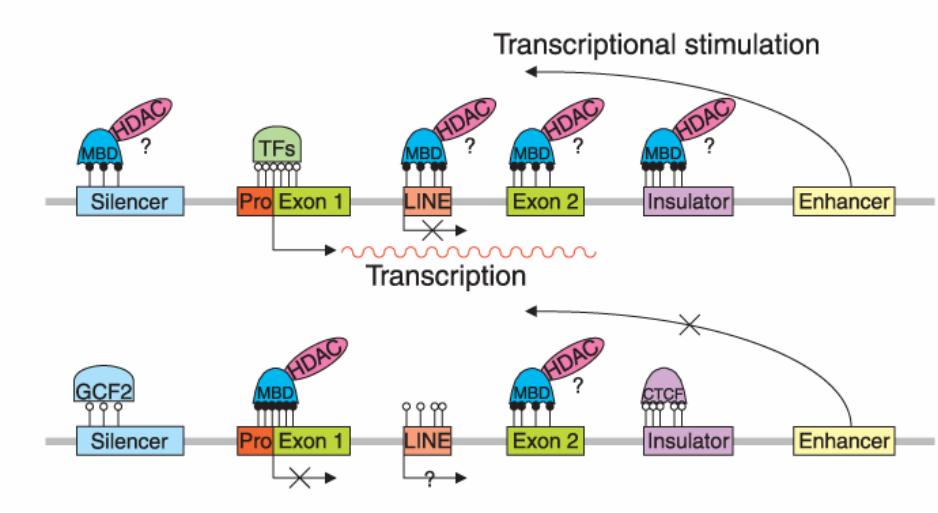


Figure 1.2.3: DNA methylation regulates expression by influencing factor binding to regulatory regions. (Top) The enhancer is functional because the silencer and insulator are methylated (filled lollipops) and, thus, not occupied by their respective cognate factors. Methylation here is permissive for expression. Transcripts are not initiated at the potentially active LINE element because it is methylated. (Bottom) Lack of methylation (open lollipops) in a silencer and insulator can lead to the binding of the cognate protein (i.e. GCF2 and CTCF at the silencer and insulator, respectively), thus preventing the enhancer from functioning. Reproduced from Jones and Takai [39]

Models cited here represent only a part of a growing body of evidence towards the fully understanding of the epigenetic regulations. Taken together, these models suggest that probably there is no such universal mechanism for epigenetic gene silencing. In contrast, a group of several molecular interactions taking part coordinately at particular tissular, developmental or physiological stages appears as the most probably scenario.

1.3. Methods to detect and measure DNA methylation

Epigenetic models drawn in the seventies were mainly theoretical. Continuous technological advances have allowed the development of methods and procedures to precisely determine the methylation status of a particular CpG position, a particular locus or even the genome-wide analysis of methylation content and variations (Table 1.3.1)

Table 1.3.1: Techniques for the detection of DNA methylation

Application	*Technique*	*Methylation detection*	*Platform*	*Key reference*
Analysis of genome-wide methylation content	HPLC	Enzymatic hydrolysis	High performance separation	[40]
	HPCE	Enzymatic hydrolysis	High performance separation	[41]
	SssI acceptance assay	Enzymatic methylation	Blotting. Radioactive detection	[42]
	Immunochemical analysis	Anti- methyl-cytosine antibody	Microscopy	[43]
Analysis of methylation status at a particular locus	MSRE-Southern	Enzymatic digestion	Blotting. Radioactive detection	[44]
	MSRE-PCR	Enzymatic digestion	Electrophoresis	[45]
	MSP	Bisulfite treatment	Electrophoresis	[46]
	COBRA	Bisulfite treatment and enzymatic digestion	Electrophoresis	[47]
	Direct /cloned bisulfite sequencing	Bisulfite treatment	Sequencing	[48]
	MS-SNuPE	Bisulfite treatment	Electrophoresis	[49]
	MethyLight	Bisulfite treatment	Real time PCR	[50]
	MS-MCA	Bisulfite treatment	Real time PCR	[51]
	MS-DGGE	Bisulfite treatment	Electrophoresis	[52]
	MS-SSCA	Bisulfite treatment	Electrophoresis	[53]
	Methylation specific microarray	Bisulfite treatment	Microarray	[54]
Discovery of novel DNA methylation sites	Differential Methylation Hybridization	Enzymatic digestion	Microarray	[55]
	RGLS	Enzyme digestion	2D-electrophoresis	[56]
	MCA-RDA	Enzyme digestion	Subtractive hybridization. Electrophoresis	[57]
	MS-APPCR	Enzyme digestion	Electrophoresis	[58]
	MBD-SPM	Methyl binding protein	Chromatography. Electrophoresis	[59]
	ICEAMP	Methyl binding protein	Subtractive hybridization. Sequencing	[60]
	Chip on chip	Methyl binding protein	Microarray	[61]

Besides differences in detection platforms and further applications, described methods in the literature share common features. Methods can be grouped according to their scope and applications. Hence, some methods are focused on a global scale to determine the overall content of cytosine methylation (i.e. HPLC, HPCE), while others are focused on detecting the methylation status of a particular locus (i.e. bisulfite sequencing, MSP). Interestingly, other methods have been developed to discover differences in methylation profiles at genome-wide scale (i.e. RGLS, MS-APPCR, DMH).

A second grouping category is the principle applied to detect methylation differences. Methods to quantify the overall content of methylated cytosines require the initial complete hydrolysis of the genome either by chemical or enzymatic treatments (i.e. HPLC, HPCE). Methylated cytosines display different peaks to unmethylated ones, enabling therefore the calculation of overall ratio of methylated/unmethylated cytosines. Global methylation analyses do not allow determining the methylation status at a particular locus. In contrast, methods based on methylation–sensitive restriction enzymes (i.e. MSRE-PCR, DMH) and methyl-cytosine binding proteins (i.e. MBD-SPM, ICEAMP, ChIP-Chip) take advantage of biological properties allowing the qualitative and quantitative methylation studies at particular locus. Isoschizomers of bacterial restriction endonucleases display different sensitivities to methylated cytosines. Despite sharing the same recognition site with the insensitive isoschizomer, sensitive restriction enzymes will cut only when cytosines are unmethylated. A popular isoschizomer pair for methylation analysis is *HpaII*/*MspI*. Both enzymes cleave CCGG sites, but *HpaII* is unable to cleave if the internal cytosine is methylated. On the other hand, proteins containing methyl-DNA binding domains bind specifically methylated CpG positions, as explained in section 1.2. Some methylation detection methods make use of this property and generally combines MBDs to chromatin immunoprecipitation (ChIP) or binding columns to obtain a fraction of the genome enriched in methylated- cytosines followed for the detection by quantitative PCR or microarray analysis. Methods based on methylation specific restriction enzymes and methyl-cytosine binding proteins suffer the limitation that they provide information only about CpGs within

the cleavage or binding site. Treatment of genomic DNA with sodium bisulfite overcomes this limitation and allows the analysis of virtually any CpG position within the genome. This method depends on the reaction of bisulfite with cytosines in single-stranded DNA, which are converted to uracil, whereas methyl-cytosines are unreactive [62] (Figure 1.3.1). Based on this chemical property, Frommer and colleagues developed a sequencing method that allows the detection of methylation positions at single-base resolution level [48]. The modified DNA strands are then amplified by PCR and further sequenced, directly or after subcloning of the amplified fragment. Direct sequencing gives information about the average methylation of a CpG position in a sample, while sequencing of cloned DNA allows the analysis of CpG sites on individual half-strand DNA molecules. Other methods relies on the single nucleotide polymorphisms (SNPs) introduced by the bisulfite treatment to design PCR primers specific for methylated and unmethylated states (i.e. MSP) or methylation specific probes for real time PCR detection (i. e. MethyLight). The major limitation of bisulfite treatment based method arises from the possibility of incomplete conversion of unmethylated cytosines to uracil. Consequently, conversion rates must be calculated and results corrected accordingly.

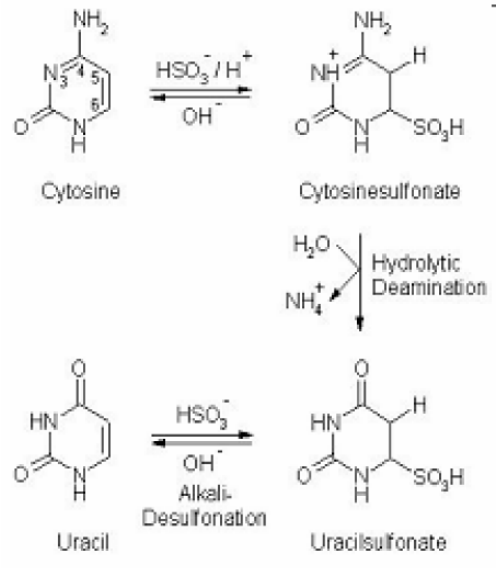


Figure 1.3.1: Conversion of unmethylated cytosine to uracil by bisulfite treatment. Unmethylated cytosines add a HSO_3^- to form cytosine sulfonate in alkaline medium, which after hydrolytic deamination is converted to uracil sulfonate. Final desulfonation in alkaline medium yields uracil.

Different platforms to detect differential methylation allow several applications at different levels of resolution. High performance separation methods are used for global methylation analysis (i.e. HPLC, HPCE). Originally, methods relying on restriction enzymes were resolved by Southern blotting. The combination of enzymatic restriction or bisulfite treatment and PCR allows the detection of differences by gel electrophoresis (i.e. MSRE-PCR, MSP, COBRA, etc). High quantitative methods, such real-time PCR (i.e. MethyLight) or microarrays (i.e. MSO, DMH), are currently the method of choice in the discovery, development and validation of methylation markers. In addition, microarray platforms allow the analysis of many genes in parallel (MSO) or even the genome-wide discovery of previously unknown methylation sites (DMH). Optimization and automation of sequencing platforms have allowed the sequencing of the whole genomes at single-based nucleotide resolution [63]. In the same manner, such advances have allowed the realization of the Human Epigenome Project, a joint effort to unravel the tissue-specific DNA methylation status in human beings, as detailed in the next section.

1.4. The Human Epigenome Project

The publishing of the complete sequence of the human genome [63] constituted a yardstick in biological research. The published data provided the background knowledge not only for further genomic research but also for the study of different layers of control of the encoded information.

Epigenetics constitute one of the most important layers of control in the expression of the genes and it is unique in explaining how environmental stimuli might affect phenotypes. In this regard, the need for a human epigenome project has been long recognized [64], [65]. Recently, a workshop sponsored by the American Association of Cancer Research (AACR) attempted to coordinate the efforts towards an international collaboration. In this work, researchers stated the guiding principle for such collaboration: "The goal of the Human Epigenome Project is to identify all the chemical changes and relationships among chromatin constituent that provide function to the DNA code, which will allow a fuller understanding of normal

development, aging, abnormal gene control in cancer, and other diseases as well as the role of the environment in human health" [64].

As part of an European Consortium (www.epigenome.org), our group took part in the deciphering of DNA methylation profiles in normal tissues and primary cells. In a pilot phase, authors demonstrated the feasibility of the project by determining the methylation patterns in the Major Histocompatibility Complex (MHC) on human chromosome 6 [66]. Completing the first phase of the full-scale HEP, we have published the methylation profiles of human chromosomes 6, 20 and 22 [67].

Similarly to the Human Genome Project data, HEP data require exhaustive interpretation to understand the phenotypic consequences of the established DNA methylation profiles. In some parts of this work, I have taken the produced data in HEP as baseline knowledge for further practical work in the laboratory. Accordingly, based on discovered individual T-DMRs, I have studied the expression of their associated genes by semiquantitative RT-PCR (section 2.1.1) and the methylation status of orthologous genes in mouse (section 2.3.1). The results of my experiments presented in these sections are included in the publication of the results of the first phase of HEP [67]. Furthermore I selected discovered T-DMRs corresponding to pseudogenes and determined the methylation status of their parental genes by direct sequencing of bisulfite-converted DNA (section 2.3.3.2).

1.5. DNA methylation role in biological processes in healthy tissues

1.5.1. Gene silencing

Hypermethylation of CpG dense 5′ untranslated regions (5′-UTR) frequently correlates with the silencing of the cognate transcript. Some models relate this effect to a blocking of binding site for transcription factors, while other attributes the silencing effect to an tight chromatin condensation in which methylated DNA interacts with modified histones, reviewed in part 1.2. Besides the responsible mechanism, hypermethylation of particular 5′-UTR has a phenotypic effect and are frequently referred as "epimutations".

Experimental approaches to investigate the functional consequences of DNA methylation are usually based on the utilization of analogous of deoxycytidine, which allow the pharmacological inhibition of DNMTs. Advantages and drawbacks for this experimental approach are discussed in section 3.2 in the Discussion chapter.
The correlation between hypermethylation of the 5'- UTR is not so clear when tissue-specific DNA methylation profiles are taken in account. In section 2.1 of this work, I show in a group of 55 T-DMRs that methylation of 5'-UTR correlates with a decreased expression of the associated transcript in 37% of the cases.

1.5.2. *De novo* DNA methylation of integrated foreign DNAs

DNA methylation is associated to the silencing of mobile elements inserted into the genome. Transgenes very often become inactive upon insertion, despite they were artificially introduced to the cells (i.e. by genetic engineering procedures) [68] or in naturally occurring viral infections [69] and genome recombination [70]. Promoter methylation has been noted as one of the mechanisms responsible for the inactivation of mobile elements [71]. Accordingly, silenced transgenes can be subsequently reactivated by 5- azacytidine treatment [72].
Upon insertion, some viral genomes become *de novo* methylated in specific patterns related to their transcriptional activity in the transformed cells [73]. The site of initiation of this *de novo* methylation appears to be variable. Starting from this point, DNA methylation spreads subsequently in both directions [74].
In section 2.3.3, I have studied the dynamics of DNA methylation in three gene duplication events: functional gene families, unprocessed pseudogenes and processed pseudogenes. Differently to functional gene families and unprocessed pseudogenes, processed pseudogenes are thought to be evolutionary relics of transposon insertion [75]. By comparative analysis of 17 pseudogene-parental gene pairs, I determined that 76% displayed tissue- specific methylation in at least one of the tissues analyzed and 24% hypermethylation in all analyzed tissues. In contrast to the unprocessed pseudogenes, I did not find a conservation of tissue-specific methylation between any of the analyzed processed pseudogene - parental gene pairs.

1.5.3. Gene duplications and tissue- specific DNA methylation profiles

Several recent reports have recognized the widespread occurrence of T-DMRs in healthy tissues [67] [76] [77] [78]. Extrapolating from these results, about 5 to 25% of all human genes might be tissue specifically methylated. T-DMRs frequently map to the 5' regions of annotated genes, however many are located in exon, introns or intergenic regions. In addition, differential methylation is not limited to coding genes, but has been observed in pseudogenes as well [67] [79]. Currently, little is known how T-DMRs evolve and are maintained and the signals that regulate them. One possibility is that differentially methylated loci might bear an intrinsic genetic or epigenetic signature that triggers differential methylation. Another possibility, not mutually exclusive, is that differential methylation is the result of other *cis*- or *trans*-acting factors or sequences that regulate the methylation profile of a particular locus. In this regard, gene duplication events represent a well-suited model to address the dynamics of (differential) methylation.

Gene duplications are a source of genomic novelties leading to gene families and unprocessed pseudogenes. Using comparative analysis, Lynch and Conery [80] estimated that gene duplications arise at a high frequency of about 0.01 per gene per million years. The classical model by Ohno [81] predicts that upon gene duplication, one copy retains the original function under strong surveillance by negative selection while the other copy becomes free of selective constraints and evolves mainly in a neutral fashion.

Two possible fates are therefore possible for the new duplicates. If the mutation is advantageous, it is fixed and the duplicate will develop a new function. In that way, functional divergence may lead to genes either with a reduced functional capacity (subfunctionalization) or a new function (neofunctionalization). Groups of functionally related genes are frequently recognized as gene families. On the other hand, if the mutation is disadvantageous, the duplicate will degenerate to an unprocessed pseudogene.

As most of the mutations are expected to be deleterious, pseudogenization constitutes the most likely fate of a new gene duplication event. However, little is known on the mechanisms that prevent degradation into a non-functional gene.

I have selected two gene families that differ greatly in their evolutionary age, the ancient TBX family that exists in vertebrate and invertebrates and the more recently evolved PLG family that is only present in the hominoid lineage. The unifying feature of the TBX transcription factor family is the presence of the T-domain that confers DNA binding and dimerization. In mammals, 17 distinct genes have been identified that can be grouped into 5 subfamilies (the T, TBX1, TBX2, TBX6 and Tbr1 subfamilies) based on their evolutionary distance [82]. TBX transcription factors are crucial in regulating a plethora of processes such as craniofacial development, limb outgrowth and patterning, as well as T-cell differentiation [82]. The second gene family I investigated is the plasminogen precursor, PLG, family consisting of three known members (*PLG*, *PLGLA* and *PLGLB*). While the PLG gene itself is located on chromosome 6q25.3 within the IGF2R imprinting cluster, the two plasminogen-related genes, *PLGLA* and *PLGLB* map to chromosome 2q12.2 and 2q11-p11, respectively [83] [84]. *PLG* encodes plasminogen that circulates in the plasma as a pro-enzyme [85] and in the presence of a fibrin clot, is converted to plasmin by the tissue plasminogen activator (tPA). Upon this activation, the proteolytic plasmin digests the insoluble fibrin clot, playing a key role in tissue repair and wound healing. However, the functions of *PLGA* and *PLGB* remain largely elusive.

In section 2.3.3.1, I have analyzed the conservation of T-DMRs in the PLG and TBX families in healthy human tissues. I found that methylation and sequence are more conserved in the more recently PLG family.

Moreover, I have studied the conservation of T-DMRs in pseudogenes (section 2.3.3.2). As mentioned above, pseudogenes are divided in two groups according to their origin [75]. Unprocessed pseudogenes derive from a gene duplication mechanism and generally maintain the exon/intron structure of their parental genes. In turn, processed pseudogenes arose due to the reverse transcription of the parental gene mRNA and typically lack both regulatory sequences and an exon/intron structure. While some unprocessed pseudogene display T-DMR conservation similarly to that observed in gene families, none of the processed pseudogenes do.

1.5.4. Development

It is generally accepted that the developmental program from fertilized egg to adult is under epigenetic control. Knockout mice for DNA methyltransferases undergo genome-wide DNA methylation alterations that lead to embryo lethality or developmental defects [86], [10].

In mammals, methylation patterns are reprogrammed in vivo in germ cells and early embryos, reviewed in [87]. Experiments in mouse showed that primordial germinal cells (PGCs) are highly methylated in early development but undergo genome-wide demethylation before embryonic day 13-14. Remethylation takes place later, earlier in the male germ line (from day 15 onwards) and after birth in the female germ line (Figure 1.5.1, panel A). In preimplantation embryos (Figure 1.5.2, panel B) the paternal genome is demethylated by an active mechanism after fertilization while maternal genome is demethylated by a DNA replication-dependent passive mechanism. At the time of implantation both genomes are remethylated to different extents in embryonic and extraembryonic lineages. While reprogramming of germ line cell erases parental imprints and acquired epigenetic modifications, postzygotic reprogramming are thought to erase those acquired during gametogenesis and to be linked to first-lineage decisions during mammalian development.

Pioneer models stated that specific DNA methylation patterns are imposed in the genome at defined developmental time points in precursor cells and maintained by DNMT1. This maintenance would lead to predetermined patterns of gene expression during development in descendants of precursor cells and specific demethylation events could reactivate genes in differentiated tissues [88], [89]. These models, however, are essentially theoretical. Experimental evidence is contradictory and how DNA methylation is related to gene expression during normal development remains to be elucidated [90], [91], [92].

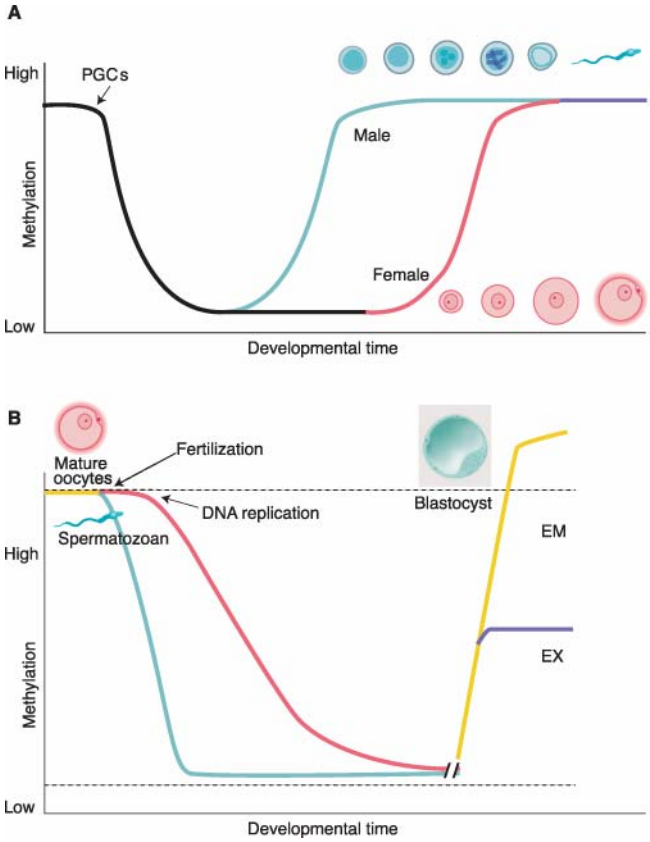


Figure 1.5.1: DNA methylation reprogramming in the germ line and preimplantation embryos. Primordial germ cells (PGCs, panel A) become demethylated in early development. Remethylation begins at E16 in male mouse cells and after birth in growing oocytes. In preimplantation embryos (panel B) the paternal genome (blue) is demethylated by an active mechanism immediately after fertilization. The maternal genome (red) is demethylated by a passive mechanism depending on DNA replication. Both genomes are remetylated around the time of implantation to different extents in embryonic (EM) and extraembryonic (EX) lineages. Methylated imprinted genes and some repeat sequences (dashed lines) do not become demethylated. Unmethylated imprinted genes (dashed line) do no become methylated. Reproduced from Reik et al [87]

In section 2.2 of the present work, I determined genome-wide expression profiles in fetal and healthy adult lung samples. 99 genes appeared as highly expressed in fetal lung and 354 as highly expressed in healthy adult lung. Out of them, I selected 43 genes displaying CG-rich 5'UTRs and studied their DNA methylation status by direct sequencing of bisulfite-converted DNA. Highly methylated 5'-UTR regions correlated with decreased expression levels in 19% of the cases.

1.5.5. Coexpression in "head-to-head" genes

Noteworthy, CpG islands are frequently located in regions known as "bi-directional promoters". Genes in these regions, also denominated "head-to-head" genes, are transcribed in opposite directions and their transcription start sites (TSSs) are closely located. Little is known about the methylation status of these regions and how it might influence in the regulation of the associated genes. Recently, Ikeda and colleagues reported hypermethylation of the bi-directional promoter region of *COL4A5/COL4A6* and subsequent loss of expression in colorectal cancer [93]. Moreover, Shu and colleagues showed that levels of mRNA were inversely correlated with the degree of methylation in three bi-directional promoters and reactivated by 5'-aza-2'-deoxycytidine treatment in several cell lines [94]. Whether differential methylation of "bi-directional" promoters is a common feature in lung development remains to be determined. In section 2.2.5, I have determined the methylation profiles of 20 putative "bi-directional" promoters in fetal and healthy adult lung. Expression profiles of the paired genes suggested co-expression when compared to random-pairs of genes in different chromosomes. In addition, the majority (17 bi-directional promoters) was unmethylated and their respective "head-to-head" genes highly expressed. In contrast, 3 hypermethylated regions showed lower expression of the associated transcripts, giving experimental support for co-regulation.

1.5.6. Imprinting

Genomic imprinting refers to the parent-of-origin-specific gene silencing. Since the discovery of this phenomenon more than 20 years ago, [95], [96], more than 80 imprinted genes have been reported. Imprinting status of orthologous genes are frequently discordant between human and mouse. Furthermore, imprinted genes are frequently noncoding RNAs or genes derived from retrotransposition in many cases [97]. Among the most common examples of protein coding imprinted genes can be mentioned the insulin-like growth factor 2, *IGF2,* and the alpha unit of the Gs protein, *GNAS1* [97]. In mammals, imprinted genes are especially implicated in the

regulation of fetal growth, development and function of the placenta, and postnatal behaviors, reviewed in [98].

Most imprinted genes are clustered in large chromosomal domains [97]. Common regions referred as "Imprinting Controlling Regions" (ICRs) generally direct the expression in such clusters. ICRs are regulated epigenetically, being usually one of the parental alleles methylated. Parental DNA methylation imprints at ICRs are established during gametogenesis. After fertilization, imprints are maintained throughout the development (Figure 1.5.2). In addition to DNA methylation, monoallelic silencing of some genes require the interaction of histone modifications [99] or non-coding RNAs [100], [101].

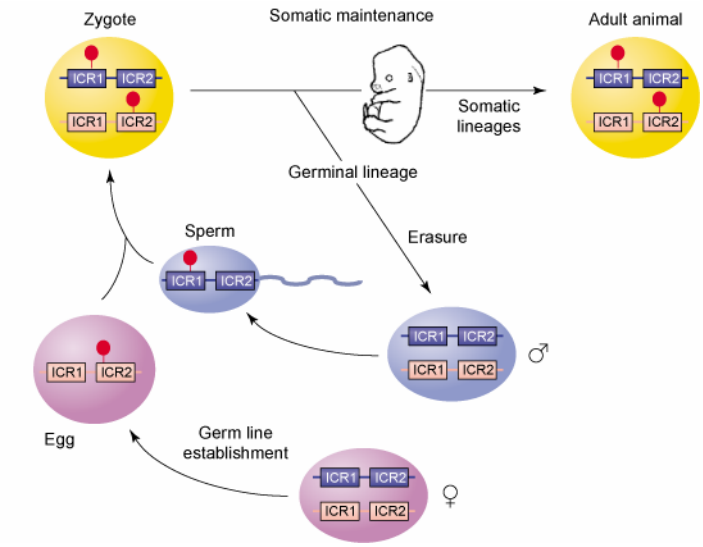


Figure 1.5.2: Parental imprints are established during oogenesis or spermatogenesis at ICRs and maintained thereafter. After fertilization the imprints are maintained throughout development. DNA methylation (red lollipops) is the most consistent hallmark of imprints. Two examples are depicted: an ICR with paternally derived (ICR1), and one with maternally derived DNA methylation (ICR2). Reproduced from Delaval and Feil [102]

Several developmental and behavioral disorders in human are associated to alterations in imprinted genes. In many of them, the origin of the disorder is the inheritance of two chromosomal copies of the locus from a parent, in an alteration denominated uniparental disomy (UPD). The best-characterized imprinting syndromes are the Beckwith-Wiedemann syndrome (BWS) and Prader-Willi/Angelman syndrome (PWS and AS, respectively). In BWS, a paternal UPD of

chromosome 11p15.5 leads to the biallelic expression of the *IGF2* gene and loss of expression of the *CDKN1C* gene. BWS phenotypic features are exomphalos, macroglossia, and gigantism in the neonate. Moreover, BWS patients show a predisposition to early tumors, reviewed in [103].

PWS is caused by deletion or disruption of a gene or several genes on the chromosome 15p15.5 or maternal UPD, because the gene(s) on the maternal chromosome(s) 15 are virtually inactive through imprinting. Correspondingly, the paternal UPD in the same region causes AS. Phenotypic features of PWS patients include diminished fetal activity, obesity, muscular hypotonia, mental retardation, short stature, hypogonadotropic hypogonadism, and small hands and feet. AS features includes mental retardation, movement or balance disorder, characteristic abnormal behaviors, and severe limitations in speech and language, reviewed in [104].

In this work, I have studied the role of DNA methylation in the imprinting in healthy adult tissues of the IGF2R cluster in human chromosome 6, as presented in sections 2.1.2 and 2.3.2

1.6. Impact of DNA methylation in cancer

1.6.1. General considerations

Alterations of DNA methylation profiles have been recognized as an important, and in many cases essential, component of cancer development. When compared to normal cells with the same tissue of origin, transformed cells display aberrant profiles, which have an application in diagnosis and molecular classification of the tumor [105]. In this regard, it has been proposed that epigenetic alterations in stem cells might precede the differences that clearly distinguish tumor types [106].

Nowadays, aberrant methylation profiles constitute a widely accepted feature in cancer. Aberrant methylation profiles in more than 120 genes have been related to malignant phenotypes in cancer. Some of these epigenetic alterations were detected in a particular cancer or tumor type, while others appear as more universal. As an

example table 1.6.1 lists 20 genes, which are frequently reported as aberrantly methylated in cancer.

Table 1.6.1: Examples of genes aberrantly methylated in cancer

Gene name	*Gene Name*	*Aberrant methylation*	*Key reference*
CCNA1	Cyclin A1	Head and neck squamous cell carcinoma	[107]
CCND1	Cyclin D1	B-cell malignancy	[108]
CDH1	E-cadherin	Cervical cancer	[109]
CDKN1A	p21CIP1	Prostatic cancer	[110]
CDKN1B	Cyclin-dependent kinase inhibitor 1B (p27, Kip1)	Lymphoid malignancies	[111]
CDKN1C	Cyclin-dependent kinase inhibitor 1C (p57, Kip2)	Pancreatic ductal neoplasms / Imprinting	[112]
ESR1	Estrogen receptor 1 (alpha)	Colorectal cancer	[113]
FHIT	Fragile histidine triad gene	Leukemia	[114]
GSTP1	Glutationine S-transferase	Prostate cancer	[115]
H19	H19	Colorectal cancer	[116]
HIC1	Hypermethylated in cancer 1	Invasive ductal breast carcinoma	[117]
IGF2	Insulin-like growth factor 2 (somatomedin A)	Hepatoblastoma	[118]
MLH1	mutL homologue 1, colon cancer, nonpolyposis type 2 (*E. coli*)	Colorectal carcinoma	[119]
PTEN	Phosphatase and tensin homologue (mutated in multiple advanced cancers 1)	Breast Cancer	[120]
PTGS2	Prostaglandin-endoperoxide synthase 2 (prostaglandin G/H synthase and cyclooxygenase)	Prostate cancer	[121]
RARB	Retinoic acid receptor, beta	Uterus carcinoma	[122]
RASSF1	Ras association (RalGDS/AF-6) domain family 1	Lung Small Cell Carcinoma	[123]
S100A2	S100 calcium binding protein A2	Prostate cancer	[124]
TP53	Tumor protein p53 (Li-Fraumeni syndrome)	Ductal carcinoma	[125]
VHL	Von Hippel-Lindau tumor suppressor	Renal carcinoma	[126]

1.6.2. Hyper- and hypomethylation in cancer

At global level, the methyl-cytosine content was shown to be lower in transformed cells [127]. In normal cells, repetitive sequences originated by ancestral retrotransposon insertions (e.g. LINEs, SINEs) are mainly hypermethylated, as well as coding regions and intronic CpG positions. These non-promoter regions contain approximately 95 % of the methyl-cytosines of the human genome. Many of these regions became hypomethylated in cancer cells leading to the registered "global hypomethylation" values [128]. In mouse models, the overall degree of

hypomethylation of satellite or LINE-1 sequences correlated with the frequency of chromosomal aberrations and increased tumor frequency [129].

In parallel to the observed global hypomethylation, cancer cells also display aberrant hypermethylation of normally unmethylated CpG islands [3]. Hypermethylation mainly occurs in tumor-suppressor genes and is often linked to silencing of the cognate gene. Promoter hypermethylation appears as an early event in tumorigenesis. Epigenetic alterations can produce the early losses of cell cycle control, altered regulation of gene transcription factors, disruption of cell-cell and cell-substratum interaction and even multiple types of genetic instability, characteristic of human cancer, reviewed in [130], [131]. Epigenetic silencing of tumor suppressor genes occurs at least as frequently as mutations or deletions in human cancers [3].

According to the "Two-hits" hypothesis of carcinogenesis proposed by Knudson [132] loss of function of both alleles is required for malignant transformation. A mutation of a critical gene typically constitutes the first hit, followed by loss of the wild-type allele throughout deletion or loss of heterozygosity. Hypermethylation of tumor suppressor genes can constitute the initial hit in many somatic cancers, with subsequent alterations at DNA sequence level eliminating the second gene copy [133]. Furthermore, epigenetic alterations have also been reported to cause the second hit in familial cancers [134], [135].

In contrast, promoter hypomethylation affects frequently genes, known as "oncogenes", whose products are known to have oncogenic properties that over-expressed lead to the malignant transformation of the cell [128]. Several genes that are hypomethylated in cancer fall into the class of cancer-testis antigens (e.g. MAGE-A, -B and –C) [136], [137]). In addition, other hypomethylated genes in cancer are related to cellular stress and immune response (e.g. S100P [138]), cell cycle (e.g. CDKN2A [139]) and homeobox genes (e.g. HOX11 [140])

1.6.3. Loss of imprinting in cancer

Loss of imprinting (LOI) is the disruption of imprinted epigenetic marks through gain or loss of DNA methylation, or simply the loss of normal allele-specific gene

expression [141]. LOI of a growth-promoting imprinted gene leads to activation of the normally silent allele and results in abnormally high expression of the gene product, which gives cells a growth advantage. LOI of the *IGF2* locus is the most common LOI event across a wide range of tumor types, including colon, liver, lung and ovarian cancers as well as Wilms' tumor [142].

1.6.4. Potential applications for DNA methylation in clinical practice

DNA methylation has a great potential for translation into clinical practice. Aberrant methylation has been shown to have potential in risk-assessment, early detection, disease classification and prognosis prediction in a variety of cancers [143], [144]. In addition, some research in the field points towards the therapeutic application of targeting DNA methylation. As observed in early embryogenesis, DNA replication in absence of DNMTs leads to hypomethylation. Therefore, pharmacological inhibition of DNMTs could reactivate tumor-suppressor genes silenced by hypermethylation in cancer cells. Two of these inhibitors, 5-azacytidine (azacytidine) and 5-aza-2'-deoxycytidine (decitabine), demonstrated encouraging antileukemic activity but little activity in solid tumors, reviewed in [145]. However, when the molecular action of these drugs is considered, the simple inhibition of DNMTs appears as an over-simplification, as discussed in section 3.2.

1.6.5. DNA methylation in Lung Cancer

Lung Cancer is the second most common cancer and the most common cause of cancer-related death in both men and women in developed countries. The American Cancer Society estimates more than 213,000 new cases and 160,000 new deaths in the United States for 2007 [146].

There are two main types of lung cancer, small lung cell cancer and non-small cell lung cancer. Classically, tumor types are identified by pathological examination of cell morphology from patients' biopsies.

Small cell lung cancer. Classification, prognosis and treatment alternatives

Small cell carcinoma is the most aggressive type of lung cancer. Without treatment, patients show a median survival of only 2 to 4 months. Because patients with small cell lung cancer tend to develop distant metastases, localized forms of treatment, such as surgical resection or radiation therapy, rarely produce long-term survival [147]. Chemotherapy regimens prolong the survival 4- to 5- fold, being the overall survival at 5 years 5-10 % [148].

Currently, the cellular classification of subtypes of small cell lung cancer comprises small cell carcinoma, mixed small cell/large cell carcinoma and combined small cell carcinoma [149].

Small cell cancer patients are separated in two groups, according to the extension of the location of the tumors at the time of diagnosis. Patients with "limited-stage" disease have tumors confined to the hemithorax, mediastinum or supraclavicular lymph nodes. Under current treatments, patients with limited stage disease have a median survival of 16 to 24 months [150]. Patients with tumors that have spread beyond the supraclavicular areas belong to the "extensive-stage" disease group and have a worse prognosis.

Regardless of the stage, the current prognosis for patients with small cell lung cancer is unsatisfactory and the National Cancer Institute (U.S.A) recommends considering all patients for the inclusion in clinical trials at the time of diagnosis [151]

Non-small cell lung cancer. Classification, prognosis and treatment alternatives

Three main histological subtypes are differentiated in non-small cell lung cancer (NSCLC): squamous carcinoma (or epidermoid), adenocarcinoma and large cell carcinoma. Approaches to diagnosis, staging, prognosis and treatment are similar for these subtypes.

Factors that have correlated with adverse prognosis include the following: presence of pulmonary symptoms, large tumor size (>3 cm), nonsquamous histology, metastases to multiple lymph nodes and vascular invasion [152].

Possibility to surgically remove the tumor correlates with its staging and will mark the treatment approach. The most frequent treatment for patients with resectable

disease in NSCLC is surgery. In many cases, surgery is complemented with adjuvant chemotherapy. Local control can be achieved with radiation therapy in a large number of patients with unresectable disease. Patients with locally advanced, unresectable disease may have long-term survival with radiation therapy combined with chemotherapy. Patients with advanced metastatic disease may achieve improved survival and palliation of symptoms with chemotherapy [149].
Because the treatment is not satisfactory for almost all patients with NSCLC, the National Cancer Institute (USA) recommends considering eligible patients for clinical trials [151].

The need of a precise diagnosis for lung cancer – Application of DNA methylation markers

The application of effective and specific therapies requires precise diagnosis and classification of tumors. Particularly in lung cancer, small cell carcinomas, which respond well to chemotherapy and are generally not treated surgically, can be confused on microscopic examination with non-small cell carcinoma [149].
Currently, several genomic and proteomic methods allow the molecular characterization of lung cancers based on physiological processes during oncogenic transformation, reviewed in [153]. Among those processes, epigenetic phenomena are widely recognized to play a fundamental role in carcinogenesis and tumor progression [154]. In particular, DNA methylation alterations (e.g. inappropriate gene silencing through promoter hypermethylation) affect the normal physiological and morphological state of the cell towards malignancy. Hence, changes in methylation patterns appear as very promising markers for lung cancer in tissue-based diagnosis [155], as well as screening in body fluids [156].

Similarities between tumor ontogeny and embryonic development. Potential role of embryonic biomarkers in cancer

Reya and colleagues suggested that lung tumor ontogeny is determined by the consequences of gene transcription activation and/or repression events that recapitulate events important in embryonic development [157]. Several experimental

works in lung cancer support this hypothesis in terms of expression profiling [158], [159], [160] and sequence alterations [161]. Despite the accepted role of DNA methylation in oncogenesis and mammalian development, it is not known whether DNA methylation profiles in cancer cells also resemble those observed in fetal cells.
To further investigate this possibility, in section 2.2.7 I have determined the expression profiles of fetal lung tissue samples relative to healthy adult lung tissue samples and explored the methylation status of differentially expressed genes in fetal, healthy adult and tumor lung. Using this approach, I have identified four genes (*MEOX2*, *MDK*, *LAPTM5* and *FGFR3*) that are differentially methylated in lung cancer. *MEOX2* is uniformly hypermethylated in lung cancer compared to healthy adult and fetal lung. Furthermore, I provide evidence that the methylation is not uniform in lung cancer but is rather correlated with either the differentiation state (*MDK*, *LAPTM5*) or the lung cancer type (*FGFR3*). The results presented in section 2.2.7 indicate that some differentially methylated regions during lung morphogenesis may act as molecular markers in lung cancer.

1.7. Impact of DNA methylation in non-oncogenic diseases

1.7.1. DNA methylation in complex non- oncogenic diseases

Besides the overwhelming evidence linking aberrant DNA methylation with cancer, DNA methylation is suspected to play a role in other non-oncogenic human pathologies as well. Among them, experimental work has related aberrant DNA methylation profiles to a wide range of complex diseases, such diabetes [162], arteriosclerosis [163], inflammatory bowel disease [164], inflammatory arthritis [165], schizophrenia [166] and autoimmune diseases [167].

1.7.2. Interaction of genetic and epigenetic alterations in monogenic diseases

DNA methylation represents the cause in several monogenic diseases interacting with alterations at DNA sequence level. LOI in the *GNAS1* locus was proposed as the molecular mechanism that underlies sporadic and familial types of

pseudohypoparathyroidism type Ib (PHP-Ib) [168]. Similarly, LOI in a CpG-island-associated ICR leading to an increased expression of the *PLAGL1* gene constitutes one of the mechanisms proposed to give rise to transient neonatal diabetes mellitus, TNDM [169].

DNA methylation defects have also been linked to human diseases that are associated with repeat instability. For example, Fragile X syndrome (FRAXA), an X-linked disorder, is a common cause of inherited mental retardation. The molecular basis of FRAXA lies in the highly polymorphic CGG repeat within the 5'-UTR of the *FMR1* gene. Normally, 6-52 copies of the repeat are present, which increase to more than 200 copies in individuals with FRAXA. Affected individual show *de novo* methylation of the expanded CGG repeat and silencing of *FMR1* [170]. Further evidence suggests that CGG repeat and hypermethylation in other folate-sensitive fragile sites across the genome [171]. Hypomethylation seems also to play a role in the development of non-oncogenic human diseases associated to repeat instability. In Fascioscapulohumeral muscular dystrophy (FSHD), some patients display the contraction of a repeat in the *D4Z4* gene, whereas others do not such this alteration. Despite the presence of such a contraction, all FSHD display aberrant hypomethylation of the D4Z4 gene, suggesting that hypomethylation of D4Z4 is a key event in the cascade of epigenetic events causing FSHD [172].

1.7.3. Interaction of genetic and epigenetic alterations in non-oncogenic complex diseases

Although it represents an inherent idea of epigenetic regulation, little is known about how the epigenetic layer of control interacts with the genetic background in the development of a diseased phenotype in a complex disease, discussed in [173] [174]. In this regard, progeroid syndromes represent an excellent model to study such interaction.

Mutations within the *LMNA* gene cause many diseases with different phenotypes, such as different muscular dystrophies, progeroid syndromes, mandibuloacral dysplasia, dilated cardiomyopathies, restrictive dermopathy and lipodystrophy syndromes. Among laminopathies, the mutation position within the *LMNA* sequence

can partially predict the phenotype [175], but the underlying mechanisms for the development of these different phenotypes are still unclear. A combined effect of a mutation in the gene encoding the lamin protein and altered gene expression of genes encoding proteins that interact with the nuclear lamina might constitute one of the mechanisms leading to different phenotypes [176].

Familial partial lipodystrophy type 2, FPLD2, (OMIM ID: 151660) is an autosomal dominant syndrome, whose phenotypic manifestations include gradual lack of subcutaneous adipose tissue in extremities and gluteal and troncal regions at the onset of puberty. Simultaneously, adipose tissue accumulates on face and neck. Affected patients are insulin-resistant and may develop glucose intolerance and diabetes mellitus after 20 years of age. Several mutations in the *LMNA* gene have been mapped in patients with FPLD2 [177], [178].

Progeroid syndromes are characterized by a premature ageing of the affected patients (OMIM ID: 176670). In particular, the Hutchison- Gilford progeria syndrome is a rare disorder characterized by precocious senility. Death from coronary artery disease is frequent and may occur before 10 years of age. As in FPLD2, mutations in the *LMNA* gene are mapped in Hutchison- Gilford syndrome patients [179]. Concordantly, variations in DNA methylation have been widely reported as correlating with age [180].

In section 2.4, I investigate of the methylation status of 10 candidate genes in patients affected with FPLD2 and Progeria bearing different mutations in the *LMNA* gene and healthy control individuals. In general, genes displayed unmethylation. However, the *RARB* gene showed a significant higher methylation in all 6 FPLD2 patients when compared to the Progeria patients as well as the healthy control individuals. In addition, *LMNA* gene was clearly, though not significantly, differentially methylated in two different FPLD2 families bearing distinct LMNA mutations. The results presented in section 2.4 allow postulating that alterations at epigenetic level modulate the different phenotypes observed in progeroid syndromes upon a similar genetic background.

1.8. Structure of the thesis

1.8.1. Topics

The general focus of the present work is the study of differential DNA methylation influencing phenotypes in healthy and diseased tissues. It brings valuable information about several biological processes occurring in healthy tissues, such as tissue-specific gene silencing, imprinting and epigenetic control in human development. In addition, I applied the produced knowledge and the established methods to study DNA methylation biomarkers for an oncogenic disease, Lung Cancer, and a group of non-oncogenic complex diseases, Laminopathies.
Specifically, the contained topics can be listed as it follows:

- Correlation of DNA methylation and transcription profiles in healthy human tissues
- Conservation of tissue-specific DNA methylation profiles throughout the evolution.
- Interaction of genetic and epigenetic alterations in modulating phenotypes in laminopathies

1.8.2. Hypothesis

The main hypothesis under test in the present work is:

- Patterns of differential methylation influence biological processes in tissues or cells where they were observed.

Complementarily, secondary hypotheses are under test along the work:

- Specific methylation in healthy tissues and primary cells correlates with the expression of the cognate transcript
- DNA methylation at bi-directional promoters drives the co-expression of the corresponding “head-to-head” genes
- Differentially expressed genes between fetal and adult lung tissue samples are candidates for differential methylation in lung cancer

- Tissue specific methylation observed in human genes is conserved in their murine orthologues
- Tissue specific methylation is conserved upon genome duplications
- Differential DNA methylation modulates phenotypic expression upon similar genetic backgrounds

1.8.3. Objectives

General objective:

- To study the influence of differential methylation in biological processes occurring in healthy individuals as well as patients with oncogenic and non-oncogenic diseases.

Specific objectives:

- To determine the degree of correlation between hypermethylation of the 5'-UTR and gene silencing in healthy adult and fetal human tissues as well as in primary cells.
- To estimate the degree of evolutionary conservation of tissue-specific DNA methylation profiles by studying orthologous and paralogous genes.
- To study the interaction of DNA methylation and alterations at DNA sequence level in establishing the different phenotypes characteristic of laminopathies.

1.8.4. Relevance of the presented results

Results on correlation between hypermethylation and gene silencing in healthy human tissues bring experimental evidence supporting a long-standing theory of epigenetic regulation of gene expression, as discussed in [32]. The study of such correlation in human fetal tissues gives an insight on the role of DNA methylation in organogenesis during lung development. In addition, the establishment of high quality transcriptional profiles of fetal lung provides valuable information on functional genomics during lung development beyond epigenetics.

Results on conservation of tissue-specific DNA methylation profiles in human and murine orthologue genes provide very valuable base-line knowledge for further experimentation of the epigenetic phenomena using animal models.

The discovery of differential methylation in four genes in Lung cancer, *MEOX2*, *LAPTM5*, *MDK* and *FGFR3* nominate them as putative novel biomarkers for this disease. Evaluation of these markers in a larger sample panel would lead to a future application in diagnosis or screening.

Finally, results the study of the synergetic effects of mutations and aberrant methylation help explaining the different phenotypes observed in distinct laminopathies. Additionally, their relevance goes beyond laminopathies and progeroid syndromes. It supports a role for DNA methylation in the modulation of different phenotypes upon similar genetic backgrounds.

2. Results

2.1. Tissue specific DNA methylation and expression profiles

2.1.1. Candidate gene approach

It is widely accepted that methylation of the promoter region is related to gene silencing. However, the available experimental evidence supporting this concept is scarce.

To further investigate the functional role of tissue-specific DNA methylation regions in human healthy adult tissues, I have selected 55 tissue-specific differentially methylated regions (T-DMRs) from the results of the Human Epigenome Project (HEP) [67] and studied the expression of their associated genes by semiquantitative RT-PCR. The results of my experiments presented in this section were published as part of the first phase of HEP [67]

My experiments showed that hypermethylation of the 5' region correlates with transcriptional silencing in 37 % of the investigated genes, but I did not find such correlation when the T-DMRs were located either in the coding region or downstream to the gene.

I have studied the expression profiles of the selected genes in the same samples, where methylation profiles were determined. In that manner, I included in the study total RNA samples from three major human healthy tissues (Heart Muscle, Liver and Skeletal Muscle) and primary cells (CD4+ and CD8+ lymphocytes, keratinocytes, melanocytes and fibroblasts). Unlike established cell lines, studies using tissue and primary cells samples avoid changes in DNA methylation and expression profiles introduced by culture and repeated cell passage [181]

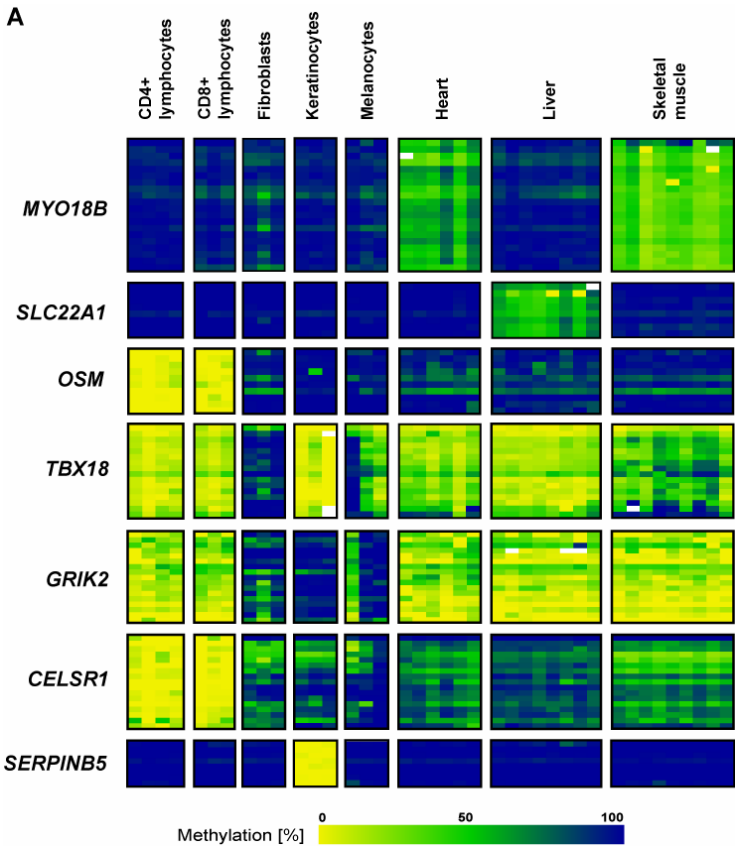
Figure 2.1.1 shows the results of single gene semiquantitative RT-PCR and bisulfite sequencing results for some example genes in the studied tissues. Expression levels were compared to the expression of a housekeeping gene, *ACTB1*, which displayed high levels of expression in all studied tissues and cells.

MYO18B, *SLC22A1* and *OSM* are examples of genes where hypermethylation of the 5' region correlates with gene silencing. In that manner, *MYO18B* was expressed only

in heart and skeletal muscle (50 % methylation), *SLC22A1* only in liver (50% methylation) and *OSM* in CD4+/CD8+ lymphocytes (0% methylation). These three genes were silenced in the rest of studied tissues and primary cells showing 100 % methylation.

TBX18, *GRIK2* and *CELSR1* are examples of genes whose methylation status does not correlate with the expression profiles. *TBX18* was highly expressed in fibroblasts despite hypermethylation, while cells showing no methylation (CD4+/CD8+ lymphocytes, keratinocytes) did not express the gene. Fibroblasts, melanocytes and keratinocytes displayed hypermethylation of *GRIK2*, however this gene was highly, slightly and no expressed in these cells, respectively. *CELSR1* was unmethylated and slightly expressed in CD4+/CD8+ lymphocytes, however it was highly expressed in keratinocytes, melanocytes, heart and liver, which showed 100% methylation.

Furthermore, I analyzed the expression of *SERPINB5* in the included tissues and cells. In agreement with that reported in the literature [31], I found the gene being expressed only in keratinocytes, which showed no methylation, but not in any other tissue (all showing 100% methylation).



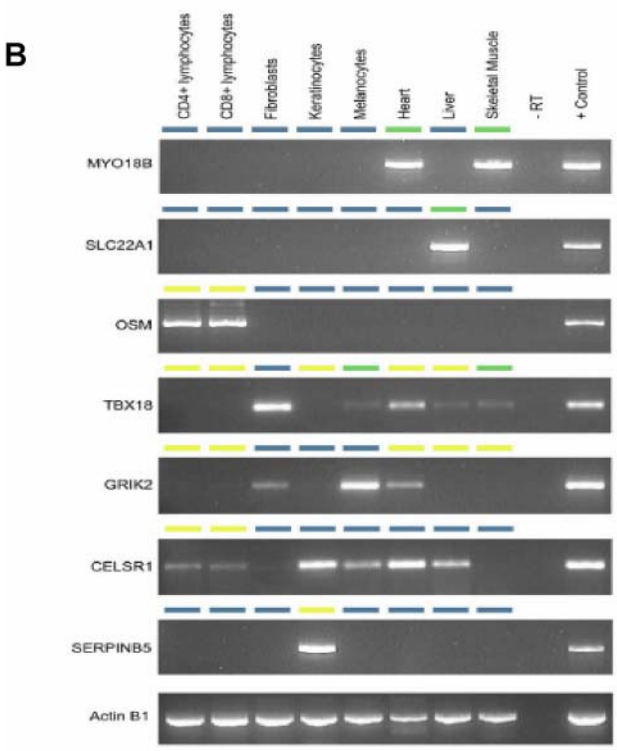


Figure 2.1.1: Examples of DNA methylation and expression profiles. A) DNA methylation profiles for six example genes and the positive control *SERPINB5* in primary cells and healthy adult tissues. Samples are displayed column wise with rows representing individual CpGs of the PCR fragment. Quantitative methylation analysis results are shown in a color scale ranging from yellow (~0% methylation), green (~50% methylation) to dark blue (~100% methylation). B) Correlation between 5'-UTR methylation and mRNA expression. RT-PCR results for genes in A) are shown as well as the housekeeping gene *ACTINB1.* While *MYO18B*, *SLC22A1* and *OSM* displayed a correlation between hypermethylation and gene silencing, *TBX18*, *GRIK2* and *CELSR1* did not show any correlation. I have used *SERPINB5* as positive control, as correlation was reported for this gene [31]. Mean of methylation at the 5'UTR are displayed as a colored bar above each gene. Color code corresponds to that in panel A). Shown RT-PCR results are representative of three independent samples per tissue type. Pool of adult and fetal RNAs was used as positive control for the RT-PCR (+ control)

Table 2.1.1 lists the studied genes, the position of the found T-DMR and the relation to the expression of the cognate transcript.

Table 2.1.1: Studied genes, position of T-DMR and relation to expression

Gene Name	*Ensembl ID*	*T-DMR Position*	*T-DMR / Expression*
Group 1			
CTA-299D3.6-005	OTTHUMG00000030140	5'-UTR	Inverse correlation
OSM	ENSG00000099985	5'-UTR	Inverse correlation
SOX10	ENSG00000100146	5'-UTR	Inverse correlation
SLC22A1	ENSG00000175003	5'-UTR	Inverse correlation
MYOD18B	ENSG00000133454	5'-UTR	Inverse correlation
IHPK3	ENSG00000161896	5'-UTR	Inverse correlation
PARVG	ENSG00000138964	5'-UTR	Inverse correlation
A4GALT	ENSG00000128274	5'-UTR	Inverse correlation
AP006358.2-002	OTTHUMG00000030241	5'-UTR	Inverse correlation
MDFI	OTTHUMG00000014681	5'-UTR	Inverse correlation
TGM3	ENSG00000125780	5'-UTR	Inverse correlation
PLG	ENSG00000122194	5'-UTR	Inverse correlation
PRAME	ENSG00000185686	5'-UTR	Inverse correlation
PIB5PA	ENSG00000185133	5'-UTR	Inverse correlation
Q8N7A7	ENSG00000183964	5'-UTR	Inverse correlation
FILIP1	ENSG00000118407	5'-UTR	Inverse correlation
SLC7A4	OTTHUMG00000030129	5'-UTR	No correlation
CTA-243E7.3	OTTHUMG00000030167	5'-UTR	No correlation
TBX18	ENSG00000112837	5'-UTR	No correlation
SSTR3	ENSG00000183473	5'-UTR	No correlation
RP4-756G23.1	OTTHUMG00000030852	5'-UTR	No correlation
RP3-398D13.4-001	OTTHUMG00000014188	5'-UTR	No correlation
RP11-174C7.4-001	OTTHUMG00000015553	5'-UTR	No correlation
RP3-43804.2-002	OTTHUMG00000030922	5'-UTR	No correlation
SUSD2	ENSG00000099994	5'-UTR	No correlation
TFAP2A	OTTHUMG00000014235	5'-UTR	No correlation
RIN2	ENSG00000132669	5'-UTR	No correlation
RP1-47A17.8	OTTHUMG00000030878	5'-UTR	No correlation
RP1-32B1.4-001	OTTHUMG00000015628	5'-UTR	No correlation
ENSESTG00000027621	ENSESTG00000027621	5'-UTR	No correlation
CMAH	ENSG00000168405	5'-UTR	No correlation
GRAP2	ENSG00000100351	5'-UTR	No correlation
GAR22/GAS2L1	ENSG00000185340	5'-UTR	No correlation
GABRR1	ENSG00000146276	5'-UTR	No correlation
PACSIN1	ENSG00000124507	5'-UTR	No correlation
PLA2G3	ENSG00000100078	5'-UTR	No correlation
PKHD1	ENSG00000170927	5'-UTR	No correlation
HMGA1	ENSG00000137309	5'-UTR	No correlation
GRIK2	ENSG00000164418	5'-UTR	No correlation
SEPT5	ENSG00000184702	5'-UTR	No correlation
RAET1E	ENSG00000164520	5'-UTR	No correlation
RASSF2	ENSG00000101265	5'-UTR	No correlation
VNN1	ENSG00000112299	5'-UTR	No correlation
Group 2			
RP1-165D5.4-004	OTTHUMG00000030772	Coding region	No correlation
GAR22/GAS2L1	ENSG00000185340	Coding region	No correlation
RP3-355C18.2	OTTHUMG00000030072	Coding region	No correlation
NP_955377	ENSG00000183506	Coding region	No correlation
HMG2L1	ENSG00000100281	Coding region	No correlation
CELSR1	ENSG00000075275	Coding region	No correlation
SMTN	ENSG00000183963	Coding region	No correlation
PIPPI	ENSG00000172346	Coding region	No correlation
FER1L4	OTTHUMG00000032354	Coding region	No correlation
CECR1	ENSG00000093072	Coding region	No correlation
Group 3			
PMM1	ENSG00000100417	Intergenic	No correlation
SA50	ENSG00000100347	Intergenic	No correlation

Studied genes can be separate in three groups. The first group consisted of 43 genes displaying tissue-specific DNA methylation in the 5' region. I defined this 5' region as the area between 2 kB upstream and 1 kB downstream of the Transcription Start Site (TSS). Within this group, 16 genes (37.21 %) showed a correlation between the DNA methylation status and gene silencing. Among the remaining 27 genes (62.79%), I did not found DNA methylation correlating neither with decreased nor with increased expression.

The second group consisted of 10 genes containing T-DMRs within their coding regions. T-DMRs in these genes were located in both, exons and introns. In no case, I found this differential methylation correlating with the expression profiles of the genes, neither positive nor inverse.

The third group consisted of 2 genes that showed T-DMRs located downstream of their coding regions. Investigated regions were located in intergenic areas and displayed a clear tissue-specific DNA methylation. Although there is no examples in the literature linking functionally these two T-DMRs to the investigated genes, I have chosen these genes because they are the ones located immediately upstream of the T-DMRs. In none of the two genes, the observed hypermethylation of the upstream region correlated with the transcriptional silencing of the gene.

2.1.2. Case of study: Are *SLC22A1* and *PLG* genes imprinted in healthy adult tissues?

Among those genes showing a correlation between hypermethylation of their 5' region and gene silencing, *SLC22A1* and *PLG* were worth noting. Both genes are located in chromosome 6q25.3 in a cluster containing also the *IGF2R* gene, which is a well-studied locus for imprinting in mouse [182]. By analyzing the allelic profiles for DNA methylation and expression in these two genes, I concluded that *SLC22A1* and *PLG* are not imprinted in healthy adult human liver.

First, I directly sequenced the bisulfite-converted DNA PCR products and analyzed the expression by semiquantitative RT-PCR in matched DNA/RNA samples derived from human liver, skeletal muscle and heart muscle. *SLC22A1* and *PLG* showed 50% methylation and expression in liver, while 100% methylation and silencing in all the

other studied tissues and cells. Figure 2.1.2 shows the methylation and expression values observed in these genes.

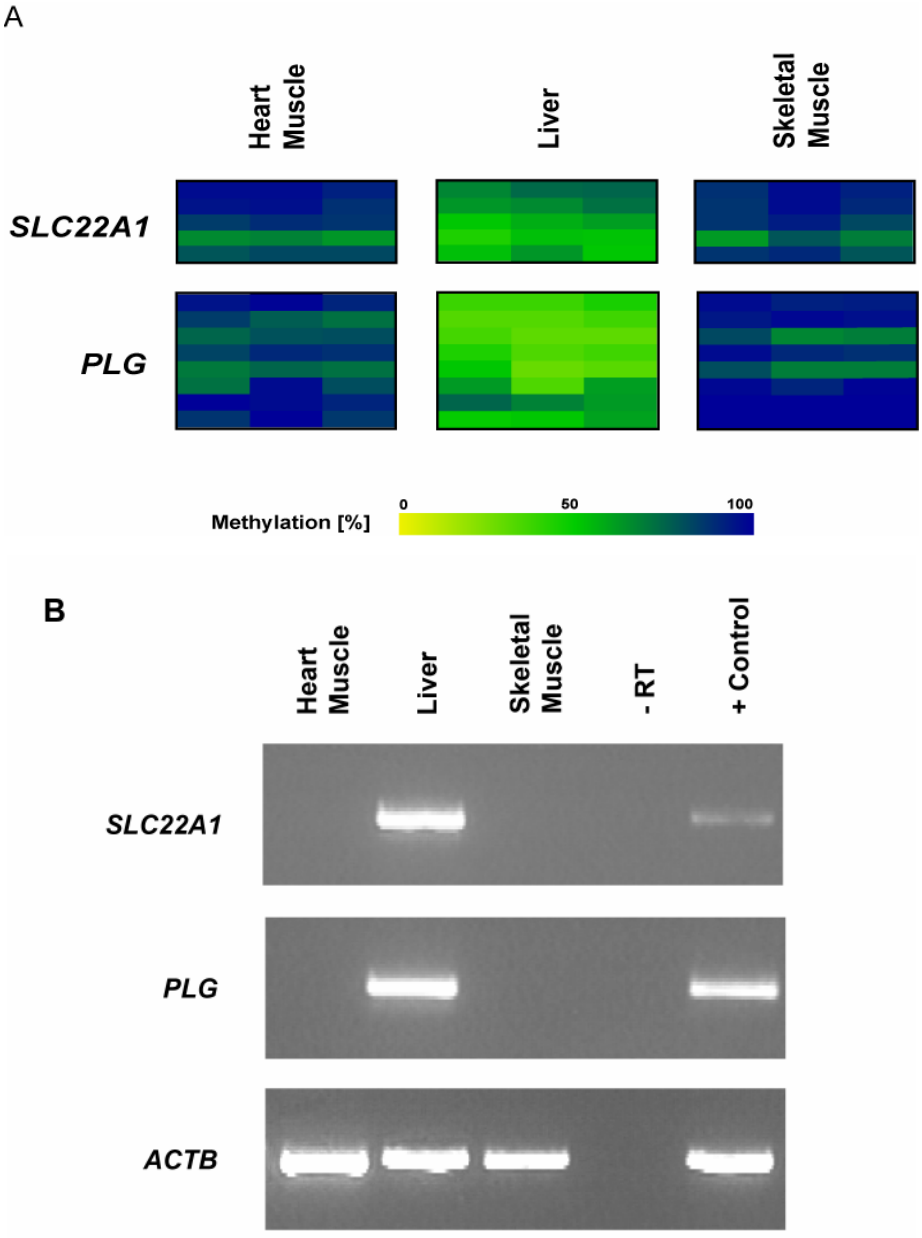


Figure 2.1.2: *SLC22A1* and *PLG* are expressed only in tissues showing 50% methylation. Direct bisulfite sequencing (A) and RT-PCR (B) results for the methylation and expression profiles of *SLC22A1* and *PLG* in healthy adult heart, liver and skeletal muscle. Three matched RNA/DNA samples from the same individuals were used. Methylation results in A) are shown as in figure 2.1.1. Expression results of only one representative sample are shown in B). The housekeeping gene *ACTB1* was used as a RNA integrity control. A pool of adult and fetal RNAs was used as positive control for the RT-PCR (+ control)

The observed 50% methylation in liver might constitute a hint for allelic specific methylation. When taken together, fully methylated alleles (i.e. maternal allele) and unmethylated alleles (i.e. paternal allele) will give a quantitative signal of 50%. Thus, genes whose imprinting is regulated by DNA methylation will show this kind of mosaic methylation pattern in an allelic-specific manner segregating from parents to their offspring. Alternatively, genes showing a composite pattern of methylation with consecutive CpG sites in the same amplicon being both, methylated and unmethylated will give also a 50% methylation value in direct PCR fragment sequencing.

Hence, I further analyzed the distribution of the observed 50% methylation in *SLC22A1* and *PLG* by sequencing the subcloned PCR products. Very strikingly, I observed that the 50% methylation in liver in both genes is distributed in a mosaic manner, with the CpG sites in the amplicon being either all methylated or all unmethylated in the same subcloned fragment. However, the observed mosaic methylation did not segregate with an annotated SNP neither for *SLC22A1* (rs1867351) nor for *PLG* (rs1950562). These results suggest that the observed tissue-specific methylation in *SLC22A1* and *PLG* is not allelic-specific and, if these genes are imprinted in human healthy liver, DNA methylation is not the responsible epigenetic mechanism.

Figure 2.1.3 shows the allelic distribution of methylation in *SLC22A1* and *PLG*. Moreover, this figure shows the allelic distribution in two regions used to control the selected experimental strategy. *GNAS1* has been reported as imprinted in several tissues [183] [184]. I found that methylation in the T-DMR for this gene shows a mosaic distribution of DNA methylation in liver, with an annotated SNP (rs1800905) segregating with the methylated allele. Unlike, an intergenic region in Chromosome 20p11.23 showed also 50% methylation in direct PCR product sequencing experiments but it is distributed in a composite pattern throughout the CpG sites and subcloned fragments.

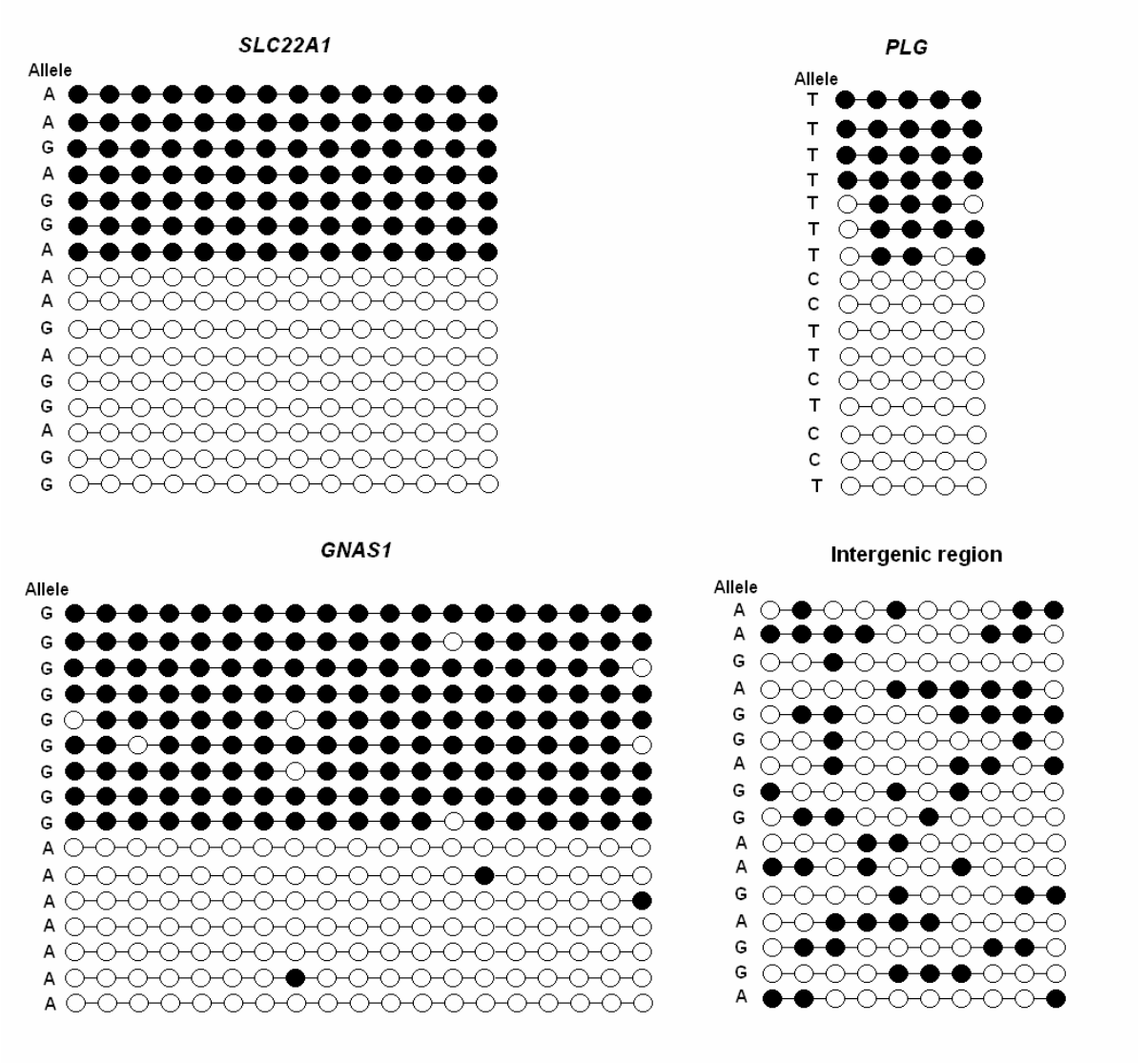


Figure 2.1.3: DNA methylation in *SLC22A1* and *PLG* is distributed in mosaic, but is not allelic-specific in healthy liver. Bisulfite sequencing results of subcloned PCR fragments corresponding to the 5'-UTR of the *SLC22A1*, *PLG* and *GNAS1* gene as well as an intergenic region on chromosome 20p11.23. All PCR fragments displayed 50% methylation by direct bisulfite sequencing (not shown). *SLC22A1*, *PLG* and *GNAS1* display a mosaic distribution while the intragenic region displays a composite pattern. Methylation in *SLC22A1* and *PLG* does not segregate with annotated SNPs indicating that it is not allelic-specific. Unlike, methylation in *GNAS1* segregates with the SNP, being the alleles G and A methylated and unmethylated, respectively. *GNAS1* has been reported as imprinted in healthy liver [183].

To verify that *SLC22A1* and *PLG* are effectively not imprinted in human liver, I analyzed the allelic profile of expression for these genes. I directly sequenced the amplified cDNA fragment in the RT-PCR. By analyzing the expressed genotype of two annotated SNPs contained in these fragments (rs2282143 for *SLC22A1* and

rs1130656 for *PLG*, respectively), I determined that both genes are biallelical expressed in healthy human liver. Figure 2.1.4 shows the obtained trace files for *SLC22A1* and *PLG* in one of the liver sample heterozygous for the studied SNPs at DNA level (genomic PCR fragment, upper panel) and in the expressed mRNA (cDNA fragment, lower panel).

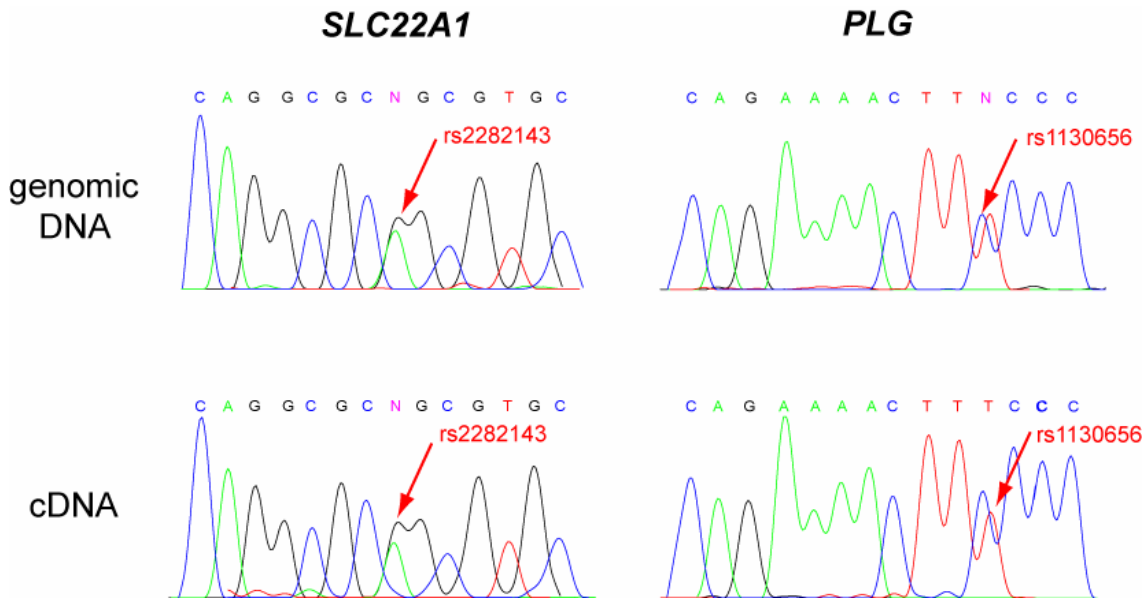


Figure 2.1.4: Biallelic expression of *SLC22A1* and *PLG* in healthy adult liver. Sequencing traces showing the presence of an annotated SNP in genomic DNA (upper panel) and messenger RNA (lower panel). Healthy adult liver samples were heterozygous for the rs2282143 and rs1130656 SNPs in the *SLC22A1* and *PLG* gene, respectively. Matched DNA/RNA samples from three individuals were used.

2.2. Determination of differential transcriptional and DNA methylation profiles in fetal and healthy adult lung

2.2.1. Experimental design

In the previous part, I have studied the extent of the correlation between DNA methylation and gene expression in healthy adult tissues in a candidate gene approach. Complementarily, genome-wide expression profiles using microarrays offer a different approach to this question, by selecting genes that are differentially expressed and subsequently studying the methylation status at their 5' regions.

First, I determined the expression profiles in 4 individual fetal lung and 4 individual healthy adult lung samples using Affymetrix microarrays. After strict upfront quality controls of the samples under study, as well as the produced raw data, I analyzed the produced profiles using the Robust Microarray Analysis (RMA) method. For a gene being considering as differentially expressed, I fixed a selection criterion of absolute log2 fold changes (Log2Fc) higher than 2. In that manner, genes at least 4-times highly expressed in the corresponding tissue were above the cutoff value. In total, 99 genes were highly expressed in Fetal Lung and 354 were highly expressed in Healthy Adult Lung.

Next, I determined the methylation status of 43 differentially expressed genes to study the putative role of DNA methylation in establishing the obtained transcriptional profiles. I selected the genes according to their measured differential expression either in fetal or healthy adult (Log2Fc>2, table 2.2.1) and high CG content of their 5' UTR (>50%). I designed PCR fragments covering these CG rich regions and determined their methylation status in unrelated fetal (n=3) and healthy adult lung (n=3) samples by direct bisulfite sequencing. Highly methylated 5'-UTR regions correlated with decreased expression levels in 19% of the cases.

The results presented in this section were submitted for publication and are currently under editorial review [Cortese R. et al, *Correlative gene expression and DNA methylation profiling in lung development nominate new biomarkers in lung cancer*, submitted to The International Journal of Biochemistry and Cell Biology]

Table 2.2.1. Expression and DNA methylation in 43 differentially expressed genes during lung development

Gene Symbol	*Probeset*	*Upregulation*	*Log2Fc*[1]	*Methylation FL*[2]	*Methylation HAL*[2]	*Wilcoxon test*[3]
PLK2	201939_at	Fetal	2,52	0%	0%	p>0.05
SOX11	204914_s_at	Fetal	3,98	0%	0%	p>0.05
COL11A1	37892_at	Fetal	3,37	0%	20%	p<0.05
GRB14	206204_at	Fetal	2,53	10%	10%	p>0.05
MEOX2	206201_s_at	Fetal	3,04	0%	20%	p<0.05
SERPINE2	212190_at	Fetal	2,37	10%	30%	p<0.05
SOX9	202935_s_at	Fetal	2,36	0%	20%	p<0.05
COL2A1	213492_at	Fetal	2,59	10%	10%	p>0.05
TRO	211700_s_at	Fetal	2,66	40%	30%	p>0.05
ROR1	211057_at	Fetal	2,3	10%	10%	p>0.05
FBN2	203184_at	Fetal	3,66	20%	50%	p<0.05
MEST	202016_at	Fetal	3,08	40%	40%	p>0.05
MDK	209035_at	Fetal	2,63	40%	70%	p<0.05
FGFR3	204379_s_at	Fetal	2,2	90%	100%	p>0.05
GOLGIN67	208798_x_at	Fetal	2,59	90%	90%	p>0.05
C5orf13	201310_s_at	Fetal	3,83	0%	0%	p>0.05
IGF2BP3	203819_s_at	Fetal	3,82	0%	0%	p>0.05
NID2	204114_at	Fetal	3,28	10%	10%	p>0.05
RAMP	218585_s_at	Fetal	4,04	10%	10%	p>0.05
SOX4	213668_s_at	Fetal	2,47	10%	10%	p>0.05
TOP2A	201291_s_at	Fetal	2,73	0%	0%	p>0.05
COL21A1	208096_s_at	Fetal	2,39	20%	20%	p>0.05
FANCL	218397_at	Fetal	2,41	10%	10%	p>0.05
GRP	206326_at	Fetal	3,38	10%	10%	p>0.05
COL1A1	202311_s_at	Fetal	3,6	20%	40%	p<0.05
IGF2	202409_at	Fetal	2,86	90%	90%	p>0.05
CYP1A1	205749_at	Healthy Adult	3,65	10%	10%	p>0.05
HCK	208018_s_at	Healthy Adult	3,27	0%	0%	p>0.05
SLC39A8	209267_s_at	Healthy Adult	4,47	0%	0%	p>0.05
FBP1	209696_at	Healthy Adult	3,17	10%	10%	p>0.05
TU3A	209074_s_at	Healthy Adult	3,76	20%	40%	p<0.05
WISP2	205792_at	Healthy Adult	3,07	30%	30%	p>0.05
PRELP	204223_at	Healthy Adult	2,94	60%	40%	p>0.05
SERPINA1	202833_s_at	Healthy Adult	5,8	40%	40%	p>0.05
MARCO	205819_at	Healthy Adult	3,44	80%	60%	p<0.05
PTGDS	211748_x_at	Healthy Adult	3,39	60%	70%	p>0.05
SCNN1A	203453_at	Healthy Adult	4,31	70%	90%	p>0.05
LAPTM5	201720_s_at	Healthy Adult	3,28	90%	50%	p<0.05
SFTPD	214199_at	Healthy Adult	4,78	50%	50%	p>0.05
BHLHLB3	221530_s_at	Healthy Adult	2,81	0%	0%	p>0.05
MT1M	217546_at	Healthy Adult	3,56	10%	10%	p>0.05
SLC15A3	219593_at	Healthy Adult	2,2	30%	30%	p>0.05
MSR1	214770_at	Healthy Adult	3,49	100%	100%	p>0.05

[1]: Results from one representative probeset corresponding to the detailed gene.
[2]: Methylation values at the 5′ UTR. Methylation values represent the amplicon mean methylation.
HAL: Healthy Adult Lung (n=3); **FL**: Fetal Lung (n=3)
[3]: p-values obtained with Wilcoxon test comparing fetal and healthy adult lung samples

In an expression profiling experiment using Affymetrix GeneChip Arrays the sources of variability in descending order are: biological, sample preparation (total RNA isolation as well as labeling), and system (instrument and arrays) [185]. Quality of the total RNA prior to starting the experiment and of the products during the labeling procedure was thoroughly tested as explained in further sections. In the other hand, system noise is negligible when using GeneChip Array and there was no need for technical replicates [186] in the well-calibrated systems at Epigenomics.

To address the biological variability, I have faced several constraints regarding the number of the available samples. Fetal human tissues are very difficult to obtain, especially in quality for RNA isolation; therefore, number of available samples was limited. All used fetal lung RNA samples were obtained from 20-25 weeks old fetuses. At this fetal age, lungs are in the Canicular and Saccular stages of development when pre-alveolar saccules expand and are subdivided in alveoli [187]. Both, female and male, individuals were included in both groups.

In order to obtain statistical significant results, I used 4 independent samples for each group in the microarray experiment (biological replicates). I assumed that the intra-group variation was lower than the inter-group variation, since both are non-cancerous tissues expected to be homogeneous in terms of cellular composition. Retrospective unsupervised clustering analysis considering all the probesets in the array demonstrated that this assumption was correct (Figure 2.2.1). Furthermore, I validated the obtained results by single gene quantitative RT-PCR in 8 independent samples for each group, which gives universality to the discovered differences.

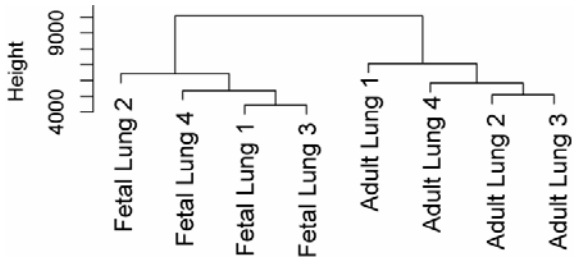


Figure 2.2.1: Inter-group variation in the expression profiles is higher than intra-group variation. Unsupervised clustering of all normalized signals. Fetal and healthy adult samples clustered separately showing higher inter- than intra-group variation. Dendrogram was drawn using Manhattan distance and average linkage hierarchical clustering. The height measures represent the distance between individual samples and groups of samples.

2.2.2. Determination of expression profiles in Fetal and Healthy Adult Lung

Upfront quality control of the samples under study

As described in Material and Methods, I started the experiment by controlling the quality of the total RNA samples used for the microarray analysis and the gene candidate approach. Quality of the starting material is an essential feature in expression experiments and samples, which do not pass upfront quality controls, should be excluded from the analysis. Conservation of the biological material prior to RNA isolation, i.e. snap freezing or chemical treatments as RNA later ™, is crucial to preserve the integrity of the RNA, as well as the transcriptional status of the analyzed genes. Working with RNA samples isolated from tissues instead of cultured cells, make this point very challenging and therefore, I have followed very strict standards to evaluate the quality of the procured samples after receiving them. All 24 total RNA samples included in the analysis showed very good quality in terms of RNA integrity and purity.

Target preparation, hybridization and scanning. First Order quality control

After checking the quality of the included samples, I prepared cRNA from 4 fetal lung and 4 healthy adult lung samples according to the protocols stated in Material and Methods.

After scanning, an image file (.dat file) was generated that allowed a first-hand estimation of the hybridization quality. First, arrays did not present image artifacts, i.e. High/low intensity spots, scratches, high regional, or overall background, etc., under ocular inspection of the image. Moreover, Oligo B2 spiked in the hybridization cocktail served as a positive hybridization control. Its hybridization was highlighted on the image by the pattern of intensities on the corners and edges, as well as the array name printed on the array.

To further control the quality of the hybridization procedure, a Hybridization Report (.rpt file) was generated containing data regarding the efficiency of hybridization of

positive controls contained in the arrays as well as spiked controls added to the hybridization cocktail.

Table 2.2.2 contains the more relevant values of the Hybridization Report for the analyzed samples. All samples showed very good performances for all the controls, indicating that produced microarray data was of very good quality and truly represent the biological status of the studied samples.

Table 2.2.2: Hybridization report values

	Fetal 1	*Fetal 2*	*Fetal 3*	*Fetal 4*	*Adult 1*	*Adult 2*	*Adult 3*	*Adult 4*
Noise (RawQ)	2.59	2.04	2.44	1.87	2.52	1.87	2.33	2.21
Background	61.68	58.22	55.34	57.16	63.01	56.84	57.21	55.54
Present calls	13128 (58.9%)	11831 (53.1%)	13231 (59.4%)	11616 (52.13 %)	12513 (56.2%)	11573 (51.9%)	12977 (58.2 %)	11831 (53.1%)
Housekeeping Controls[1]								
GAPDH	11450.2 (P)	12762 (P)	9700.3 (P)	13175.3 (P)	16527.6 (P)	9815.1 (P)	8657.7 (P)	9138.8 (P)
ACTB1	11523.9 (P)	18704.2 (P)	13368.4 (P)	14811.9 (P)	13513.7 (P)	18369.2 (P)	16354.5 (P)	12123.4 (P)
Ratios (3'/5')								
GAPDH	1.10	1.21	1.17	1.34	1.04	0.92	1.02	1.19
ACTB	2.03	1.67	1.75	1.68	1.83	1.80	1.68	2.87
Spiked Controls[1]								
Probe Set								
BIOB	581.9 (P)	727.2 (P)	480.5 (P)	676.9 (P)	504.9 (P)	440 (P)	482.7 (P)	738.9 (P)
BIOC	2466.1 (P)	2495.5 (P)	2063.2 (P)	2559.9 (P)	2016.2 (P)	2284 (P)	1803.6 (P)	2571 (P)
BIODN	5550.9 (P)	6011.5 (P)	4181.5 (P)	5304.7 (P)	3853.0 (P)	4278 (P)	3431.2 (P)	5153.8 (P)
CREX	25258.8 (P)	20528.3 (P)	17027.9 (P)	27224.2 (P)	17759.6 (P)	21835.6 (P)	12835.9 (P)	23833.6 (P)

[1]: Only data corresponding to 5′ probesets is shown. (P): Present signal

Several values are relevant in the evaluation of the Hybridization Report and they must display values within accepted ranges [185]. Noise (RawQ) values represent the distribution of intensity of signal across the array. This value depends on particular factors for each scanner (i.e. electrical properties) and may vary from an instrument to the other, even in the same lab. Data used in the analysis should show comparable Noise values. Background values depend on the efficiency of washing and staining procedure and must range from 20 to 100 in expression arrays. Number of present calls will depend on the type of analyzed sample. However, very low values may

indicate poor sample quality and number of present calls must be comparable among the samples of the same biological group. Calls for all probesets corresponding to two housekeeping gene (*GAPDH* and *ACTB*) must be present in high intensity. Furthermore, the ratio between intensities corresponding to their 5′ and 3′ probesets must be lower than 3, otherwise indicating sample degradation. Finally, controls BIOB, BIOC, BIODN and CREX are spiked in increasing concentrations and must display therefore increasing signal intensities, indicating right sensitivity of hybridization and probe detection.

Taken together, results of the first order quality control analysis demonstrated that the produced data was of very high quality, indicating that the analysis could be followed without discarding any microarray of the set.

Quality controls of the produced raw data

As a first step in the RMA procedure, raw data must be controlled in order to determine whether it fits to the assumptions of the model [188] [189]. All methods used were implemented in the *affy*, *affyPLM* and *simpleaffy* packages of the Bioconductor project (www.bioconductor.org)

All analyzed arrays displayed very good results for all the quality controls and passed all the quality criteria, indicating that no array must be excluded from the analysis.

First, I controlled that the produced data fits to the proposed "probe-level model". This is a linear model in which probeset level parameters are usually factor variable for each probe. Chip level parameters could be factor variable grouping the arrays into treatment/status groups. In that manner, standard errors between the expected and observed values for each probeset are represented as box plots for each array (Figure 2.2.2, panel A). The distributions of the standard errors were comparable between all the arrays, indicating that the data set fits to the proposed model. Moreover, I represented the weights from the linear robust linear model fit as pseudo chip images (Figure 2.2.2, panel B). This graphical representation allows detecting spatial artifacts on the chip. As no microarray displayed poor quality areas bigger than 10% than the total surface, I continued the analysis including all the arrays.

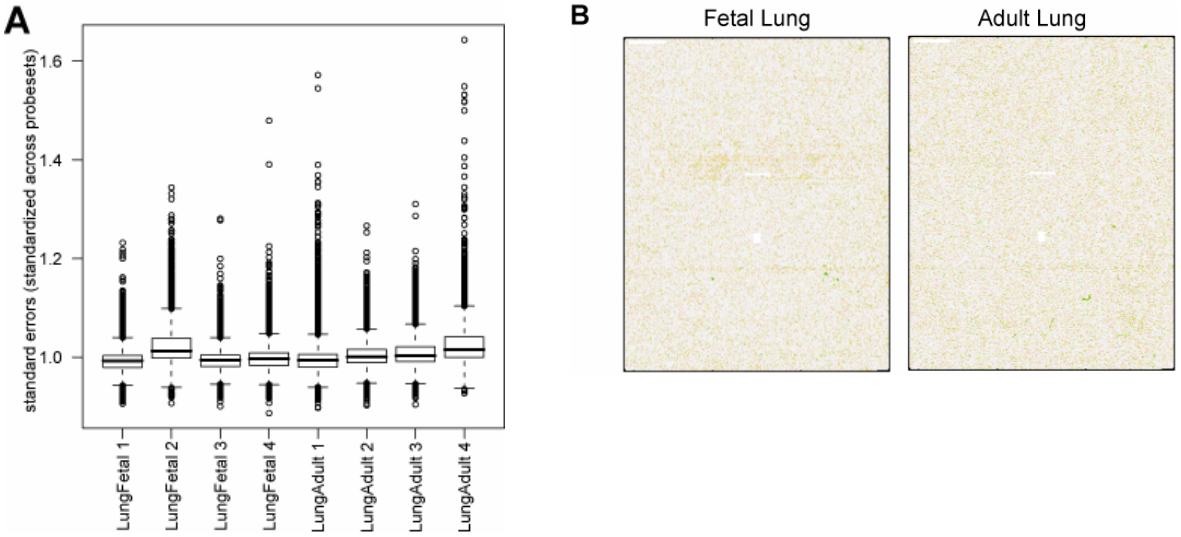


Figure 2.2.2: Produced expression data fits to the "probe-level model". A) Box plot of the model standard errors (standardized across probesets). The distributions are comparable between arrays. Narrow distributions suggest good quality of the produced data. B) Pseudo chip image of weights from the robust linear model fit. Areas of low weight are greener (poor quality), high weights are light gray. The absence of large green areas (more than 10% of the array surface) depicts good quality of the produced data. Only one representative array per group is shown.

RMA analysis and determination of expression profiles

After the quality control of the produced raw data, I determined the expression profiles of Fetal and Adult Lung. All method used were implemented in the *affy*, *simpleaffy* and *multtest* packages of the Bioconductor project.

First, I normalized the obtained raw data, according to the RMA method [189]. After normalization, distribution of the data for each data set displayed a similar distribution of the standard errors (data not shown) and Fetal and Adult arrays cluster separately in the unsupervised hierarchical clustering (Figure 2.2.1).

Plot of average intensity (A) versus estimated change (M), "MA-plot", showed the majority of the probe sets scattering around zero (Figure 2.2.3). Several probesets located above or below a cutoff line with corresponding to 2 fold Log2Fc. These probesets represented the genes highly expressed in Fetal Lung (Log2Fc>2) or in Adult Lung (Log2Fc<-2), respectively.

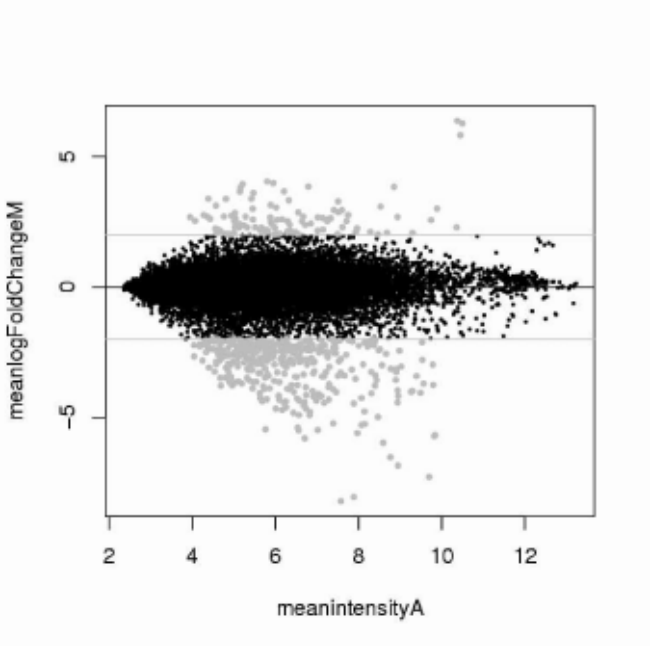


Figure 2.2.3: Cutoff at Log2Fc separates differentially expressed genes. Average intensity A plotted vs. estimated log-fold change M (log (fetal) - log (Adult)). Grey dots represent probesets selected after fitting the cutoff at Log2Fc=2.

In order to assign a statistical significance to the selected candidate genes, I applied multiple testing procedures, as assigned in the *multtest* package. Applied tests included Westfal &Yound maxT algorithm, Benjamini and Hochberg false discovery rates (FDR) and Bonferroni corrected two-sample Welch t-statistics for unequal variances, reviewed in [190].

As selection criteria for the differentially expressed genes, I have chosen a minimum absolute Log2Fc of 2. According to these criteria, 99 genes were highly expressed in Fetal Lung (Log2Fc<-2) whereas 354 were highly expressed in Healthy Adult Lung (Log2Fc>2).

Figure 2.2.4 graphically represents the differential expression profiles obtained for fetal and healthy adult lung. Hierarchical clustering of the differentially expressed 458 probesets demonstrated that expression values clearly distinguish both groups.

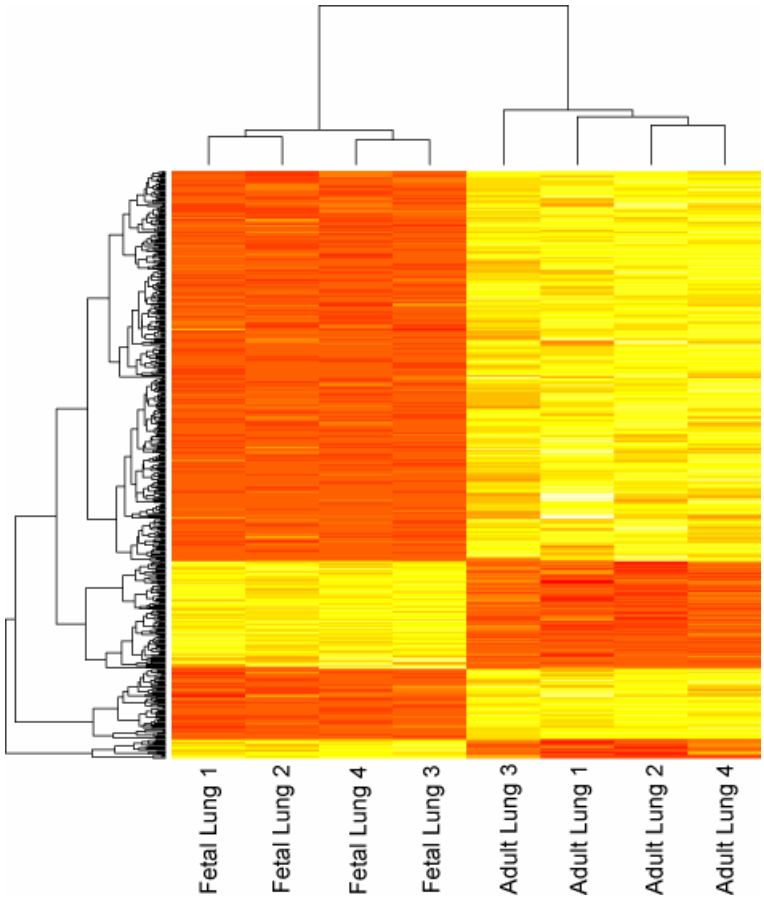


Figure 2.2.4: Expression profiles of fetal and healthy adult lung. Hierarchical clustering of 458 probesets found to be differentially expressed between fetal and healthy adult lung (Log2Fc >2 after multiple testing, see text for details). Columns represent each analyzed sample whereas probesets are arranged in rows. Expression values were obtained by RMA [189]. Expression values are represented in a color code ranging from orange (downregulation) to yellow (upregulation)

2.2.3. Validation of expression profiles by qRT-PCR

In order to validate the obtained profiles, I analyzed the expression of 21 differentially expressed genes in 8 independent samples for each group by quantitative reverse-transcription PCR (qRT-PCR).

Figure 2.2.5 summarizes the result of the validation experiment by showing the Log2Fc values obtained in the microarray and qRT-PCR experiments. 15 genes (71%) displayed coincident expression profiles in fetal and adult tissues with both methods. In no case, I found a gene that displayed significant higher expression in one tissue in the microarray experiment, while significant higher expression in the opposite tissue by qRT-PCR. Observed differences in the expression profiles might be due to

technical differences between the methods or different sensitivities of the statistical analysis used in each case. However, as samples used in each setup were independent, these results might also indicate differences in the expression of a particular gene in each included individual.

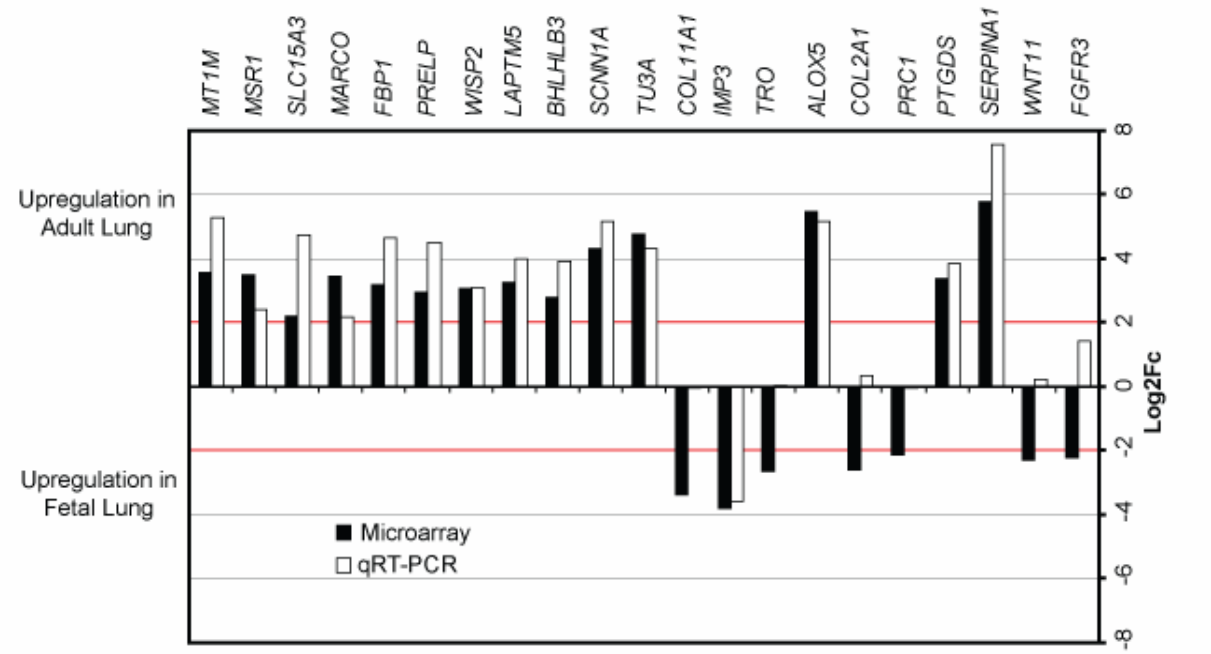


Figure 2.2.5: Validation of expression profiles by qRT-PCR. Comparison of expression values for single gene measurement as obtained by microarray (black bars) and quantitative reverse transcription analysis (qRT-PCR) (white bars). 15 out of 21 genes (71%) displayed coincident expression profiles in fetal and adult tissues with both methods. No gene displayed significantly higher expression in opposite groups when both measurements are compared. Microarray results correspond to one representative probeset per gene. qRT-PCR values were calculated in 8 independent samples per group and processed with the Q-Gene algorithm [191].

Taken together these results indicate that the determined expression profiles in the microarray experiment are representative of the expression status of more than the half of the analyzed genes in fetal and adult lung in general. Furthermore, the fact that no gene displayed significant upregulation in opposite tissues when using two different methods in independent samples, suggests that the determined patterns in the microarray experiment are consistent with the biological expression profiles of these tissues in general.

The chosen method for qRT-PCR involves the relative quantification of transcribed mRNA for a particular gene in two biological groups. In order to normalize expression differences between tissues, expression results of the target gene must be calibrated against the expression of a reference gene, which is equally expressed in both tissues. To identify a gene that does not display significant differences in expression between fetal and adult lung, I screened a panel of 7 putative housekeeping genes, *GAPDH*, *ACTB*, *UBC*, *RPL13A*, *PSMB6*, *ATP50* and *POLR2F*, with the BestKeeper algorithm proposed by Pfaffl and colleagues [191]. Although all seven genes displayed equivalent variances in the analyzed samples (data not shown), I decided to use *GAPDH* as a calibrator. This gene displayed the most robust results in the real time PCR amplification, in terms of standard variation between replicates and shape of the dissociation curves.

2.2.4. Pathway analysis

To gain further insight into the biological function of the expression profiles, I analyzed the data with the MAPPFinder software [192] thereby relating the expressed genes to each term of the Gene Ontology hierarchy. Table 2.2.3 shows the MAPPFinder results for 10 most over-represented groups ranked by their corresponding Z- score groups. Groups over-represented in genes upregulated in fetal lung were related to extracellular matrix components: "fibrillar collagen" group (GO ID: 5583, 6/9 changed genes, Z-score: 25.735); "collagen type I" (GO ID: 5584, 3/3 changed genes; Z-score: 22.348); "collagen" (GO ID: 5581, 8/27 changed genes, Z-score: 19.603); "extracellular matrix structural constituent" group (GO ID: 5201, 11/69 changed genes Z- score: 15.942). Groups identified as highly over-represented in genes upregulated in adult lung were related to immunological processes: the "Complement activation, classical pathway" (GO ID: 6958, 6/23 changed genes, Z-score: 8.125), the "antigen processing\, exogenous antigen via MHC class II" (GO ID: 19886, 10/13 changed genes, Z-score: 19.023) and "antimicrobial humoral response" (GO ID: 19735, 8/82 changed genes, Z-score: 6.275).

Table 2.2.3: Over-represented GO groups in genes upregulated in fetal and healthy adult lung

GO Name	GOID	GO Type[1]	Number Changed	Number Measured	Number in GO	Percent Changed	Percent Present	Z Score
Groups over-represented in genes upregulated in fetal lung								
fibrillar collagen	5583	C	6	9	9	66.67	100.00	25.74
collagen	5581	C	8	27	32	29.63	84.38	19.60
extracellular matrix structural constituent	5201	F	11	69	85	15.94	81.18	16.60
phosphate transport	6817	P	9	61	89	14.75	68.54	14.39
inorganic anion transport	15698	P	9	114	146	7.89	78.08	10.17
anion transport	6820	P	9	139	175	6.47	79.43	9.05
extracellular matrix (sensu Metazoa)	5578	C	13	307	377	4.23	81.43	8.39
extracellular matrix	31012	C	13	311	383	4.18	81.20	8.32
sensory perception of mechanical stimulus	50954	P	6	83	97	7.23	85.57	7.87
sensory perception of sound	7605	P	6	83	97	7.23	85.57	7.87
Groups over-represented in genes upregulated in healthy adult lung								
response to biotic stimulus	9607	P	98	764	912	12.83	83.77	21.74
defense response	6952	P	94	730	870	12.88	83.91	21.32
immune response	6955	P	87	657	769	13.24	85.44	20.83
Antigen processing\. exogenous antigen via MHC class II	19886	P	10	13	22	76.92	59.09	19.02
MHC class II receptor activity	45012	F	9	12	21	75.00	57.14	17.81
Antigen presentation\. exogenous antigen	19884	P	9	12	21	75.00	57.14	17.81
response to external biotic stimulus	43207	P	61	468	544	13.03	86.03	17.10
Response to pest\. pathogen or parasite	9613	P	60	458	512	13.10	89.45	17.01
response to external stimulus	9605	P	66	617	713	10.70	86.54	15.56
response to stimulus	50896	P	110	1552	2053	7.09	75.60	15.07

[1]: GO group type according to GO nomenclature: C= cellular location; F= molecular function; P= biological process

2.2.5. Correlation between DNA methylation and gene silencing in fetal and adult lung

To study a possible epigenetic regulation of the identified differentially expressed genes, I obtained methylation profiles of the 5'- UTR for 43 genes using direct bisulfite sequencing. I have selected these genes according to their measured differential expression either in fetal or healthy adult (Log2Fc>2, table 2.2.1) and high CG content of their 5' UTR (>50%). Figure 2.2.6 displays examples of the 5'-UTR methylation for several genes while table 2.2.1 displays the obtained methylation values in combination with the calculated Log2Fc for each gene. Out of the 43 analyzed genes, 8 (19%) exhibited a correlation between DNA methylation of the 5' UTR and downregulation of the cognate gene. Among these genes, 6 (*COL11A1*, *MEOX2*, *SERPINE2*, *SOX9*, *SERPINE2*, *FBN2* and *MDK*) were highly expressed and hypomethylated in fetal lung and 2 (*LAPTM5* and *MARCO*) highly expressed and hypomethylated in adult lung (figure 2.2.6 and table 2.2.1).

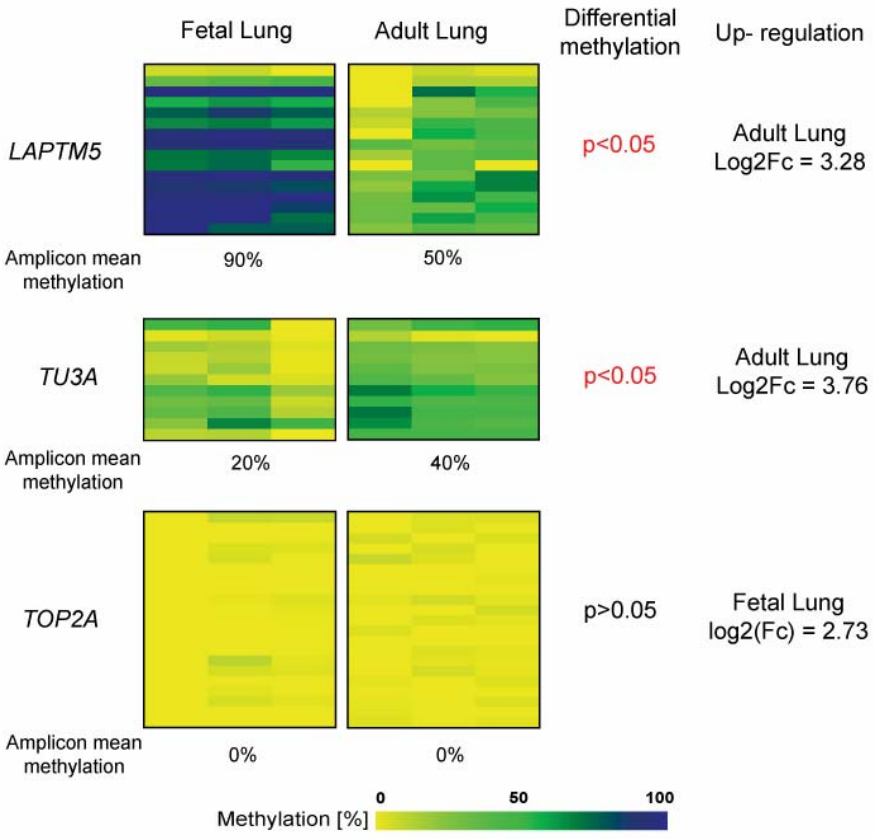


Figure 2.2.6: DNA methylation profiles in fetal and healthy adult lung. 8 out 43 (19%) exhibited a correlation between DNA methylation of the 5′ UTR and downregulation of the gene. Representative results for three genes are shown. *LAPTM5* is highly expressed (Log2Fc=3.28) and displayed significantly lower methylation in adult lung ($p<0.05$). *TU3A* showed higher expression (Log2Fc=3.76) and significantly higher methylation in some CpG positions in adult lung ($p<0.05$ compare to figure). *TOP2A* is highly expressed in fetal lung (Log2Fc=2.73) and is unmethylated in healthy fetal and adult lung ($p>0.05$). Statistical differences were assessed using the Kruskal-Wallis test [194].

For example, the lysosome-associated protein, transmembrane-5 gene (*LAPTM5*) is highly expressed (Log2Fc=3.28) in adult lung and displayed lower methylation in adult lung (90% and 50% amplicon mean methylation in fetal and adult lung respectively, figure 2.2.6). Interestingly, 2 genes (5%) (*SCNN1A* and *TU3A*, figure 2.2.6 and table 2.2.1) showed higher methylation in their respective 5′-UTRs in tissues where the cognate genes were also highly expressed. This finding may point to a positive regulation of expression by DNA methylation. The majority of the genes (33 genes (76%), table 2.2.1) displayed no changes in their methylation profiles, being hypermethylated or unmethylated in fetal and healthy adult lung.

2.2.6. DNA methylation and expression profiling of bi-directional promoters

The distribution of the genes throughout the genome might influence their transcriptional profiles. By aligning full length cDNA sequences, Trinklein and colleagues [193] identified a class of divergently transcribed pairs, whose TSSs are separated by less than 1000 bp. The authors also reported that these "head-to-head" genes represent more than 10% of the transcriptome, thus any phenomena affecting their regulation might influence the phenotype. Authors showed that some of these genes are co-regulated, whereas others exhibit complementary expression profiles.
In order to study a putative role for DNA methylation in the regulation of bi-directional promoters, I selected 20 pairs of reported "head-to-head" genes, studied their co-expression out of the results of the genome-wide expression profiles from fetal and healthy adult lung and analyzed the methylation profiles of the shared

regions at their 5'-UTRs by direct bisulfite sequencing. Out of the 20 regions, 17 contained an annotated CpG island (Ensembl v41) and 3 did not, although they exhibited a high CG content. Table 2.2.4 summarizes the results obtained for DNA methylation analysis of these regions in fetal and adult lung samples and the expression values for the annotated "head-to-head" genes.

Table 2.2.4: DNA methylation and expression in 20 pairs of "head-to-head" genes

Gene Symbol (- strand)	*Log2Fc*[1]	*Gene Symbol (+ strand)*	*Log2Fc*[1]	*CpG island*	*Methylation Adult Lung*[2]	*Methylation Fetal Lung*[2]
ARFRP1	0.19	*ZGPAT*	0,37	Yes	25%	25%
NTHL1	0.39	*TSC2*	0,5	Yes	0%	0%
THYN1	0.01	*ACAD8*	0,16	Yes	0%	0%
DERL2	0.01	*MIS12*	0,09	Yes	0%	0%
MATN4	0.08	*RBPSUHL*	0,04	No	100%	100%
TAX1BP3	0.23	*TMEM93*	0,02	Yes	0%	0%
CEBPA	0.54	*FLJ12355*	0,01	Yes	0%	0%
ZCWPW1	0.16	*BCDIN3*	0,18	Yes	0%	0%
CXorf2	0.06	*TKTL1*	0,21	No	100%	100%
LSM1	0	*BAG4*	0,11	Yes	0%	0%
THAP10	0.18	*LRRC49*	0,8	No	0%	0%
MOAP1	0.29	*C14orf130*	0,5	Yes	0%	0%
PSMA7	0.27	*SS18L1*	0,67	Yes	0%	0%
SLC25A11	0.17	*RNF167*	0,08	Yes	0%	0%
SUPT7L	0.44	*SLC4A1AP*	0,02	Yes	0%	0%
EME2	0.28	*NUBP2*	0,01	Yes	0%	0%
RECQL	0.53	*GOLT1B*	0,33	Yes	0%	0%
CABP7	0	*UCRC*	0,13	Yes	0%	0%
ARFIP2	0.51	*FXC1*	0,28	Yes	0%	0%
GGA3	0.29	*MRPS7*	0,12	Yes	0%	0%

[1] Only the results for one representative probeset per gene are shown
[2] Results are binned in four groups (0, 25, 50 and 100%)

Most regions were unmethylated (0-25%) and only two pairs, pair 5 (*MATN4-RBPSUHL* genes) and pair 9 (*CXORF2-TKTL1* genes), were hypermethylated. Noteworthy these two pairs did not contain a CpG island. The third pair not containing a CpG island, pair 11 (*THAP10- LRRC49* genes), displayed unmethylation (0%) similarly to CpG island- contained regions. I did not find any differential methylation between fetal and adult lung samples in these regions. Figure 2.2.7 illustrates the obtained results showing the position of the studied regions and their methylation status in fetal and adult lung tissues for regions containing (panel A) or not containing (panel B) a CpG island.

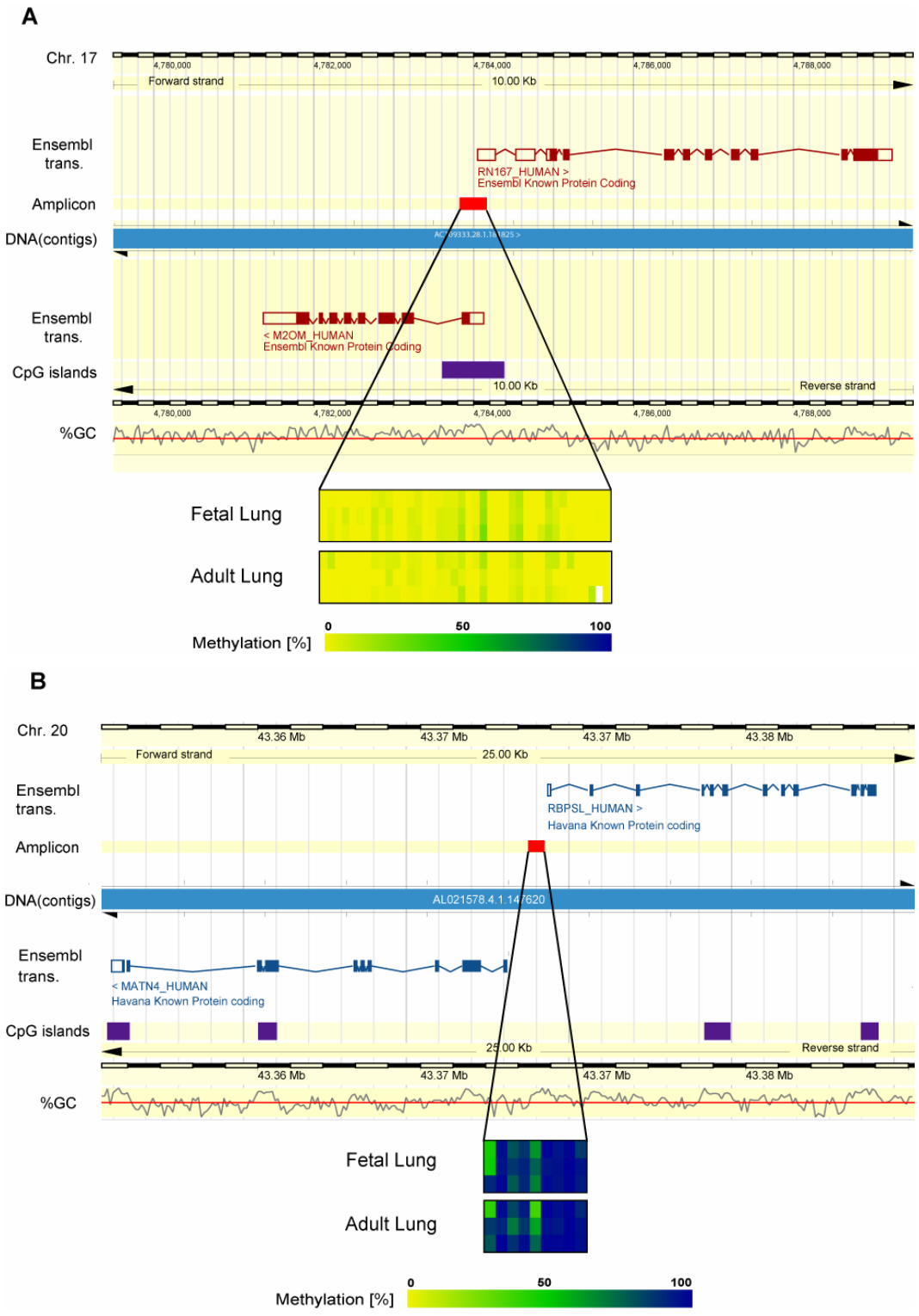
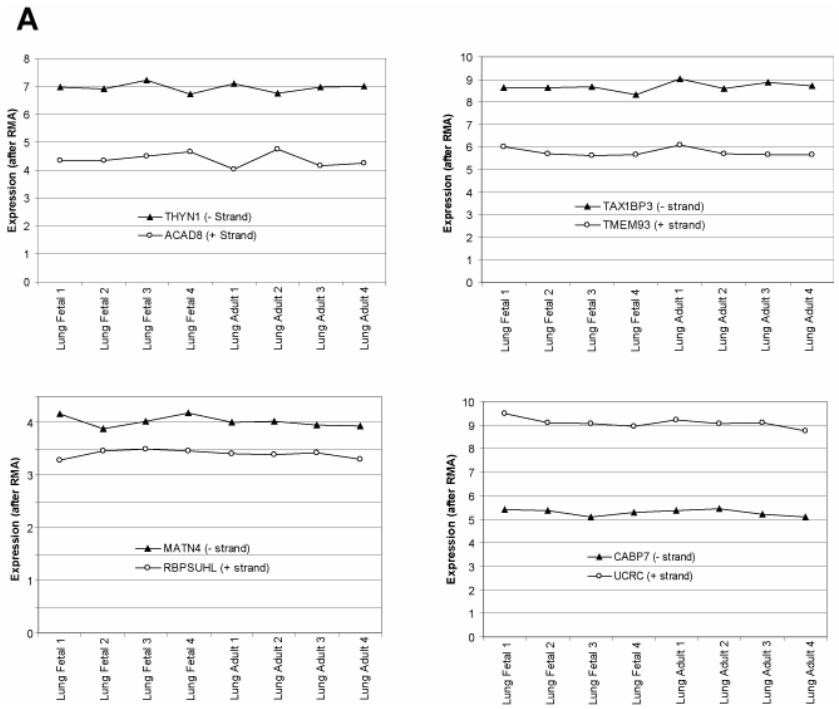


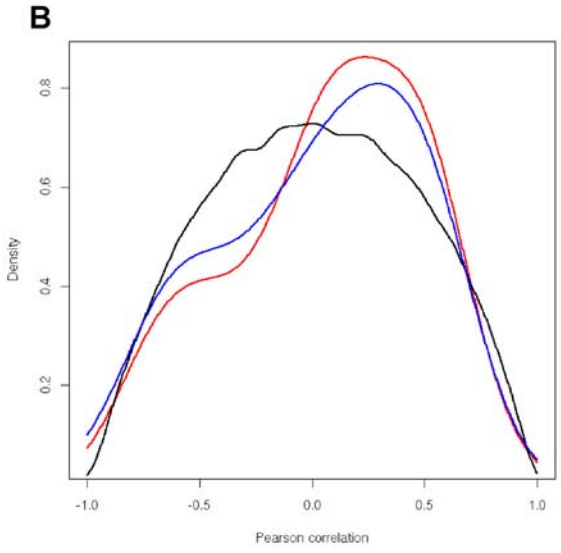
Figure 2.2.7: DNA methylation of bi-directional promoters in fetal and healthy adult lung. A) Bi-directional promoters containing an annotated CpG island are unmethylated (<25%) in fetal and healthy adult lung. Representative result for a "head-to-head" gene pair is shown (*RN167* and *M2OM* in the + and – strands, respectively). B) 2 out of 3 bi-directional promoters not containing a CpG island are hypermethylated (>90%) in fetal and healthy adult lung. Representative result for a "head-to-head" gene pair is shown (*RBPSL* and *MATN4* in the + and – strands, respectively).

To look for trends suggesting co-expression of the selected "head to head" genes, I analyzed their expression using the microarray expression profiles obtained from fetal and healthy adult lung samples in the previous part of this work. Paired genes were similarly expressed in both tissue groups showing Log2Fc values between 0 and 0.8. Genes in "head-to-head" gene pairs did not show major variances between their expression values when compared individually. Figure 2.2.8 A illustrates the comparison in fetal and healthy adult lung by showing 4 "head-to-head" pairs. Expression values were similar for genes in opposite strands, despite differences attributable to differences in hybridization efficiencies for individual probesets, indicating that "head-to-head" genes may co-express in fetal and healthy adult lung. To address this possibility, I calculated the Pearson correlation coefficient for the 20 "head-to-head" genes contained in table 2.2.4 and compared to randomly generated pairs of genes in different chromosomes (n=100,000). Figure 2.2.8 B shows that expression in studied pairs (blue line, median = 0.4) is more correlated than in random pairs (black line, median = 0). In addition, I calculated the Pearson correlation coefficients for all bi-directional genes listed by Trinklein and colleagues [193] included in the HG-U133A microarray (red line, median= 0.38, n=57 pairs). Similarities in distribution of correlation coefficients between the selected 20 pairs and those from Trinklein and colleagues suggested that studied pairs are representative of the overall expression of "head-to-head" genes.

Comparing the expression values for each "head-to-head" gene pair with the methylation of the respective 5'-UTR, I found that genes containing a hypermethylated putative promoter exhibited the lowest expression values in both strands (Figure 2.2.8 C). Genes corresponding to hypermethylated promoters were downregulated, regardless the strand where they were located (left panel) and the tissue type (right panel). This last observation constitutes a strong evidence for role of DNA methylation controlling the co-expression of "head-to-head" genes.

Taken together the results of my experiments suggest that "head-to-head" genes are co-expressed in fetal and healthy adult lung and that methylation of the bi-directional promoter may influence this co-expression.


A
Expression (after RMA)
THYN1 (- Strand)
ACAD8 (+ Strand)
TAX1BP3 (- strand)
TMEM93 (+ strand)
MATN4 (- strand)
RBPSUHL (+ strand)
CABP7 (- strand)
UCRC (+ strand)
Lung Fetal 1
Lung Fetal 2
Lung Fetal 3
Lung Fetal 4
Lung Adult 1
Lung Adult 2
Lung Adult 3
Lung Adult 4


B
Density
Pearson correlation

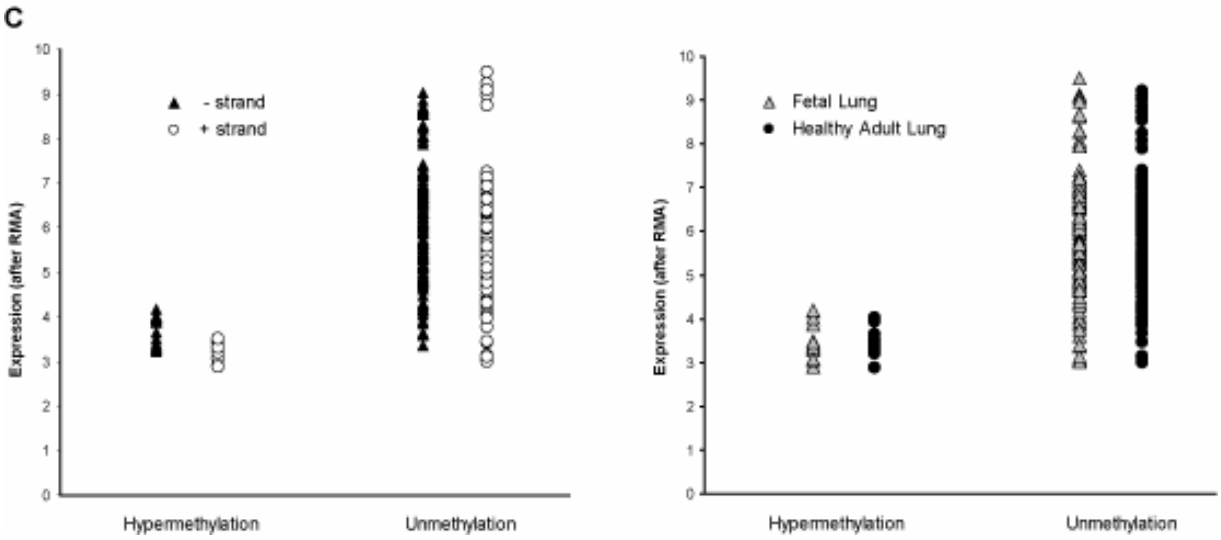

C
Expression (after RMA)
- strand
+ strand
Fetal Lung
Healthy Adult Lung
Hypermethylation
Unmethylation

Figure 2.2.8: Methylation and expression data suggest co-expression of "head-to-head" genes in healthy adult and fetal lung. A) "Head-to-head" genes are equally expressed in fetal and healthy adult lung. Representative results for four pairs are shown. Black triangles represent individual expression values for genes in the – strand and white circles those for genes in the + strand. As lines run parallel across individual sample measurements, no evidence against co-regulation is detected. B) Expression values in lung are correlated in genes in opposite strand for "head-to-head" gene pairs but not in randomly paired genes. Pearson coefficients higher than 0 indicate positive correlation while coefficients equal to 0 indicate no correlation. Blue line represent "head-to-head" gene pairs listed on table 2.2.4 (n=20), red line correspond to other head-to-head genes reported by Trinklein and colleagues [193] also present in the HG-U133A array (n=57) and black line represents randomly paired genes in opposite strand and different chromosomes (n=100,000). C) "Head-to-head" genes corresponding to hypermethylated bi-directional promoters are downregulated, regardless coding strand and tissue type. Expression values for hypermethylated and unmethylated bi-directional promoters are shown. Genes in + and – strand are highlighted in left-side plot (black triangles and white circles, respectively). Values for genes in fetal (gray triangles) and healthy adult lung (black circles) are shown in the right-side plot.

2.2.7. Differentially expressed genes in lung development as methylation marker in lung cancer

Current knowledge suggests that lung cancer cell ontogeny is determined by the consequences of gene transcriptional activation and/or repression events that recapitulate events important in embryonic lung development [157]. However, nothing is known about the role of the epigenetic control of gene expression on this regulation.

I studied the DNA methylation status of the 43 candidate genes, already studied in fetal lung, in lung cancer. I profiled a panel of non-related 15 lung small cell lung carcinoma samples (12 adenocarcinomas and 3 squamous cell carcinomas) and 5 healthy adult lung DNA samples. Table 2.2.5 summarizes the results of the DNA methylation analysis obtained in this analysis together with the methylation data obtained for fetal samples (section 2.2.5)

Table 2.2.5: DNA methylation values for candidate genes in healthy adult and tumor lung samples

Gene Symbol	*Methylation HAL*[1]	*Methylation MDT*[1]	*Methylation PDT*[1]	*Methylation FL*[1]	*Kruskal-Wallis test* [2]
PLK2	0%	0%	0%	0%	p>0.05
SOX11	0%	0%	0%	0%	p>0.05
COL11A1	20%	20%	20%	0%	p>0.05
GRB14	10%	0%	0%	10%	p>0.05
SERPINE2	30%	50%	50%	10%	p>0.05
SOX9	20%	50%	50%	0%	p>0.05
COL2A1	10%	0%	0%	10%	p>0.05
TRO	30%	0%	0%	40%	p>0.05
ROR1	10%	20%	20%	10%	p>0.05
FBN2	50%	50%	50%	20%	p>0.05
MEST	40%	50%	50%	40%	p>0.05
MDK	70%	80%	50%	40%	p<0.05
FGFR3	100%	100%	100%	90%	p>0.05
GOLGIN67	90%	100%	100%	90%	p>0.05
C5orf13	0%	0%	0%	0%	p>0.05
IGF2BP3	0%	0%	0%	0%	p>0.05
NID2	10%	0%	0%	10%	p>0.05
RAMP	10%	0%	0%	10%	p>0.05
SOX4	10%	0%	0%	10%	p>0.05
TOP2A	0%	0%	0%	0%	p>0.05
COL21A1	20%	20%	20%	20%	p>0.05
FANCL	10%	20%	20%	10%	p>0.05
GRP	10%	20%	20%	10%	p>0.05
COL1A1	40%	50%	50%	20%	p>0.05
IGF2	90%	100%	100%	90%	p>0.05
CYP1A1	10%	0%	0%	10%	p>0.05
HCK	0%	0%	0%	0%	p>0.05
SLC39A8	0%	0%	0%	0%	p>0.05
FBP1	10%	20%	20%	10%	p>0.05
TU3A	40%	50%	50%	20%	p>0.05
WISP2	30%	30%	30%	30%	p>0.05
PRELP	40%	50%	50%	60%	p>0.05
SERPINA1	40%	50%	50%	40%	p>0.05
MARCO	60%	50%	50%	80%	p>0.05
PTGDS	70%	70%	70%	60%	p>0.05
SCNN1A	90%	100%	100%	70%	p>0.05
LAPTM5	50%	50%	80%	90%	p<0.05
SFTPD	50%	50%	50%	50%	p>0.05
BHLHLB3	0%	0%	0%	0%	p>0.05
MT1M	10%	0%	0%	10%	p>0.05
SLC15A3	30%	30%	30%	30%	p>0.05
MSR1	100%	100%	100%	100%	p>0.05

[1]: Methylation values at the 5′ UTR. Methylation values represent the amplicon mean methylation. **HAL**: Healthy Adult Lung (n=5); **MDT**: Moderately differentiated tumor (n=6); **PDT**: Poorly differentiated tumor (n=4); **FL**: Fetal Lung (n=3)

[2]: p-values obtained with Kruskal-Wallis test comparing healthy adult and differentially differentiated tumor samples

In general, all tumor and healthy lung samples displayed low methylation values at their 5′-UTR, ranging from 0 to 50 % methylation and only for 5 genes (*IGF2*, *MDK*, *GOLGIN67*, *FGFR3* and *MARCO*), I found hypermethylation (>70%) of the analyzed region. Methylation profiles were, in general, unchanged between fetal, healthy adult

and tumor samples. Noteworthy, 4 genes displayed patterns of differential methylation among those samples, arising as possible biomarkers for lung cancer: *MEOX2*, *MDK*, *LAPTM5*, and *FGF3R*.

When comparing all lung cancer samples to either healthy fetal or adult lung, *MEOX2* displayed statistically significant differential methylation ($p<0.002$, Kruskal-Wallis test [194]) in lung cancer (Figure 2.2.9). Specifically, *MEOX2* gene displayed mean methylation values of approximately 50%, 20% and 0% for tumor, healthy adult and fetal lung, respectively (Figure 2.2.9).

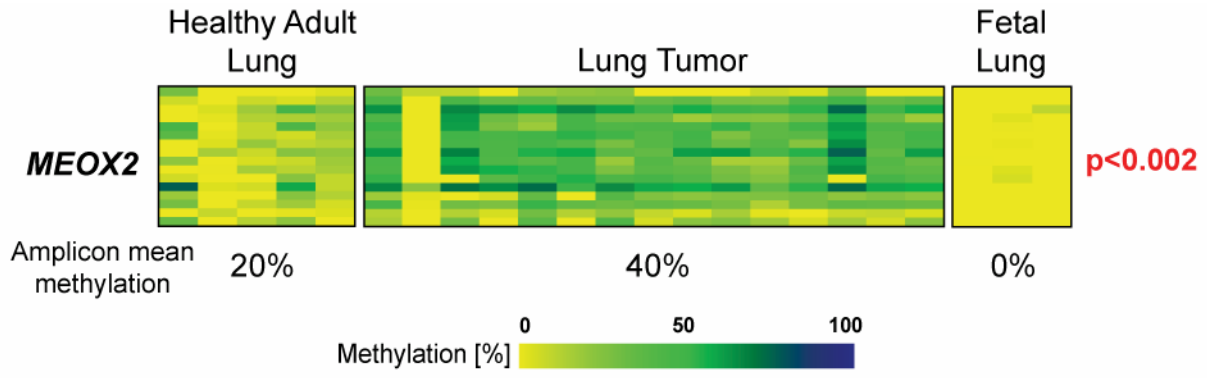


Figure 2.2.9: *MEOX2* is differentially methylated in lung cancer. *MEOX2* displayed higher methylation (mean methylation 40%) in lung cancer samples when compared to healthy fetal and adult lung (amplicon mean methylation 20% and 0%, for healthy fetal and adult lung, respectively) ($p<0.002$). Statistical differences were assessed using the Kruskal-Wallis test [194]

Although several genes were not aberrantly methylated in all lung tumor samples, I observed statistical significant methylation patterns when analyzing the differential methylation in respect to tumor differentiation and tumor type (as assessed by pathological review). For two genes (*MDK* and *LAPTM5*, Figure 2.2.10), less differentiated lung tumors displayed methylation values similar to those observed in fetal samples ($p<0.005$, Kruskal-Wallis test) and differentiated tumors were more similar in their methylation to healthy adult lung tissues. *MDK* exhibited mean methylation values of 40% and 50% in fetal lung and poorly differentiated tumors, respectively, while being 70% and 80% methylated in healthy adult lung and moderately differentiated tumors respectively (Figure 2.2.10, upper panel). *LAPTM5* exhibited mean methylation values of 90% and 80% in fetal lung and poorly

differentiated tumors, respectively, and displayed 50% methylation in both healthy adult lung and moderately differentiated tumors (Figure 2.2.10, lower panel).

The region of differential methylation in *MDK* was located outside of a CpG island at the 5'-UTR (950 bp upstream from TSS). Differentially methylated region in *LAPTM5* was located surrounding the TSS. For *MDK* and *LAPTM5* patterns of differential methylation in lung cancer resemble those observed in fetal lung. The differential methylation observed in less differentiated tumors has a putative application in lung cancer diagnosis, as lack of cellular differentiation is indicative of increased tumor aggressiveness [195].

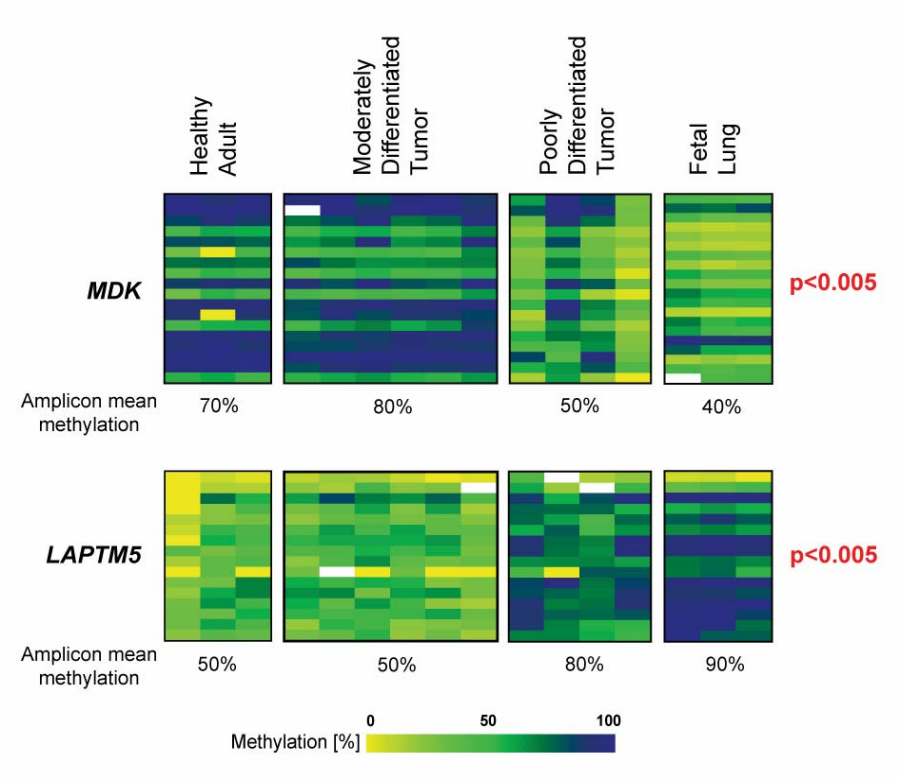


Figure 2.2.10: DNA methylation profiles of *MDK* and *LAPTM5* in poorly differentiated lung tumors resemble those observed in fetal lung. *MDK* (upper panel) displays higher methylation in healthy adult lung similar to moderately differentiated tumors (amplicon mean methylation 70% and 80%, respectively), whereas lower methylation in poorly differentiated tumors and fetal lung (amplicon mean methylation 50% and 40%, respectively). *LAPTM5* (lower panel) displays lower methylation in healthy adult lung and moderately differentiated tumors (amplicon mean methylation 50% for both groups) while being higher methylated in poorly differentiated tumors and fetal lung (amplicon mean methylation 80% and 90%, respectively). Fetal lung profiles in both genes are similar to those in poorly differentiated tumors and significantly different to those in moderately differentiated tumors (p=0.00499, and p=0.00536 in *MDK* and *LAPTM5*, respectively). Statistical differences were assessed using the Kruskal-Wallis test [194]

The *FGFR3* gene was hypermethylated in fetal, healthy adult and most lung cancer samples (Figure 2.2.11 and table 2.2.5). However, when comparing the methylation between different tumor types, I observed that squamous cell carcinoma samples were significantly hypomethylated (mean methylation value 70%; p<0.015, Kruskal-Wallis test) compared to the rest of the analyzed samples (mean methylation values 100%, 90% and 100% for healthy adult lung, fetal lung and adenocarcinoma, respectively).

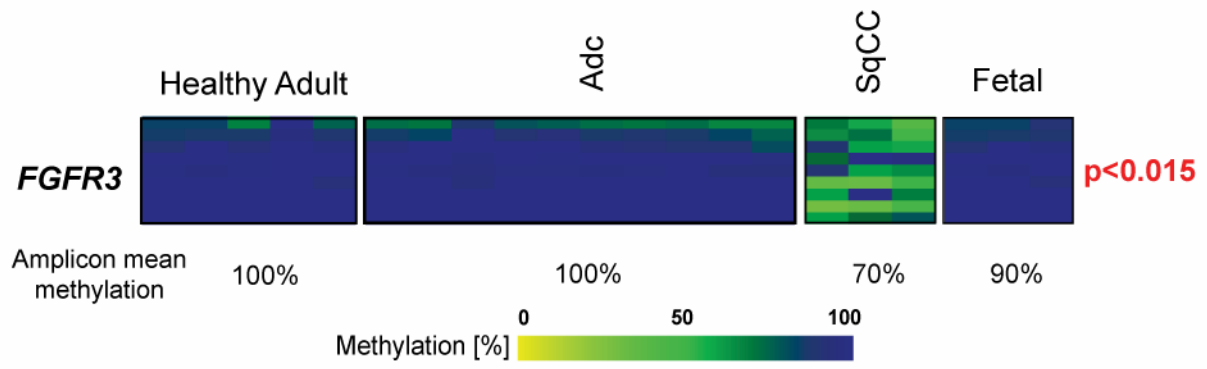


Figure 2.2.11: *FGFR3* is less methylated in squamous cell carcinoma than in healthy adult and fetal lung, as well as adenocarcinoma. *FGFR3* is significantly (p<0.015, Kruskal-Wallis test) hypomethylated in squamous cell carcinoma (SqCC) compared to healthy fetal and adult lung, as well as adenocarcinoma (Adc). The examined region displayed a mean methylation of approximately 70% in SqCC and a mean methylation of 100%, 90% and 100% in healthy adult lung, fetal lung and Adc tumors, respectively.

2.3. Conservation of tissue-specific DNA methylation profiles throughout the evolution

2.3.1. Conservation of tissue-specific DNA methylation patterns in human / mouse orthologues

Syntenic regions between human and mouse chromosomes are mapped in genome browsers such Ensembl (www.ensembl.org). Syntenic blocks display strong conservation of sequence, however little is known about the conservation of tissue-specific DNA methylation profiles. To study their conservation, I have studied the methylation status in 61 human/mouse orthologue regions. By direct sequencing of bisulfite-treated DNA, I determined the methylation profiles in four adult mouse tissues and compared these values with those retrieved from the Human Epigenome Project [67]. Consequently, the results of my experiments presented in this section were published as part of the first phase of HEP [67]

I found that the majority (69.4%) of the studied regions differed by less than 20% methylation. This result suggests a significant conservation of DNA methylation between human genes and their murine orthologues.

Table 2.3.1 shows the studied genes and their methylation values in human and mouse. I included in the study 61 regions with different sequence homology (55-100%) and designed PCR fragments for bisulfite-converted DNA. Among the selected regions, 24 corresponded to 5'-UTR of annotated genes, while 37 corresponded to Conserved Non-Genic regions (CNGs). I determined the methylation status of those regions in four tissues in human and mouse, heart muscle, liver, skeletal muscle and skin. In general, regions that showed hypermethylation and no methylation in all four tissues, displayed similar patterns in human and mouse. Unlike, tissue-specific methylation was conserved only in four orthologous regions.

Table 2.3.1: Conservation of DNA methylation profiles in human and murine orthologues

Position	*Gene Human*	*Heart*	*Liver*	*Sk. Muscle*	*Skin*	*Gene Mouse*	*Heart*	*Liver*	*Sk. Muscle*	*Skin*	*Sequence homology*
5'-UTR	*MYO18B*	50	100	25	100	*Q8C3H0*	100	100	75	100	62.5%
5'-UTR	*GP1BB*	0	0	100	0	*Gp1bb*	25	25	25	25	69.6%
5'-UTR	*HIRA*	25	25	25	25	*Hira*	0	0	0	0	56.7%
5'-UTR	*TOMM22*	0	0	0	0	*Tomm22*	0	0	0	ND	69.4%
5'-UTR	*HORMAD2*	75	75	75	75	*Hormad2*	75	75	75	75	66.7%
5'-UTR	*TIMP3*	100	100	100	100	*Timp3*	100	100	100	100	77.3%
5'-UTR	*CSDC2*	50	75	50	50	*Csdc2*	100	50	50	ND	73.0%
5'-UTR	*OSM*	100	100	10	50	*Osm*	50	75	75	75	61.8%
5'-UTR	*PSCD4*	50	50	50	50	*Pscd4*	50	75	75	75	74.1%
5'-UTR	*CRYBB1*	100	100	100	100	*Crybb1*	100	100	75	75	63.8%
5'-UTR	*SUSD2*	50	25	25	50	*Susd2*	25	25	25	25	69.6%
5'-UTR	*Q6ZRW2*	50	100	25	50	*Zc3h12d*	0	0	0	25	71.8%
5'-UTR	*CCDC28A*	0	0	0	0	*Ccdc28a*	0	0	0	0	63.2%
5'-UTR	*LACE1*	0	0	0	0	*Lace1*	0	0	0	0	65.7%
5'-UTR	*PKHD1*	50	50	50	50	*Pkhd1*	100	50	100	100	65.7%
5'-UTR	*GRIK2*	25	25	25	50	*Grik2*	0	0	25	0	73.2%
5'-UTR	*WISP3*	100	100	100	100	Unannotated	100	100	75	75	59.8%
5'-UTR	*VNN1*	100	50	50	100	*Vnn1*	50	50	50	50	67.9%
5'-UTR	*SLC22A1*	100	50	100	100	*Slc22a1*	100	50	100	100	82.7%
5'-UTR	*TBX18*	25	0	50	25	*Tbx18*	25	25	50	25	59.4%
5'-UTR	*TFAP2A*	25	25	25	25	*Tcfap2a*	25	25	25	25	53.1%
5'-UTR	*TGM3*	75	100	50	100	*Tgm3*	75	50	50	25	81.7%
5'-UTR	***RIN2***	**0**	**100**	**0**	**0**	***Rin2***	**0**	**100**	**0**	**0**	**82.3%**
5'-UTR	*NCOA5*	0	0	0	0	*Ncoa5*	0	0	0	0	57.9%
Intergenic	CNG	75	100	25	75	CNG	100	100	75	75	83.2%
Intergenic	CNG	100	100	50	75	CNG	75	100	50	50	72.9%
Intergenic	**CNG**	**100**	**50**	**50**	**100**	**CNG**	**100**	**50**	**50**	**100**	**93.6%**
Intergenic	CNG	100	100	100	100	CNG	100	100	100	100	88.6%
Intergenic	CNG	100	75	50	50	CNG	100	100	50	50	86.1%
Intergenic	CNG	100	75	50	50	CNG	100	100	75	50	83.9%
Intergenic	CNG	100	50	100	75	CNG	100	100	100	100	86.1%
Intergenic	CNG	100	100	100	100	CNG	75	75	75	75	81.1%
Intergenic	CNG	0	100	0	0	CNG	100	0	25	0	66.8%
Intergenic	CNG	100	100	100	100	CNG	50	50	50	50	63.3%
Intergenic	CNG	75	75	50	75	CNG	75	75	75	75	76.0%
Intergenic	CNG	0	0	0	0	CNG	100	100	100	100	87.4%
Intergenic	CNG	0	0	0	0	CNG	0	0	0	0	77.2%
Intergenic	CNG	0	0	0	0	CNG	0	0	0	0	69.2%
Intergenic	CNG	75	100	50	100	CNG	75	75	50	75	74.5%
Intergenic	**CNG**	**75**	**100**	**50**	**100**	**CNG**	**75**	**75**	**50**	**100**	**77.2%**
Intergenic	CNG	25	100	25	25	CNG	50	50	50	50	71.5%
Intergenic	CNG	25	75	0	25	CNG	50	100	25	50	84.3%
Intergenic	CNG	100	100	100	100	CNG	100	100	100	100	68.5%
Intergenic	CNG	100	100	100	100	CNG	100	100	100	50	83.7%
Intergenic	CNG	100	100	100	100	CNG	100	100	100	100	88.3%
Intergenic	CNG	100	100	100	100	CNG	75	75	75	75	81.8%
Intergenic	CNG	100	100	100	100	CNG	75	75	75	75	84.0%
Intergenic	CNG	100	100	100	100	CNG	100	100	100	100	79.9%
Intergenic	CNG	100	100	100	100	CNG	100	100	100	100	92.6%
Intergenic	CNG	100	100	100	100	CNG	100	100	100	100	95.1%
Intergenic	CNG	75	75	75	75	CNG	75	75	75	75	92.7%
Intergenic	CNG	100	100	100	100	CNG	100	100	75	100	90.7%
Intergenic	CNG	100	100	100	100	CNG	100	100	100	100	94.0%
Intergenic	CNG	100	100	50	75	CNG	100	100	100	100	89.5%
Intergenic	CNG	50	100	50	50	CNG	100	75	50	75	88.9%
Intergenic	CNG	100	100	100	100	CNG	100	100	100	100	72.2%
Intergenic	CNG	100	100	100	100	CNG	100	100	100	100	65.9%
Intergenic	**CNG**	**75**	**100**	**50**	**50**	**CNG**	**75**	**100**	**50**	**50**	**79.6%**
Intergenic	CNG	0	0	0	0	CNG	0	0	0	0	67.8%
Intergenic	CNG	25	0	25	25	CNG	0	0	0	0	74.3%
Intergenic	CNG	0	0	0	0	CNG	0	0	0	0	71.0%

Methylation values were binned in four groups (0, 25, 50 and 100%). ND: Not determined. CNG: Conserved non-genic region. Paired regions displaying conserved tissue-specific methylation are in bold.

I found conservation of the tissue-specific DNA methylation in the 5′ UTR of *RIN2*/*Rin2* genes and in 3 CNGs (Figure 2.3.1). Remarkably, regions displaying conservation of methylation between mouse and human also showed high conservation of sequence (homology higher than 75%).

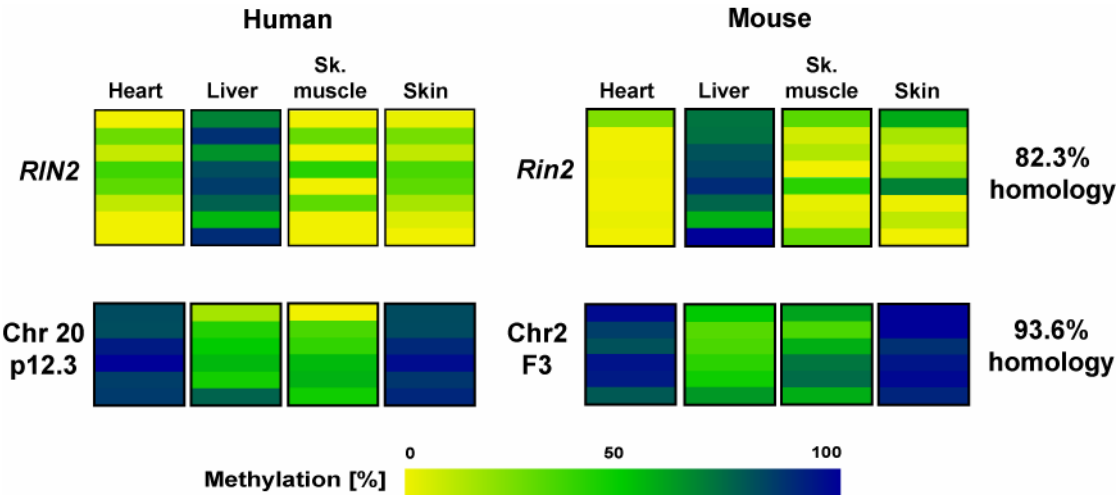


Figure 2.3.1: Conservation of tissue-specific DNA methylation profiles in human and mouse. Conservation of tissue-specific methylation profiles was observed in some gene-associated 5′-UTR and conserved non-genic regions. Two examples are shown: *RIN2*/*Rin2* and Chr20p12.3/Chr2F3 display 82.3% and 93.6% sequence homology, respectively. Samples are displayed column wise with rows representing individual CpGs of the PCR fragment. Quantitative methylation analysis results are shown in a color scale ranging from yellow (~0% methylation), green (~50% methylation) to dark blue (~100% methylation).

Next, I analyzed the data altogether to detect common trends that might apply to the conserved regions. Three different clusters were clearly distinguishable in the correlation plots (Figure 2.3.2). Regions showing hypermethylation and no methylation are clustered and clearly differentiable, however regions of intermediate methylation are much more dispersed (Panel A), suggesting a poorer conservation rate in regions of heterogeneous methylation. In addition, human-mouse pairs with high sequence homology displayed low methylation differences (Panel B). I did not find any difference in conservation neither between the different tissues analyzed (Panel A) nor between regions located in the 5′-UTR and CNGs (Panel B), as they are dispersed similarly in the plot. 1000X resampling with randomization of the human-

mouse pairs demonstrated that the observed trends are not stochastic (Panel C). Furthermore, histogram of frequencies for the differences in methylation showed that low differences in methylation between human and mouse are the most frequently observed (Panel D)

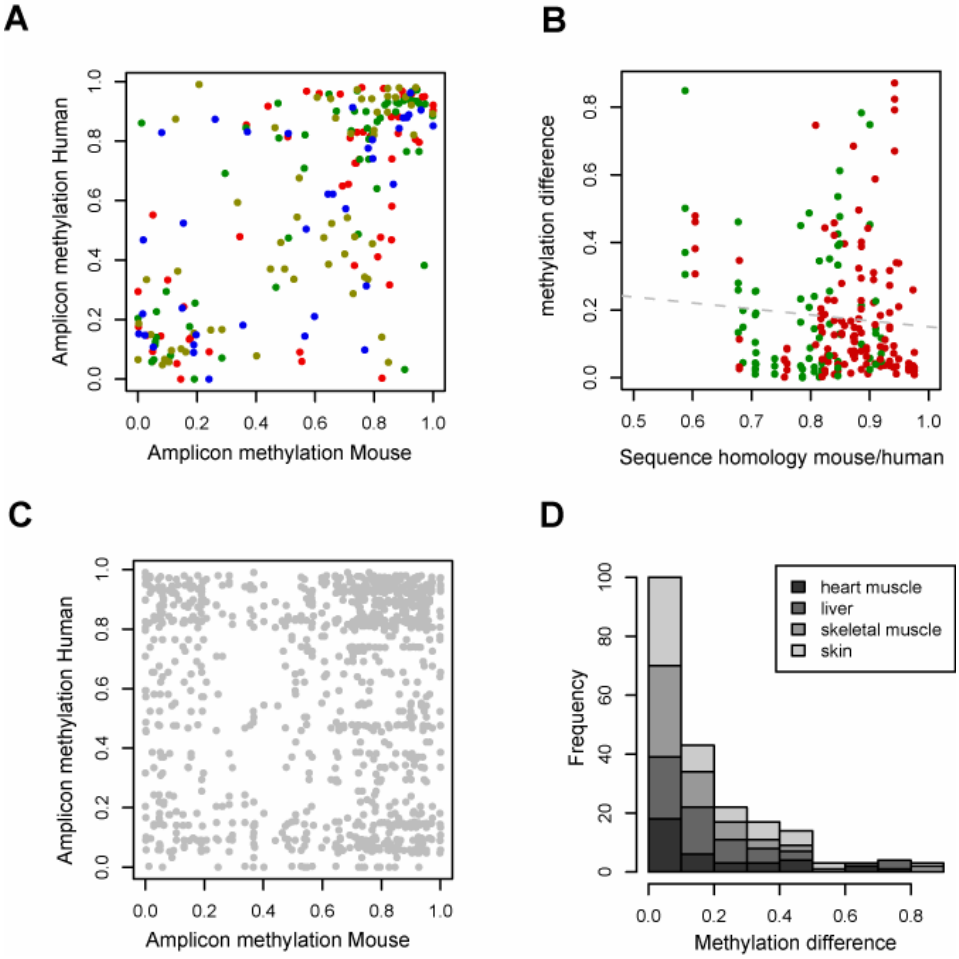


Figure 2.3.2: DNA methylation is highly conserved between human and mouse. Correlation of methylation of CpGs in mouse and human matched by local alignment of conserved regions. A) High and low methylated regions were more conserved than regions with heterogeneous methylation. All studied tissues displayed similar rates of conservation, as they were equally dispersed in the plot Color code: heart muscle (red), liver (green), skeletal muscle (yellow), skin (blue). B) Human-mouse pairs with high sequence homology displayed low methylation differences, with no differences regarding to amplicon location Color code: 5'-UTR (green), CNG (red). (C) Data was resampled 1000 x with replacement randomizing pairs from mouse and human, showing that observed distributions were not stochastic. (D) Histogram of frequencies shows that lower methylation differences are the most frequent in all studied tissues.

2.3.2. Case of study: Methylation and expression in the *IGF2R* locus in human and murine orthologous loci

IGF2R locus is highly conserved between human and mouse. Genes are arranged in identical manner (Figure 2.3.2) and share a high sequence homology. In a previous section, I have determined that two genes in the human locus (*SLC22A1* and *PLG*) displaying novel T-DMRs are not imprinted in healthy adult human tissues (section 2.1.2). To gain more knowledge about the methylation status of genes within this locus and its conservation in their murine orthologues, I analyzed their methylation profiles by direct sequencing of bisulfite-treated DNA in human and mouse tissue samples.

I found that DNA methylation is highly conserved between human and mouse in *SLC22A1*, *PLG*, *SLC22A2*, *SLC22A3* in agreement with the observed high homology of DNA sequence. Unlike, methylation and DNA sequence are low conserved in an imprinting controlling region (ICR) of the *IGF2R* gene. Very interestingly indeed, the tissue-specific differential methylation observed in *PLG* and *SLC22A1* in section 2.1.2 was conserved in mouse.

Figure 2.3.3 is a schematic view of the *IGF2R* locus in human (upper panel) and the *Igf2r* locus in mouse (lower panel) and the location of the designed bisulfite-converted DNA PCR fragments for each gene.

Genes in the *IGF2R* locus in human are arranged in the same manner than in the murine orthologous locus *Igf2r* on chromosome 17A2. I designed bisulfite-converted DNA PCR fragments for a reported ICR on intron 2 of the *IGF2R* gene in human, as well as the 5' regions of the *SLC22A1*, *SLC22A2*, *SLC22A3* and *PLG*. Moreover, I designed PCR fragments for bisulfite-converted DNA in their murine orthologues. The examined regions displayed a sequence homology of 57.5% for *IGF2R/Igf2r*, 77.9% for *SLC22A1/Slc22a1*, 80.4% for *SLC22A2/Slc22a2*, 72.9% for *SLC22A3/Slc22a3* and 65.5% for *PLG/Plg*.

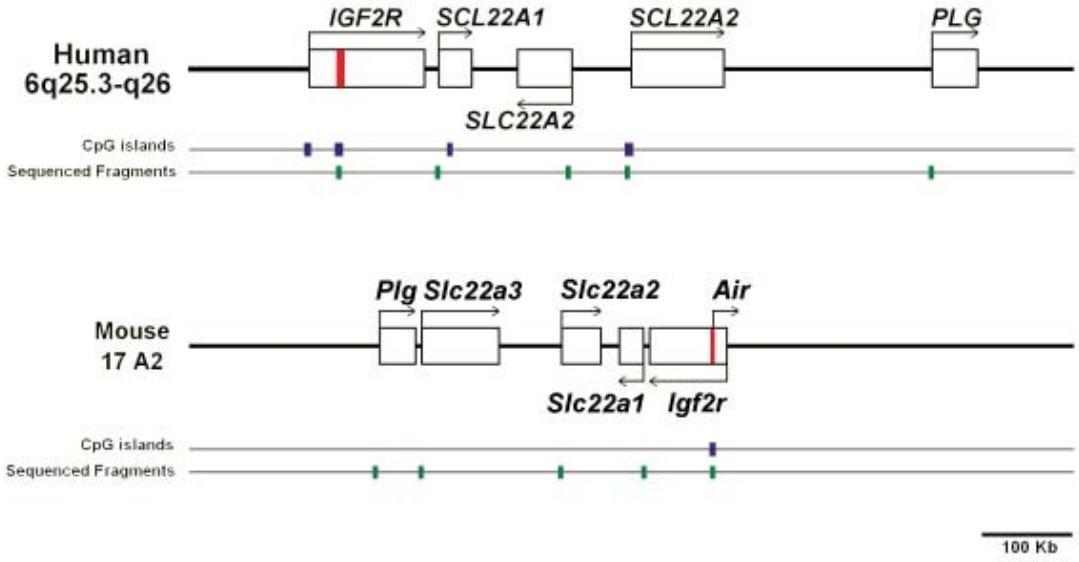


Figure 2.3.3: Genomic structure of the human and mouse IGF2R locus. Schematic representation of the IGF2R locus in human (upper panel) and mouse (lower panel). Red Boxes within the *IGF2R* and *Igf2r* genes represent the annotated imprinted control region (ICR) in intron 2. Purple boxes represent the annotated CpG islands (Ensembl v.40) and the orthologous regions analyzed for DNA methylation are shown as green boxes. The examined regions displayed a sequence homology of 57.5% for *IGF2R*/*Igf2r*, 77.9% for *SLC22A1*/*Slc22a1*, 80.4% for *SLC22A2*/*Slc22a2*, 72.9% for *SLC22A3*/*Slc22a3* and 65.5% for *PLG*/*Plg*.

Figure 2.3.4 shows the methylation status of the studied genes in human (panel A) and mouse (panel B) in heart muscle, liver and skeletal muscle. Observed tissue-specific methylation in *SLC22A1* and *PLG* in human tissues is conserved in mouse. In the same manner, both human and mouse tissues showed hypermethylation in the 5´ regions of *SLC22A2* and no methylation in the 5' regions of *SLC22A3*, respectively. Unlike, ICR in intron 2 of the human *IGF2R* gene was hypermethylated in all the studied tissues, while the murine orthologue, *Igf2r*, displayed 50% methylation in all the studied tissues.

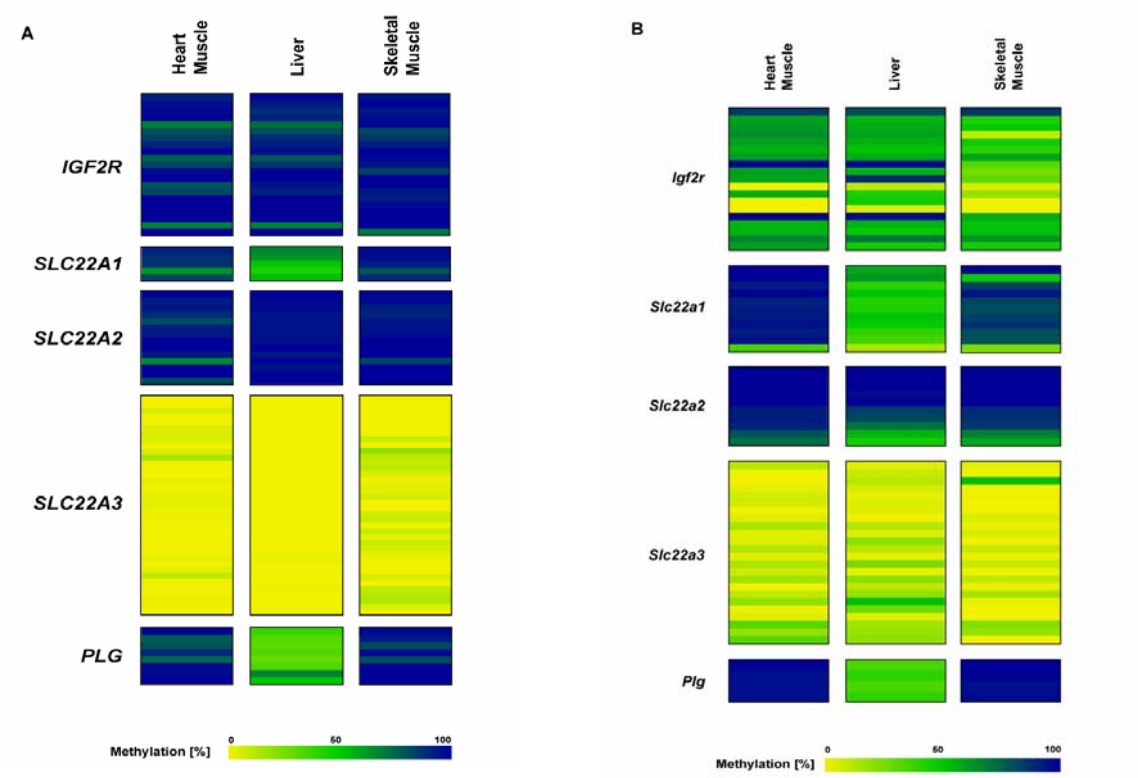


Figure 2.3.4: DNA methylation analysis for genes in the IGF2R locus in human (A) and mouse (B). Tissue specific methylation was only observed in *PLG* and *SLC22A1*. The observed methylation profiles in human are conserved in mouse for *Plg*, *Slc22a1*, *Slc22a2* and *Slc22a3* but not for *Igf2r*. Samples are displayed column wise with rows representing individual CpGs of the PCR fragment. Quantitative methylation analysis results are shown in a color scale ranging from yellow (~0% methylation), green (~50% methylation) to dark blue (~100% methylation).

Moreover, I studied the expression of *IGF2R*, *SLC22A2* and *SLC22A3* genes in human skeletal muscle, liver and heart muscle. While *IGF2R* displayed a highly methylated ICR2 and high expression, *SLC22A2* displayed high methylation at the 5'-UTR in all three tissues and low expression only in skeletal muscle. In turn, *SLC22A3* was highly expressed and displayed unmethylation at the 5'-UTR in all studied tissues (Figure 2.3.5, panel A). In addition, I confirmed the lack of imprinting of these genes by detecting biallelic expression, similarly to that observed in *SLC22A1* and *PLG*. (Figure 2.3.5, panel B)

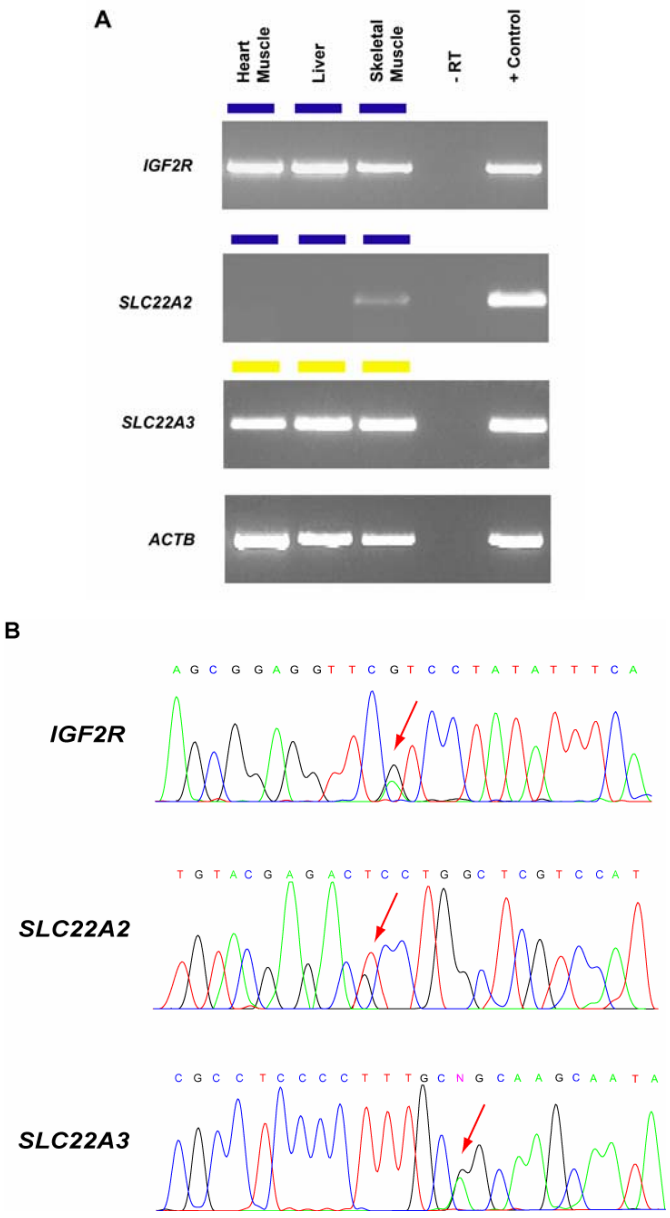


Figure 2.3.5: Biallelical expression of *IGF2R*, *SLC22A2* and *SLC22A3* in healthy adult tissues. A) Results of RT-PCR showing the expression of *IGF2R*, *SLC22A2* and *SLC22A3* genes. Similar amounts of cDNA were used as indicated by control amplification of actin ß (*ACTB*). Representative results for 3 independent samples are shown. Pooled total RNAs derived from tissues and cell lines were used as positive control. Colored bars above each gel represent the methylation status, as indicated in figure 2.3.4. B) Biallelic expression was analyzed by sequencing of the amplified cDNA and identification of heterozygous annotated SNPs. Red arrows mark the position of the annotated SNP in each gene. Annotated SNPs (Ensembl v.40) were rs1570070, rs624249 and rs2292334 for *IGF2R*, *SLC22A2* and *SLC22A3* respectively. Presence of both alleles in the sequenced cDNA indicates biallelical expression

2.3.3. Conservation of DNA methylation upon gene duplications

The results presented in this section have been sent for publication and are currently under editorial review [Cortese R. et al, *DNA Methylation in gene duplications leading to gene families and pseudogenes,* sent to Genomics]

2.3.3.1. Conservation of DNA methylation in the PLG and TBX gene families

PLG gene is a member of a gene family. In order to study whether the observed tissue-specific methylation is conserved in the other members of the family, I analyzed the methylation status of *PLGLA* and *PLGLB*. I found that the 50% methylation in liver and its mosaic-like distribution are conserved in both *PLGLA* and *PLGLB* genes.

Comparative sequence analysis of *PLG* in humans showed high homology to three variants of the *PLG* gene family. The *PLGLA* gene (ENSG00000169659) is located on chromosome 2q12.2 and shows 95.8% homology with the *PLG* gene in its 5' region and exon 1. In turn, two variants of the *PLGLB1* gene (ENSG00000125551 and ENSG00000183281) are located on chromosome 2p11.2 with a homology of 95% and 96%, respectively. Figure 2.3.6 shows the alignment of the 5' regions of the *PLG* gene and the mentioned variants. Primers for the designed bisulfite-converted DNA PCR fragment bound to the regions showing no mismatches between the variants. However, the fragment contained several mismatches that allowed the identification of the variants by sequencing the subcloned PCR fragment (Figure 2.3.6).

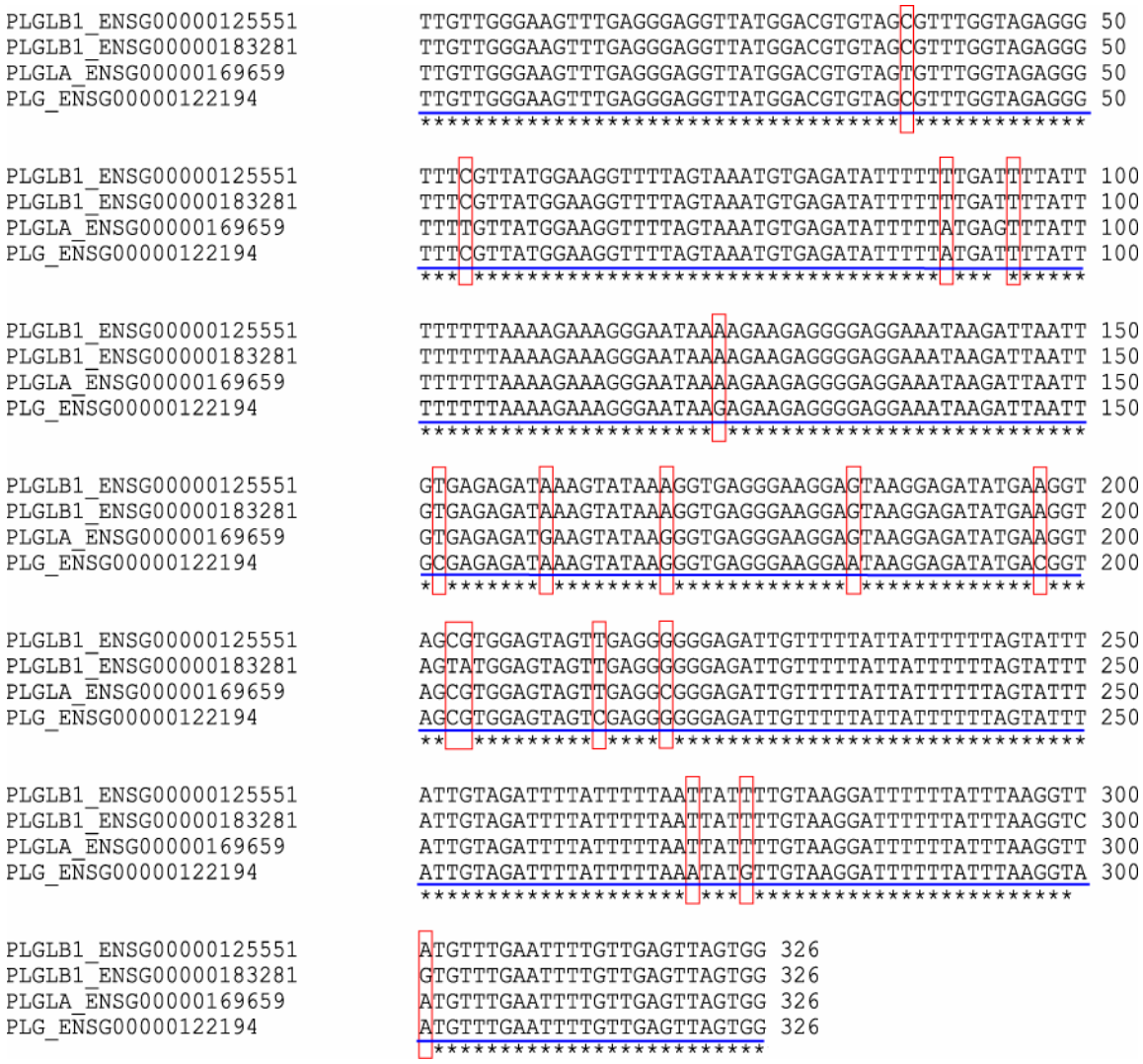


Figure 2.3.6: High sequence conservation among members of the PLG family. *PLG* and its paralogous genes display high sequence homology as evidenced by genomic sequence alignment of the *PLG*, *PLGLA* and *PLGLB* 5′ regions. Sequences analyzed by bisulfite sequencing are underlined. Red rectangles highlight the mismatches between the aligned sequences.

Figure 2.3.7 shows the distribution of DNA methylation for *PLG*, *PLGLA* and *PLGLB* in subcloned PCR products from liver samples. As observed in the parental gene, methylation in *PLGLA* and *PLGB* was distributed in a mosaic manner. Methylation in PLG family clones did not correlate with an annotated SNP in this region (rs4252059), indicating that these genes are not allele-specifically methylated.

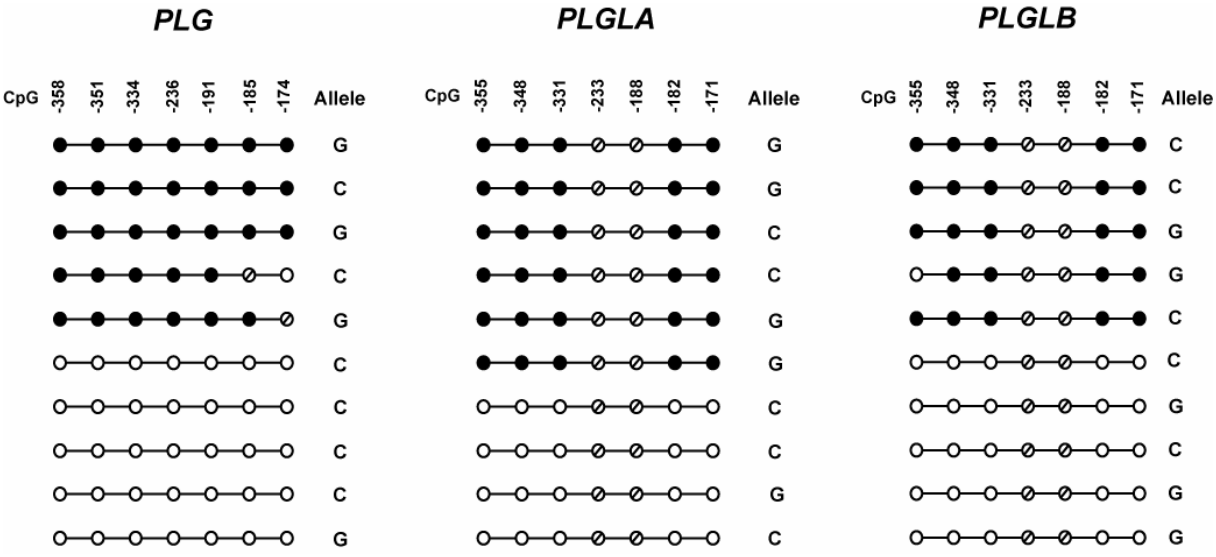


Figure 2.3.7: Mosaic distribution of DNA methylation in *PLG*, *PLGLA* and *PLGLB* in liver. CpG sites in the subcloned PCR products were either all methylated or all unmethylated. Numbers indicate the position of each CpG relative to the corresponding transcription start site (TSS) (Ensembl NCBI 40). Filled and empty circles represent methylated and unmethylated CpGs, respectively. Dashed circles indicate CpG positions lost due either to sequence polymorphism (in *PLG* gene) or sequence mismatches with *PLG* (*PLGLA* and *PLGLB* genes). Alleles were identified by an annotated SNP (rs4252059) within the amplified sequence for *PLG* or the equivalent polymorphism in *PLGLA* and *PLGLB*.

Expression analysis in matched RNA samples by RT-PCR and further sequencing of the cDNA fragments (Figure 2.3.8) revealed that, like *PLG*, *PLGLA* and *PLGLB* are expressed in liver (showing 50% methylation) but silenced in skeletal muscle and heart muscle (showing 100% methylation). The high sequence homology of the transcribed mRNAs made impossible to distinguish between them simply by designing different RT-PCR fragments. To approach this issue, I first designed a fragment, which allowed the amplification of all three variants (Figure2.2.8, panel A). This fragment amplified only in liver samples, suggesting that all three variants were expressed only in liver. The presence of 12 mismatches between the *PLG*, *PLGLA* and *PLGLB* transcripts allowed the verification of the expression of all variants by sequencing the amplified cDNA. Figure 2.2.8 panel B, shows the possible polymorphisms for each studied mismatch. In addition, it shows an example of the obtained trace file. Two mismatches, MM6 (A/G) and MM7 (A/C), confirming the expression of *PLG* and, at least, *PLGLA* or *PLGLB*. Combinatorial analysis with other

mismatches (data not shown) demonstrated that all three variants are expressed in liver.

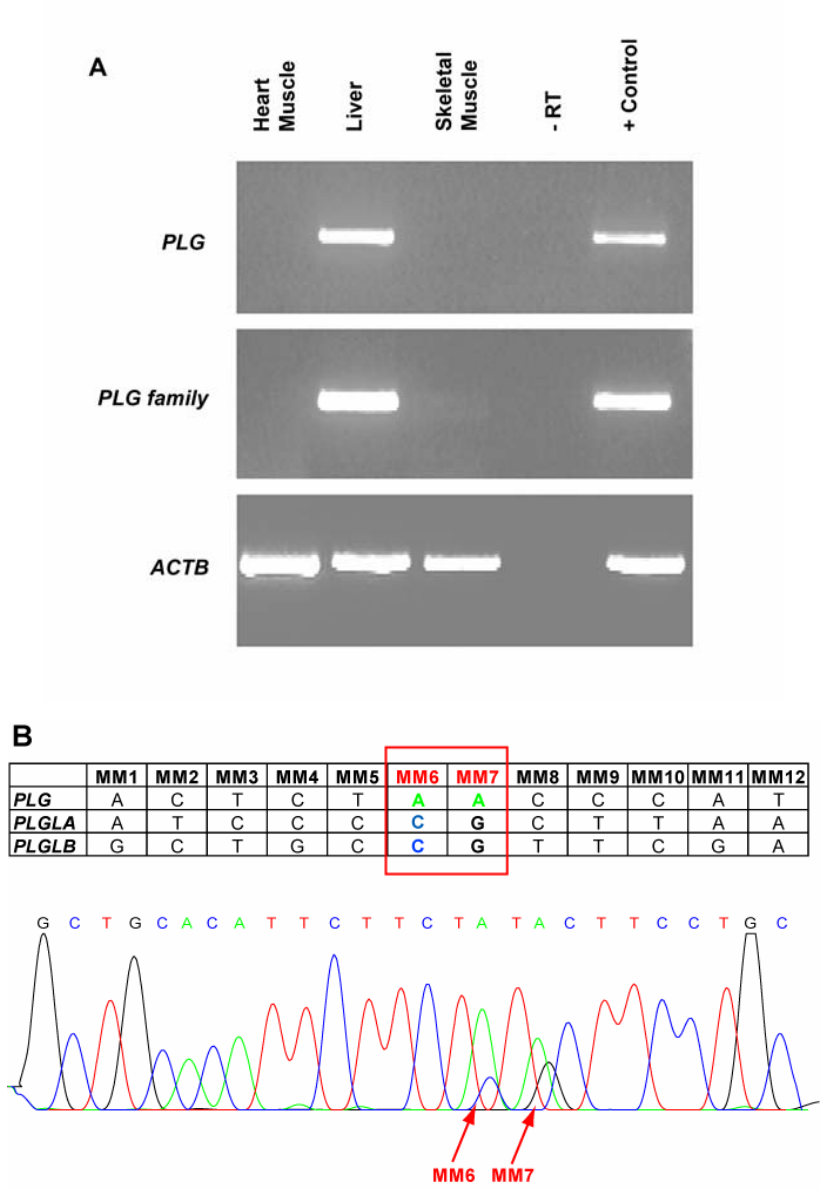


Figure 2.3.8: *PLG* and its paralogous genes are expressed only in liver. A) A fragment allowing the amplification of all three variants (*PLG*, *PLGLA* and *PLGLB*) display the same expression than a fragment specific for *PLG*. B) The combinatorial analysis of 12 mismatches between *PLG*, *PLGLA* and *PLGLB* allowed the verification of the expression of all three variants. Example of only two proximal mismatches (MM6 and MM7, red square and arrows) is shown. Amplified RT-PCR fragment were directly sequenced as detailed in materials and methods.

Hence, all variants displayed a correlation between hypermethylation of the 5' region and the silencing of the gene. The presence of several mismatches in the alignment of

the studied sequences allowed inferring the methylation and expression status of the three variants as well as the identity of the expressed transcript (Figures 2.3.7 and 2.3.8). However, because the two *PLGLB1* variants differ by only one mismatch in the coding region with no homology to *PLG* (data not shown), I could not assess whether there are differences in methylation and expression between both *PLGLB1* variants, ENSG00000125551 and ENSG00000183281.

In contrast to the *PLG* gene duplicates that arose recently in the hominoid lineage, the TBX family arose early during evolution and is present in vertebrates, invertebrates and protostomes (e.g. *Drosophila melanogaster* and *Caenorhabditis elegans*). To further examine the methylation profiles in this large gene family I selected 15 genes from four subfamilies (Tbr1 subfamily: *TBX21*, *EOMES* and *TBR1*; Tbx1 subfamily: *TBX1*, *TBX10*, *TBX22*, *TBX18*, *TBX15* and *TBX20*; Tbx2 subfamily: *TBX2*, *TBX3*, *TBX4* and *TBX5*; Tbx 6 subfamily: *MGA* and *TBX6*) and performed DNA methylation profiling of their 5'-UTR in 8 different tissues and primary cells (Figure 2.3.9). Most of the genes and tissues were unmethylated with *TBX10* being the most prominent exception. Out of these genes, 7 (*TBX10*, *TBX1*, *TBX18*, *TBX15*, *TBX4*, *TBX5* and *TBX21*) were differentially methylated in at least one of the tissues, but none of the genes showed an identical methylation pattern. In particular, differential methylation was evident in genes displaying a shorter evolutionary distance. For example, *TBX1* was hypermethylated in melanocytes, fibroblasts and skeletal muscle while being unmethylated in the remaining tissues. The closely related *TBX10* gene was hypermethylated in most of the tissues but four CpGs (93, 104, 119 and 139 bp upstream from the TSS) were unmethylated. Similarly, genes of the Tbx 2 subfamily were differentially methylated as well. Genes such as *MGA* and *TBR1* that are more distant from other family members displayed no differential methylation.

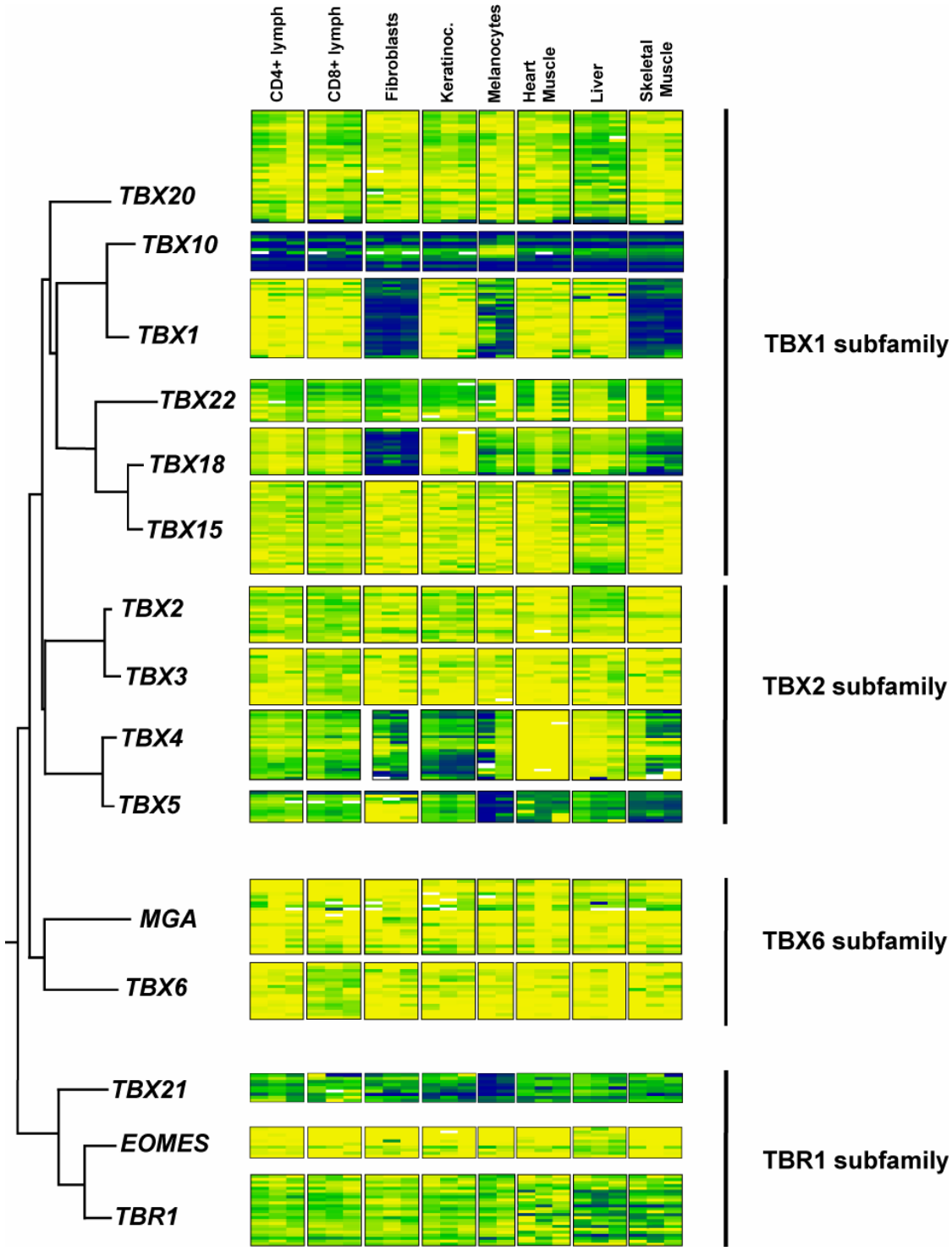


Figure 2.3.9: Methylation profiling of TBX gene family members. Methylation profiles of 15 genes of the TBX transcription factor family and their evolutionary relation are shown. The 15 genes are from four subfamilies (Tbr1 subfamily: *TBX21*, *EOMES* and *TBR1*; Tbx1 subfamily: *TBX1*, *TBX10*, *TBX15*, *TBX18*, *TBX20* and *TBX22*; Tbx2 subfamily: *TBX2*, *TBX3*, *TBX4* and *TBX5*; Tbx 6 subfamily: *MGA* and *TBX6*). Closely related TBX genes display complementary methylation profiles (e.g. *TBX1*/*TBX10* and *TBX2*/*TBX3*). The evolutionary more distant *MGA* and *TBR1* are not differentially methylated in the tissues examined. Phylogenetic relations within the TBX family were previously described in [196].

2.3.3.2. Conservation of DNA methylation in pseudogenes

To further study the conservation of DNA methylation upon gene duplication, I examined the methylation profiles of processed and unprocessed pseudogenes. In agreement with the observed in the PLG gene family, I found that tissue-specific DNA methylation observed in the parental genes is conserved in some unprocessed pseudogenes, which derive from gene duplications. Unlike, I did not find such conservation in processed pseudogenes, which derive from the insertion of retrotransposons in the genome.

Table 2.3.2 summarizes the observed values for DNA methylation in processed and unprocessed pseudogenes and their respective parental genes, as well as the sequence homology in the studied regions.

Table 2.3.2: Methylation values for pseudogenes and their respective parental genes.

Pseudogene Name	Sample 1 Methylation[1]	Sample 2 Methylation	Parental Gene	Sample 1 Methylation	Sample 2 Methylation	Sequence Homology	Distance Ψ–IN	Type Pseudogene
AP000357.2	100%	0%	*ACTR2*	100%	100%	57.10%	0.61995	Processed
AP000357.3	100%	25%	*ARL5A*	0%	0%	92.30%	0.0279	Processed
RP11-758C21.1	100%	50%	*BDH2*	100%	100%	96.60%	0.0285	Processed
RP1-181C9.1	100%	25%	*ANP32B*	0%	0%	71.40%	0.2568	Processed
AP000358.2	100%	25%	*FN3K*	100%	50%	87.70%	0.0525	Processed
CTA-229A8.2	100%	50%	*GAPDH*	50%	50%	62.50%	0.11345	Processed
AC007050.7	100%	100%	*POM121L3*	100%	100%	96.80%	0.0693	Processed
RP3-412A9.4	100%	50%	*SNRPN*	100%	100%	88.10%	0.1189	Processed
KB-1269D1.3	100%	25%	*MAD1L1*	100%	75%	95.10%	0.021	Processed
AC000078.2	100%	25%	*RPL8*	100%	75%	65.90%	1.17195	Processed
AC004019.3	100%	50%	*LOC376522*	75%	50%	84.00%	0.2314	Processed
CTA-373H7.4	100%	0%	*HBLD1*	0%	0%	99.50%	0.0375	Processed
CTA-373H7.4	100%	0%	*HBLD1*	75%	75%	75.80%	0.16795	Processed
RP1-47A17.8	100%	25%	*ADAMTS7*	100%	100%	84.50%	0.13465	Processed
RP1-106I20.2	100%	100%	*NDUFA9*	100%	100%	94.10%	0.07255	Processed
AC004471.4	100%	100%	NM_032028.2	100%	100%	82.80%	0.15665	Processed
RP3-405J24.1	100%	100%	*RPL12*	100%	50%	92.20%	0.06145	Processed
CTA-150C2.8	100%	50%	*APOBEC3G*	100%	100%	64.30%	0.27425	Unprocessed
CTA-246H3.2	0%	0%	*LRP5*	0%	0%	89.10%	0.13055	Unprocessed
RP4-539M6.7	100%	25%	*SLC39A1*	100%	25%	84.80%	0.1258	Unprocessed
KB-1592A4.6	100%	100%	*BCR*	100%	100%	97.20%	0.0212	Unprocessed
KB-1995A5.12	100%	50%	*Q6UW61*	100%	50%	94.40%	0.0183	Unprocessed
RP11-34P13.1	100%	100%	*DDX11*	75%	75%	71.30%	0.192	Unprocessed
RP11-223J15.2	100%	100%	*EEF1A2*	50%	50%	64.30%	0.5787	Unprocessed
RP4-732G19.2	100%	100%	*CYP4Z1*	100%	100%	88.80%	0.147	Unprocessed
RP11-552J9.1	100%	100%	*XAGE2*	100%	100%	72.60%	0.307	Unprocessed
BMS1LP6	100%	100%	*BMS1L*	100%	100%	89.10%	0.12695	Unprocessed
CTSLL3	75%	75%	*CTSL*	75%	75%	89.30%	0.11065	Unprocessed
RP11-432I13.3	75%	100%	*CUBN*	100%	75%	89.40%	0.0983	Unprocessed
RP11-453N3.6	75%	75%	*ABCD1*	75%	75%	95.40%	0.04085	Unprocessed
RP11-392A23.3	50%	50%	*GSTA1*	100%	100%	83.70%	0.1901	Unprocessed
NM_002688.4	0%	100%	*SEPT5*	0%	0%	53.50%	0.6047	Unprocessed

[1] DNA methylation values corresponding to two paired samples being tissues (liver, skeletal muscle and heart muscle) or primary cells (keratinocytes, fibroblasts, melanocytes and CD4+ lymphocytes) per parental gene-pseudogene pair. The selection was based on pseudogenes, for which our group had previously detected tissue specific methylation [67]. The values shown are the median of all CpGs position within the PCR fragment rounded to 0, 25%, 50%, 75% and 100%, respectively

Pseudogenes are distributed through the genome and, in many cases, the paralogous parental gene have been described. I selected 32 pseudogenes displaying differential methylation in human healthy adult tissues from a large scale study [67]. Next, I selected the corresponding parental genes according to its annotation in the Vertebrate Annotation Database (VEGA) [197] and determined their methylation status by direct sequencing of bisulfite-converted DNA in regions of high sequence homology. In total, I have included the in the analysis 15 unprocessed pseudogenes and 17 processed pseudogenes and their parental genes. To search for conservation of tissue specific methylation, I have analyzed the methylation status of each pair in two healthy adult human tissue (skeletal muscle, liver and heart muscle) or primary cells (fibroblasts, melanocytes, keratinocytes and CD4+ lymphocytes).

Unprocessed Pseudogenes

Unprocessed pseudogenes derive from the parental gene after genome duplication and further rearrangements and mutations [75]. Among the 15 unprocessed pseudogenes analyzed, five (33%) showed tissue-specific methylation in at least one examined tissue. For two pseudogenes, RP4-539M6.7 and *KB-1995A5.12*, this differential methylation was conserved in the respective parental genes, *SLC39A1* and *Q6UW61_HUMAN*. Six (40%) unprocessed pseudogenes and their respective parental genes were hypermethylated in both tissues, while one (7%) pseudogene-parental gene pair was unmethylated in the analyzed tissues. The remaining 3 unprocessed pseudogenes (20%) and their respective parental genes displayed heterogeneous methylation values ranging from 25% to 75%. In no case, I found a parental gene displaying tissue-specific methylation, while the respective unprocessed pseudogene had lost the differential methylation.

<u>Pairs displaying tissue specific methylation</u>

The parental gene *SLC39A1* and its unprocessed pseudogene RP4-539M6.7 showed 100 % and 25 % methylation in liver and keratinocytes, respectively. Alike, the gene

coding for the *Q6UW61_HUMAN* protein and its unprocessed pseudogene KB-1995A5.12 showed 100 % methylation in liver while 50 % methylation in skeletal muscle.

Figure 2.3.10 illustrates the conservation of T-DMRs in parental genes and unprocessed pseudogenes, taking as an example the *Q6UW61_HUMAN*/KB-1995A5.12 pair, for the unprocessed pseudogene (panel A) and the parental gene (panel B). In addition, the figure also shows the position of both analyzed fragments. Despite the very high sequence homology (> 94 %), the fragment corresponding to the parental gene is located in the intron 2, while the corresponding to the unprocessed pseudogenes is located in its 5'-UTR. Further analysis of the sequences surrounding the studied regions in both, parental gene and pseudogene, revealed that they were highly homologous (>95 %). Moreover, the parental exon 3 is highly similar to the exon 1 of the pseudogene, and the exon-intron architecture of the unprocessed pseudogene resembles that observed in the parental gene. These observations suggest that the accumulated mutations after duplication might have affected critical sequence motifs within the coding region (i.e. splicing signals) leading to the generation of new transcripts, despite the high sequence homology in general.

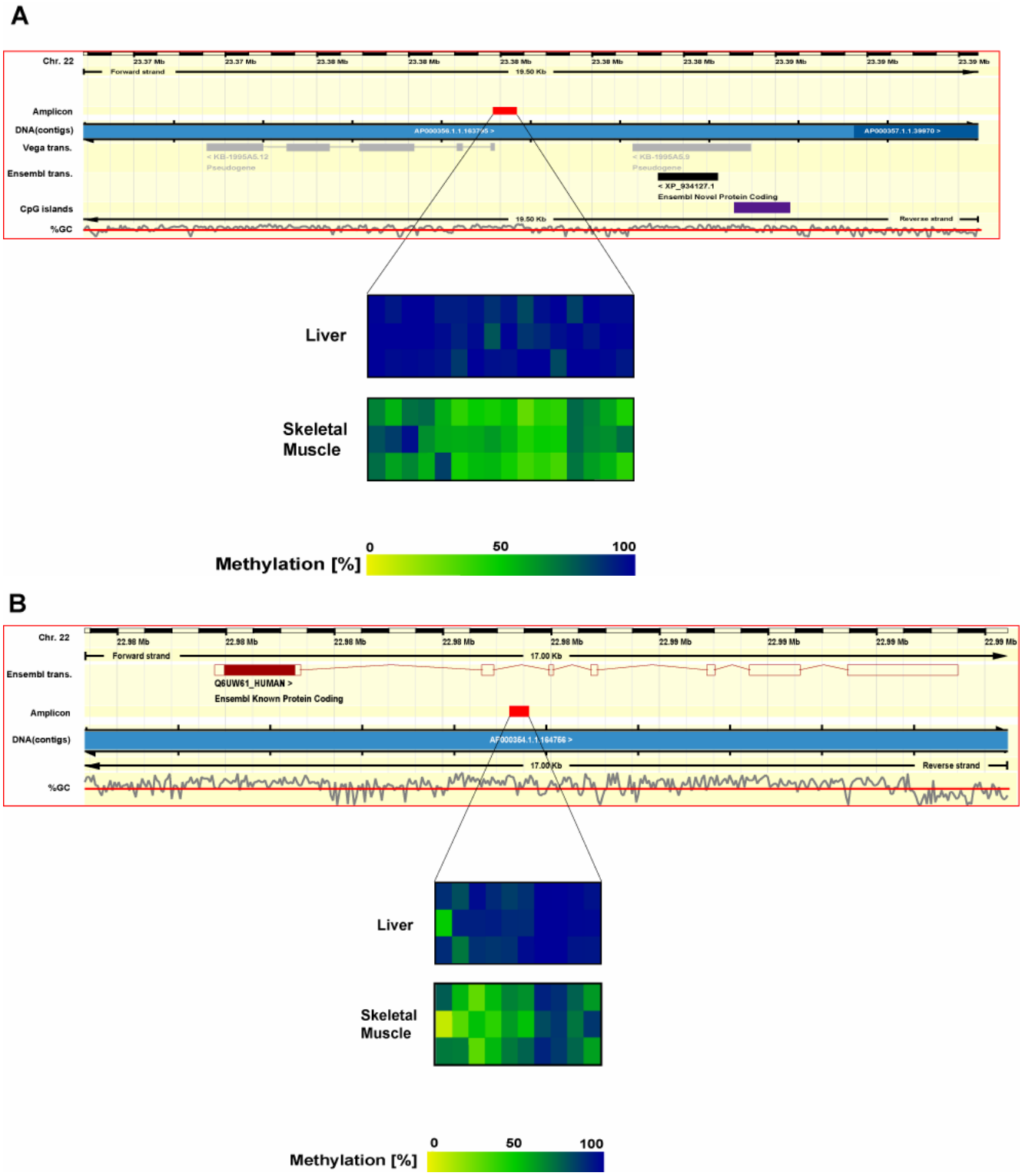


Figure 2.3.10: Conservation of tissue specific methylation in an unprocessed pseudogene / parental gene pair. A and B) The T-DMR of the unprocessed pseudogene KB-1995A5.12 (A) is conserved in its parental gene *Q6UW61_HUMAN* (B). The positions of the analyzed regions are shown. Liver samples are hypermethylated in both regions, while skeletal muscle samples display 50% methylation. Rows represent individual samples with columns representing individual CpGs positions along the PCR fragment. Quantitative methylation analysis results are shown in a color scale ranging from yellow (~0% methylation), green (~50% methylation) to dark blue (~100% methylation). The position of each analyzed fragment is depicted according Ensembl v40

SLC39A1 and its unprocessed pseudogene displayed also conservation of the tissue specific methylation. Similarly to the observed in *Q6UW61_HUMAN*, regions of high sequence homology are located in the 5'-UTR in the unprocessed pseudogene and in the coding region in the parental gene. However, in this case there were not a relation between analyzed fragment and the architecture of both transcripts.

The unprocessed pseudogene CTA-150C2.8 displayed tissue specific methylation (hypermethylation in liver, 50 % methylation in skeletal muscle), but its parental gene, *APOBEC3G*, was hypermethylated. Studied regions were located in the 5'-UTR for the unprocessed pseudogene and in the coding region for the parental gene. These regions showed relatively low sequence homology (57%). In the same manner, the unprocessed pseudogene NM_002688.4 displayed tissue specific methylation (unmethylation in liver, hypermethylation in skeletal muscle) in its 5'-UTR, while its parental gene, *SEPT5*, was unmethylated in intron 1 in both analyzed tissues. Analyzed regions also showed low sequence homology (54%). Very interestingly, parental gene and pseudogene are located consecutively on chromosome 22 and both analyzed regions are within the same CpG island.

The RP11-432I12.3 unprocessed pseudogene displayed tissue specific methylation, 75% in liver and 100% in skeletal muscle. Noteworthy, its parental gene, *CUBN*, displayed also specific methylation in the studied tissues, but in an inverted manner (100% in liver and 75% in skeletal muscle). Studied regions showed high sequence homology (89%) and were located in the last intron of both genes (intron 6 and intron 66, for the unprocessed pseudogene and parental gene, respectively).

<u>Pairs displaying hypermethylation</u>

In all cases, studied regions were located in the coding region of both, unprocessed pseudogene and parental gene. Out of 6 studied pairs, 4 displayed conservation of the hypermethylation in both tissues for the unprocessed pseudogene and the parental gene. These 4 pairs showed also high sequence homology (>72 %). Unlike, 2 pairs not showing conservation of the hypermethylation were those with the lowest sequence homology (71% and 64%).

The studied region in the RP11-34P12.1 pseudogene contained a part of intron 2, the whole exon 3 and a part of the intron 4. It showed hypermethylation in both studied tissues, liver and skeletal muscle. In turn, the studied region on its parental gene, *DDX11*, contained a part of intron 23, the whole exon 24 and a part of the intron 24. The registered methylation values were lower (75%) than in the unprocessed pseudogene and no tissue-specific methylation was found.

Differences were still more pronounced between the unprocessed pseudogene RP11-223J15.2 and its parental gene, *EEF1A2*. Studied region in the unprocessed pseudogene consisted of the whole exon 2 and part of the intron 2, while that for the parental gene contained part of intron 1 and exon 2. Both regions shared relatively low sequence homology (64 %)

Pairs displaying unmethylation

In only one pair both, unprocessed pseudogene and parental gene, showed unmethylation. The studied region in the unprocessed pseudogene CTA-246H2.2 was located in an annotated CpG island in the 5' end of the transcript. For its parental gene, *LRP5*, I studied also a CpG island located in the 5' end of the respective transcript. Both regions showed high sequence homology (89 %) and were unmethylated in liver and skeletal muscle.

Pairs displaying heterogeneous methylation

Among the remaining 3 analyzed pairs, *CTSL* and its unprocessed pseudogene, *CTSLL3*, showed 75% methylation in both analyzed tissues (liver and skeletal muscle). Both studied regions were located in the coding region (exon 2-3 and exon 3-4, respectively) and shared high sequence homology (89%)

Likewise, the parental gene *ABCD1* and its unprocessed pseudogene, RP11-453N.6, showed 75% methylation in liver and skeletal muscle. In this case, studied regions were located also in coding regions for both genes, exons 2-3 for the pseudogene and exons 7-8 for the parental gene and shared high sequence homology (95%)

Unlike, the unprocessed pseudogene RP11-392A22.3 displayed 50% methylation in liver and skeletal muscle, but its parental gene, *GSTA1*, was hypermethylated in these tissues. For this pair, studied regions were also located in the coding regions, part of intron 5 and exon 6 for both genes, and were highly homologous (83 %).

Processed Pseudogenes

Processed pseudogenes derive from retrotransposition of the parental gene and incorporation into the genome [75]. Among the 17 analyzed processed pseudogenes, 13 (76%) displayed tissue- specific methylation, while 4 (24%) were hypermethylated. I did not found conservation of the tissue- specific methylation between the processed pseudogenes and their respective parental gene.

Pairs displaying tissue- specific methylation

Despite the high number of pairs analyzed, I did not find any pair where the differential methylation found in the processed pseudogene was conserved in the respective parental gene.

Figure 2.3.11 shows an example of the lack of conservation of the tissue- specific methylation observed in processed pseudogenes. Panel A shows the position of the analyzed region in the processed pseudogene AP000357.3 and the observed methylation. The analyzed fragment was located at the 5' end of the transcript and tissue-specific methylation was observed between fibroblasts (0-25%) and keratinocytes (75-100%). Such differential methylation was not observed in the respective parental gene, *ARL5*, amid the high homology of both analyzed sequences (92%). Panel B shows the position of the analyzed region in the parental gene and the observed unmethylation in both cell types, fibroblasts and keratinocytes.

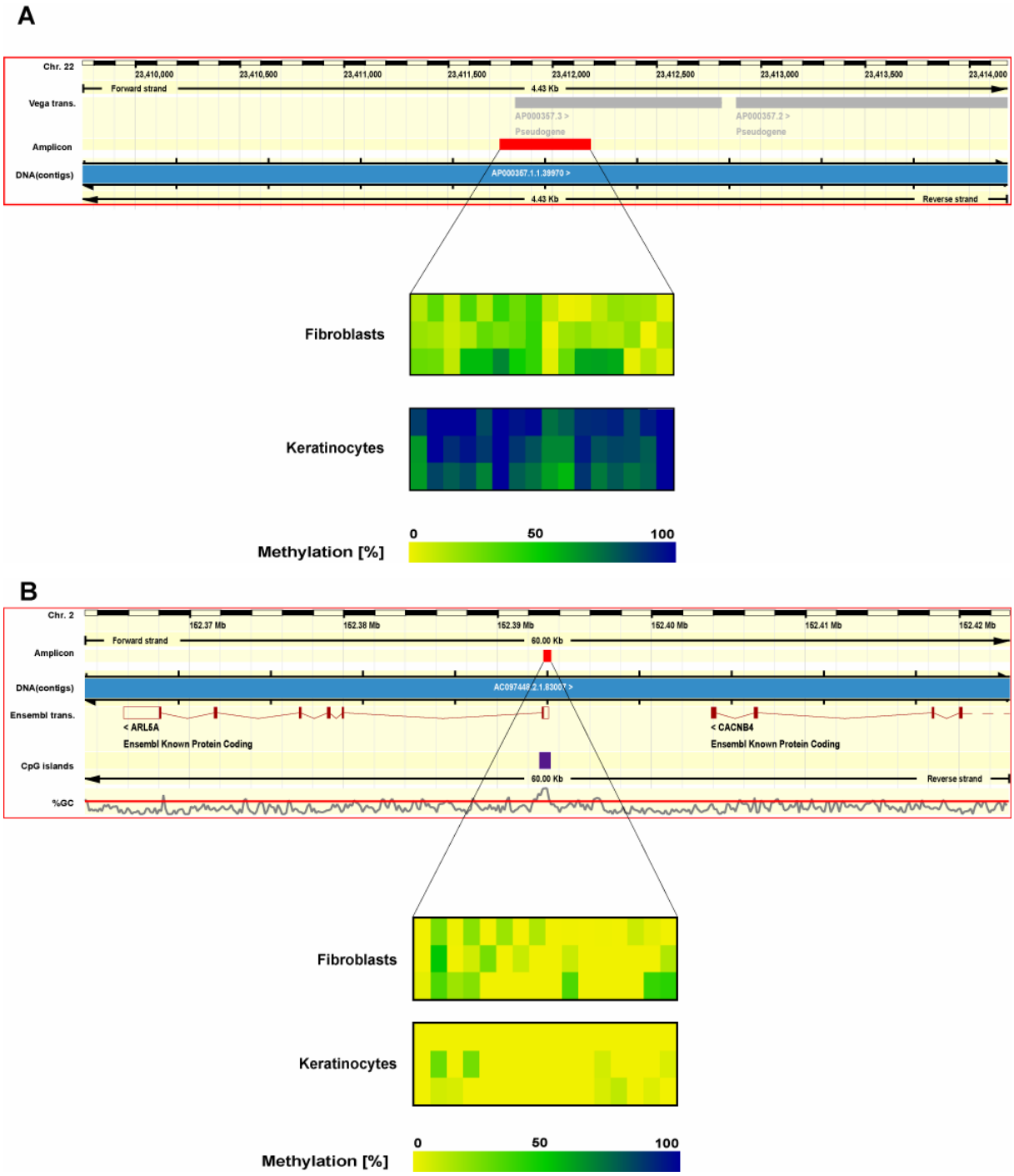


Figure 2.3.11: Tissue-specific methylation is not conserved in processed pseudogene / parental gene pairs. A and B) The T-DMR located within the processed pseudogene AP000357.3 (A) is not conserved in its parental gene *ARL5A* (B). The processed pseudogene is unmethylated (<20%) in fibroblasts and hypermethylated (>80%) in keratinocytes. In contrast, the parental gene is unmethylated in both cell types. Methylation values and fragment positions are shown as in figure 2.3.10.

The studied region in the CTA-373H7.4 processed pseudogene surrounded the TSS. I found differential methylation between CD4+ lymphocytes (hypermethylated) and fibroblasts (unmethylated). Besides the very high sequence homology (99%), the

studied region in the parental gene, *HBLD1*, was located within an annotated CpG island. The parental gene did not display differential methylation being both cell types unmethylated.

CTA-229A8 displayed hypermethylation in CD4+ lymphocytes and 50% methylation in skeletal muscle. Very interestingly, its annotated parental gene was the housekeeping gene *GAPDH*, which showed 50% methylation in both tissues. Sequences of the studied regions were 62 % homologous.

AP000357.2 displayed hypermethylation in liver while unmethylated in sperm at the 5′ end of the transcript. The studied region displayed low sequence homology (57%) with the respective parental gene, *ACTR2*, which showed hypermethylation in both tissues.

Similarly, the processed pseudogene RP11-758C21.1 showed tissue specific methylation at the 5′ end of the transcript (100% methylation in CD4+ lymphocytes and 50% methylation in fibroblasts). However, the region of high homology (96%) of sequence in its parental gene, *BDH2*, displayed hypermethylation in both analyzed tissues.

A region comprising the exon 1 and part of the 5′ UTR of the processed pseudogene RP1-181C9.1 displayed hypermethylation in liver and 25 % methylation in keratinocytes. The region 5′ region of its parental gene, *ANP32B*, displayed high sequence homology (71%), however was unmethylated in both tissues.

The whole transcript and a part of the 5′UTR of the AP000358.2 pseudogene showed differential methylation between melanocytes (25%) and CD4+ lymphocytes (hypermethylation). Its parental gene, *FN3K*, shared high sequence homology within the coding region (88%). CD4+ lymphocytes showed also hypermethylation of this region, but melanocytes displayed 50% methylation.

RP3-412A9.4 processed pseudogene displayed hypermethylation in its 5′ region in CD4+ lymphocytes, while 50% in skeletal muscle. The 5´ region in its annotated parental gene, *SNRPN*, was highly homologous (88%), but displayed hypermethylation in both samples.

KB-1269D1.3 displayed tissue specific methylation in the 5′-UTR. Studied region was hypermethylated in CD4+ lymphocytes and 50% methylated in keratinocytes. The annotated parental gene for this processed pseudogene was *MAD1L1*. Both studied

regions displayed 95% sequence homology. The parental gene showed hypermethylation in lymphocytes but 75% methylation in keratinocytes.

Similarly, *AC000078.2* displayed differential methylation in its 5'-UTR that was lost in its parental gene, *RPL8*. Studied regions were 66% identical. Although both genes displayed hypermethylation in liver, the processed pseudogene showed 25% methylation in fibroblasts and the parental gene 75 %in same samples.

Regions surrounding the TSS were highly homologous (84%) between the processed pseudogene *AC004019.3* and its annotated parental gene *LOC376522*. The pseudogene displayed differential methylation in liver (100%) and fibroblasts (50%). The parental gene showed 75% methylation in liver and 50% in fibroblasts in the studied region.

RP1-47A17.8 showed hypermethylation in liver and 25 % methylation in fibroblasts at the 5'end. Region of high homology (84%) in the parental gene was located in the coding region of its parental gene NP_055087.2. This region lost the differential methylation being hypermethylated in both studied tissues.

Pairs displaying hypermethylation

4 processed pseudogenes showed hypermethylation in both analyzed tissues. In 3 of them, I also observed hypermethylation in the respective parental gene. Very interestingly however, RP3-405J24 showed hypermethylation in liver and keratinocytes, but its parental gene, *RPL12*, showed tissue specific methylation (100% in liver and 50% in keratinocytes, respectively). Studied regions were highly homologous (92%)

Studied regions in AC007050.7 and its parental gene, *POM121L3*, were hypermethylated in liver and fibroblasts. They were 96% homologous.

Similarly, an intragenic region showing high sequence homology (94%) in RP1-106I20.2 pseudogene and its parental gene *NDUFA9* was hypermethylated in liver and skeletal muscle.

AC004471.4 was hypermethylated in CD4+ lymphocytes and fibroblasts. Alike, its parental gene *NM_032028.2* was hypermethylated in both tissues. Studied regions shared an 82% homology in their sequences.

2.3.3.3. DNA methylation profiles in duplicated genomic sequences and their relation to evolutionary distances

To investigate the relation of the DNA methylation in duplicated sequences and the time of duplication, I determined the evolutionary distance between the pseudogenes and they respective parental genes.

Figure 2.3.12 shows examples of the obtained radial trees and the calculated DNA distances for PLG and TBX families, as well as pairs of parental gene with processed or unprocessed pseudogenes.

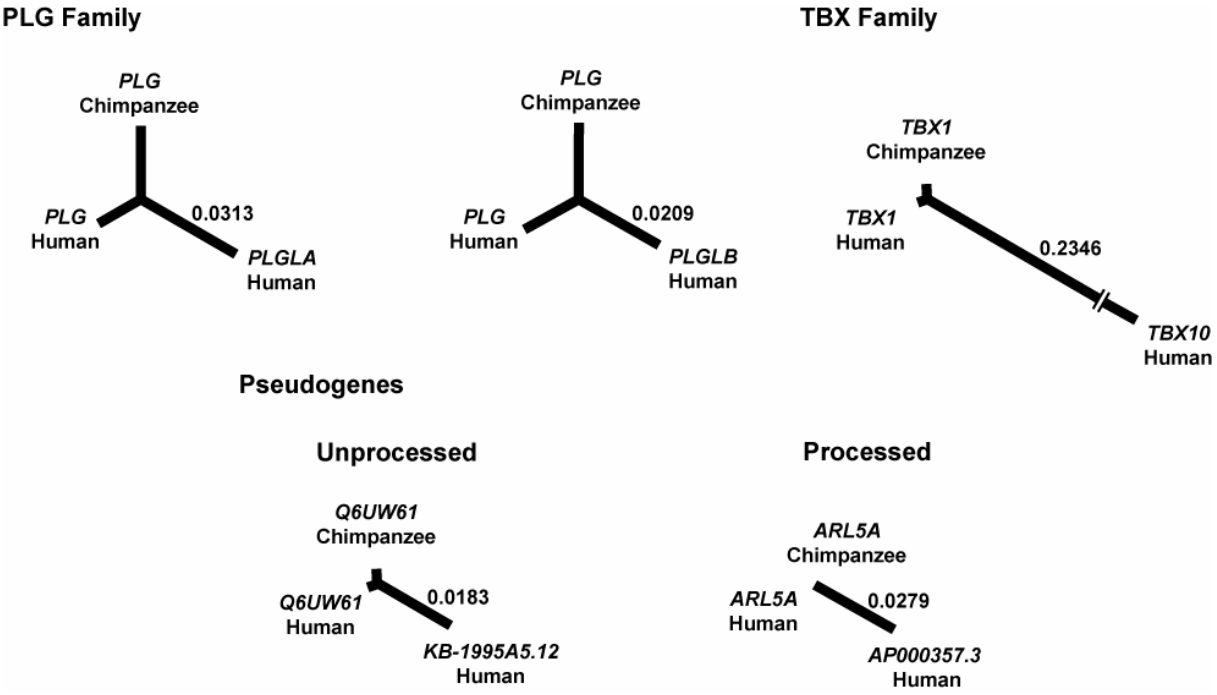


Figure 2.3.12 Examples of DNA distances between gene variants and pseudogene / parental gene pairs. DNA distances were calculated using the PHYLIP software and results are represented as a radial tree for each gene pair. The distance between the internal node and the duplicated gene or pseudogene is indicative for the evolutionary distance to the parental gene. Examples of distances from the internal node are shown for the *PLG* variants (0.0313 and 0.0209, for *PLGLA* and *PLGLB*, respectively, upper panel), the *TBX1/TBX10* (0.2346) pair and unprocessed (0.0183) and processed pseudogenes (0.0856, lower panel). The DNA distance observed for *TBX1/TBX10* (0.2346) is about 10 times larger than those observed in the PLG family.

Processed and unprocessed pseudogenes displayed a wide range of distances, regardless their type (Table 2.3.2). I did not find any differences in distribution of

DNA distance values between processed and unprocessed pseudogene (p=0.86532 One-way ANOVA Test – Bonferroni Corrected) (Figure 2.3.13)

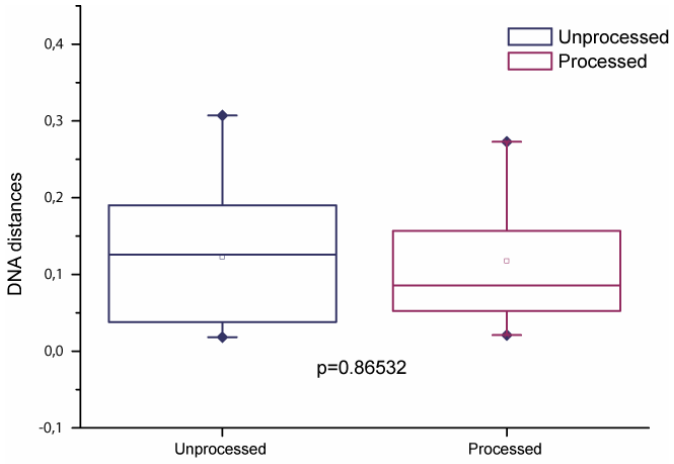


Figure 2.3.13: DNA distances to the parental gene are similar in processed and unprocessed pseudogenes. Box plots comparing the calculated DNA distances obtained using the PHYLIP algorithm. Both populations are not significantly different (p=0.86532, One-way ANOVA Test – Bonferroni Corrected)

DNA distances for the variants of the *PLG* gene showed similar distances, 0.0313 and 0.0209 for *PLGLA* and *PLGLB*, respectively. As *PLGLB* orthologues exist in Chimpanzee (*Pan troglodytes*) but not in old world monkey (*Maccaca mulatta*) and within the *Eutheria* infraclass, it can be estimated that the PLG duplication occurred after the hominid-cercopithecoid divergence, some 29-35 million years ago [198]. In contrast, the evolutionary distance of the *TBX1* and *TBX10* pair (0.2346) is about 10 times larger than the one observed for *PLG* and these genes are present as well in other mammals such as mice indicating that this duplication occurred at least 80-130 million years ago [199].

I have calculated the DNA distances according to the Jukes-Cantor model [200]. This model assumes that all substitutions, transitions and transversions, are equally likely to happen. Therefore, the calculated distance is an estimate of the divergence time between both sequences. In the Jukes-Cantor model the obtained estimate is a maximum likelihood estimate.

To calculate the distances between the analyzed sequences, I computed DNA distance matrices for each pair using the chimpanzee orthologue of the parental gene

as an "outgroup". Human and chimpanzee genomes have a very close evolutionary relationship, with an average nucleotide divergence of 1.2% [201]. Although the chimpanzee genome is not as well annotated as the human genome, it offered the chance to control whether the studied pairs had similar mutation rates. Consequently, pairs included in the study were only those from confirmed orthologues (as retrieved from Ensembl Database v40) and showing similar distances between the human and chimpanzee genes.

I have chosen to use unrooted phylogenetic trees to represent the results. Since only three sequences are under study, a triangular configuration with a determined internal node (IN) is the only one possible. Hence, distances from each sequence to the IN can be easily calculated (see details in Materials and Methods). This configuration allows comparing the distances to the internal node between trios of human/chimpanzee/duplicated sequence. Due to this simple configuration and the high homology of the analyzed sequences, e.g. *PGL* human/*PLG* chimpanzee/*PLGLA* human trio displayed 20 mismatches in 289 nucleotides, resampling controls were not necessary. In such controls, e.g. bootstrapping, random dataset are created out of the original dataset and the entire possible configuration tested [202].

Next, I analyzed the conservation of DNA methylation in each tissue in its dependency to the evolutionary distance of the *PLG* variants, the *TBX1*/*TBX10* pair and the pseudogene - parental gene pairs (Figure 2.3.14). In this analysis, conserved methylation was assigned if the methylation values from the respective gene pair differed by less than 20% between measurements in the same tissues. For both, unprocessed and processed pseudogenes, I observed a lack of methylation conservation for most of the tissues and pseudogenes examined. This lack of conserved methylation was evident as well for pseudogene - parental gene pairs that arose fairly recently as evidenced by a short evolutionary distance (Figure 2.3.14, panel A). Conserved T-DMRs were only observed for two unprocessed pseudogenes/parental gene pairs (see above) and for the PLG variants. Notably, these pseudogenes displayed a rather short evolutionary distance to their parental genes (0.1258 and 0.0183 for the RP4-539M6.7/*SLC39A1* and LL22NC03-31F3.7/ *Q6UW61_HUMAN* pairs, respectively). In addition, I analyzed if pseudogenes are

preferentially hypermethylated or hypomethylated compared to their parental genes finding no evidence that the analyzed pseudogenes became persistently hypermethylated (Figure 2.3.14, panel B). I found pseudogenes that were both hypomethylated and hypermethylated when compared to the cognate parental gene. Although the analyzed number of gene pairs is too small to generalize this observation, I observed a trend indicating that hypermethylated processed pseudogenes were evolutionarily more distant to their parental genes than unmethylated processed pseudogenes.

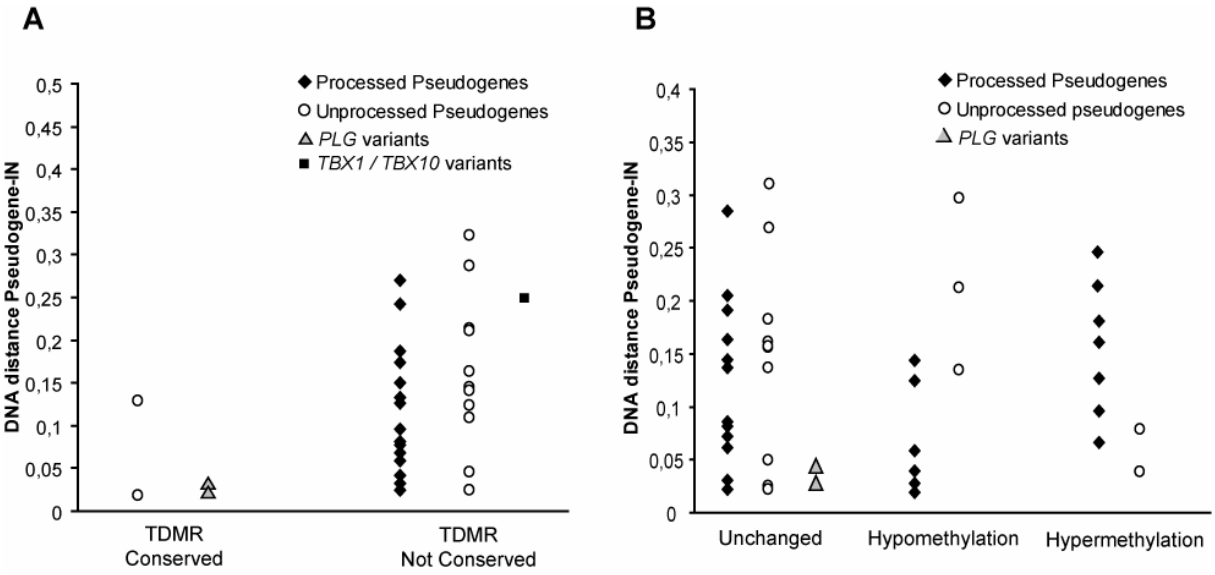


Figure 2.3.14: Relations between the conservation of DNA methylation and evolutionary distances. A) Relation between tissue-specific DNA methylation conservation and DNA distance in gene families and unprocessed and processed pseudogenes. Pairs displaying conservation of methylation are among those showing the shorter distances. B) DNA methylation transitions in gene duplications. The methylation profiles from the parental genes were compared to the respective gene variants and pseudogenes. Many of the parental genes displayed the same methylation values compared to their respective pseudogenes in each individual tissue. Hypermethylated processed pseudogenes exhibit larger DNA distances to the parental genes than unmethylated processed genes. Black rhombuses, white circles, gray triangles and black squares represent unprocessed pseudogenes, processed pseudogenes, *PLG* gene variants and the *TBX1/TBX10* pair, respectively. Y- Axis: DNA distances to the internal node calculated by PHYLIP algorithm.

2.4. Interaction of Genetic and Epigenetic variations in disease

2.4.1. Case of study: DNA Methylation profiles in Progeria and FPLD patients

Although it represents an inherent idea of epigenetic regulation, little is known about how the epigenetic layer of control interacts with the genetic background in the development of a diseased phenotype, discussed in [173] [174]. In this regard, progeroid syndromes represent an excellent model to study such interaction.

As described in the introduction, mutations in the *LMNA* (lamin A/C) gene are recognized as an etiological factor for several laminopathies displaying progeroid features [175]. However, the underlying mechanisms are not clear. Moreover, patients displaying distinct phenotypes, as different as to be considered different syndromes, bear the very same mutation. Thus, I studied the methylation profiles of 10 genes selected from the literature to investigate whether they are connected to the distinct phenotypes observed in Familial Partial Lipodystrophy (FPLD) and Hutchinson-Gilford syndrome (Progeria). The results of my experiments presented in this section have been published in a peer-reviewed scientific journal [266].

Progeroid syndromes are characterized by a premature ageing of the affected patients (OMIM ID: 176670). Interestingly, variations in DNA methylation have been widely reported as correlating with age [180]. In this regard, I selected from the literature a panel of 9 age-related candidate genes plus the *LMNA* gene itself, to look for methylation variations on them.

The panel of analyzed genes consisted of the following genes: Chondroitin Sulfate Proteoglycan 2 (Versican) (*CSPG2*), Estrogen Receptor 1 (*ESR1*), Insulin-like Growth Factor 1 Receptor (*IGF1R*), Insulin-like Growth Factor 2 Receptor (*IGF2R*), Lamin A/C (*LMNA*), MutL Homologue 1 (*MLH1*), RAN binding protein 1 (*RANBP1*), Retinoic Acid Receptor, Beta (*RARB*), Zinc Metallopeptidase (STE24 homologue, yeast) (*ZMPSTE24*), and Transforming Growth Factor, Beta Receptor I (activin A receptor type II-like kinase, 53kDa) (*TGFBR1*).

Individuals included in the study belonged to three groups, as described in Material and Methods, the “FPLD group”, the “Progeria group” and the “Control group”.

DNAs were isolated from cultured B- Lymphoblastoid cells from these individuals (see culture description in Material and Methods). Table 4.1.1 in Materials and Methods section summarizes the phenotypes and genotypes of the included individuals.

The ten investigated genes showed low methylation in all samples, regardless of the group they belong, with methylation rates ranging between 0 and 30 % (Figure 2.4.1). However, two results were remarkable. The *RARB* gene showed a significantly higher methylation in all 6 patients of the "FPLD group" when compared to the patients of the "Progeria group" as well as the "Control group". In addition, *LMNA* was clearly but not significantly lower methylated in patients 1 – 3 (Family A) than in patients 4 – 6 (Family B) within the "FLDP group".

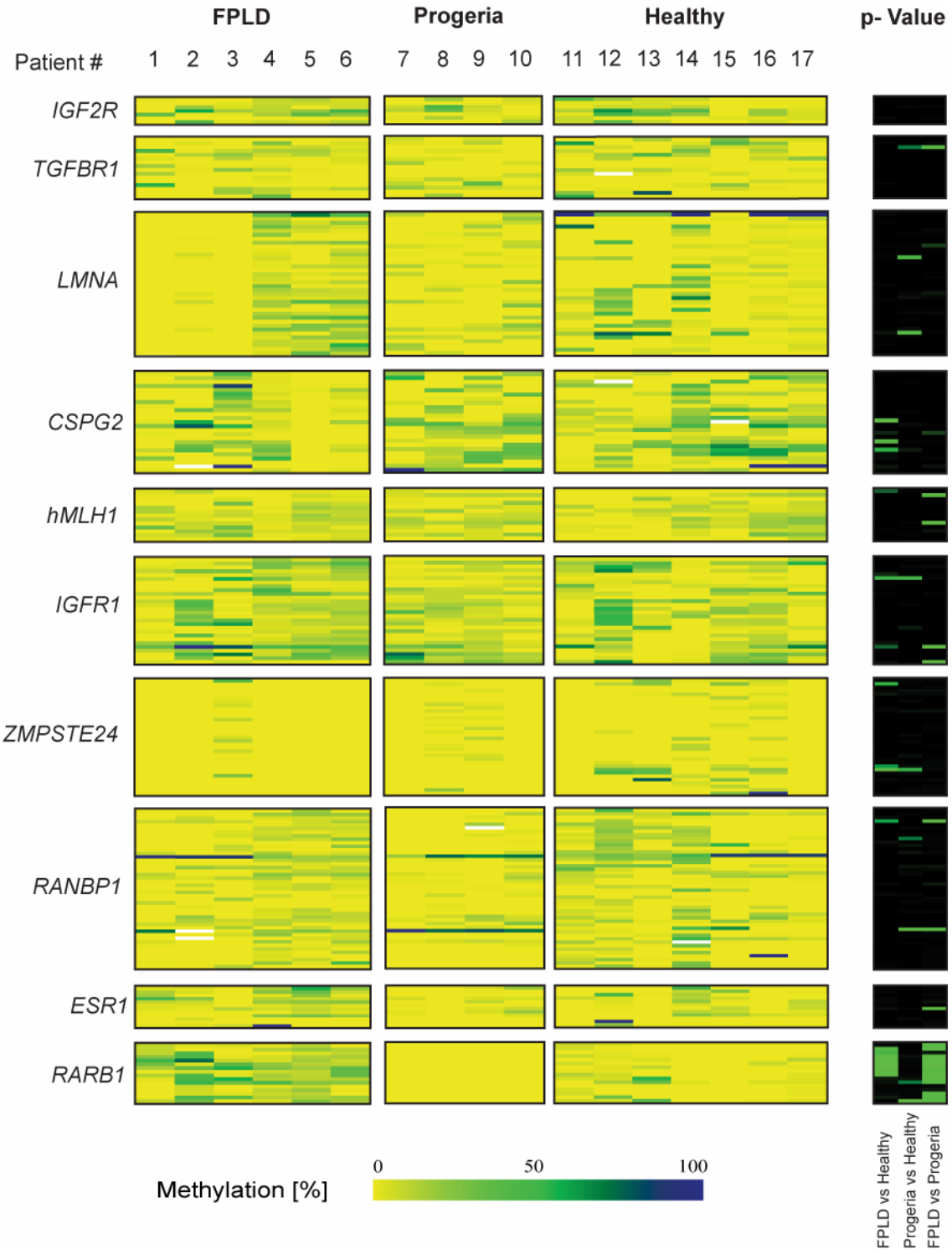


Figure 2.4.1: DNA methylation profiles for 10 candidate genes in all included individuals. In general genes displayed low DNA methylation values. *RARB1* was highly methylated in FPLD patients than in Progeria patients or healthy individuals. *LMNA* was differentially methylated in FPLD patients members of two families bearing different mutations. Color code for methylation values ranges from yellow (~0%) to dark blue (~100%). p-values represent the results of Wilcoxon test and are shown in a color code. Green cells corresponds to $p<0.05$ meaning significant differences and black cells corresponds to $p>0.05$, meaning non-significant differences.

Figure 2.4.2 shows the methylation profiles for the *RARB* gene in more detail. The *RARB* gene encodes 2 different protein isoforms that arise by alternative usage of two promoters: *RARB1* and *RARB2*. Three transcripts for this gene are annotated in the Ensembl Database, a long transcript (ENST00000264330) and two shorter transcripts (ENST00000330688 and ENST00000383772), probably regulated by promoter 1 (P1) and promoter 2 (P2), respectively. Differential methylation of P2 correlates with gene silencing of the *RARB2* isoform in several cancer cell lines [203] and represents one of the best studied T-DMRs in epigenetic studies.The region I have studied was located in P2, upstream to the short transcript and in intron 1 of the long transcript.

Patients in the FPLD group displayed 50% methylation, while those in the "Progeria goup" and the "Control goup" displayed unmethylation of the studied region. 13 out of 16 investigated CpG dinucleotides showed significantly higher methylation values in FPLD patients than in Progeria patients ($p<0.05$, Wilcoxon Test). Similarly, 15 out of 16 investigated CpG dinucleotides displayed higher methylation values in FPLD patients compared to healthy individuals.

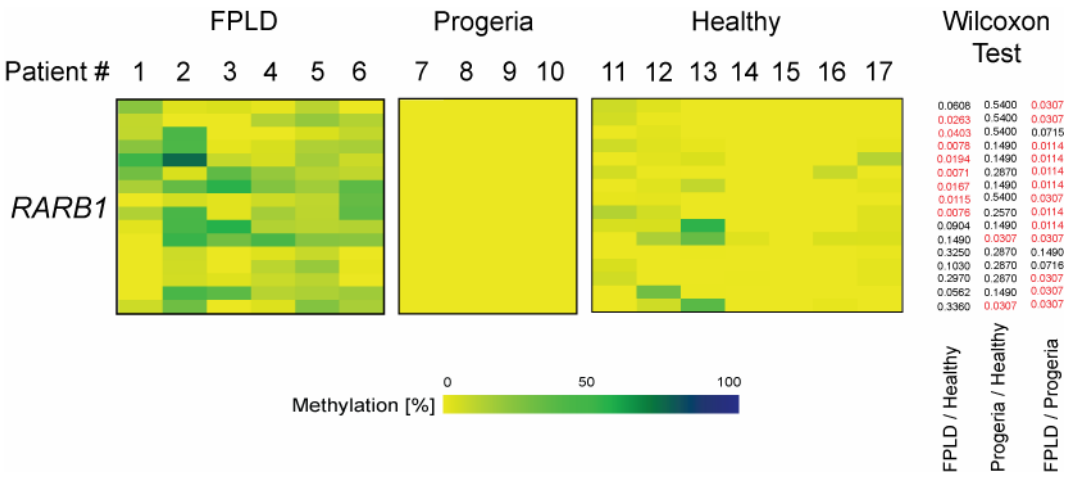


Figure 2.4.2: *RARB* is highly methylated in FPLD patients compared to progeria patients and healthy individuals. FPLD patiens displayed 50% methylation in P2 promoter of the *RARB* gene, whereas this region was unmethylated in progeria patients and healthy individuals.

Although sharing the same clinical signs, affected members of both families in the "FPLD group" beared different mutations. Interestingly, I found also methylation differences when I compared the members of both FPLD families. Studied region in the *LMNA* gene displayed 50 % methylation in the affected members of the Family B

(carrying the R471G mutation) and unmethylation in affected members in Family A (carrying the R482L mutation) (Figure 2.4.3). The studied region was a CpG island comprising the exon 1 and part of the 5′ upstream region of the *LMNA* gene. Though clearly visible, this difference was not statistically significant (p>0.05, Wilcoxon test), probably due to the low number patients in each group.

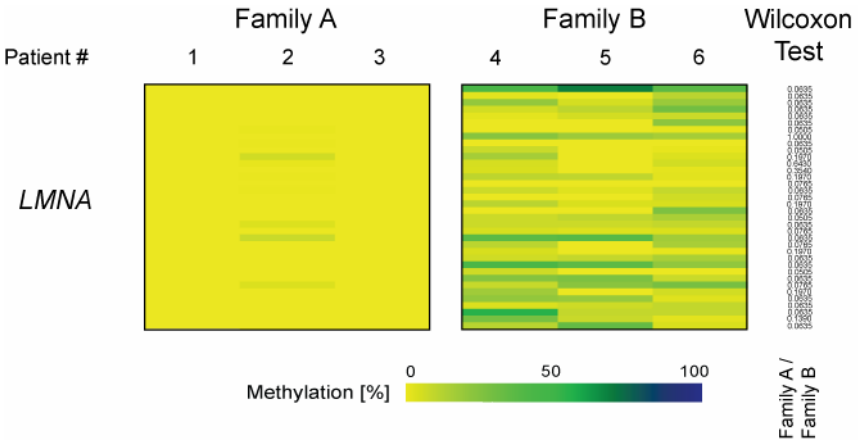


Figure 2.4.3: FPLD patients bearing different mutations in the *LMNA* gene show differential DNA methylation in the 5′-UTR of this gene. Studied region in the *LMNA* gene displayed 50% methylation in the affected members of the Family B (carrying the R471G mutation) and unmethylation in affected members in Family A (carrying the R482L mutation). Although differences were clearly visible, they were not statistically significant (p>0.05, Wilcoxon test)

Noteworthy, I detected all methylation changes in B-lymphoblastoid cells isolated from whole blood samples. This fact enables the study with minimal intervention, which is a very desirable feature in a diagnostic procedure. In order to control possible changes in the methylation patterns introduced during immortalization, culture and thawing of the cells, I compared DNA from freshly prepared lymphocytes and the corresponding imortalized lymphocytes of three healthy probands (Figure 2.4.4).

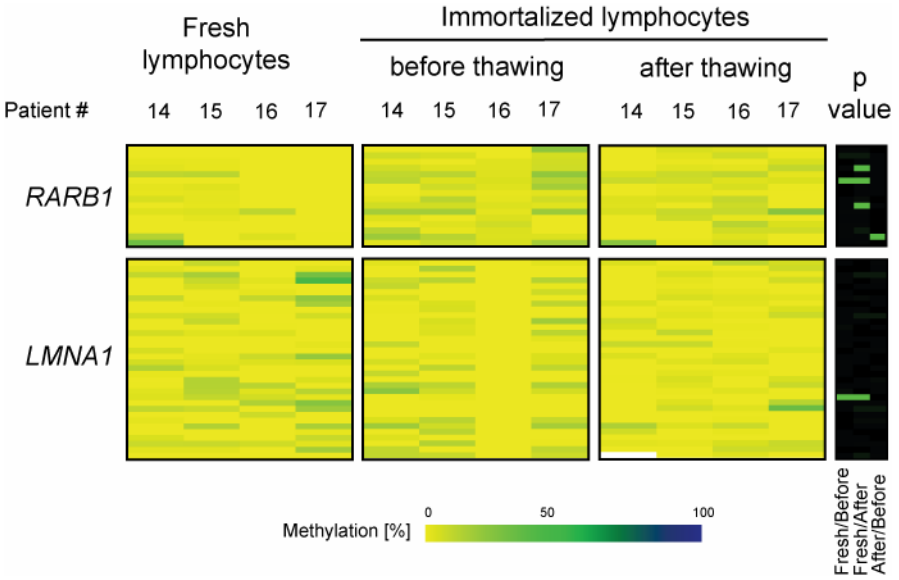


Figure 2.4.4: Immortalization and culture procedures did not affect the methylation profiles of *RARB1* and *LMNA1*. Observed unmethylation in healthy controls in *RARB1* and *LMNA* remained unchanged after immortalization, culture and thawing of the lymphoblastoid cells. p-values represents results of Wilcoxon test and are shown in a color code. Green cells corresponds to $p<0.05$ meaning significant differences and black cells correspond to $p>0.05$, meaning non-significant differences.

3. Discussion

3.1. General considerations

Epigenetic phenomena in general and DNA methylation in particular play a major role in the regulation of gene expression, contributing to the modeling of cellular phenotypes. Differences in DNA methylation can be observed at very different levels, such tissue- and developmental stage- specific variations in healthy tissues but also related to oncogenic transformation or in other non-cancer diseases as well.
In this work, I have studied the functional aspects of the differential DNA methylation. I have focused the study on its correlation with RNA expression in healthy tissues, the evolutionary conservation of tissue-specific DNA methylation profiles and the putative influence of DNA methylation in modifying phenotypes despite the same genetic background.

3.2. Relations between DNA methylation and gene expression – Theory and experimental approaches

It is widely accepted that methylation of the promoter region is related to gene silencing. However, experimental evidence on this topic tends to be ambiguous. Pioneer works in this area by A.D. Riggs [88] and Holliday and Pough [89] proposed theoretical models based on DNA methylation to explain chromosome X inactivation and other events in cell differentiation during development. Since then, several models have been drawn to explain the epigenetic regulation of gene expression, not only by DNA methylation but also by histone modifications and RNA interference (as detailed in Introduction). Current knowledge supports the idea that there is no such universal mechanism. Instead, different mechanisms might regulate particular genes in determined physiological states and they might have a synergic action, as demonstrated in the maintenance of the X-chromosome inactivation [35].
Methylation patterns are established during development and tissue- and cell-specific patterns. Nevertheless, whether this differential methylation is related to tissue-specific gene expression still remains as an open question. Changes of

methylation patterns in cancer cells are a frequent feature, reviewed in [204] [106]. After the discovery of methylation differences, researchers attempt to establish correlations with the expression profiles in the involved genes in order to understand the functional consequences of the findings.

The use of analogous of deoxycytidine allows the pharmacological inhibition of the enzymes responsible for DNA methylation in eukaryote cells, DNA methyltransferases (DNMTs). Among these analogous, 5-azacytidine (azacytidine) and 5-aza-2'-deoxycytidine (decitabine) are the most used. Cells treated with these agents display expression profiles different to those observed in untreated cells and therefore, a reactivation of genes potentially silenced by DNA methylation is concluded. These agents are widely employed to discover genes regulated by DNA methylation and they were also tested clinically as potential anti-cancer drugs, reviewed in [145]. Recently, microarray-based expression profiles of cells treated with methylation inhibitors were employed to detect differentially methylated genes in different human cell lines [205] [206] [207] [208].

Pharmacological inhibition of DNMTs represented a huge advance in the understanding of epigenetic regulation of gene expression. However, some authors discussed its fundamentals and applications. By using a microarray-based approach, Gius and colleagues [206] reported that the effects of decitabine in expression profiles were more similar to those of an inhibitor of histone acetylation, Trichostatin A (TSA), than to single or double knockouts of DNMTs. In addition, the same study reported as many genes up- as downregulated after demethylation. These observations, though controversial, raise questions about the mechanisms of action and effects of deoxycytidine analogous. The vast variety of epigenetic mechanisms thought to regulate gene expression and a complementary functionality among DNMTs might explain these observed differences. In addition, a myriad of different laboratory protocols are available for the treatment of cultured cells with DNA methylation inhibitors. Doses and exposure times that have no effect in some experiments are highly cytotoxic in others. Response differences related to cell type and culture conditions explain these differences and protocols must be adjusted accordingly.

Its inapplicability when tissues samples are used instead of cell lines, e.g. patient samples, represents another drawback for this approach. Not surprisingly, the inverse correlation between DNA methylation and gene expression become more diffuse when tissue-based studies are conducted [90] [209]. Very few clear examples of tissue-specific control of gene expression by hypermethylation of the promoter region were known prior to the results of my experiments, i.e. the maspin gene, *SERPINB5* [31].

3.3. RNA expression related to tissue- specific differentially methylated regions

Methylation profiles produced by direct sequencing of bisulfite converted DNA have the highest possible resolution (single-base pair) and allow the simultaneous analysis of several CpGs in the regions of interest. In addition, the use of tissue samples adds valuable experimental evidence. It gives *in vivo* evidence for functionally relevant tissue-specific differentially methylated regions (T-DMRs) and opens the possibility to directly study their involvement in human pathologies. I made use of these advantages to approach a long-standing question: do tissue-specific DNA methylation patterns determine gene expression profiles that define tissue phenotypes?

Among genes displaying T-DMRs in healthy tissues and primary cells, I found a 37% correlation rate between decreased expression levels and methylation of the 5'-UTR (16 out of 43 genes).

I have defined the 5'-UTR as the region ranging between 2 kb upstream and 1 kb downstream of the TSS. Despite the fact that I have not confirmed the presence of annotated regulatory features in each of the investigated areas, CpG-rich regions surrounding the TSS are expected to have functional consequences in the expression of the gene. Kim and colleagues defined a high-resolution map of active promoters in the human genome by analyzing the *in vivo* binding of the RNApol II preinitiation complex (PIC). The authors reported that 87% of the TFIID-binding sites were within 2.5 kb of annotated 5' ends of known messenger RNA and suggested that binding sites beyond these limits correspond to unannotated transcriptional units [210]. Accordingly, hypermethylation of the 5'-UTR correlates with silencing of the *SERPINB5* gene in human tissues [31]. The results presented here extend the list of

genes, which rely on methylation as a mean to restrict gene expression to different tissues and cell-types, and constitute baseline information for further experimentation in this area. When compared to the germ line data, where the majority of 5'-UTR appears as unmethylated [211], tissue-specific methylation profiles suggest that DNA methylation is acquired throughout the development silencing particular genes and modeling tissue phenotypes. Mouse models, in which lack of DNA methyltransferases causes embryonic lethality, have highlighted the importance of DNA methylation in normal development [10] [86].

In addition, I did not find any correlation when the T-DMRs were located in the coding region or downstream to an annotated transcript. Few examples in the literature locate functional promoters within the coding regions of the genes [212] [213] [214] [215] and probably they represent alternative promoters for unannotated transcripts. In addition, a molecular function has been described for intragenic DNA methylation: it may decrease gene expression by altering the chromatin structure and reducing the efficiency of Pol II elongation [216]. In turn, intergenic regions may regulate the expression of neighboring genes either directly through the folding of the coding DNA strand during transcription [217] or indirectly by creating open /close chromatin structures that enable the access of the transcriptional machinery, reviewed in [218]. Alternatively, these T-DMRs can actually be promoters of non-coding RNA transcripts that regulate in *trans* the expression of related genes, as observed in the *Air* transcript in murine *Igf2r* [100] and other imprinted genes.

Taken together, the data presented here support the idea of several layers of epigenetic controls regulating gene expression in different tissues. These data also suggest that DNA methylation could not only avoid expression by blocking the assembly and action of the transcription machinery, but also by blocking enhancers and insulators located outside of the promoter and coding regions. The lack of correlation with expression in several T-DMRs supports the interplay with other molecular mechanisms, such histone modifications and non-coding RNAs.

3.4. DNA methylation and transcriptional profiles in lung development

It is widely accepted that DNA methylation has an important impact during mammalian development, for recent reviews see [219] [220]. However, little is known about stage-specific differential methylation and its functional consequences in lung morphogenesis. To study the role of the epigenetic phenomena in the establishment of transcriptional profiles during lung development, I investigated DNA methylation status of differentially expressed genes between fetal and healthy adult lung.

Genome-wide expression profiles and their biological interpretation

First, I have determined the expression profiles of human fetal lung tissue samples during the canalicular and saccular stage (20-25 weeks) of development. Differentially expressed genes allowed the unsupervised molecular classification of the samples according to their tissue type (Figure 2.2.1). In addition, I have validated the expression of 15 out of 21 studied genes (71%) in independent samples by a different method (qRT-PCR).

Analysis of gene ontology (GO groups) showed that groups related to components of the extracellular matrix (ECM) are over-represented in fetal lung. During canalicular and saccular stages in lung development, tissular structures suffer dramatic morphologic changes where ECM plays a key role [221]. As part of the ECM, collagen proteins play a key role in morphogenesis and epithelial differentiation [222]. Unlike, in healthy adult lung tissular structures are less dynamic and lower turnover of ECM components is expected. The results presented in this work suggest that the increased requirements of ECM components, especially collagen proteins, in these developmental stages are fulfilled, at least in part, by transcriptional activation of the corresponding genes.

In turn, groups over-represented in adult lung are related to immunological processes. The postnatal immune system is a product of regulated events towards the acquisition of definitive functional properties, reviewed in [223]. The results of this work support a role for an increased transcriptional activity for the effector molecules (e.g. complement components and major histocompatibility complex (MHC) antigens) either by infiltrated cells into the lung tissues or by lung cells themselves. In this regard, NF-kappa B pathway activation in epithelial airway cells

was reported in inflammation [224]. Activation of NF-kappa B causes enhanced expression of genes encoding inflammatory cytokines, acute phase proteins, immunoreceptors, and chemokines, suggesting that mature pulmonary cells are able to express effectors molecules upon proinflammatory stimuli.

Correlation between methylation and expression in fetal lung

I studied the methylation status of 43 genes differentially expressed genes and found that 19% of them displayed a correlation between decreased expression and methylation of the 5′ UTR. I selected the genes according to their differential expression (Log2Fc>2) and a high CpG content at the 5′ region. Notably, most of the studied regions were unmethylated in both, fetal and healthy adult samples according with the notion of extensive hypomethylation (<20%) in 5′ CpG islands and regions of high CpG content [67]. As I focused the study on 5′ regions, obtained results may underestimate the regulatory features located along the coding region or downstream to the encoded transcript that might also regulate expression by DNA methylation at this developmental stage.

3.5. DNA methylation and regulation of bi-directional promoters

Recently, Trinklein and colleagues reported a relative abundance of bi-directional promoters in the human genome. They constitute a common, though poorly understood, feature in mammalian genomes [193]. I analyzed the methylation status of 20 bi-directional promoters in fetal and healthy adult lung and did not found differential methylation in any of them. Promoters located within an annotated CpG island (Ensembl v41) exhibited unmethylation in both groups, while two out of three promoters, which were not located within a CpG island, were hypermethylated. By comparing the expression profiles of "head-to-head" genes located in opposite strands, I found that they were more frequently co-expressed than randomly paired genes located in different chromosomes. Noteworthy, genes corresponding to the hypermethylated promoters were downregulated suggesting that DNA methylation play a role in the regulation by bi-directional promoters. Other groups reported

differential methylation of particular bi-directional promoters in colorectal cancer [93] and in several cell lines [94]. Here, I present experimental evidence showing that DNA methylation might also control co-expression of "head-to-head" genes during lung development.

3.6. DNA methylation profiles in lung development and lung tumorigenesis

It has been suggested that the transcriptome of non-small cell carcinomas resemble those of lung fetal cells [157]. Bonner and colleagues [160] identified 62 genes as being involved in lung development, for which the expression was markedly increased in mouse lung carcinoma samples. However, such relation for DNA methylation profiles has not been determined yet. Using differentially expressed genes in healthy fetal and adult lung as a guidepost, I have identified four genes (*MEOX2*, *MDK*, *LAPTM5* and *FGFR3*) that are differentially methylated in lung cancer. I found that *MEOX2* was uniformly hypermethylated in lung cancer compared to healthy adult and fetal lung. Furthermore I have provided evidence that the methylation is not uniform in lung cancer but is rather correlated with either the differentiation state (*MDK*, *LAPTM5*) or the lung cancer type (*FGFR3*).

MEOX2 (also known as growth arrest specific homeobox gene, *GAX*) was uniformly hypermethylated in lung cancer tissues compared to healthy adult and fetal lung specimens. Besides its role in somite formation and patterning [225] *MEOX2* is expressed in vascular endothelial cells and inhibits angiogenesis [226]. Hence, the observed hypermethylation of the *MEOX2* 5'-UTR, combined with the inverse correlation of gene expression and 5'-UTR methylation in healthy fetal and adult lung might indicate that *MEOX2* is silenced in tumors via DNA methylation and thus allows for neo-vascularization required for tumor growth and spread.

The methylation status of *MDK* and *LAPTM5* in poorly differentiated lung tumors was significantly altered when compared to moderately differentiated lung tumors. For *MDK*, I observed hypomethylation in poorly differentiated tumors. *MDK* exerts pleiotropic biological activities related to tumorigenesis (i.e. fibrinolytic, anti-apoptotic, mitogenic, transforming, angiogenic, and chemotactic activities) [227] and it has been proposed as a prognostic marker in gastrointestinal stromal tumors

(GISTs) where expression of *MDK* was correlated with a worse outcome compared to MDK-negative GISTs [228]. In contrast, *LAPTM5* was hypermethylated in poorly differentiated lung tumors compared to moderately differentiated lung tumors. This gene is located on chromosome 1p34, a region frequently involved in chromosomal rearrangements in lung cancer [229]. Similar to multiple myeloma cell lines, where *LAPTM5* is often silenced by hypermethylation [230], silencing of this gene by DNA methylation may therefore contribute to the development of more aggressive cancer phenotypes as reflected by the poorly differentiated histology. As histological grade represents an independent prognostic factor for survival in non-small cell lung cancer [195], it is conceivable that both *MDK* and *LAPTM5* methylation profiles can be used as marker for tissue differentiation and might represent prognostic markers for lung cancer. Further studies using more samples will help to elucidate the prognostic power of these markers.

Differential methylation in *FGFR3* allows the differentiation of squamous cell carcinoma samples from adenocarcinoma. Squamous cell carcinoma tumors are usually slow growing, with late metastasis predominately to the liver, adrenal glands, kidneys, and bones [231]. Recently developed therapies based on antiangiogenic anti-VEGF antibodies, i.e. bevacizumab (Avastin, Roche), are associated to adverse effects in the circulatory system (hemorrhage, hypertension, tromboembolic events and proteinuria) in squamous cell carcinoma patients, and only patients without predominant squamous cell histology are eligible to receive bevacizumab [232]. Currently, classification of non-small cell lung carcinomas (NSCLC) is based on histopathological review, but lung tumor heterogeneity hinders the precise classification in many cases [149]. For that reason, molecular markers that allow the precise classification of NSCLC are highly desired. In this regard, Dosaka-Akita and colleagues reported a low expression of GnT-V gene significantly associated to squamous cell carcinomas [233]. My finding of differential methylation in *FGFR3* nominates this new biomarker for classifying squamous cell carcinoma at DNA level.

3.7. Conservation of DNA methylation between human and mouse

The use of animal models is a driving force in biological research. Particularly in epigenetic studies, methylation status of murine genes is determined and frequently, these results are extrapolated to human orthologues [157] [159] [160]. However, no systematic analyses have been performed so far to determine the degree of conservation of DNA methylation profiles in these two species. In this study, I have determined that the majority of studied regions (69.4%) differed by less than 20% in their methylation profiles in mouse and human. This finding suggests that DNA methylation is evolutionary conserved between human and mouse in the studied tissues. Bernstein and colleagues showed that histone modifications are strongly conserved, even though many of the corresponding sites are not conserved at DNA level [234]. Unlike, I found that syntenic regions displaying conservation of methylation between mouse and human showed high conservation of sequence (homology higher than 75%).

Additionally, I did not find differences in conservation rates neither between 5'-UTR and conserved non-genic regions (CNGs) nor between tissue types. The global conservation of methylation patterns despite its location supports the possibility of other physiological functions for DNA methylation rather than only controlling gene expression in *cis*. Conserved non-genic regions might represent indeed controlling elements for the expression of currently undiscovered non-coding RNAs, which regulate in turn expression of coding genes in *trans*. Antisense transcripts are a common regulatory element in imprinting, as described for mouse *Igf2r* [100] and *Kcnq1* [101], as well as *GNAS1* [235] in human. Alternatively, they might also represent relics of silenced retrotransposons incorporated in a common ancestor, as shown in genome-wide analysis in Arabidopsis [236].

3.8. Methylation of the IGF2R cluster but not imprinting status is conserved in mouse and human orthologues

Imprinting of the *IGF2R* gene has been thoroughly studied in human and mouse. However, the polymorphic occurrence of this phenomenon in healthy adult human tissues and the absence of an orthologue for the antisense *Air* transcript in human

make it still an unresolved issue. Moreover, little is known about the imprinting status of other genes in the cluster and the role of DNA methylation remains unclear. I have found that the methylation profiles of the 5′ UTRs of genes located in the IGF2R locus are conserved between human and mouse. I have shown that *SLC22A2* and *SLC22A3*, for which the imprinting status in adult tissues was previously not known, are not imprinted in healthy human tissues. In addition, I have not found imprinting of human *IGF2R* in these tissues, which is consistent with previous reports showing that human *IGF2R* is only imprinted in some individuals and in some fetal tissues [237]. In mice, *Igf2r*, *Slc22a2* and *Slc22a3* are maternally imprinted [182] [238] and silencing of the paternal allele depends on the expression of the antisense transcript *Air* [100]. In turn, the expression of *Air* is regulated in part by DNA methylation of the *Air* promoter with the maternal allele being hypermethylated and the paternal allele being unmethylated. Therefore, absence of imprinting for all genes within the human IGF2R locus is further supported by the lack of an allele-specifically methylated putative *Air* promoter. The comparative methylation profiling presented in this work revealed that the murine *Air* promoter is about 50% methylated but the orthologous region in human is homogenously hypermethylated. Of all regions profiled within this locus, the *Air* promoter and its homologous region in human were the only regions for which the DNA methylation was not conserved.

3.9. Conservation of DNA methylation upon genome duplications

The mechanisms and signals that lead to tissue-specific methylation are currently not known. Here, I have analyzed the fate of T-DMRs in gene duplication events. For the functional *PLG* gene family, T-DMRs are conserved in all known members independent of the genomic location of each variant. For all three variants, the methylation status of the T-DMRs is inversely correlated with the respective gene expression suggesting that the function of these T-DMRs is conserved as well. Similarly, I found conserved T-DMRs in two pairs of parental/unprocessed pseudogenes, *SLC39A1*/RP4-539M6.7 and *Q6UW61_HUMAN*/KB-1995A5.12. These results suggest that the sequence itself might contain the signal conferring tissue

specific methylation independent of the genomic location. A possibility I currently cannot rule out is that the differential methylation of e.g. *PLGA* and *PLGB* arose independently and re-occurred after gene duplication.

The T-DMRs observed in the TBX family and the vast majority of T-DMRs found in both unprocessed and processed pseudogenes are not conserved suggesting that other mechanisms leading to T-DMRs must exist as well. Possibly, once the gene duplication is transmitted through the germ line, DNA methylation associated with e.g. tissue development may override the existing methylation mark of the duplicated gene and thereby generate a new methylation profile for this gene. I have observed a bias for recently evolved processed pseudogenes being rather unmethylated compared to more distant processed pseudogenes. This finding, if confirmed by a larger gene panel, may indicate that new processed pseudogenes become preferentially integrated into open chromatin structures that are generally associated with unmethylated DNA. A similar bias has been reported for retroviruses such as the human immunodeficiency virus (HIV) that preferentially integrate in open chromatin structures [239].

The function of T-DMRs located within pseudogenes is not known. They could be non-functional evolutionary relics of abortive gene duplications but may as well confer stage- and tissue- specific expression of pseudogenes. Some pseudogenes, although not coding for a functional protein, are transcribed and have a regulatory function [240]. For example, the *NOS* [241] and the *Markorin* [242] pseudogenes regulate expression of the respective parental genes, although others have disputed these results [243] [31].

Rodin and Riggs [70] proposed an epigenetic complementation model (ECM) to predict the fate of gene duplicates. In this model, stage and tissue-specific epigenetic silencing/activation help to maintain negative selection on both copies of the duplicated gene. This epigenetic complementation would thus lead to a complementary expression of the original gene and its twin copy and consequently, complementary gene expression would expose both copies to a purifying selection and may prevent degradation into a pseudogene. In this aspect, it is of interest to note that the methylation profiles observed for the functional *TBX* gene family were very gene-specific and each gene had a distinct methylation profile. In this study, I

have analyzed a limited number of different tissues and cells and it is likely that a more comprehensive analysis would reveal further specific T-DMR profiles. For example, our group has recently shown that *TBX21* is specifically unmethylated in CD4+ Th1 and CD8+ memory lymphocytes but not in CD4+ and CD8+ naïve lymphocytes [244]. In contrast, the very recently evolved *PLG* family members displayed similar expression and methylation profiles that may indicate that some *PLG* genes are destined to pseudogenization. Thus, my findings provide experimental evidence supporting an epigenetic complementation in gene family evolution and point to a prominent role of DNA methylation in shaping the plasticity of the human genome.

3.10. DNA methylation in non-cancer diseases. Modulating the phenotypic features upon similar genetic backgrounds

Variations in DNA methylation profiles between affected and healthy individuals are a hallmark in cancer [204] [106]. Nevertheless, although a role for CpG methylation has been proposed (as detailed in introduction), experimental evidence in other human pathologies is far from conclusive. Because the epigenetic status is more dynamic than changes at DNA sequence level, it has been hypothesized a coordinated role of genetic and epigenetic factors in the etiology of complex disease and suggested that they should be investigated in parallel [173]. Beyond mutations in coding genes for proteins involved in DNA methylation mechanisms that may impair their functions, e.g. mutation in *MECP2* gene in Rett syndrome [245], very few examples are known for phenotypic modeling at epigenetic level upon a common genetic background. In this regard, preliminary evidence supports a model that incorporates both genetic and epigenetic contributions in the causation of autism [246].

Lipodystrophies are familial or acquired disorders characterized by variable loss of fat tissue. In familial partial lipodystrophy (FPLD) loss of fat occurs at puberty. In a FPLD subgroup, Dunnigan type or FPLD2, subcutaneous fat is lost from the limbs and partially from the trunk. Later on, the patients may show acanthosis nigricans, hirsutism, polycystic ovaries, insulin-resistant diabetes, dyslipidemia and liver

steatosis. Although, FPDL2 is caused by mutations in the *LMNA* gene located in chromosome 1 [177], its manifestation is markedly dependent on sex with female carriers being more severely affected [247]. Mutations in the *LMNA* gene are proposed to cause many other diseases including muscular dystrophies [248], progeroid syndromes [249], mandibuloacral dysplasia [250], dilated cardiomyopathy [251] and restrictive dermopathy [252]. The mechanisms by which the different mutations in the *LMNA* gene lead to the very different tissue-specific disease phenotypes are still unclear, but they might be attributable to epigenetic variation of affected patients.

In the present study, I have investigated the relation of DNA methylation and the underlying mutation as a reason for the development of the differential phenotypes in patients bearing *LMNA* mutations. For this purpose, I analyzed the methylation patterns of 10 candidate genes in patients with FPLD2, with other laminopathies and controls. One gene, *RARB*, is significantly highly methylated in patients with FPLD2 compared to other laminopathies. In addition, *LMNA* is highly methylated in FLDP2 patients carrying the *LMNA* R471G mutation. .

The *RARB* gene encodes the retinoic acid receptor beta and mediates retinoid effects in controlling cell growth, differentiation, apoptosis, and carcinogenesis. *RARB* has been shown to be differentially methylated in several cancers, such as malignant melanoma [253], testicular tumors [254] and carcinoma of cervix [122]. Nevertheless, the function of *RARB* in the etiology of FLPD2 is currently not known.

Protease inhibitor therapy of HIV is frequently associated with a FPLD-like condition characterized by lipoatrophy of the limbs, fat accumulation of abdomen, breasts and cervical region ("buffalo hump"), hyperlipidemia, and insulin resistance. This HAART (Highly Active Anti-Retroviral Therapy) related lipodystrophy syndrome may be a result of the inhibition of two proteins involved in lipid metabolism that have significant homology to the catalytic site of HIV protease, namely cytoplasmic retinoic acid binding protein type 1 and low density lipoprotein-receptor-related protein [255]. The cellular retinoic acid binding protein II, *CRABP2*, is an intracellular protein involved in the transmission of the vitamin A-derived signal, which regulates genes responsible for lipid metabolism and adipocyte differentiation [256]. This data suggest a close link between the retinoic acid pathway and the development to

lipodystrophy and support a role for changes in the methylation of the retinoic acid receptor gene in the development of lipodystrophic phenotypes in the investigated patients.

The differential DNA methylation of the *LMNA* gene between both FPLD families may play a role synergic to the different mutation in development of the different FPLD phenotype. Although the difference is visible in the methylation plot, p-values after Wilcoxon test do not show statistical significance. As only three patients per group were available for comparison, marginal p-values are expected. Thus, these results would suggest a trend towards differential methylation of *LMNA* gene between both families carrying different mutations.

The results of my experiments suggest a relation between *LMNA* mutations phenotypes and DNA methylation. The combined effects of *LMNA* mutations and DNA methylation of functionally related genes might explain therefore the variety of phenotypes observed in the different laminopathies.

3.11. DNA methylation markers found in blood. A feature enabling low-invasive diagnostic procedures and screening

Biomarkers at DNA levels, i.e. mutations and SNPs, are generally present in a mosaic manner in affected individuals. Unlike, functional biomarkers such differentially expressed genes, proteins and DNA methylation markers are mainly associated to a tissue or cellular type. Because of this fact, diagnostic procedures based on these markers require invasive procedures (e.g. biopsies) for sampling or the use of related remote samples (e.g. sputum or bronchial lavage in lung cancer). Although they allow screening in general population, biomarkers in remote samples are not always as sensitive and specific as those found in the tissue of origin. Thus, sensitive and specific biomarkers based on easily accessible samples are preferred.

To investigate the methylation status of candidate genes in lipodystrophies, I based the study on cultured lymphocytes isolated from patients' blood samples. However, cell culturing may induce changes in methylation values [181]. To confirm that the discovered markers were not an artifact of culturing procedures, I proved that their

methylation profiles did not change through passages in lymphocytes of three healthy controls (Figure 2.7.4)

The fact that I found methylation differences in cultured lymphocytes from peripheral blood makes them very valuable. Next steps would require the study of their sensitivity and specificity in larger sample sets.

3.12. Concluding remarks

The relation between hypermethylation and the silencing of the associated gene represent one of the most accepted concepts in tissue-specific gene regulation, despite the lack of experimental evidence supporting it. Here, I have approached this issue in two directions. First, using a semiquantitative RT-PCR approach I studied the expression of associated genes to 55 regions discovered as tissue-specifically methylated in healthy human tissues and primary cell lines. The obtained correlation rate (39% of the studied regions at the 5'-UTR) is quite different to that usually reported in the literature obtained by treating cell lines with de-methylating agents and subsequent gene reactivation. The observed differences support the idea of a co-existence of epigenetic mechanisms regulating expression, which might be masked by the overwhelming action of the de-methylating drugs.

On the other hand, I have determined expression profiles for healthy adult and fetal lung identifying the differentially expressed genes. By looking at the methylation status of the 5'-UTR of 43 differentially expressed genes, I have determined a correlation rate of 19%. Besides variations introduced by the different technologies applied, differences in the observed rates (obtained by looking at distinct genes) give support to the idea of several regulation mechanisms acting rather in a gene-specific manner.

As tumor ontogeny often recapitulates transcriptional profiles of embryonic development, it is conceivable to use fetal expression and DNA methylation profiles to guide a biomarker discovery in tumor samples. Here, I have made use of the comparative expression profile and DNA methylation profile analysis obtained from fetal and adult lung tissues to identify novel biomarker candidates for lung tumors. I have identified four genes (*MEOX2*, *MDK*, *LAPTM5* and *FGFR3*) that are

differentially methylated in lung cancer. Whereas I found that *MEOX2* was uniformly hypermethylated in lung cancer compared to healthy adult and fetal lung, I have provided evidence that the methylation is not uniform in lung cancer but is rather correlated with either the differentiation state (*MDK*, *LAPTM5*) or the lung cancer type (*FGFR3*). If confirmed in a larger sample panel, differential methylation on *MEOX2*, *MDK*, *LAPTM5* and FGFR3 will turn to biomarkers with an application in lung cancer molecular characterization.

4. Materials and methods

4.1. Samples and Patients

4.1.1. Tissue samples and primary cells from healthy human individuals

Healthy human tissues samples for DNA isolation

Healthy human tissue samples were obtained from one of the following sources: Asterand, (Detroit, US), Pathlore Plc. (Nottingham, UK), Tissue Transformation Technologies (T-cubed, Edison, US), Northwest Andrology (Missoula, US), NDRI (Philadelphia, US) and Biocat GmBH (Heidelberg, Germany). Only anonymized samples were used and ethical approval was obtained for the study. The approving institutions were the Ethical Commission of the Medical Association in Berlin (Ärztekammer Berlin, Friedrichstrasse 16, 10969 Berlin, Germany) and the Cambridge Local Research Ethics Committee (Box 148, Addenbrooke's NHS Trust, Hills Road, Cambridge CB2 2QQ, UK, approval ID: LREC-03/094). Contamination by blood cells is estimated to be low as blood specific methylation profiles were not detected in the tissues.

Human primary cells for DNA and RNA isolation

Human primary cells were obtained from Cascade Biologics (Mansfield, UK), Cell Applications Inc. (San Diego, US), Analytical Biological Services Inc. (Wilmington, US), Cambrex Bio Science (Verviers, Belgium) and from the DIGZ (Berlin, Germany). Dermal fibroblasts, keratinocytes and melanocytes were cultured according to the supplier's recommendations up to a maximum of 3 passages reducing the risk of aberrant methylation due to extended culturing. As an additional control, the average methylation of selected amplicons obtained from dermal fibroblasts, keratinocytes and melanocytes was compared with the methylation of the same loci in human skin samples. No significant deviation between the methylation of the primary cells and tissues was detected, indicating that cell culturing for a limited

number of passages does not change DNA methylation (data not shown). CD4+ T-lymphocytes were isolated from fresh whole blood by depletion of CD4+ monocytes followed by a negative selection. CD8+ cells were isolated from fresh whole blood by positive selection. Subsequent FACS analysis confirmed a purity of CD4+/CD8+ T-lymphocytes greater than 90%.
Aliquots of the same samples of the human melanocytes, keratinocytes, fibroblasts, CD4+ and CD8+ cells that were used for methylation analysis were used for RNA analysis. Primary cell cultures were harvested and kept at -80 °C until RNA isolation.

Matched DNA/ RNA samples

For correlation studies, matched total RNA/DNA samples (heart muscle, skeletal muscle and liver) were purchased from BioCat GmbH (Heidelberg, Germany).

Isolated RNA samples from human tissues

Isolated total RNA samples from healthy human tissues (heart muscle, skeletal muscle and liver) were purchased from Ambion (Austin, US). Samples were kept at -80°C until used for reverse transcription or microarray analysis.

4.1.2. Tissue samples from mouse

Mouse DNA samples from liver, heart, skin and skeletal muscle were acquired from BioCat GmbH (Heidelberg, Germany).

4.1.3. Patients and controls for the study of DNA methylation profiles in Progeria and FPLD syndromes

Dr. Thomas Brune (University of Magdeburg, Germany) coordinated the recruitment and diagnosis of patients and controls for this study.

Patients and controls group:

Ten patients from two independent families bearing different mutations in the *LMNA* gene were enrolled in the study. Five healthy adults and one healthy member of the one FPLD2 family served as controls. The patients and the controls were evaluated according the following groups (Table 4.1.1):

FPLD Group: Patients 1-3 belong to the same family (Family A) and exhibit the typical phenotype of familial partial lipodystrophy type 2 (FPLD2) including acanthosis nigricans, partial lipodystrophy with sparing of the face, severe hypertriglyceridemia, insulin-resistant diabetes, and severe hypertriglyceridemia. All three carry the mutation R482L in the *LMNA* gene. One healthy member of the family with no mutation serves as family intern control. Patients 4-6 belong to the same family (Family B) and carry the R471G mutation in the *LMNA* gene. Patient 4 is a 14 years old girl affected by an overlapping syndrome including partial lipodystrophy, insulin-resistant diabetes, acanthosis nigricans, liver steatosis, muscle weakness and contractures. Patient 5 is her 18-year-old sister and exhibits only a partial lipodystrophy. Patient 6 is their father carrying the same mutation but does not show a lipodystrophy phenotype. Their healthy mother does not carry any mutation in the *LMNA* gene and serves as a family intern control.

Progeria Group: Patients 7 and 8 are a 1-year-old girl and a 5-years-old boy, respectively. They are affected by Hutchinson-Gilford syndrome (Progeria) reported by De Busk and colleagues [257]. Both patients carry the progeria-typical Cl960T mutation in the *LMNA* gene [179]. Patient 9 is a 5 years old girl with an overlapping syndrome including early-onset myopathy with progeroid features. She carries a *de novo* heterozygous mutation (S143F) in the *LMNA* gene. Patient 10 is a 3.5-years-old boy suffering from scleroderma-like lesions of the skin occurred at 3 months of age. Poor weight gain with loss of subcutaneous fat, prominent scalp veins, progressive hair loss and lentigo senilis were noted shortly after. X-rays revealed the absence of the lateral regions of clavicles as well as acroosteolysis of the fingers and myopathy and cardiomyopathy were absent. Sequencing of *LMNA* gene revealed a heterozygous stop mutation leading to a protein that is elongated by 7 additional aminoacids at the C-terminal. The mother and a healthy brother of the patient carry the same mutation. Sequencing of the *ZMPSTE24* gene showed a homozygous

deletion of an acceptor splice site (IVS9-Ex10) leading to the complete loss-of-function of the mutated ZMPSTE24 [258].

Control group: 5 healthy females (24- 33 years old) individuals served as controls.

Table 4.1.1. Phenotypes and genotypes of the enrolled patients

	Patient number	Sex	Age (years)	Phenotype	Lamin-Mutation	Additional Mutation
FPLD-Group	Patient 1			Familial Partial Lipodystrophy	LMNA A/C R482L	none
	Patient 2			Familial Partial Lipodystrophy	LMNA A/C R482L	none
	Patient 3			Familial Partial Lipodystrophy	LMNA A/C R482L	none
	Patient 4	Female	14	Familial Partial Lipodystrophy	LMNA A/C R471G	none
	Patient 5	Female	18	Familial Partial Lipodystrophy, mild phenotype	LMNA A/C R471G	none
	Patient 6	Male		No phenotype	LMNA A/C R471G	none
Laminopathies-Group	Patient 7	Female	1	Hutchinson-Gilford-Syndrome	LMNA A/C Cl960T	none
	Patient 8	Male	7	Hutchinson-Gilford-Syndrome	LMNA A/C Cl960T	none
	Patient 9	female	5	Early onset myopathy with progeria like symptoms	LMNA A/C S143F	none
	Patient 10	Male	3	Progeria like Symptoms, myopathy and cardiomyopathy were absent	LMNA A/C c1960 CT	ZMPSTE24 IVS9-Ex10
Healthy Control group	Patient 11	female	30	Healthy Control	none	none
	Patient 12	female	32	Healthy Control	none	none
	Patient 13	female	29	Healthy Control	none	none
	Patient 14	female		Healthy Control	none	none
	Patient 15	female		Healthy Control	none	none

Cell culture and DNA isolation from patients and controls:

10 ml venous blood was drawn from each patient and controls into a heparin blood collection tube. The blood samples were fractionated by standard Ficoll-Hypaque method (Ficoll, 1,077 g/ml Nycomed Pharma AS, Oslo, Norway). After 3 times washing with phosphate buffered saline (PBS) the cells were cultured in tissue culture medium RPMI 1640 supplemented with 10% fetal bovine serum and essential aminoacids. The culture was infected with EBV supernatant. After 30 min. incubation, further tissue culture medium (RPMI 1640 supplemented with 12% fetal calf serum (FCS) containing Cyclosporin-A was added. The cells usually started to show morphological changes after 3 to 4 days when dividing cells can be seen as dumbbell shaped structures under an inverted microscope. Six to eight weeks after

showing typical manifestation of big cellular masses an aliquot of the cells was used for further investigation. To exclude an impact of the immortalization process on the methylation pattern of the investigated genes I compared DNA from freshly prepared lymphocytes (A) and the corresponding immortalized lymphocytes (B) of three healthy probands (Figure 2.4.4)

To isolate DNA from B-lymphoblastoid cells, they were pelleted by centrifugation, washed two times in 1 x PBS and resuspended in 2 ml OLD-T buffer (40 mM Tris/HCl, 150 mM NaCl, 25 mM EDTA, pH 7.5). After addition of 50 µl proteinase K and 100 µl 20 % SDS, incubation was carried out at room temperature over night, followed by phenol/chloroform extraction. The DNA was precipitated using 2 volumes of absolute ethanol.

4.1.4. Fetal and Healthy Adult Lung tissue samples

Total RNA samples from fetal and healthy adult lung were acquired from Biocat GmbH (Heidelberg, Germany). Sample set consisted of 24 unrelated samples (12 fetal lung and 12 healthy adult lung samples) including male and female individuals. All fetal lung samples were derived from 20-25 week all fetuses. Samples were kept at – 80 °C until they were used.

Before proceeding with microarray analysis and qRT-PCR, RNA samples passed a very strict upfront quality control in terms of RNA integrity and purity. Integrity of the total RNA was confirmed by running the samples on denaturing urea-agarose gels. Before loading, 1 µg total RNA was mixed with 3 volumes of loading buffer (8 M Urea, 10 mM Tris-HCl, 1 mM EDTA, 1% NP 40, 0.05% Bromophenolblue) and denatured by heating 2 minutes at 95°C. Samples were run on a 2% agarose-1M Urea gel and stained with Ethidium Bromide prior to UV- visualization. Non-degraded total RNAs display two characteristic bands on the gel, corresponding to 18S and 28S ribosomal RNAs. Degraded RNAs would show smears instead of clearly visible band on the denaturing agarose electrophoresis. Moreover, all 24 samples showed A260/280 ratios within the 1.8-2.0 range when diluted in RNAse-free TE buffer, indicating than they were not contaminated with an excess of proteins, which will impair the cDNA synthesis and *in vitro* transcription.

4.1.5. Patients and controls for the study of DNA methylation profiles in Lung Cancer

DNA samples from fetal lung were purchased from Biocat GmbH (Heidelberg, Germany). Tissue samples for DNA isolation from healthy adult and tumor lung were acquired from Integrated Laboratory Services (Bethesda, US) and Asterand (Detroit, US). Lung cancer sample panel consisted of 15 non- small cell lung carcinoma samples (12 adenocarcinomas and 3 squamous cell carcinomas). Samples were pathologically reviewed before DNA isolation and clinical information about the patients was available from the supplier. In all cases, only anonymized samples were used.

4.2. DNA isolation

DNA was extracted using the DNeasy kit (Qiagen, Hilden, Germany) according to the manufacturer's recommendation. Briefly, up to 25 mg of tissue or 1 x 10^6 cells were incubated with Proteinase K (Qiagen, Hilden, Germany) at 37 °C until total disintegration. Next, lysis solution was loaded into silica columns and washed with buffers included in the kit. DNA was finally eluted with the corresponding buffer (AE buffer, included in the kit).
Isolated DNA was quantified spectrophotometrically and purity was determined calculating A260/A280 ratios. Integrity of the DNA was confirmed by running the samples in a 1% agarose gel.

4.3. RNA isolation

Total RNA was isolated using the RNeasy kit (Qiagen, Hilden, Germany) according to supplier's recommendations. Briefly, up to 25 mg of tissue or 1 x 10^6 cells were mechanically disrupted with a rotor-stator device (DIAX100) in a lysis buffer included in the kit, which contained ß-mercaptoethanol. Lysis solution was loaded

into silica columns and washed with buffers included in the kit. Total RNA was eluted with RNAse free water.
Isolated total RNA was quantified spectrophotometrically. Isolated samples also had to pass the same strict quality controls described for the purchased samples. Samples were stored at -80°C until they were used in expression profiling experiments.

4.4. PCR amplification

4.4.1. Amplification of bisulfite-treated DNA

DNA was bisulfite converted and PCR amplified following Epigenomics' proprietary protocols [259]. Bisulfite-specific primers with a minimum length of 18bp were designed using a modified Primer3 program. The target sequence of the designed primers contained no CpGs allowing an unbiased amplification of both hypomethylated and hypermethylated DNAs. Primers that gave rise to an amplicon of the expected size using non-bisulfite treated DNA as a template were discarded, thus ensuring the specificity for bisulfite-converted DNAs. Primers were also tested for specificity by electronic PCR (ePCR).

4.4.2. Amplification of genomic DNA

Genomic DNA amplification was carried out using the HotStartTaq DNA polymerase kit (Qiagen, Hilden, Germany) with 10 ng of genomic DNA and gene-specific primers. Primers for genomic DNA amplification were designed using the Primer3 software [260]. Amplification conditions for the genomic DNA were: 15', 95°C followed by 40 cycles of 92°C for 60 sec., 72°C for 60 sec., 72°C for 60 sec. and a final extension step of 10 minutes at 72°C. Genomic PCR fragments used for the genotyping of the matched DNA/RNA samples contained at least one reported SNP.

4.5. Reverse transcription PCR

4.5.1. Semiquantitative reverse transcription PCR

cDNA was synthesized from total RNA using the Omniscript RT kit (Qiagen, Hilden, Germany) and random hexamers. PCR (92°C for 1 minute, 60°C for 1 minute, 72°C for 1 minute for 40 cycles) was performed using the HotStartTaq DNA polymerase kit (Qiagen) with 3 µl of the prepared cDNA and gene-specific primers. All kits were used according to the manufacturer's recommendations. PCR products were analyzed by electrophoresis on 2.5 % agarose gels. All RT-PCR fragments were designed spanning at least one intron to avoid amplification of contaminating genomic DNA. Universal RNA (BioCat, Heidelberg, Germany) was used as positive control. RT-PCR fragments used to determine allelic expression contained at least one reported SNP.

4.5.2. Quantitative reverse transcription PCR

cDNA was synthesized using the Omniscript RT kit (Qiagen, Hilden, Germany) and a mixture of poly T primer (1 µM) and random hexamers (10 µM). This feature allowed the transcription of long transcripts avoiding at the same time the overestimation of mRNA copy number produced by random hexamers [261]. PCR was performed using the SYBR Green PCR Master Mix (Applied Biosystems, Foster City, US) with 2 µl of a 1:100 dilution of the prepared cDNA and gene-specific primers. All kits were used according to the manufacturer's recommendations. All qRT-PCR fragments had a maximum length of 150 bp and were designed spanning at least one intron to avoid amplification of contaminating genomic DNA. Universal RNA (BioCat, Heidelberg, Germany) was used as positive control.

4.6. Bisulfite sequencing

4.6.1. Direct sequencing strategy

Direct sequencing was carried out as previously described [262] [66] [67]. PCR amplicons from bisulfite treated DNA, genomic DNA and semiquantitative RT- PCR products were quality controlled by agarose gel electrophoresis. Amplicons showing

single and intense bands were purified with ExoSAP-IT (USB Corporation, Cleveland, US) to remove any excess nucleotides and primers, and sequenced directly in forward and reverse directions.

Sequencing was performed on an ABI 3730 capillary sequencer using 1/20th dilution of ABI Prism BigDye terminator V3.1 sequencing chemistry after hotstart (96°C for 30 seconds) thermocycling (92°C for 5 seconds, 50°C for 5 seconds, 60°C for 120 seconds x 44 cycles). Before injection, products were purified on DyeEx plates (Qiagen, Hilden, Germany). PCR and RT-PCR fragments were directly sequenced with the same primers as in the PCR reaction. The obtained sequencing chromatograms were used to quantify the methylation at a given CpG as described below. Samples and expressed alleles were genotyped by identification of annotated SNPs in the trace files.

4.6.2. Sequencing of subcloned fragments

For the analysis of allele specific methylation, PCR products from bisulfite-treated DNA were cloned into a TA-cloning plasmid according to the manufacturer's instruction (pGEM- T-Easy cloning kit, Promega) and DNA was isolated using a Qiaprep Spin Plasmid Miniprep kit (Qiagen, Hilden, Germany) according to manufacturer's instructions. Cloned PCR products were sequenced in forward and reverse direction using M13 primers (M13-F: TGTAAAACGACGGCCAGT, M13-R: CAGGAAACAGCTATGACC). Sequencing was performed using the same conditions described for direct sequencing. Methylation in particular CpG positions was determined by visual inspection of the trace files.

4.7. Expression profiling using Affymetrix GeneChip™ microarrays

Figure 4.7.1 illustrates the workflow followed for the target preparation and hybridization using Affymetrix GeneChip system. All steps were performed using kits and chemicals acquired from Affymetrix (Santa Clara, USA) according to manufacturer's instructions.

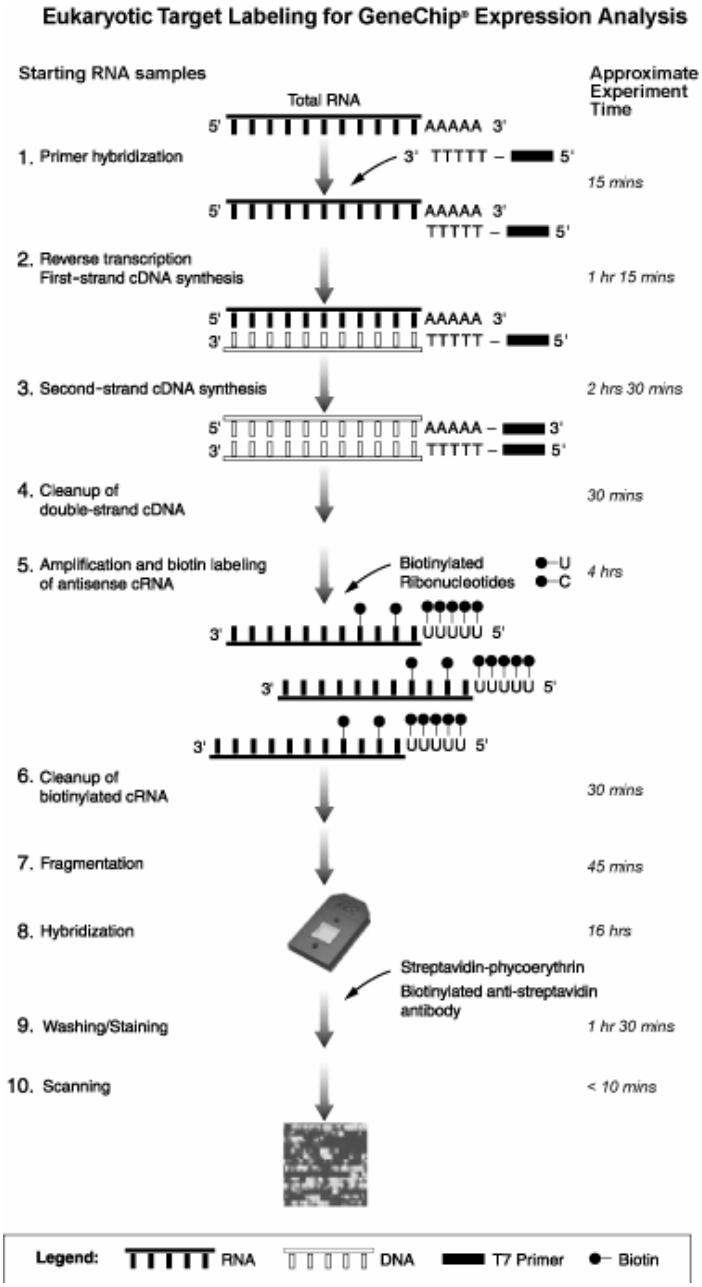


Figure 4.7.1: Experimental workflow for genome wide expression profiling. Double-stranded cDNA is produced by reverse transcription. Amplification of double stranded DNA with biotin labeled nucleotides produces labeled antisense copy RNA (cRNA). cRNA is purified, fragmented and hybridized to microarrays. After hybridization, microarrays are washed and stained with streptoavidin-phycoerithrin conjugated and scanned. Reproduced from www.affymetrix.com

Briefly, double-stranded cDNA was prepared from 5 µg of Total RNA in a One-Cycle cDNA synthesis reaction. After clean-up, biotin-labeled cRNA was produced by *in vitro* transcription, purified and fragmented prior to hybridization. The quality of all the steps was controlled by electrophoresis on agarose gels or by spectrophotometrical measurement, accordingly. Fragmented cRNA was hybridized

to HG-U133A GeneChip microarrays (Affymetrix, Santa Clara, USA) according to manufacturer's recommendations. After hybridization, microarrays were washed and stained with Streptoavidin Phycoerythrin (SAPE) conjugate. Microarrays were scanned and raw data produced for further analysis, as described below.

4.8. Analysis and statistical methods

4.8.1. DNA methylation studies

Sequencing results were analyzed using the ESME program as previously described [262] [66] [67]. This program allows quantifying methylation rates from electropherogram curves.

Figure 4.8.1 shows the rationale of the ESME analysis method. DNA samples are mainly heterogeneous in terms of methylation. Among the genome copies particular CpG positions are either methylated (mC) or unmethylated (C). Bisulfite treatment of genomic DNA converts unmethylated cytosine (C) nucleotides into uracil (U) [48]. Thus, differences in methylation after bisulfite treatment produce single nucleotide polymorphisms, which can be detected by sequencing of PCR products. During the amplification step, uracil nucleotides are replaced by thymine (T). In a direct sequencing strategy, obtained trace files of heterogeneously methylated CpG positions will show both, C signals for methylated copies and T signals for unmethylated copies. Analysis method in ESME software includes the normalization of the obtained peaks and further quantification of methylation by comparing the areas below the curves in the corresponding peaks. Furthermore, conversion rates for unmethylated Cs are determined as a control of the performance of the bisulfite treatment. Only sequences that display a conversion rate higher than 98% were used for analysis. Methylation values for individual Cs within each amplicon and sample are calculated and visualized by color coding ranging from dark blue (100% methylation) to yellow (0% methylation). Differential methylation between the studied groups was tested for statistical significance using Wilcoxon's rank sum test [263] and Kruskal-Wallis test [194].

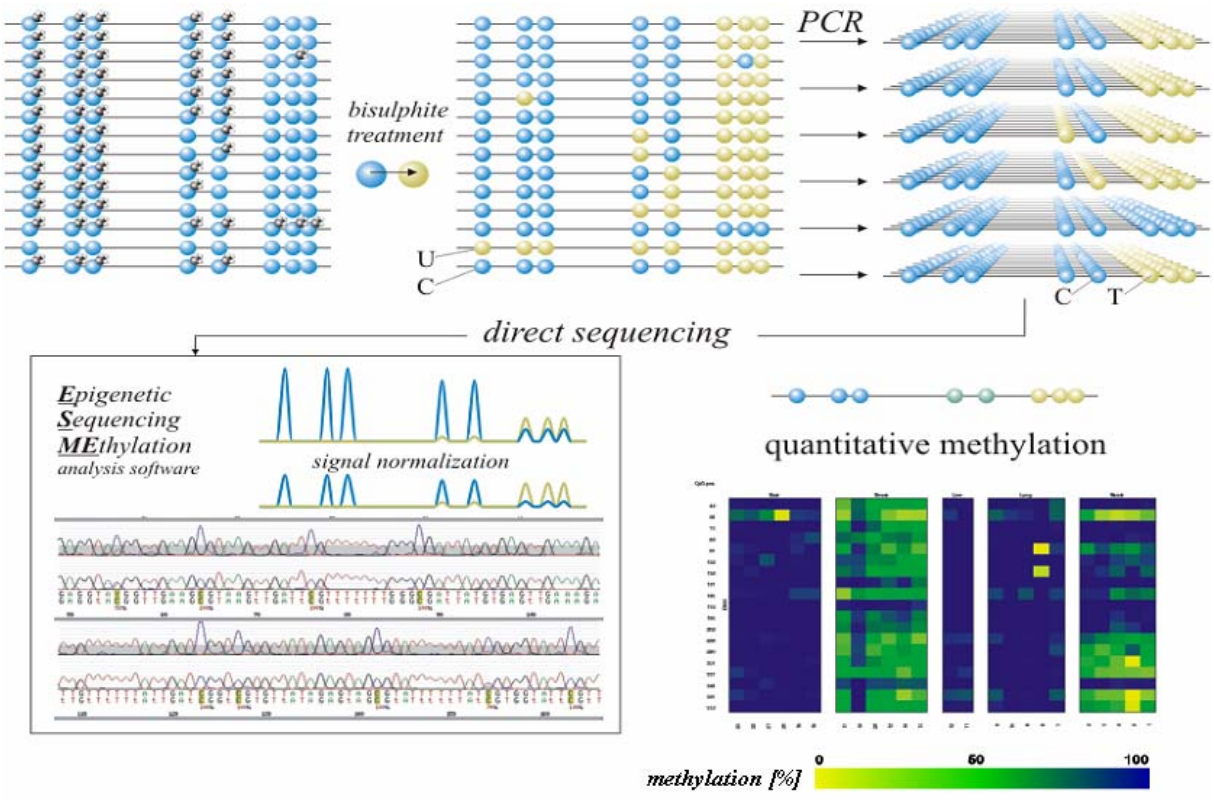


Figure 4.8.1: The sodium bisulfite treatment and direct bisulfite sequencing strategy. Genomic DNA containing unmethylated (blue spheres) and methylated (gray lollypops) cytosines is treated with sodium bisulfite as indicated [259]. Unmethylated cytosines are converted to uracils (orange spheres) whereas methylated cytosines remain unchanged. Converted DNA is amplified by PCR. Produced uracils are replaced by thymines. PCR products are directly sequenced and raw trace files are analyzed using the ESME software [262]. Analysis method includes the normalization of the obtained peaks and further quantification of methylation by comparing the areas below the curves in the corresponding peaks. Quantitative methylation results are displayed as color-coded matrices ranging from yellow (~0% methylation) to green (~50% methylation) and dark blue (~100% methylation)

4.8.2. DNA distance matrices and radial trees

DNA distance matrices and radial trees were constructed with the PHYLIP software package [264]. Distances between the internal node and the pseudogene sequence d(HΨ,IN) were computed according to the following equation: d(HΨ,IN)= [d(H,HΨ) + d(HΨ,C) – d(H,C)] /2, where d(H,HΨ) is the distance between the parental gene and pseudogene sequences, d(HΨ,C) the distance between the pseudogene and chimpanzee gene sequences and d(H,C) the distance between the parental human

and chimpanzee gene sequences, as obtained in the DNA distance matrix for each gene group. To determine whether methylation values were conserved between parental and derivative sequences, I assigned the methylation values in two tissues per pair. I binned the obtained methylation values to 0%, 25%, 50%, 75% and 100%, as shown in Table 2.3.2. Pairs displaying methylation differences lower than 25% were considered unchanged.

4.8.3. Microarray expression profiling

After microarray scanning, probeset intensities were calculated from the raw data according to the Robust Multiarray Analysis (RMA) method [189]. All used methods are implemented in the *affy*, *affyPLM* and *simpleaffy* packages of the Bioconductor project (www.bioconductor.org). Applied tests for statistical analysis and determination of differentially expressed genes included Westfal &Yound maxT algorithm, Benjamini and Hochberg false discovery rates (FDR), Bonferroni corrected two-sample Welch t-statistics for unequal variances, reviewed in [190]. All methods used are implemented in the *multtest* package of the Bioconductor project. As a selection criterion in this experiment, I have chosen a minimum absolute (log)-fold change of 2, Log2Fc>2. This cutoff allows selecting genes that are, at least, four-times highly expressed in the corresponding tissues.

4.8.4. q-RT-PCR expression analysis

Expression data produced by single gene qRT-PCR was processed using the qGene application [265]. Normalized gene expressions were calculated using Mean Normalized Expressions. Target gene expression was standardized against the expression of the *GAPDH* gene. This gene was chosen as a calibrator after evaluating six genes in the model system (*GAPDH*, *ACTB*, *UBC*, *RPL13A*, *PSMB6* and *ATP50*) using Repeated Pair-Wise Correlation Analysis implemented in the BestKeeper algorithm [191]. Differential expression in fetal and healthy adult groups was determined by two-tailed unpaired t-Test, implemented in the qGene application.

4.8.5. Pathway analysis

MAPPFinder software relates microarray data from differentially expressed genes and annotated pathways in GenMAPP as well as the annotations from the Gene Ontology (GO) Consortium [192]. This software calculates a cumulative total of genes changed for a pathway or GO group and assign a statistical value, the Z-score. This score represents the standard statistical test under the hypergeometrical distribution and is calculated by subtracting the expected number of genes meeting the criterion from the observed number of genes, and then dividing by the standard deviation of the observed number of genes.

5. References

1. Rakyan, V.K., *et al.*, *The marks, mechanisms and memory of epigenetic states in mammals.* Biochem J., 2001. **356**(Pt 1): p. 1-10.

2. Paz, M.F., *et al.*, *Germ-line variants in methyl-group metabolism genes and susceptibility to DNA methylation in normal tissues and human primary tumors.* Cancer Res., 2002. **62**(15): p. 4519-24.

3. Herman, J.G. and S.B. Baylin, *Gene silencing in cancer in association with promoter hypermethylation.* N Engl J Med., 2003. **349**(21): p. 2042-54.

4. Gardiner-Garden, M. and M. Frommer, *CpG islands in vertebrate genomes.* J Mol Biol., 1987. **196**(2): p. 261-82.

5. Jeltsch, A., *Molecular enzymology of mammalian DNA methyltransferases.* Curr Top Microbiol Immunol., 2006. **301**: p. 203-25.

6. Yoder, J.A., *et al.*, *DNA (cytosine-5)-methyltransferases in mouse cells and tissues. Studies with a mechanism-based probe.* J Mol Biol., 1997. **270**(3): p. 385-95.

7. Siedlecki, P. and P. Zielenkiewicz, *Mammalian DNA methyltransferases.* Acta Biochim Pol., 2006. **53**(2): p. 245-56. Epub 2006 Apr 3.

8. Okano, M., S. Xie, and E. Li, *Dnmt2 is not required for de novo and maintenance methylation of viral DNA in embryonic stem cells.* Nucleic Acids Res., 1998. **26**(11): p. 2536-40.

9. Hermann, A., S. Schmitt, and A. Jeltsch, *The human Dnmt2 has residual DNA-(cytosine-C5) methyltransferase activity.* J Biol Chem., 2003. **278**(34): p. 31717-21. Epub 2003 Jun 6.

10. Okano, M., *et al.*, *DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development.* Cell., 1999. **99**(3): p. 247-57.

11. Villar-Garea, A. and M. Esteller, *DNA demethylating agents and chromatin-remodelling drugs: which, how and why?* Curr Drug Metab., 2003. **4**(1): p. 11-31.

12. Santi, D.V., C.E. Garrett, and P.J. Barr, *On the mechanism of inhibition of DNA-cytosine methyltransferases by cytosine analogs.* Cell., 1983. **33**(1): p. 9-10.

13. Bestor, T.H. and G.L. Verdine, *DNA methyltransferases.* Curr Opin Cell Biol., 1994. **6**(3): p. 380-9.

14. Tate, P.H. and A.P. Bird, *Effects of DNA methylation on DNA-binding proteins and gene expression.* Curr Opin Genet Dev., 1993. **3**(2): p. 226-31.

15. Prendergast, G.C. and E.B. Ziff, *Methylation-sensitive sequence-specific DNA binding by the c-Myc basic region.* Science., 1991. **251**(4990): p. 186-9.

16. Mancini, D.N., *et al.*, *Site-specific DNA methylation in the neurofibromatosis (NF1) promoter interferes with binding of CREB and SP1 transcription factors.* Oncogene., 1999. **18**(28): p. 4108-19.

17. Santoro, R. and I. Grummt, *Molecular mechanisms mediating methylation-dependent silencing of ribosomal gene transcription.* Mol Cell., 2001. **8**(3): p. 719-25.

18. Takizawa, T., *et al.*, *DNA methylation is a critical cell-intrinsic determinant of astrocyte differentiation in the fetal brain.* Dev Cell., 2001. **1**(6): p. 749-58.

19. Prokhortchouk, E. and B. Hendrich, *Methyl-CpG binding proteins and cancer: are MeCpGs more important than MBDs?* Oncogene., 2002. **21**(35): p. 5394-9.

20. Bird, A.P. and A.P. Wolffe, *Methylation-induced repression--belts, braces, and chromatin.* Cell., 1999. **99**(5): p. 451-4.

21. Prokhortchouk, A., *et al.*, *The p120 catenin partner Kaiso is a DNA methylation-dependent transcriptional repressor.* Genes Dev., 2001. **15**(13): p. 1613-8.

22. Bellacosa, A., *et al.*, *MED1, a novel human methyl-CpG-binding endonuclease, interacts with DNA mismatch repair protein MLH1.* Proc Natl Acad Sci U S A., 1999. **96**(7): p. 3969-74.

23. Hendrich, B., *et al.*, *The thymine glycosylase MBD4 can bind to the product of deamination at methylated CpG sites.* Nature., 1999. **401**(6750): p. 301-4.

24. Guy, J., *et al.*, *A mouse Mecp2-null mutation causes neurological symptoms that mimic Rett syndrome.* Nat Genet., 2001. **27**(3): p. 322-6.

25. Chen, R.Z., *et al.*, *Deficiency of methyl-CpG binding protein-2 in CNS neurons results in a Rett-like phenotype in mice.* Nat Genet., 2001. **27**(3): p. 327-31.

26. Suzuki, M., *et al.*, *Direct association between PU.1 and MeCP2 that recruits mSin3A-HDAC complex for PU.1-mediated transcriptional repression.* Oncogene., 2003. **22**(54): p. 8688-98.

27. Yoon, H.G., *et al.*, *N-CoR mediates DNA methylation-dependent repression through a methyl CpG binding protein Kaiso.* Mol Cell., 2003. **12**(3): p. 723-34.

28. El-Osta, A., *et al.*, *Precipitous release of methyl-CpG binding protein 2 and histone deacetylase 1 from the methylated human multidrug resistance gene (MDR1) on activation.* Mol Cell Biol., 2002. **22**(6): p. 1844-57.

29. Cameron, E.E., *et al.*, *Synergy of demethylation and histone deacetylase inhibition in the re-expression of genes silenced in cancer.* Nat Genet., 1999. **21**(1): p. 103-7.

30. Fukuda, H., *et al.*, *Simple histone acetylation plays a complex role in the regulation of gene expression.* Brief Funct Genomic Proteomic., 2006. **5**(3): p. 190-208.

31. Futscher, B.W., *et al.*, *Role for DNA methylation in the control of cell type specific maspin expression.* Nat Genet., 2002. **31**(2): p. 175-9. Epub 2002 May 20.

32. Costello, J.F. and P.M. Vertino, *Methylation matters: a new spin on maspin.* Nat Genet., 2002. **31**(2): p. 123-4.

33. Harikrishnan, K.N., *et al.*, *Brahma links the SWI/SNF chromatin-remodeling complex with MeCP2-dependent transcriptional silencing.* Nat Genet., 2005. **37**(3): p. 254-64. Epub 2005 Feb 6.

34. Vire, E., *et al.*, *The Polycomb group protein EZH2 directly controls DNA methylation.* Nature., 2006. **439**(7078): p. 871-4. Epub 2005 Dec 14.

35. Csankovszki, G., A. Nagy, and R. Jaenisch, *Synergism of Xist RNA, DNA methylation, and histone hypoacetylation in maintaining X chromosome inactivation.* J Cell Biol., 2001. **153**(4): p. 773-84.

36. Constancia, M., *et al.*, *Deletion of a silencer element in Igf2 results in loss of imprinting independent of H19.* Nat Genet., 2000. **26**(2): p. 203-6.

37. Eden, S., *et al.*, *An upstream repressor element plays a role in Igf2 imprinting.* Embo J., 2001. **20**(13): p. 3518-25.

38. Hark, A.T., *et al.*, *CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus.* Nature., 2000. **405**(6785): p. 486-9.

39. Jones, P.A. and D. Takai, *The role of DNA methylation in mammalian epigenetics.* Science., 2001. **293**(5532): p. 1068-70.

40. Kuo, K.C., *et al.*, *Quantitative reversed-phase high performance liquid chromatographic determination of major and modified deoxyribonucleosides in DNA.* Nucleic Acids Res., 1980. **8**(20): p. 4763-76.

41. Fraga, M.F., *et al.*, *High-performance capillary electrophoretic method for the quantification of 5-methyl 2'-deoxycytidine in genomic DNA: application to plant, animal and human cancer tissues.* Electrophoresis., 2002. **23**(11): p. 1677-81.

42. Wu, J., *et al.*, *Expression of an exogenous eukaryotic DNA methyltransferase gene induces transformation of NIH 3T3 cells.* Proc Natl Acad Sci U S A., 1993. **90**(19): p. 8891-5.

43. Oakeley, E.J., A. Podesta, and J.P. Jost, *Developmental changes in DNA methylation of the two tobacco pollen nuclei during maturation.* Proc Natl Acad Sci U S A., 1997. **94**(21): p. 11721-5.

44. Southern, E.M., *Detection of specific sequences among DNA fragments separated by gel electrophoresis.* J Mol Biol., 1975. **98**(3): p. 503-17.

45. Singer-Sam, J., *et al.*, *A quantitative HpaII-PCR assay to measure methylation of DNA from a small number of cells.* Nucleic Acids Res., 1990. **18**(3): p. 687.

46. Herman, J.G., *et al.*, *Methylation-specific PCR: a novel PCR assay for methylation status of CpG islands.* Proc Natl Acad Sci U S A., 1996. **93**(18): p. 9821-6.

47. Xiong, Z. and P.W. Laird, *COBRA: a sensitive and quantitative DNA methylation assay.* Nucleic Acids Res., 1997. **25**(12): p. 2532-4.

48. Frommer, M., *et al.*, *A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands.* Proc Natl Acad Sci U S A., 1992. **89**(5): p. 1827-31.

49. Gonzalgo, M.L. and P.A. Jones, *Rapid quantitation of methylation differences at specific sites using methylation-sensitive single nucleotide primer extension (Ms-SNuPE).* Nucleic Acids Res., 1997. **25**(12): p. 2529-31.

50. Eads, C.A., *et al.*, *MethyLight: a high-throughput assay to measure DNA methylation.* Nucleic Acids Res., 2000. **28**(8): p. E32.

51. Worm, J., A. Aggerholm, and P. Guldberg, *In-tube DNA methylation profiling by fluorescence melting curve analysis.* Clin Chem., 2001. **47**(7): p. 1183-9.

52. Aggerholm, A., *et al.*, *Extensive intra- and interindividual heterogeneity of p15INK4B methylation in acute myeloid leukemia.* Cancer Res., 1999. **59**(2): p. 436-41.

53. Maekawa, M., *et al.*, *DNA methylation analysis using bisulfite treatment and PCR-single-strand conformation polymorphism in colorectal cancer showing microsatellite instability.* Biochem Biophys Res Commun., 1999. **262**(3): p. 671-6.

54. Adorjan, P., *et al.*, *Tumour class prediction and discovery by microarray-based DNA methylation analysis.* Nucleic Acids Res., 2002. **30**(5): p. e21.

55. Huang, T.H., M.R. Perry, and D.E. Laux, *Methylation profiling of CpG islands in human breast cancer cells.* Hum Mol Genet., 1999. **8**(3): p. 459-70.

56. Costello, J.F., *et al.*, *Aberrant CpG-island methylation has non-random and tumour-type-specific patterns.* Nat Genet., 2000. **24**(2): p. 132-8.

57. Toyota, M., *et al.*, *Identification of differentially methylated sequences in colorectal cancer by methylated CpG island amplification.* Cancer Res., 1999. **59**(10): p. 2307-12.

58. Gonzalgo, M.L., *et al.*, *Identification and characterization of differentially methylated regions of genomic DNA by methylation-sensitive arbitrarily primed PCR.* Cancer Res., 1997. **57**(4): p. 594-9.

59. Shiraishi, M., Y.H. Chuu, and T. Sekiya, *Isolation of DNA fragments associated with methylated CpG islands in human adenocarcinomas of the lung using a methylated DNA binding column and denaturing gradient gel electrophoresis.* Proc Natl Acad Sci U S A., 1999. **96**(6): p. 2913-8.

60. Brock, G.J., *et al.*, *A novel technique for the identification of CpG islands exhibiting altered methylation patterns (ICEAMP).* Nucleic Acids Res., 2001. **29**(24): p. E123.

61. Ballestar, E., *et al.*, *Methyl-CpG binding proteins identify novel sites of epigenetic inactivation in human cancer.* Embo J., 2003. **22**(23): p. 6335-45.

62. Hayatsu, H., *et al.*, *Reaction of sodium bisulfite with uracil, cytosine, and their derivatives.* Biochemistry., 1970. **9**(14): p. 2858-65.

63. Venter, J.C., *et al.*, *The sequence of the human genome.* Science., 2001. **291**(5507): p. 1304-51.

64. Jones, P.A. and R. Martienssen, *A blueprint for a Human Epigenome Project: the AACR Human Epigenome Workshop.* Cancer Res., 2005. **65**(24): p. 11241-6.

65. Esteller, M., *The necessity of a human epigenome project.* Carcinogenesis., 2006. **27**(6): p. 1121-5. Epub 2006 May 13.

66. Rakyan, V.K., *et al.*, *DNA methylation profiling of the human major histocompatibility complex: a pilot study for the human epigenome project.* PLoS Biol., 2004. **2**(12): p. e405. Epub 2004 Nov 23.

67. Eckhardt, F., *et al.*, *DNA methylation profiling of human chromosomes 6, 20 and 22.* Nat Genet., 2006. **38**(12): p. 1378-85. Epub 2006 Oct 29.

68. Stam, M., *et al.*, *Position-dependent methylation and transcriptional silencing of transgenes in inverted T-DNA repeats: implications for posttranscriptional silencing of homologous host genes in plants.* Mol Cell Biol., 1998. **18**(11): p. 6165-77.

69. Sutter, D. and W. Doerfler, *Methylation of integrated adenovirus type 12 DNA sequences in transformed cells is inversely correlated with viral gene expression.* Proc Natl Acad Sci U S A., 1980. **77**(1): p. 253-6.

70. Rodin, S.N. and A.D. Riggs, *Epigenetic silencing may aid evolution by gene duplication.* J Mol Evol., 2003. **56**(6): p. 718-29.

71. Linn, F., *et al.*, *Epigenetic changes in the expression of the maize A1 gene in Petunia hybrida: role of numbers of integrated gene copies and state of methylation.* Mol Gen Genet., 1990. **222**(2-3): p. 329-36.

72. Finnegan, E.J., *et al.*, *Dna Methylation in Plants.* Annu Rev Plant Physiol Plant Mol Biol., 1998. **49**: p. 223-247.

73. Orend, G., *et al.*, *The initiation of de novo methylation of foreign DNA integrated into a mammalian genome is not exclusively targeted by nucleotide sequence.* J Virol., 1995. **69**(2): p. 1226-42.

74. Toth, M., U. Muller, and W. Doerfler, *Establishment of de novo DNA methylation patterns. Transcription factor binding and deoxycytidine methylation at CpG and non-CpG sequences in an integrated adenovirus promoter.* J Mol Biol., 1990. **214**(3): p. 673-83.

75. D'Errico, I., G. Gadaleta, and C. Saccone, *Pseudogenes in metazoa: origin and features.* Brief Funct Genomic Proteomic., 2004. **3**(2): p. 157-67.

76. Bibikova, M., *et al.*, *High-throughput DNA methylation profiling using universal bead arrays.* Genome Res., 2006. **16**(3): p. 383-93. Epub 2006 Jan 31.

77. Khulan, B., *et al.*, *Comparative isoschizomer profiling of cytosine methylation: the HELP assay.* Genome Res., 2006. **16**(8): p. 1046-55. Epub 2006 Jun 29.

78. Song, F., *et al.*, *Association of tissue-specific differentially methylated regions (TDMs) with differential gene expression.* Proc Natl Acad Sci U S A., 2005. **102**(9): p. 3336-41. Epub 2005 Feb 22.

79. Grunau, C., W. Hindermann, and A. Rosenthal, *Large-scale methylation analysis of human genomic DNA reveals tissue-specific differences between the methylation profiles of genes and pseudogenes.* Hum Mol Genet., 2000. **9**(18): p. 2651-63.

80. Lynch, M. and J.S. Conery, *The evolutionary fate and consequences of duplicate genes.* Science., 2000. **290**(5494): p. 1151-5.

81. Ohno, S., *Evolution by gene duplication*. 1970, Berlin: Springer-Verlag. 160.

82. Naiche, L.A., *et al.*, *T-box genes in vertebrate development.* Annu Rev Genet., 2005. **39**: p. 219-39.

83. Frank, S.L., *et al.*, *A gene homologous to plasminogen located on human chromosome 2q11-p11.* Genomics., 1989. **4**(3): p. 449-51.

84. Lewis, V.O., *et al.*, *Homologous plasminogen N-terminal and plasminogen-related gene A and B peptides. Characterization of cDNAs and recombinant fusion proteins.* Eur J Biochem., 1999. **259**(3): p. 618-25.

85. Petersen, T.E., *et al.*, *Characterization of the gene for human plasminogen, a key proenzyme in the fibrinolytic system.* J Biol Chem., 1990. **265**(11): p. 6104-11.

86. Li, E., T.H. Bestor, and R. Jaenisch, *Targeted mutation of the DNA methyltransferase gene results in embryonic lethality.* Cell., 1992. **69**(6): p. 915-26.

87. Reik, W., W. Dean, and J. Walter, *Epigenetic reprogramming in mammalian development.* Science., 2001. **293**(5532): p. 1089-93.

88. Riggs, A.D., *X inactivation, differentiation, and DNA methylation.* Cytogenet Cell Genet., 1975. **14**(1): p. 9-25.

89. Holliday, R. and J.E. Pugh, *DNA modification mechanisms and gene activity during development.* Science., 1975. **187**(4173): p. 226-32.

90. Walsh, C.P. and T.H. Bestor, *Cytosine methylation and mammalian development.* Genes Dev., 1999. **13**(1): p. 26-34.

91. Siegfried, Z., *et al.*, *DNA methylation represses transcription in vivo.* Nat Genet., 1999. **22**(2): p. 203-6.

92. Thomassin, H., *et al.*, *Glucocorticoid-induced DNA demethylation and gene memory during development.* Embo J., 2001. **20**(8): p. 1974-83.

93. Ikeda, K., *et al.*, *Loss of expression of type IV collagen alpha5 and alpha6 chains in colorectal cancer associated with the hypermethylation of their promoter region.* Am J Pathol., 2006. **168**(3): p. 856-65.

94. Shu, J., *et al.*, *Silencing of bidirectional promoters by DNA methylation in tumorigenesis.* Cancer Res., 2006. **66**(10): p. 5077-84.

95. McGrath, J. and D. Solter, *Completion of mouse embryogenesis requires both the maternal and paternal genomes.* Cell., 1984. **37**(1): p. 179-83.

96. Surani, M.A., S.C. Barton, and M.L. Norris, *Development of reconstituted mouse eggs suggests imprinting of the genome during gametogenesis.* Nature., 1984. **308**(5959): p. 548-50.

97. Morison, I.M., J.P. Ramsay, and H.G. Spencer, *A census of mammalian imprinting.* Trends Genet., 2005. **21**(8): p. 457-65.

98. Reik, W. and J. Walter, *Genomic imprinting: parental influence on the genome.* Nat Rev Genet., 2001. **2**(1): p. 21-32.

99. Vu, T.H., T. Li, and A.R. Hoffman, *Promoter-restricted histone code, not the differentially methylated DNA regions or antisense transcripts, marks the imprinting status of IGF2R in human and mouse.* Hum Mol Genet., 2004. **13**(19): p. 2233-45. Epub 2004 Aug 4.

100. Sleutels, F., R. Zwart, and D.P. Barlow, *The non-coding Air RNA is required for silencing autosomal imprinted genes.* Nature., 2002. **415**(6873): p. 810-3.

101. Thakur, N., *et al.*, *An antisense RNA regulates the bidirectional silencing property of the Kcnq1 imprinting control region.* Mol Cell Biol., 2004. **24**(18): p. 7855-62.

102. Delaval, K. and R. Feil, *Epigenetic regulation of mammalian genomic imprinting.* Curr Opin Genet Dev., 2004. **14**(2): p. 188-95.

103. Li, M., J.A. Squire, and R. Weksberg, *Molecular genetics of Beckwith-Wiedemann syndrome.* Curr Opin Pediatr., 1997. **9**(6): p. 623-9.

104. Cassidy, S.B. and S. Schwartz, *Prader-Willi and Angelman syndromes. Disorders of genomic imprinting.* Medicine (Baltimore)., 1998. **77**(2): p. 140-51.

105. Jones, P.A. and S.B. Baylin, *The fundamental role of epigenetic events in cancer.* Nat Rev Genet., 2002. **3**(6): p. 415-28.

106. Feinberg, A.P., R. Ohlsson, and S. Henikoff, *The epigenetic progenitor origin of human cancer.* Nat Rev Genet., 2006. **7**(1): p. 21-33.

107. Tokumaru, Y., *et al.*, *Inverse correlation between cyclin A1 hypermethylation and p53 mutation in head and neck cancer identified by reversal of epigenetic silencing.* Cancer Res., 2004. **64**(17): p. 5982-7.

108. Liu, H., J. Wang, and E.M. Epner, *Cyclin D1 activation in B-cell malignancy: association with changes in histone acetylation, DNA methylation, and RNA polymerase II*

binding to both promoter and distal sequences. Blood., 2004. **104**(8): p. 2505-13. Epub 2004 Jun 29.

109. Chen, C.L., *et al.*, *E-cadherin expression is silenced by DNA methylation in cervical cancer cell lines and tumours.* Eur J Cancer., 2003. **39**(4): p. 517-23.

110. Bott, S.R., *et al.*, *p21WAF1/CIP1 gene is inactivated in metastatic prostatic cancer cell lines by promoter methylation.* Prostate Cancer Prostatic Dis., 2005. **8**(4): p. 321-6.

111. Nakatsuka, S., *et al.*, *Methylation of promoter region in p27 gene plays a role in the development of lymphoid malignancies.* Int J Oncol., 2003. **22**(3): p. 561-8.

112. Sato, N., *et al.*, *Epigenetic down-regulation of CDKN1C/p57KIP2 in pancreatic ductal neoplasms identified by gene expression profiling.* Clin Cancer Res., 2005. **11**(13): p. 4681-8.

113. Belshaw, N.J., *et al.*, *Methylation of the ESR1 CpG island in the colorectal mucosa is an 'all or nothing' process in healthy human colon, and is accelerated by dietary folate supplementation in the mouse.* Biochem Soc Trans., 2005. **33**(Pt 4): p. 709-11.

114. Iwai, M., *et al.*, *Expression and methylation status of the FHIT gene in acute myeloid leukemia and myelodysplastic syndrome.* Leukemia., 2005. **19**(8): p. 1367-75.

115. Lee, W.H., *et al.*, *Cytidine methylation of regulatory sequences near the pi-class glutathione S-transferase gene accompanies human prostatic carcinogenesis.* Proc Natl Acad Sci U S A., 1994. **91**(24): p. 11733-7.

116. Nakagawa, H., *et al.*, *Loss of imprinting of the insulin-like growth factor II gene occurs by biallelic methylation in a core region of H19-associated CTCF-binding sites in colorectal cancer.* Proc Natl Acad Sci U S A., 2001. **98**(2): p. 591-6. Epub 2000 Dec 19.

117. Parrella, P., *et al.*, *HIC1 promoter methylation and 17p13.3 allelic loss in invasive ductal carcinoma of the breast.* Cancer Lett., 2005. **222**(1): p. 75-81.

118. Eriksson, T., *et al.*, *Methylation changes in the human IGF2 p3 promoter parallel IGF2 expression in the primary tumor, established cell line, and xenograft of a human hepatoblastoma.* Exp Cell Res., 2001. **270**(1): p. 88-95.

119. Herman, J.G., *et al.*, *Incidence and functional consequences of hMLH1 promoter hypermethylation in colorectal carcinoma.* Proc Natl Acad Sci U S A., 1998. **95**(12): p. 6870-5.

120. Goel, A., *et al.*, *Frequent inactivation of PTEN by promoter hypermethylation in microsatellite instability-high sporadic colorectal cancers.* Cancer Res., 2004. **64**(9): p. 3014-21.

121. Bastian, P.J., *et al.*, *Diagnostic and prognostic information in prostate cancer with the help of a small set of hypermethylated gene loci.* Clin Cancer Res., 2005. **11**(11): p. 4097-106.

122. Narayan, G., *et al.*, *Frequent promoter methylation of CDH1, DAPK, RARB, and HIC1 genes in carcinoma of cervix uteri: its relationship to clinical outcome.* Mol Cancer., 2003. **2**: p. 24.

123. Dammann, R., T. Takahashi, and G.P. Pfeifer, *The CpG island of the novel tumor suppressor gene RASSF1A is intensely methylated in primary small cell lung carcinomas.* Oncogene., 2001. **20**(27): p. 3563-7.

124. Rehman, I., *et al.*, *Promoter hyper-methylation of calcium binding proteins S100A6 and S100A2 in human prostate cancer.* Prostate., 2005. **65**(4): p. 322-30.

125. Kang, J.H., *et al.*, *Methylation in the p53 promoter is a supplementary route to breast carcinogenesis: correlation between CpG methylation in the p53 promoter and the mutation of the p53 gene in the progression from ductal carcinoma in situ to invasive ductal carcinoma.* Lab Invest., 2001. **81**(4): p. 573-9.

126. Herman, J.G., *et al.*, *Silencing of the VHL tumor-suppressor gene by DNA methylation in renal carcinoma.* Proc Natl Acad Sci U S A., 1994. **91**(21): p. 9700-4.

127. Ehrlich, M., *DNA methylation in cancer: too much, but also too little.* Oncogene., 2002. **21**(35): p. 5400-13.

128. Hoffmann, M.J. and W.A. Schulz, *Causes and consequences of DNA hypomethylation in human cancer.* Biochem Cell Biol., 2005. **83**(3): p. 296-321.

129. Eden, A., *et al.*, *Chromosomal instability and tumors promoted by DNA hypomethylation.* Science., 2003. **300**(5618): p. 455.

130. Baylin, S.B., *et al.*, *Aberrant patterns of DNA methylation, chromatin formation and gene expression in cancer.* Hum Mol Genet., 2001. **10**(7): p. 687-92.

131. Laird, P.W., *Cancer epigenetics.* Hum Mol Genet., 2005. **14 Spec No 1**: p. R65-76.

132. Knudson, A.G., *Two genetic hits (more or less) to cancer.* Nat Rev Cancer., 2001. **1**(2): p. 157-62.

133. Galm, O., J.G. Herman, and S.B. Baylin, *The fundamental role of epigenetics in hematopoietic malignancies.* Blood Rev., 2006. **20**(1): p. 1-13. Epub 2005 Feb 23.

134. Grady, W.M., *et al.*, *Methylation of the CDH1 promoter as the second genetic hit in hereditary diffuse gastric cancer.* Nat Genet., 2000. **26**(1): p. 16-7.

135. Esteller, M., *et al.*, *DNA methylation patterns in hereditary human cancers mimic sporadic tumorigenesis.* Hum Mol Genet., 2001. **10**(26): p. 3001-7.

136. De Smet, C., *et al.*, *DNA methylation is the primary silencing mechanism for a set of germ line- and tumor-specific genes with a CpG-rich promoter.* Mol Cell Biol., 1999. **19**(11): p. 7327-35.

137. Kaneda, A., *et al.*, *Frequent hypomethylation in multiple promoter CpG islands is associated with global hypomethylation, but not with frequent promoter hypermethylation.* Cancer Sci., 2004. **95**(1): p. 58-64.

138. Sato, N., *et al.*, *Identification of maspin and S100P as novel hypomethylation targets in pancreatic cancer using global gene expression profiling.* Oncogene., 2004. **23**(8): p. 1531-8.

139. Bisogna, M., *et al.*, *Molecular analysis of the INK4A and INK4B gene loci in human breast cancer cell lines and primary carcinomas.* Cancer Genet Cytogenet., 2001. **125**(2): p. 131-8.

140. Watt, P.M., R. Kumar, and U.R. Kees, *Promoter demethylation accompanies reactivation of the HOX11 proto-oncogene in leukemia.* Genes Chromosomes Cancer., 2000. **29**(4): p. 371-7.

141. Feinberg, A.P., H. Cui, and R. Ohlsson, *DNA methylation and genomic imprinting: insights from cancer into epigenetic mechanisms.* Semin Cancer Biol., 2002. **12**(5): p. 389-98.

142. Pavelic, K., D. Bukovic, and J. Pavelic, *The role of insulin-like growth factor 2 and its receptors in human tumors.* Mol Med., 2002. **8**(12): p. 771-80.

143. Rodenhiser, D. and M. Mann, *Epigenetics and human disease: translating basic biology into clinical applications.* Cmaj., 2006. **174**(3): p. 341-8.

144. Brena, R.M., T.H. Huang, and C. Plass, *Quantitative assessment of DNA methylation: Potential applications for disease diagnosis, classification, and prognosis in clinical settings.* J Mol Med., 2006. **84**(5): p. 365-77. Epub 2006 Jan 17.

145. Goffin, J. and E. Eisenhauer, *DNA methyltransferase inhibitors-state of the art.* Ann Oncol., 2002. **13**(11): p. 1699-716.

146. American Cancer Society, *Cancer Facts & Figures 2007*, . 2007, American Cancer Society: Atlanta GA. p. 56.

147. Prasad, U.S., *et al.*, *Long term survival after pulmonary resection for small cell carcinoma of the lung.* Thorax., 1989. **44**(10): p. 784-7.

148. Fry, W.A., H.R. Menck, and D.P. Winchester, *The National Cancer Data Base report on lung cancer.* Cancer., 1996. **77**(9): p. 1947-55.

149. Travis, W., *Histological Typing of lung and pleural tumours*. 3rd ed. 1999, Berlin: Springer-Verlag. 156.

150. Johnson, B.E., *et al.*, *Patients with limited-stage small-cell lung cancer treated with concurrent twice-daily chest radiotherapy and etoposide/cisplatin followed by cyclophosphamide, doxorubicin, and vincristine.* J Clin Oncol., 1996. **14**(3): p. 806-13.

151. National Cancer Institute USA. www.cancer.gov

152. Ichinose, Y., *et al.*, *Prognostic factors obtained by a pathologic examination in completely resected non-small-cell lung cancer. An analysis in each pathologic stage.* J Thorac Cardiovasc Surg., 1995. **110**(3): p. 601-5.

153. Meyerson, M. and D. Carbone, *Genomic and proteomic profiling of lung cancers: lung cancer classification in the age of targeted therapy.* J Clin Oncol., 2005. **23**(14): p. 3219-26.

154. Baylin, S.B. and J.E. Ohm, *Epigenetic gene silencing in cancer - a mechanism for early oncogenic pathway addiction?* Nat Rev Cancer., 2006. **6**(2): p. 107-16.

155. Tsou, J.A., *et al.*, *DNA methylation analysis: a powerful new tool for lung cancer diagnosis.* Oncogene., 2002. **21**(35): p. 5450-61.

156. Belinsky, S.A., *Gene-promoter hypermethylation as a biomarker in lung cancer.* Nat Rev Cancer., 2004. **4**(9): p. 707-17.

157. Reya, T., *et al.*, *Stem cells, cancer, and cancer stem cells.* Nature., 2001. **414**(6859): p. 105-11.

158. Borczuk, A.C., *et al.*, *Non-small-cell lung cancer molecular signatures recapitulate lung developmental pathways.* Am J Pathol., 2003. **163**(5): p. 1949-60.

159. Bonner, A.E., W.J. Lemon, and M. You, *Gene expression signatures identify novel regulatory pathways during murine lung development: implications for lung tumorigenesis.* J Med Genet., 2003. **40**(6): p. 408-17.

160. Bonner, A.E., *et al.*, *Molecular profiling of mouse lung tumors: association with tumor progression, lung development, and human lung adenocarcinomas.* Oncogene., 2004. **23**(5): p. 1166-76.

161. Garnis, C., *et al.*, *Involvement of multiple developmental genes on chromosome 1p in lung tumorigenesis.* Hum Mol Genet., 2005. **14**(4): p. 475-82. Epub 2004 Dec 22.

162. Maier, S. and A. Olek, *Diabetes: a candidate disease for efficient DNA methylation profiling.* J Nutr., 2002. **132**(8 Suppl): p. 2440S-2443S.

163. Dong, C., W. Yoon, and P.J. Goldschmidt-Clermont, *DNA methylation and atherosclerosis.* J Nutr., 2002. **132**(8 Suppl): p. 2406S-2409S.

164. Kim, Y.I., *et al.*, *DNA hypomethylation in inflammatory arthritis: reversal with methotrexate.* J Lab Clin Med., 1996. **128**(2): p. 165-72.

165. Petronis, A. and R. Petroniene, *Epigenetics of inflammatory bowel disease.* Gut., 2000. **47**(2): p. 302-6.

166. Sharma, R.P., *Schizophrenia, epigenetics and ligand-activated nuclear receptors: a framework for chromatin therapeutics.* Schizophr Res., 2005. **72**(2-3): p. 79-90.

167. Richardson, B., *DNA methylation and autoimmune disease.* Clin Immunol., 2003. **109**(1): p. 72-9.

168. Liu, J., J.G. Nealon, and L.S. Weinstein, *Distinct patterns of abnormal GNAS imprinting in familial and sporadic pseudohypoparathyroidism type IB.* Hum Mol Genet., 2005. **14**(1): p. 95-102. Epub 2004 Nov 10.

169. Temple, I.K. and J.P. Shield, *Transient neonatal diabetes, a disorder of imprinting.* J Med Genet., 2002. **39**(12): p. 872-5.

170. Oostra, B.A. and P. Chiurazzi, *The fragile X gene and its function.* Clin Genet., 2001. **60**(6): p. 399-408.

171. Jones, C., *et al.*, *Co-localisation of CCG repeats and chromosome deletion breakpoints in Jacobsen syndrome: evidence for a common mechanism of chromosome breakage.* Hum Mol Genet., 2000. **9**(8): p. 1201-8.

172. van Overveld, P.G., *et al.*, *Hypomethylation of D4Z4 in 4q-linked and non-4q-linked facioscapulohumeral muscular dystrophy.* Nat Genet., 2003. **35**(4): p. 315-7. Epub 2003 Nov 23.

173. Schumacher, A. and A. Petronis, *Epigenetics of complex diseases: from general theory to laboratory experiments.* Curr Top Microbiol Immunol., 2006. **310**: p. 81-115.

174. Feinberg, A.P. and B. Tycko, *The history of cancer epigenetics.* Nat Rev Cancer., 2004. **4**(2): p. 143-53.

175. Hegele, R., *LMNA mutation position predicts organ system involvement in laminopathies.* Clin Genet., 2005. **68**(1): p. 31-4.

176. Hutchison, C.J., M. Alvarez-Reyes, and O.A. Vaughan, *Lamins in disease: why do ubiquitously expressed nuclear envelope proteins give rise to tissue-specific disease phenotypes?* J Cell Sci., 2001. **114**(Pt 1): p. 9-19.

177. Cao, H. and R.A. Hegele, *Nuclear lamin A/C R482Q mutation in canadian kindreds with Dunnigan-type familial partial lipodystrophy.* Hum Mol Genet., 2000. **9**(1): p. 109-12.

178. Speckman, R.A., *et al.*, *Mutational and haplotype analyses of families with familial partial lipodystrophy (Dunnigan variety) reveal recurrent missense mutations in the globular C-terminal domain of lamin A/C.* Am J Hum Genet., 2000. **66**(4): p. 1192-8.

179. Eriksson, M., *et al.*, *Recurrent de novo point mutations in lamin A cause Hutchinson-Gilford progeria syndrome.* Nature., 2003. **423**(6937): p. 293-8. Epub 2003 Apr 25.

180. Liu, L., *et al.*, *Aging, cancer and nutrition: the DNA methylation connection.* Mech Ageing Dev., 2003. **124**(10-12): p. 989-98.

181. Paz, M.F., *et al.*, *A systematic profile of DNA methylation in human cancer cell lines.* Cancer Res., 2003. **63**(5): p. 1114-21.

182. Barlow, D.P., *et al.*, *The mouse insulin-like growth factor type-2 receptor is imprinted and closely linked to the Tme locus.* Nature., 1991. **349**(6304): p. 84-7.

183. Hayward, B.E., *et al.*, *The human GNAS1 gene is imprinted and encodes distinct paternally and biallelically expressed G proteins.* Proc Natl Acad Sci U S A., 1998. **95**(17): p. 10038-43.

184. Hayward, B.E., *et al.*, *Bidirectional imprinting of a single gene: GNAS1 encodes maternally, paternally, and biallelically derived proteins.* Proc Natl Acad Sci U S A., 1998. **95**(26): p. 15475-80.

185. Affymetrix, *GeneChip Expression Analysis - Data Analysis fundamentals*, . 2004, Affymetrix Inc: Santa Clara CA, USA. p. 108.

186. Affymetrix, *Manufacturing Quality Control and Validation Studies of GeneChip Arrays*, . 2002, Affymetix Inc: Santa Clara CA, USA.

187. Perl, A.K. and J.A. Whitsett, *Molecular mechanisms controlling lung morphogenesis.* Clin Genet., 1999. **56**(1): p. 14-27.

188. Bolstad, B.M., *et al.*, *A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.* Bioinformatics., 2003. **19**(2): p. 185-93.

189. Irizarry, R.A., *et al.*, *Summaries of Affymetrix GeneChip probe level data.* Nucleic Acids Res., 2003. **31**(4): p. e15.

190. Dudoit S, S.J., Boldrick JC, *Multiple Hypothesis Testing in Microarray Experiments*, . 2002, University of California, Berkeley. U.C. Berkeley Division of Biostatistics: Berkeley, CA, USA. p. 58 pp.

191. Pfaffl, M.W., *et al.*, *Determination of stable housekeeping genes, differentially regulated target genes and sample integrity: BestKeeper--Excel-based tool using pair-wise correlations.* Biotechnol Lett., 2004. **26**(6): p. 509-15.

192. Doniger, S.W., *et al.*, *MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data.* Genome Biol., 2003. **4**(1): p. R7. Epub 2003 Jan 6.

193. Trinklein, N.D., *et al.*, *An abundance of bidirectional promoters in the human genome.* Genome Res., 2004. **14**(1): p. 62-6.

194. Kruskal WH, W.W., *Use of Ranks in One-Criterion Variance Analysis. Journal of the American Statistical Association*, 1952. **47**(260): p. 583-621.

195. Sun, Z., *et al.*, *Histologic grade is an independent prognostic factor for survival in non-small cell lung cancer: an analysis of 5018 hospital- and 712 population-based cases.* J Thorac Cardiovasc Surg., 2006. **131**(5): p. 1014-20.

196. Papaioannou VE, G.S., *Introduction to the T-box genes and their roles in developmental signaling pathways*, in *Inborn errors of development. The molecular basis of*

clinical disorders of morphogenesis, E.R. Epstein CJ, Wynshaw-Boris A, Editor. 2003, Oxford University Press: Oxford. p. 686-98.

197. Ashurst, J.L., *et al.*, *The Vertebrate Genome Annotation (Vega) database.* Nucleic Acids Res., 2005. **33**(Database issue): p. D459-65.

198. Steiper, M.E., N.M. Young, and T.Y. Sukarna, *Genomic data support the hominoid slowdown and an Early Oligocene estimate for the hominoid-cercopithecoid divergence.* Proc Natl Acad Sci U S A., 2004. **101**(49): p. 17021-6. Epub 2004 Nov 30.

199. Glazko, G.V., E.V. Koonin, and I.B. Rogozin, *Molecular dating: ape bones agree with chicken entrails.* Trends Genet., 2005. **21**(2): p. 89-92.

200. Jukes TH, C.C., *Evolution of protein molecules*, in *Mammalian protein metabolism*, M. HN, Editor. 1969, Academic Press: New York. p. 21-132.

201. Ebersberger, I., *et al.*, *Genomewide comparison of DNA sequences between humans and chimpanzees.* Am J Hum Genet., 2002. **70**(6): p. 1490-7. Epub 2002 Apr 30.

202. Good, P.I., *Introduction to statistics through resampling methods and R/S-PLUS®.* Vol. XXII. 2005, Hoboken, NJ: Wiley. 229.

203. Youssef, E.M., M.R. Estecio, and J.P. Issa, *Methylation and regulation of expression of different retinoic acid receptor beta isoforms in human colon cancer.* Cancer Biol Ther., 2004. **3**(1): p. 82-6. Epub 2004 Jan 9.

204. Baylin, S.B., *DNA methylation and gene silencing in cancer.* Nat Clin Pract Oncol., 2005. **2 Suppl 1**: p. S4-11.

205. Suzuki, H., *et al.*, *A genomic screen for genes upregulated by demethylation and histone deacetylase inhibition in human colorectal cancer.* Nat Genet., 2002. **31**(2): p. 141-9. Epub 2002 May 6.

206. Gius, D., *et al.*, *Distinct effects on gene expression of chemical and genetic manipulation of the cancer epigenome revealed by a multimodality approach.* Cancer Cell., 2004. **6**(4): p. 361-71.

207. Lodygin, D., *et al.*, *Functional epigenomics identifies genes frequently silenced in prostate cancer.* Cancer Res., 2005. **65**(10): p. 4218-27.

208. Shames, D.S., *et al.*, *A genome-wide screen for promoter methylation in lung cancer identifies novel methylation markers for multiple malignancies.* PLoS Med., 2006. **3**(12): p. e486.

209. Warnecke, P.M. and S.J. Clark, *DNA methylation profile of the mouse skeletal alpha-actin promoter during development and differentiation.* Mol Cell Biol., 1999. **19**(1): p. 164-72.

210. Kim, T.H., *et al.*, *A high-resolution map of active promoters in the human genome.* Nature., 2005. **436**(7052): p. 876-80. Epub 2005 Jun 29.

211. Rollins, R.A., *et al.*, *Large-scale structure of genomic methylation patterns.* Genome Res., 2006. **16**(2): p. 157-63. Epub 2005 Dec 19.

212. Cawley, S., *et al.*, *Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs.* Cell., 2004. **116**(4): p. 499-509.

213. Kleinjan, D.A., *et al.*, *Aniridia-associated translocations, DNase hypersensitivity, sequence comparison and transgenic analysis redefine the functional domain of PAX6.* Hum Mol Genet., 2001. **10**(19): p. 2049-59.

214. Zhou, Y., P.A. Overbeek, and K.E. Bernstein, *Tissue specific expression of testis angiotensin converting enzyme is not determined by the -32 nonconsensus TATA motif.* Biochem Biophys Res Commun., 1996. **223**(1): p. 48-53.

215. Dale, T.C., *et al.*, *High-level expression of the rat whey acidic protein gene is mediated by elements in the promoter and 3' untranslated region.* Mol Cell Biol., 1992. **12**(3): p. 905-14.

216. Lorincz, M.C., *et al.*, *Intragenic DNA methylation alters chromatin structure and elongation efficiency in mammalian cells.* Nat Struct Mol Biol., 2004. **11**(11): p. 1068-75. Epub 2004 Oct 3.

217. West, A.G. and P. Fraser, *Remote control of gene transcription.* Hum Mol Genet., 2005. **14 Spec No 1**: p. R101-11.

218. Nemeth, A. and G. Langst, *Chromatin higher order structure: opening up chromatin for transcription.* Brief Funct Genomic Proteomic., 2004. **2**(4): p. 334-43.

219. Haaf, T., *Methylation dynamics in the early mammalian embryo: implications of genome reprogramming defects for development.* Curr Top Microbiol Immunol., 2006. **310**: p. 13-22.

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

221. Wasowicz, M., *et al.*, *Ultrastructural studies on selected elements of the extracellular matrix in the developing rat lung alveolus.* Folia Histochem Cytobiol., 1998. **36**(1): p. 3-13.

222. Olsen, C.O., *et al.*, *Extracellular matrix-driven alveolar epithelial cell differentiation in vitro.* Exp Lung Res., 2005. **31**(5): p. 461-82.

223. Holladay, S.D. and R.J. Smialowicz, *Development of the murine and human immune system: differential effects of immunotoxicants depend on time of exposure.* Environ Health Perspect., 2000. **108 Suppl 3**: p. 463-73.

224. Poynter, M.E., C.G. Irvin, and Y.M. Janssen-Heininger, *A prominent role for airway epithelial NF-kappa B activation in lipopolysaccharide-induced airway inflammation.* J Immunol., 2003. **170**(12): p. 6257-65.

225. Mankoo, B.S., *et al.*, *The concerted action of Meox homeobox genes is required upstream of genetic pathways essential for the formation, patterning and differentiation of somites.* Development., 2003. **130**(19): p. 4655-64.

226. Patel, S., A.D. Leal, and D.H. Gorski, *The homeobox gene Gax inhibits angiogenesis through inhibition of nuclear factor-kappaB-dependent endothelial cell gene expression.* Cancer Res., 2005. **65**(4): p. 1414-24.

227. Kadomatsu, K. and T. Muramatsu, *Midkine and pleiotrophin in neural development and cancer.* Cancer Lett., 2004. **204**(2): p. 127-43.

228. Kaifi, J.T., *et al.*, *Midkine as a prognostic marker for gastrointestinal stromal tumors.* J Cancer Res Clin Oncol, 2007. **13**: p. 13.

229. Testa, J.R., *et al.*, *Cytogenetic analysis of 63 non-small cell lung carcinomas: recurrent chromosome alterations amid frequent and widespread genomic upheaval.* Genes Chromosomes Cancer., 1994. **11**(3): p. 178-94.

230. Hayami, Y., *et al.*, *Inactivation of the E3/LAPTm5 gene by chromosomal rearrangement and DNA methylation in human multiple myeloma.* Leukemia., 2003. **17**(8): p. 1650-7.

231. Patz, E.F., Jr., *Imaging bronchogenic carcinoma.* Chest., 2000. **117**(4 Suppl 1): p. 90S-95S.

232. Herbst, R.S., *Toxicities of antiangiogenic therapy in non-small-cell lung cancer.* Clin Lung Cancer., 2006. **8 Suppl 1**: p. S23-30.

233. Dosaka-Akita, H., *et al.*, *Expression of N-acetylglucosaminyltransferase v is associated with prognosis and histology in non-small cell lung cancers.* Clin Cancer Res., 2004. **10**(5): p. 1773-9.

234. Bernstein, B.E., *et al.*, *Genomic maps and comparative analysis of histone modifications in human and mouse.* Cell., 2005. **120**(2): p. 169-81.

235. Hayward, B.E. and D.T. Bonthron, *An imprinted antisense transcript at the human GNAS1 locus.* Hum Mol Genet., 2000. **9**(5): p. 835-41.

236. Zhang, X., *et al.*, *Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis.* Cell., 2006. **126**(6): p. 1189-201. Epub 2006 Aug 31.

237. Xu, Y., *et al.*, *Functional polymorphism in the parental imprinting of the human IGF2R gene.* Biochem Biophys Res Commun., 1993. **197**(2): p. 747-54.

238. Zwart, R., *et al.*, *Bidirectional action of the Igf2r imprint control element on upstream and downstream imprinted genes.* Genes Dev., 2001. **15**(18): p. 2361-6.

239. Tsyba, L., *et al.*, *Distribution of HIV-1 in the genomes of AIDS patients.* Cell Mol Life Sci., 2004. **61**(6): p. 721-6.

240. Balakirev, E.S. and F.J. Ayala, *Pseudogenes: are they "junk" or functional DNA?* Annu Rev Genet., 2003. **37**: p. 123-51.

241. Korneev, S.A., J.H. Park, and M. O'Shea, *Neuronal expression of neural nitric oxide synthase (nNOS) protein is suppressed by an antisense RNA transcribed from an NOS pseudogene.* J Neurosci., 1999. **19**(18): p. 7711-20.

242. Hirotsune, S., *et al.*, *An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene.* Nature., 2003. **423**(6935): p. 91-6.

243. Gray, T.A., *et al.*, *The putatively functional Mkrn1-p1 pseudogene is neither expressed nor imprinted, nor does it regulate its source gene in trans.* Proc Natl Acad Sci U S A., 2006. **103**(32): p. 12039-44. Epub 2006 Aug 1.

244. Ivascu C, *et al.*, *DNA methylation profiling of transcription factor genes in normal lymphocyte development and lymphomas.* Int J Bioch Cell Biol, 2007. **In press**.

245. Moretti, P. and H.Y. Zoghbi, *MeCP2 dysfunction in Rett syndrome and related disorders.* Curr Opin Genet Dev., 2006. **16**(3): p. 276-81. Epub 2006 May 2.

246. Jiang, Y.H., *et al.*, *A mixed epigenetic/genetic model for oligogenic inheritance of autism with a limited role for UBE3A.* Am J Med Genet A., 2004. **131**(1): p. 1-10.

247. Vigouroux, C., *et al., Lamin A/C gene: sex-determined expression of mutations in Dunnigan-type familial partial lipodystrophy and absence of coding mutations in congenital and acquired generalized lipoatrophy.* Diabetes., 2000. **49**(11): p. 1958-62.

248. Bonne, G., *et al., Mutations in the gene encoding lamin A/C cause autosomal dominant Emery-Dreifuss muscular dystrophy.* Nat Genet., 1999. **21**(3): p. 285-8.

249. Cao, H. and R.A. Hegele, *LMNA is mutated in Hutchinson-Gilford progeria (MIM 176670) but not in Wiedemann-Rautenstrauch progeroid syndrome (MIM 264090).* J Hum Genet., 2003. **48**(5): p. 271-4. Epub 2003 Apr 3.

250. Novelli, G., *et al., Mandibuloacral dysplasia is caused by a mutation in LMNA-encoding lamin A/C.* Am J Hum Genet., 2002. **71**(2): p. 426-31. Epub 2002 Jun 19.

251. Fatkin, D., *et al., Missense mutations in the rod domain of the lamin A/C gene as causes of dilated cardiomyopathy and conduction-system disease.* N Engl J Med., 1999. **341**(23): p. 1715-24.

252. Navarro, C.L., *et al., Lamin A and ZMPSTE24 (FACE-1) defects cause nuclear disorganization and identify restrictive dermopathy as a lethal neonatal laminopathy.* Hum Mol Genet., 2004. **13**(20): p. 2493-503. Epub 2004 Aug 18.

253. Furuta, J., *et al., Promoter methylation profiling of 30 genes in human malignant melanoma.* Cancer Sci., 2004. **95**(12): p. 962-8.

254. Kawakami, T., *et al., Multipoint methylation analysis indicates a distinctive epigenetic phenotype among testicular germ cell tumors and testicular malignant lymphomas.* Genes Chromosomes Cancer., 2003. **38**(1): p. 97-101.

255. Behrens, G.M., *et al., Lack of mutations in LMNA, its promoter region, and the cellular retinoic acid binding protein II (CRABP II) in HIV associated lipodystrophy.* Eur J Med Res., 2003. **8**(5): p. 221-5.

256. Salazar, J., *et al., Two novel single nucleotide polymorphisms in the promoter of the cellular retinoic acid binding protein II gene (CRABP-II).* Mol Cell Probes., 2003. **17**(1): p. 21-3.

257. DeBusk, F.L., *The Hutchinson-Gilford progeria syndrome. Report of 4 cases and review of the literature.* J Pediatr., 1972. **80**(4): p. 697-724.

258. Denecke, J., *et al., A homozygous ZMPSTE24 null mutation in combination with a heterozygous mutation in the LMNA gene causes Hutchinson-Gilford progeria syndrome (HGPS): insights into the pathophysiology of HGPS.* Hum Mutat., 2006. **27**(6): p. 524-31.

259. Berlin, K., Ballhause, M., Cardon, K., *Improved bisulfite conversion of DNA*, in *PCT/WO/2005/038051*. 2005.

260. Rozen S, S.H., *Primer3 on the WWW for general users and for biologist programmers*, in *Bioinformatics Methods and Protocols: Methods in Molecular Biology*, M.S. Krawetz S, Editor. 2000, Humana Press: Totowa, NJ. p. 365-386.

261. Wong, M.L. and J.F. Medrano, *Real-time PCR for mRNA quantitation.* Biotechniques., 2005. **39**(1): p. 75-85.

262. Lewin, J., *et al.*, *Quantitative DNA methylation analysis based on four-dye trace data from direct sequencing of PCR amplificates.* Bioinformatics., 2004. **20**(17): p. 3005-12. Epub 2004 Jul 9.

263. Wilcoxon, F., *Individual Comparisons by Ranking Methods.* Biometrics, 1945. **1**: p. 80-83.

264. Felsenstein, J., *PHYLIP - Phylogeny Inference Package (Version 3.2).* Cladistics, 1989. **5**: p. 164-166.

265. Simon, P., *Q-Gene: processing quantitative real-time RT-PCR data.* Bioinformatics., 2003. **19**(11): p. 1439-40.

266. Cortese, R., et al, *The retinol acid receptor B gene is hypermethylated in patients with familial partial lipodystrophy.* J Mol Endocrinol, 2007. **38:** p. 663–671