Functional analysis of aberrant DNA methylation patterns during metastatic breast cancer progression



Dissertation zur Erlangung
des Doktorgrades der Naturwissenschaften (Dr. rer. nat.)
der Fakultät für Biologie und vorklinische Medizin der Universität
Regensburg

vorgelegt von

Jan Bartel

aus

Weißenburg in Bayern

im Jahr 2024

Promotionsgesuch eingereicht am: 21. November 2024

Promotion angeleitet von: Prof. Dr. Michael Rehli

Unterschrift:

_____

Abstract

Aberrant DNA methylation in primary breast cancer tumors plays an important role in gene regulation and transcriptome diversity and influences metastatic behavior. Despite multiple hypotheses on how malignant cells spread and become metastatic, there is a lack of comprehensive analysis of how the methylome contributes to the process of dissemination. Previous studies have shown that metastatic cells accumulate hypermethylation at promoters and demethylation events in intra- and intergenic regions. These epigenetic changes contribute to increased chromosomal instability and the downregulation of tumor suppressor genes. Furthermore, analysis conducted by our laboratory has revealed the presence of epigenetic modifications in metastasis, indicating distinct recurrent patterns of promoter methylation and demethylation. However, it is unclear whether these epigenetic changes arise as side effects of cell aging and increased proliferation or through the acquisition of epigenetic alterations by single cells in early primary tumor lesions, which are then selected to become metastatic. To address this complex issue, a novel approach to analyze diverse methylation patterns throughout primary tumor progression was developed, which is the topic of this thesis. Specifically, a barcoded bisulfite amplicon deep sequencing (BBA-seq) technique was used to characterize the methylome on a single allelic level within breast cancer in the Her2-driven Balb-NeuT mouse model. Computational processing, clustering, and precise representation of the sequenced reads according to their methylation composition led to a clear resolution of the methylome in seven primary tumors and their corresponding metastases within 9 DNA regions. These regions include demethylation,

hypermethylation, and intermediate conditions of DNA methylation. The findings revealed diverse methylation patterns, ranging from a constant increase in DNA methylation to the identification of sub-clones with an already high methylation character within primary tumors. These observations highlight the complexity and variability of methylation patterns during tumor progression. The results suggest that further investigation into the role of these sub-clones is critical in understanding how epigenetic changes drive metastasis. In addition to identifying and characterizing DNA methylation patterns, a comprehensive protocol was established to alter the DNA methylome via a gene knockdown of the de-novo DNA (cytosine-5)-methyltransferase 3a using a CRISPR/Cas9 nickase approach in hard-to-transfect primary mouse mammary gland cells. Overall, this thesis comprehensively analyzes DNA methylome remodeling during primary tumor progression and metastasis in breast cancer. The novel methodology developed in this study may also be applied to investigate other research questions related to DNA methylation in cancer biology and beyond.

Zusammenfassung

Anormale DNA-Methylierung in primären Brusttumoren spielt eine entscheidende Rolle bei der Genregulation, der Vielfalt des Transkriptoms und beeinflusst das metastatische Verhalten. Trotz zahlreicher Hypothesen darüber wie maligne Zellen sich ausbreiten und metastasieren, fehlen umfassende Analysen wie das Methylom zum Prozess der Dissemination beiträgt. Frühere Studien haben gezeigt, dass Metastasen Hypermethylierung in Promotoren und Demethylierungsereignisse in intra- und intergenen Regionen akkumulieren. Diese epigenetischen Veränderungen tragen zur erhöhten chromosomalen Instabilität und zur Hemmung von Tumorsuppressorgenen bei. Darüber hinaus hat eine vorhergehende Studie unseres Labors das Vorhandensein epigenetischer Modifikationen in Metastasen aufgezeigt, welche wiederkehrende Muster von Promotor-Methylierung und Demethylierung aufweisen. Es ist jedoch unklar, ob diese epigenetischen Veränderungen als Nebenprodukte der Zellalterung und erhöhten Proliferation entstehen oder ob einzelne Zellen im Primärtumor steigende DNA Methylierung akkumulieren, was zu einem Vorteil im Metastasierungsprozess führen könnte. Um dieses komplexe Problem zu adressieren, wurde in dieser Arbeit ein neuartiger Ansatz zur Analyse diverser Methylierungsmuster im Verlauf der primären Tumorprogression entwickelt. Dabei wurde eine barcodierte Bisulfit-Amplicon-Sequenzierungstechnik genutzt, um das Methylom auf Einzel-Allel-Ebene im Her2-getriebenen Balb-NeuT-Maus-Brustkrebsmodell zu charakterisieren. Durch computergestützte Verarbeitung, Clusterbildung und präzise Darstellung der sequenzierten DNA-Fragmente anhand ihrer Methylierungszusammensetzung konnte das Methylom

von sieben primären Tumoren und gepaarten Metastasen in neun DNA-Regionen klar aufgelöst werden. Diese Regionen umfassen Demethylierung, Hypermethylierung und intermediäre Zustände der DNA-Methylierung. Die Ergebnisse zeigten vielfältige Methylierungsmuster, die von einem kontinuierlichen Anstieg der DNA-Methylierung bis zur Identifikation von Subklonen mit bereits hohem Methylierungscharakter innerhalb der primären Tumoren reichten. Diese Beobachtungen unterstreichen die Komplexität und Variabilität der Methylierungsmuster während der Tumorprogression. Die Resultate legen nahe, dass eine genauere Untersuchung dieser Subklone wichtig ist, um zu klären, ob und wie anormale DNA-Methylierung die Metastasierung antreibt. Neben der Identifikation und Charakterisierung von DNA-Methylierungsmustern wurde in dieser Arbeit ein Protokoll entwickelt, um die Zusammensetzung des DNA-Methyloms durch Gen-Knockdown der de-novo DNA (cytosine-5)-methyltransferase 3a mithilfe eines CRISPR/Cas9-Nickase-Ansatzes in schwer zu transfizierenden primären Maus-Brustdrüsenzellen zu verändern. Insgesamt analysiert diese Arbeit die Umgestaltung des DNA-Methyloms während der primären Tumorprogression und Metastasierung bei Brustkrebs umfassend. Die in dieser Studie entwickelte neuartige Methodologie kann auch auf andere Forschungsfragen im Zusammenhang mit der DNA-Methylierung in der Krebsbiologie und darüber hinaus angewendet werden.

Contents

Abstract iii

1. Introduction 1

1.1. Epigenetics: A brief overview . . . 1
1.2. DNA methylation . . . 2
1.3. Mechanism and regulation of DNA methylation . . . 4
1.4. Other epigenetic mechanisms . . . 9
1.5. DNA methylation in tumorgenesis . . . 12
1.6. Epigenetic alterations during metastatic breast cancer progression . . . 14
1.7. Morphogenesis of the breast . . . 17
1.8. Breast cancer . . . 18
1.9. Metastasis . . . 20
1.10. Early versus late dissemination of cancer cells . . . 22
1.11. The Balb-NeuT mouse model . . . 24
1.12. Aim of work . . . 26

2. Material 29

2.1. Equipment . . . 29
2.2. Consumables . . . 31
2.3. Chemicals and reagents . . . 33
2.4. Enzymes and kits . . . 35
2.5. Western blot antibodies . . . 36

2.6. Custom oligonucleotides . . . 36
2.6.1. CRISPR . . . 36
2.6.2. BBA-seq . . . 37

3. Methods **39**

3.1. General cell culture methods . . . 39
3.1.1. Cell culture conditions and passaging . . . 39
3.1.2. Assessing cell number and viability . . . 40
3.1.3. Freezing and thawing cells . . . 40
3.1.4. Primary cultures . . . 40
3.1.5. Creating poly-HEMA and low-attachment plates . . . 41
3.1.6. Sphere culture . . . 41
3.1.7. Counting cell spheres in suspension culture . . . 42
3.1.8. Mycoplasma assay . . . 43
3.1.9. Transient transfection of mammalian cells . . . 43
3.2. Animal experiments . . . 43
3.2.1. Maintenance . . . 43
3.2.2. Surgery . . . 43
3.2.3. Mice dissection . . . 44
3.3. Methods in molecular biology . . . 44
3.3.1. DNA preparation and analysis . . . 44
3.3.2. Agarose gel electrophoresis . . . 44
3.3.3. Quantification of DNA . . . 46
3.3.4. Polymerase Chain Reaction (PCR) . . . 46
3.3.5. RNA isolation . . . 47
3.3.6. Reverse transcription PCR (RT-PCR) . . . 48
3.3.7. Quantitative real time PCR (qPCR) . . . 48
3.4. Protein biochemical methods . . . 49
3.4.1. Preparation of whole cell lysates . . . 49

3.4.2. SDS-PAGE (sodium dodecyl sulfate polyacrylamide gel electrophoresis) . . . 50
3.4.3. Analysis through Western blot and immunostaining . . . 50
3.4.4. Delivery of ribonucleoprotein complexes into Balb-NeuT mammary gland cells . . . 51
3.4.5. Analysis of Cas9 gene editing with T7 endonuclease assay . . . 52
3.5. Analysis of DNA methylation . . . 52
3.5.1. Bisulfite conversion . . . 52
3.5.2. Primer desing for BBA-seq approach . . . 52
3.5.3. PCR amplification . . . 53
3.5.4. Measurement of mean methylation values . . . 55
3.5.5. Measurement of bisulfite conversion rate . . . 55
3.5.6. From Bismark Methylation extractor to aligned `*.fasta` file . . . 56

4. Results I: Phyloepigenetic analysis of metastatic breast tumor progression 59
4.1. Target regions determination . . . 59
4.2. Identification of DNA methylation changes with bisulfite padlock probes 62
4.3. Identification of DNA methylation changes with BBA-seq . . . 64
4.4. Data processing . . . 66
4.5. Isolating tumor-specific DNA methylation signatures by eliminating non-tumor cell contamination . . . 69
4.6. General methylation analysis . . . 72
4.7. Phyloepigenetic analysis of BBA-seq-derived data and tree generation . 76
4.8. Methylation analysis in region B . . . 80
4.9. Methylation analysis in region C . . . 90
4.10. Methylation analysis in region D . . . 94
4.11. Methylation analysis in region E . . . 101
4.12. Methylation analysis in region F . . . 107
4.13. Methylation analysis in region G . . . 109

4.14. Methylation analysis in region J 113
4.15. Methylation analysis in region L 115
4.16. Identifying differentially methylated cytosines 118

5. Results II: Altering de-novo DNA methylation capabilities 123
5.1. Expression kinetics of *Dnmt3a* and *Dnmt3b* 123
5.2. In silico analysis - *Dnmt3a* 124
5.3. Functional domain prediction and target site design 125
5.4. Determining Cas9 nickase knockdown efficiency 128
5.5. Generating spheres from Balb-NeuT mammary gland tissue 130

6. Discussion and Outlook 133
6.1. Discussion 133
6.1.1. Stroma cells in tumor microenvironment 139
6.1.2. Strengths and potential limitations of the BBA-seq approach . . 142
6.1.3. DNA methylation in other cancers 145
6.2. Outlook 147

A. Code snippets 151

B. Supplementary figures 155

Bibliography 171

List of Figures

1.1. Waddington's classical epigenetic landscape 2
1.2. The mechanism of DNA methylation 3
1.3. DNA methylation changes in cancer 15
1.4. The mammary gland 17
1.5. Dissemination models 24

4.1. Target regions 61
4.2. Bisulfite padlock probes overview 62
4.3. Bisulfite padlock probes results 65
4.4. DNA methylation analysis with BBA-seq 66
4.5. Data processing overview 68
4.6. Relative frequencies of methylated CpGs in read per sample 71
4.7. General methylation analysis 73
4.8. DNA methylation resolution of 5 phylogenetic algorithms in region B . 78
4.9. DNA methylation resolution of 5 phylogenetic algorithms in region C . 79
4.10. DNA methylation resolution of 5 phylogenetic algorithms in region D . 80
4.11. Example tree 81
4.12. Subtraction of unmethylated reads in sample 3145 82
4.13. General methylation analysis region B 84
4.14. Phyloepigenetic tree for sample 3145 in region B 85
4.15. Logos and cluster composition of sample 3145 in region B 85
4.16. Distribution of reads from primary tumor samples analysed at region B 89

4.17. Arrangement of reads in region B according to their mCpG count . . . 90
4.18. General methylation analysis region C 91
4.19. Distribution of reads from primary tumor samples analysed at region C 93
4.20. General methylation analysis region D 94
4.21. Phyloepigenetic tree for sample 3145 in region D 97
4.22. Logos and cluster composition of sample 3145 in region D 97
4.23. Distribution of reads from primary tumor samples analysed at region D 99
4.24. Arrangement of reads in region D according to their mCpG count . . . 100
4.25. General methylation analysis region E 101
4.26. Phyloepigenetic tree for sample 3145 in region E 104
4.27. Sequence logos and composition of clusters 104
4.28. Distribution of reads from primary tumor samples analysed at region E 105
4.29. Arrangement of reads in region E according to their mCpG count . . . 106
4.30. General methylation analysis region F 107
4.31. Arrangement of reads in region F according to their mCpG count 109
4.32. General methylation analysis region G 110
4.33. Arrangement of reads in region G according to their mCpG count . . . 112
4.34. General methylation analysis region J 113
4.35. Distribution of reads from primary tumor samples analysed at region J 115
4.36. General methylation analysis region L 116
4.37. Distribution of reads from primary tumor samples analysed at region L 117
4.38. Standard deviation in region B 120
4.39. Standard deviation in region D 120

5.1. Domain structure of DNMT3A 125
5.2. T7 endonuclease assay 127
5.3. Overview of exon 16 of *Dnmt3a* 128
5.4. Balb-NeuT *Dnmt3a* knockout 129
5.5. Balb-NeuT mammary spheres 131

B.1. PCA Screeplots 156
B.2. Methylation analysis in region B before subtraction 157
B.3. Methylation analysis in region B after subtraction 158
B.4. DNA methylation analysis in region C at single base resolution 159
B.5. Methylation analysis in region D before subtraction 160
B.6. Methylation analysis in region D after subtraction 161
B.7. Methylation analysis in region E before subtraction 162
B.8. Methylation analysis in region E after subtraction 163
B.9. Methylation analysis in region F before subtraction 164
B.10. Methylation analysis in region F after subtraction 165
B.11. Methylation analysis in region G before subtraction 166
B.12. Methylation analysis in region G after subtraction 167
B.13. DNA methylation analysis in region J at single base resolution 168
B.14. DNA methylation analysis in region L at single base resolution 169

List of Tables

2.1. Technical devices and equipment 31
2.2. Consumables 33
2.3. Chemicals and reagents 34
2.4. Enzymes and kits 36
2.5. Antibodies for Western Blot 36
2.6. Primer sequences CRISPR/Cas9 gene knockdown 36
2.7. Primer sequences used for qPCR 37
2.8. Sequences CRISPR guideRNAs with PAMs 37
2.9. Primer sequences BBA-seq first PCR 38
2.10. Primer sequences BBA-seq second PCR 38

3.1. Agarose gel electrophoresis 45
3.2. Buffers for gel electrophoresis 45
3.3. Standard PCR reaction 47
3.4. Conditions for PCR 47
3.5. Reaction setup reverse transcription 48
3.6. Conditions for first BBA-seq PCR 53
3.7. Reaction setup for first BBA-seq PCR. 54
3.8. Conditions for second BBA-seq PCR 54

4.1. Location and characteristics of BBA-seq PCR products 67
4.2. Cut-offs and metrics of read subtraction 70

4.3. Quality metrics of analyzed samples 72
4.4. DMCs in region B 119
4.5. DMCs in region D 121

List of File Excerpts

3.1. Bismark Methylation Extractor `*.txt` output file 57
3.2. Intermediate `*.fasta` file 58
3.3. Final `*.fasta` all file 58
3.4. Final `*.fasta` count file 58

A.1. Code snippet `convertFastaI.R` 152
A.2. Code snippet `convertFastaII.R` 153

1. Introduction

1.1. Epigenetics: A brief overview

Epigenetics is the science of studying how gene expression is regulated beyond the DNA sequence itself. It has been defined as the investigation of heritable alterations in gene expression or cellular phenotype that do not involve modifications to the underlying DNA sequence [1]. Epigenetic modifications, which can be tremendously influenced by a variety of environmental factors, alter the architectural structure of DNA and, therefore, directly affects gene expression. The best-studied epigenetic mechanisms are the direct methylation of DNA base cytosine, histone modifications, and the regulation of gene expression by non-coding RNA [2]. The history of epigenetics can be traced back to the 1940s when Conrad Waddington introduced the concept of the "epigenetic landscape" (Figure 1.1), which described how various factors influence gene expression during development [3]. Especially in recent years, new technologies, such as next-generation sequencing have enabled researchers to investigate epigenetic mechanisms in greater detail, leading to significant breakthroughs in fields ranging from developmental biology to cancer research.

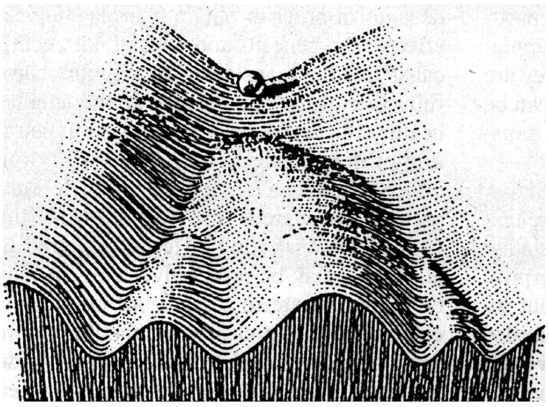


Figure 1.1.: Waddington's Classical Epigenetic Landscape. The idea of an epigenetic landscape was introduced by Conrad Waddington in 1942 to illustrate how cells make decisions during development. According to this dynamic visual representation, depicted as a landscape, the cell (symbolized by a ball) can follow different pathways at various stages, ultimately leading to distinct outcomes or cell fates [3].

1.2. DNA methylation

The best-studied and most prevalent epigenetic modification is the methylation of the DNA base cytosine. The basic concept is that a methyl group is covalently added to the fifth carbon of the cytosine pyrimidine ring. Methylation of cytosine mostly occurs when it is followed by a guanine, termed CpG dinucleotide [4]. DNA methyltransferases (DNMTs) are the key enzymatic effectors for this process, transferring the methyl group from S-adenosylmethionine (SAM) to the fifth carbon position of the cytosine residue [5]. Figure 1.2 shows the chemical process of active DNA cytosine methylation. The processes of DNA methylation and demethylation are dynamic and tightly regulated and play crucial roles in various cellular processes, including development, differentiation, and disease. In 1992 Li et al. demonstrated the importance of DNA methylation by disrupting the gene encoding the DNA MTase enzyme in mice embryos, leading to their death at embryonic day 10.5 [6]. Several scientific investigations have provided substantial evidence supporting the crucial involvement of DNA methylation in repressing and activating gene expression, with specific emphasis on genes located on the inactive

female X chromosome, tissue-specific genes, and genes that are parentally imprinted [4, 7]. Additionally, methylation may prevent insertional mutations by silencing proviruses, including LINES ad SINES [8, 9]. Furthermore, methylation provides additional stability to the genome by condensing the chromatin, making it less prone to recombination events [10–12]. As most CpGs are methylated, the distribution of CpGs throughout the genome is much lower than expected due to the high frequency of deamination of 5-methylcytosine (5mC) compared to normal cytosine. The spontaneous deamination of methylated cytosines leads to the conversion to thymine, pairing generally with adenine, making this a not detectable switch by DNA repair mechanisms. The CpG sequence is a mutation hotspot in the germline, accounting for 30% of all point mutations [13, 14]. DNA methylation contributes to a strong selective advantage since it poses a substantial mutagenic burden on the genome. As a result, it efficiently regulates transcription in complex genomes. A significant characteristic of DNA methylation is its reversibility, making it an ideal method for regulating developmental programs. In diseases such as cancer, disruptions in regulation can disturb the reversibility of the process and lead to the growth of tumors [15]. Understanding how the methylation of DNA is generally regulated is crucial to facilitate therapy and intervention in aberrant molecular processes like the development of cancer.

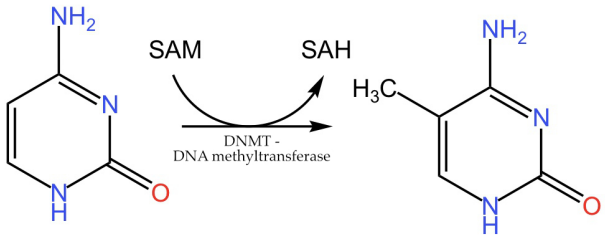


Figure 1.2.: The mechanism of DNA methylation. DNA methylation occurs through the action of DNA methyltransferases (DNMTs), which transfer a methyl group from S-adenosylmethionine (SAM) to the 5 position of cytosine residues in CpG dinucleotides, resulting in the formation of 5-methylcytosine.

1.3. Mechanism and regulation of DNA methylation

In 1982, Ehrlich et al. postulated that 5mC accounts for only 1% of all DNA bases existing in somatic cells in human beings [16]. Nevertheless, the methylation of cytosine plays a crucial role in gene regulation and takes over other important tasks depending on its location across the genome. In DNA regions between genes, namely intergenic regions, bulk methylation represses the expression of viral and transposable elements as those have integrated into the mammalian genome over time during evolution. Activating such retro-elements would lead to DNA mutations and potential gene disruption [17, 18]. The enzyme DNMT1 is accountable for the inactivation of these elements, as it was shown that deletion of this particular enzyme leads to increased de-methylation and expression of such harmful retro-transposable elements [19].

Another important DNA element in the context of methylation is called CpG island. These structures are defined as stretches within the genome with an accountable higher density of unmethylated cytosines followed by a guanine base of about 1 Kb or more [20]. Gene promoters are often found within CpG islands, especially those of housekeeping genes and those responsible for activity in stem and premature cells. It has been shown that CpG island elements are evolutionarily conserved in mammals underlining the importance of these structures for gene expression and chromatin regulation [21]. Furthermore, an increased gene expression is associated with promoters containing CpG islands, as those DNA regions are often more loosely wrapped around nucleosomes with specific histone modifications responsible for active gene production [22]. In mammals, CpG islands are often targets of active transcription factor binding, leading to increased gene expression. This is because most mammalian transcription factors have domains that bind to CpG-dense motifs. Promoters with GC-rich structures often lack other promoter-associated features like the prominent TATA box. The binding of transcription factors enhancing transcription further increases the gene expression machinery by coupling to other essential proteins involved in the transcription process,

like RNA polymerase II and the so-called mediator complex, which is a conglomerate of several proteins altering the 3D structure of the DNA enhancing transcription by bringing together enhancer and promoter structures [23]. Despite their involvement in active gene expression, CpG islands play a crucial role in the downregulation of genes when they occur methylated. In the process of ongoing cell differentiation, certain CpG-rich promoter elements become more and more methylated, leading to a decrease in the expression of the associated genes. This process could be observed in the context of gene imprinting, where the expression of the gene gets silenced on one of the two parental chromosomes due to an increase in DNA methylation [24, 25]. In summary, CpG islands represent an important feature in the mammalian genome, regulating expression in the context of their DNA methylation specificity by their ability to recruit transcription factors and other proteins that bind to methylated cytosine, therefore either enhancing or decrease the expression of the corresponding genes.

The least investigated DNA methylation presence is 5mC within the gene body. The gene body refers to the DNA segment that extends downstream from the transcription initiation site, beginning with the first exon. An increase in DNA methylation within the gene body has several positive effects on gene transcription. First, the methylation of cytosines within the gene body blocks the binding of RNA polymerase II to alternative transcription start sites, inhibiting the production of truncated proteins from intragenic transcription. This process also involves other epigenetic factors like the binding of DNMT3B and the accumulation of repressing histone marks like H3K36me3 [26]. In addition, methylation at cytosine residues plays a crucial role in the process of alternative splicing. Almost every gene undergoes this procedure where distinct exons are removed, with the extant being differentially combined. The pre-mRNA undergoes a maturation process to elucidate a version that afterward serves as a transcript for protein production. DNA polymerase II elongates on the DNA to produce a pre-mRNA and transcribes the base code toward a RNA molecule. DNA methylation in the gene body and the binding of proteins that recognize cytosine methylation serve as an

elongation speed blocker. Therefore, the polymerase can recognize DNA splice sites more efficiently, leading to alternative pre-mRNA molecules and different transcription products [27]. How the methyl group is added to the fifth carbon residue and the process of how demethylation is regulated is a topic of ongoing research. In general, the enzymes involved in this process can be divided into three subclasses. On the one hand, there are enzymes that establish and catalyze the reaction adding CH_3 to cytosine - so-called writers. Whereas on the other hand, there are proteins recognizing the methyl group (readers) directly influencing transcription and expression. Further, the last group involves enzymes that remove the covalently bound CH_3 group from the cytosine base (erasers) [28]. The enzymes catalyzing the binding from CH_3 covalently to the fifth carbon of the cytosine pyrimidine ring are members of DNA methyltransferases and comprise three proteins, namely DNMT1, DNMT3A & DNMT3B. As they are merged into the same family, they share distinct domain patterns like the regulatory domain located at the N-terminus and the important catalytic part, which is located C-terminal of the proteins. Nevertheless, they all have different tasks and functions regarding DNA methylation depending on the cell differentiation state and tissue they are transcribed. DNMT1 is responsible for maintaining and establishing DNA methylation in hemi-methylated DNA after cell replication. When cells are dividing, each daughter cell has to receive the same DNA amount and information as the mother cell. The double-stranded DNA is, therefore, cleaved between each base pairing and doubled during the process of mitosis. The newly synthesized DNA strand has to receive the same DNA methylation information as the parental single DNA strand. DNMT1 accomplishes this task. The enzyme follows up the replication fork and adds the CH_3 group according to the information it receives from the mother strand [29]. DNMT3A & B, as the name suggests, are similar in their function and structure. They were found to establish de-novo DNA methylation during mammalian cell development with no bias toward hemi-methylated newly synthesized DNA strands [30]. They differ in their expression patterns and stages; DNMT3B is mainly expressed in cells in early developmental stages, whereas DNMT3A is generally expressed in already differentiated tissues [31].

It is currently unclear how the two enzymes locate their correct target sites on the DNA and bind with their PWWP binding domain. One hypothesis suggests that RNA interference may be involved, but these studies were primarily conducted in plants and are not directly applicable to the mammalian genome [32]. Another study highlighted the interaction of DNMT3A & B with transcription factors to maintain or repress specific gene expression. In 2005 Brenner et al. published a work where they showed that the c-Myc transcription factor, which regulates the proliferation and differentiation of cells, binds to DNMT3A and therefore represses gene expression by targeting the methyltransferase to specific genomic loci [33]. In contrast, our group has previously shown that general transcription factors occupy cytosine residues at CpG islands and block the binding of DNMTs to inhibit de-novo CpG methylation, maintaining a non-methylated cytosine status [34]. Additionally, DNMT3A has been shown to specifically bind to non-methylated H3K4, selectively targeting DNMT enzyme activity to chromatin regions that lack methylated H3K4 [35]. Therefore two hypotheses exist where either specific transcription factors guide the methyltransferases toward distinct genomic loci to establish de-novo DNA methylation or the DNMTs just methylate all cytosines, which are not protected by other gene regularity proteins. The second group interacting with cytosine methylation are those reading and recognizing DNA methylation. The first discovered proteins were so-called MBD proteins, which all comprise an evolutionary conserved methyl-CpG-binding domain. There are several members of this protein family, with the most important one named MeCP2. It was shown that these proteins have a very high affinity in binding to cytosine residues, which are methylated, therefore blocking the binding of other proteins to these specific loci, which would alter gene regulation and expression. Furthermore, they contain a domain that represses transcription (TRD domain), partnering with other proteins that repress transcription initiation [36]. For example, protein MeCP2 builds a complex with DNMT1 with its TRD domain and guides the enzyme to hemimethylated DNA sites leading to 5mC orchestration [37]. Besides MBD proteins, several other proteins exist which can bind to DNA methylated at cytosines. Another protein family directly

binding to DNMT1 are ubiquitin-like enzymes named UHRF1 & 2. They possess PHD and ring finger domains interacting with the methyltransferase, actively guiding it to the DNA in the process of replication [38]. Next, proteins were found which contain a zinc-finger domain, namely ZBTB4, ZBTB38, and Kaiso. Kaiso does not contain an methyl-CpG-binding domain but is familiar with MBD proteins and has specific binding patterns recognizing one or two neighboring CpGs [39]. As they are familiar and conserved to MBD proteins, they all serve as transcription-repressing proteins by binding to methylated cytosine residues within the mammalian genome.

The last class of enzymes within the DNA methylation cycle are those which remove the methyl group from the fifth carbon of a cytosine residue; these enzymes are called erasers. It is not possible to cleave the bond between the CH_3 group and the carbon of the pyrimidine ring directly. This is because the bond appears to be very strong and has not been detected in mammalian cells so far. As the demethylation process is complex, it requires multiple intermediate steps to convert 5mC to the final cytosine, which has evolved over time. First of all, demethylation can occur actively or passively. In the process of passive demethylation, which mostly takes place during cell division, the newly synthesized DNA strand, which generally lacks cytosine methylation, fails to receive the necessary methylation information due to insufficient DNMT1 function. This process is mainly observed in malign cells [40]. Active cytosine demethylation is mainly contributed in two enzymatic ways. The AID/APOBEC (activation-induced cytidine deaminase/apolipoprotein B mRNA-editing enzyme complex) pathway deaminates the present amine group leading to the conversion to the base thymine. The resulting mispairing of guanine with thymine is recognized by the base excision repair (BER) mechanism, and finally, the thymine is exchanged with unmethylated cytosine. The BER includes an enzyme called thymine DNA glycosylase (TDG), therefore replacing the base thymine and not the guanine within the excision mechanism [41]. In line with the AID/APOBEC replacing pathway, ten-eleven translocation enzymes, namely TET1, 2 & 3, utilize another way in which 5mC is converted to the final unmodified

cytosine. Here a hydroxyl group is added to CH_3 forming 5-hydroxymethyl cytosine (5hmC). 5hmC is then further processed either via AID/APOBEC to thymine or via the TET enzymes to intermediate forms 5-formyl-cytosine (5fC) and 5-carboxy-cytosine (5caC) [42]. Those modified but malformed cytosines are then recognized by the former mentioned TDG and transformed into the final unmodified cytosine.

In summary, methylation of cytosine represents one of the best-studied and most important features in epigenetic regulation of gene expression. In the context of transcription and expression, it is involved with other proteins directly binding to or guiding other factors to DNA to establish a plethora of gene regulation mechanisms. The process of active methylation and demethylation is a complex procedure involving various enzymes and other cellular components orchestrating the transcription of genes, differentiation, and cell proliferation. Misfunction of these processes is often involved in the generation of malignancies and other disease-responsible activities disturbing the normal functions within mammalian cells. Therefore, extensive scientific research is needed to better understand cellular and epigenetic processes, particularly in the field of DNA methylation.

1.4. Other epigenetic mechanisms

DNA methylation is not the only epigenetic mechanism involved in gene regulation and the expression of specific genes. One major aspect of epigenetic regulation is how the enormous DNA molecules are arranged within the cell's nucleus. Basically, DNA is organized in two structures, hetero- and euchromatin. Within euchromatin, the DNA is loosely aggregated, and transcription can take place, whereas within heterochromatin, the DNA is tightly packed, and no transcription is possible. Chromatin, therefore, stands for a complex of the DNA molecules itself and proteins in close association. The nucleosome represents the smallest subunit of this complex. It consists of eight so-called histone proteins with the DNA helically folded around these proteins with 147 bp in

length. The eight histone proteins consist of two histone proteins H2A and two histone proteins H2B with a further tetramer of H32-H42 [43]. Between the nucleosomes, a stretch of 50bp, the linker histone H1 stabilizes the chromatin structure. The N-terminal parts of the histone tails, which stand out of the compact nucleosome, are targets of a variety of chemical modifications, namely acetylation, methylation, phosphorylation, and to a minor extent, ubiquitinylation. These alterations mainly take place at three specific amino acids, and each serves different purposes. Histone acetylation mainly contributes to softening the interaction between the histone core proteins and the adjacent DNA resulting in a higher transcriptional activity of the corresponding DNA locus [44]. This can be explained as the acetylation decreases the positive charge of the lysines at the end of the N-terminal histone tails. Enzymes involved in this process are p300, Gcn5-Related N-Acetyltransferases, and MYST-family histone acetyltransferases summarized in the group of histone acetyltransferases (HAT) [44]. Another family of enzymes, called histone deacetylases (HDAC), actively remove the acetylation bond of lysines contributing to a tighter packing of the local chromatin. The result is a decrease in transcriptional activity at these local sites. Further, it has been shown that HDAC proteins directly bind to transcription factors which repress active transcription [45]. Histone methylation can lead to either transcriptional activation or repression based on the specific methylated amino acid residue. Furthermore, the amino acid lysine has three different target sides for methylation, whereas arginine residues serve two, increasing the variability of this complex epigenetic regulation. Histone methyltransferases serve as the active fabric to bind methyl groups covalently to lysines and arginine amino acids at the histone tails, mainly of core histone H3. The single methylation of lysine 27, 9, and 79 on histone 3, lysine 20 located on histone 4, and lysine 5 on histone 2B serve as markers for increased gene transcription, whereas the triple methylation of lysines 27, 9, and 79 located on the histone tail of histone three pack the chromatin more tightly, interacting with repressing factors, and therefore, repress active transcription [46]. In addition to the direct interactions of the modified histone tails toward the DNA molecules, they interact with different proteins regulating other epigenetic pathways. It has been shown that

DNA methyltransferases not only directly influence the DNA methylation landscape by adding methyl groups to cytosine but also interact with proteins and enzymes which are involved in downregulating gene activity. In 2003 Fuks et al. demonstrated that DNMT1 and DNMT3A interact with SUV39H1 and HP1, which transfer methyl groups toward H3K4 residues. This interaction further decreases transcriptional activity [47]. Furthermore, the same group has shown previously that DNMT1 cooperates with histone deacetylases. This interplay strengthens the tight packing of the chromatin leading to further transcriptional silencing [48]. Overall, it can be seen that the enzymes responsible for epigenetic regulation don't only serve a single purpose but also show strong signs of cooperation, emphasizing the complex epigenetic regulation of gene activity within higher eukaryotic cells. This observation was further underlined by the fact that also other methyl-binding proteins, for example Ubiquitine-like proteins, containing PHD and RING finger domains 1, interact with core histone proteins increasing transcriptional repression [49]. Another protein, namely MeCP2, actively guides deacetylases toward specific DNA regions and increases the deacetylation at these loci contributing to another epigenetic feature of transcriptional repression [50]. The last and most recently found epigenetic regulation of gene regulation within eukaryotic cells is accompanied by non-coding RNA molecules, mainly miRNAs. They act in a variety of pathways, especially in silencing gene activity by degrading mature mRNA. Here small RNA stretches which can base-pair to native RNA are processed by a protein complex called Dicer and bind to the mRNA leading to a cut of the RNA and degradation of the mRNA molecule [51]. Furthermore, it has been shown that this regulation of gene activity also targets genes responsible for the transcription of epigenetic players like DNMTs [52]. This process leads to a reduction of DNA methylation at cytosine residues.

Taken together, epigenetic regulation of active gene transcription and expression is accompanied by a variety of pathways and mechanisms, underlining the complexity of this interesting field of scientific research.

1.5. DNA methylation in tumorgenesis

As previously stated, epigenetic regulation plays a vital role in altering cell-specific gene expression, deactivating the second X chromosome in females, and silencing inherited genomic sections from viruses. Furthermore, epigenetic dysfunctions have been associated with several mammalian diseases, and the research in epigenetics regarding cancer has attracted a lot of interest in recent years. It is widely recognized that mutations causing alterations in cell function play a significant role in the development and advancement of cancer [53]. There are multiple factors that can contribute to the onset and progression of this fatal disease, including inherited or rare mutations, oncogene overexpression, downregulation of tumor suppressor genes, and alterations in the epigenome that affect DNA methylation, histone modifications, and changes in the chromatin architecture. Alterations in epigenetic markers due to environmental and cellular changes have the potential to play a significant role in the onset and progression of cancer [54]. Unlike mutations that manifest in and change the DNA code itself, epigenetic aberrations are more dynamic, since chromatin modifications and DNA methylation are processes that rely on enzymes adding and removing epigenetic features. These changes can dramatically increase the onset of tumor progression when they affect promoter regions of tumor oncogenes or tumor suppressors.

DNA methylation changes and aberrant features can generally be divided into distinct types within cancer progression (1.3) [55]. On the one side, hypomethylation is defined as the global loss of DNA methylation marks at CpG sites that are generally methylated in normal cells. These CpGs are generally located in gene bodies and loci, which are commonly sparse in CpG density. Hypomethylation was the first DNA methylation aberration described in the context of cancer development [56]. In general, a loss of methylation at cytosine residues can lead to DNA architecture instability because it disturbs the compact structure of the DNA at specific sites. This can lead to aneuploidy or disrupted, segmented chromosomes and cancer progression [57]. The inactivation of

DNA methyltransferase like DNMT1 or DNMT3A/B can be the cause of a global loss of DNA methylation leading to an increase in the transcription of tumor oncogenes or the activation of virus-incorporated transposable elements [58]. Recent studies have further highlighted the importance of hypomethylation in the development of cancer. For example, Zhao et al. showed that a loss of DNA methylation at enhancer sites influencing the expression of the androgen receptor is a prominent driver of prostate cancer. They analyzed 100 human tumors with corresponding 10 metastatic samples via whole genome bisulfite sequencing and RNA-seq. With this analysis, they were able to categorize subtypes in prostate cancer and found new epigenetic subclasses [59]. Moreover, Ohshima et al. published a recent study showing that global hypomethylation is present in high-risk malignant gastrointestinal stromal tumors (GIST) with mutations in the KIT gene at exon 11. This mutation and resulting hypomethylation leads to increased malignancy and metastatic spread in this type of cancer, emphasizing the important role of this feature in the onset of tumor development and it's capability to spread throughout the body [60].

Besides a global loss of methylation, studies have reported the activation of proto-oncogenes due to a decrease of cytosine DNA methylation at their promoter sites [61]. Proto-oncogenes are genes that eventually become oncogenes with increased expression compared to normal cell behavior. In the context of this thesis, Her2 serves as a proto-oncogene. It becomes over-expressed during adolosecence because of a hormone-driven promoter leading to increased expression and therefore increased proliferation of the mammary-gland cells resulting in the development of breast cancer.

Conversely, alterations of DNA methylation in cancer often show signs of hypermethylation defined by the increase of cytosine methylation compared to normal tissue [55]. These cytosines are often located in CpG islands in promoter regions that are normally unmethylated. Besides this one finds direct changes in DNA methylation due to UV irradiation or exposure to carcinogenic substances like tobacco, radon, or asbestos [62]. In general, healthy cells exhibit unmethylated CpGs at tumor suppressor

gene promoter sites, enabling their constant transcription and expression. This leads to a consistent cell proliferation and cell fate. However, approximately 10% of CpG islands show signs of hypermethylation in cancer cells leading to a decreased expression of such genes [63]. This effect was observed in human lung cancers and lymphomas at CpG islands corresponding to the calcitonin gene (CT) [64]. Studies revealed that DNA hypermethylation is found in colon cancer, where the increase of methylation showed a correlation to BRAF mutations leading to aberrant growth and malign behavior [65]. Furthermore, recent studies have shown the importance of hypermethylation in other cancer types. In 2023 Yi et al. showed that TRIM28 hypermethylation is an important factor of the development of esophageal cancer and drives the progression of this cancer subtype via the suppression of transcription factor ZNF750 [66]. Additionally, a recent study has shown that the increase in methylation at the ALX3 promoter serves as a prognostic marker in lung cancer and therefore could lead to an early detection of this cancer subtype with a better chance of survival of the patients [67].

1.6. Epigenetic alterations during metastatic breast cancer progression

Like any other cancer, breast carcinoma is a heterogeneous disease harboring a combination of genetic and epigenetic aberrations. It has been shown that abnormalities concerning the DNA sequence itself, like copy number changes and mutations, define breast cancer subtypes.

In 2012, Stephens et al. performed an extensive study on the genomes of 100 breast cancer carcinomas, focusing on mutations in exons of protein-coding genes and copy number alterations [69]. Their research uncovered 7,421 somatic point mutations across 21,416 genes. There was notable variability in the amount of somatic mutations among

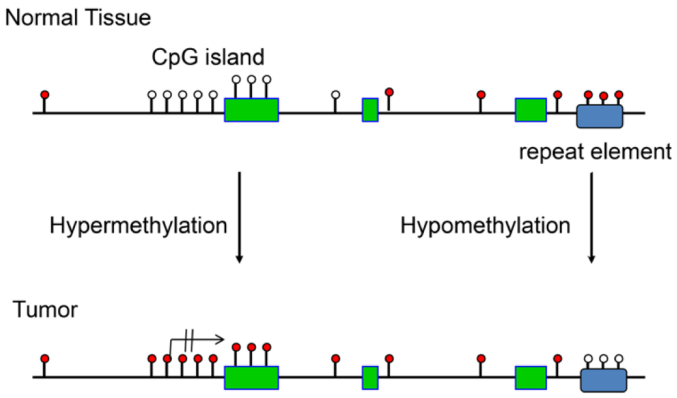


Figure 1.3.: Overview of most prominent DNA methylation changes in cancer development. Hypermethylation is definded as the gain of DNA methylation at specific sites, prominentely at CpG island at promoter sites, leading to an decrease in transcriptional activity. This phenomenon mostly affects tumor surpresssor genes which are then not further expressed. On the other hand there is DNA hypomethylation which in general is a global, genome wide occurrence leading to genome instability and tumor onset. White dots display unmethylated CpGs whereas red dots mark methylated CpGs. The arrow which is scored-out represents a transcription stop at a promoter after increased CpG methylation. Green features represent exons and boxes in blue mark position within the genebody [68].

different tumors, reflecting the heterogeneous characteristics of breast cancer. Importantly, several genes associated with the development of breast cancer, such as *BRCA1*, *RB1*,and *TP53* were found to harbor somatic mutations in the analyzed tumors. These genes are known to play critical roles in transcriptional regulation, cell-cycle regulation, and apoptosis, underscoring their potential significance in breast cancer pathogenesis.

In a different study utilizing somatic mutations for the classification of various subtypes of breast cancer, Curtis et al. performed a comprehensive analysis of copy number aberrations in approximately 2,000 human breast tumor samples [70]. The study was based on the premise that the expression characteristics of ancestral genes are shaped by the interplay of proximal cis-acting loci located within a 3000 base segment surrounding the target genes and distal trans-acting loci, which are located outside of this regions. They detected deletions in the *PPP2R2A*, *MTAP*, and *MAP2K4* genes and somatic copy number aberrations in nearly 40% of genes. Remarkably, breast cancer subtypes devoid

of copy number aberrations exhibited a more favorable prognosis compared to those with multiple copy number aberrations.

Despite these studies focusing on mutations and aberrations on the level of DNA sequence, recent research has shown that epigenetic alterations pave the way for the development of breast cancer disease. A majority of these studies have focused on elucidating the role of miRNAs in breast carcinoma development. The Cancer Genome Atlas Network investigated the levels of miRNA expression by discerning the top 3,662 genes that exhibited variable expression and identifying seven distinct subtypes of miRNA in 525 tumors, along with 22 tumor-adjacent normal tissues. The differential expression patterns of miRNA subtypes significantly correlate with the primary mRNA subtypes, which are determined based on the clinical classification of ER+, PR+, and HER2+ [71]. Other studies have linked distinct DNA methylation patterns to the different breast cancer subtypes and their metastatic behavior. In 2011 Fang et al. presented a study that investigated the epigenomic changes associated with breast cancer metastasis. The researchers performed comprehensive DNA methylation profiling of primary breast tumors, matched metastatic samples, and identified substantial dissimilarities in DNA methylation patterns between the two groups. They found that DNA methylation alterations in metastatic tumors were enriched in genes involved in cancer-related pathways, including cell adhesion, migration, and signaling. Furthermore, the study showed that specific DNA methylation changes in the primary tumors were associated with an increased risk of metastasis. The findings suggest that breast cancer metastasis is associated with distinct epigenetic changes, which may provide an epigenomic foundation for understanding the molecular mechanisms underlying metastasis and could potentially be used as prognostic markers for breast cancer patients [72]. Despite the advancements in recent research on the molecular dynamics in DNA methylation and systematic breast cancer establishment and progression, the underlying processes of methylation pattern introduction, timing, and its function remain still elusive as none of the currently available studies have shed light on this time-dependent process.

1.7. Morphogenesis of the breast

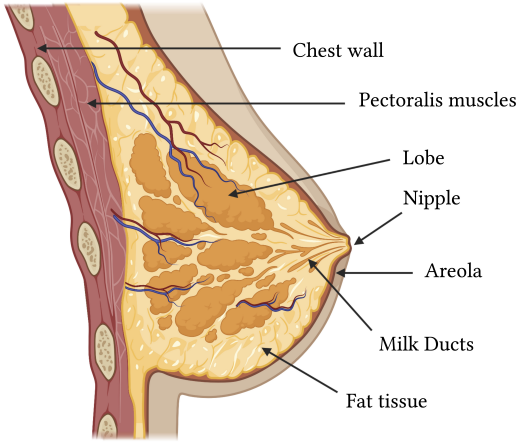


Figure 1.4.: Mammary gland, cross section of the female breast representing the structure of the major parts including the lobes where breast cancer arises.

Located at the anterior thoracic wall, women's breasts extend from the second to the sixth rib. In general, the mammalian breast consists of skin, subcutaneous tissue, and breast tissue [73]. Among its main functions are lactation and nourishment of offspring. Figure 1.4 represents a cross-section overview of the female human breast.

There are three distinct stages in the development of the mammary glands. At embryonic day (E) 10.5, bilateral milk ridges are formed. The next phase is initiated one day later (E11.5) while five sets of placodes are formed at distinct positions at the milk line. The terminal phase takes place during the period spanning from E11.5 to E12.5, during which cells originating from the placode infiltrate the subjacent mesenchyme, culminating in the formation of the characteristic bulbous-shaped mammary buds. Once this has been achieved, the buds go into quiescence until the embryo begins its final development at E15.5 to E16.5. Then, the epithelial cells of the mammary gland begin to divide and proliferate. The bud sprouts within the blastoderm, where it will grow into the developing adipose tissue of the mammary gland. During this

process, the epidermal cells covering the buds begin to differentiate into nipple skin. At birth, the mammary sprout has built a ductal tree comprising a central primary duct accompanied by ≈ 15-20 secondary branches called lobes [74, 75]. Each lobe is further divided into lobules or terminal duct lobular units (TDLUs), which are the effective components producing milk during lactation [76]. Till puberty, the mammary gland epithelial cells remain quiescence and then start to invade the mammary fat pad under the control of hormones [77, 78]. The number of substructures per TDLU and the total number of TDLUs decrease after childbearing and physiological aging. As a result of this process of involution, most breast cancer malignancies originate from TDLUs [79].

1.8. Breast cancer

Breast cancer is the most common cancer type in women and a major public health concern worldwide. The prevalence varies greatly by region and ethnicity. In developed countries, the incidence is higher than in developing countries. Age increases the risk of breast cancer, with most cases occurring in women over the age of 50. In families with a history of breast cancer, women are more likely to develop the malignancy due to a genetic predisposition [80].

Breast cancer progression begins with transformed, aberrant epithelial cells from the TDLUs proliferating and evading into the lumen. Typically, this is the first stage of non-invasive breast cancer. The process of breast cancer development can be divided into two distinct categories, namely ductal carcinoma in situ (DCIS), which is limited to the milk ducts and constitutes the majority of cases of breast cancer, and lobular carcinoma in situ (LCIS), which is located in the glands [76]. As the cancer proceeds, it becomes more invasive, causing it to penetrate through the basal membrane and invade the surrounding normal tissue that is not malignant. In the final stage of breast cancer

progression, the disease spreads to distant organs, primarily to the lungs, bone marrow, or brain [81–83].

On the basis of gene expression profiles and immunohistochemistry markers, breast cancer can be classified into five distinct subtypes. These types differ in clinical importance and are classified as luminal A, luminal B, basal, normal-like, and Her2-overexpressing [84, 85]. As our consortium wants to study early versus late dissemination in breast cancer, a Her2-overexpressing breast cancer model was chosen. This tumor subtype is defined as very aggressive with a poor prognosis. [86]. The human epidermal growth factor receptor (HER2) is overexpressed in $\approx$ 25 percent of all breast cancer cases. Its overexpression leads to an increased accumulation of HER2 receptor molecules on the cell surface, resulting in a dysregulated proliferation of cells [87]. The aberrant proliferation is mediated by the upregulation of intrinsic cellular processes and the activation of key genes such as *CDC45L* [88], a known driver of cell proliferation, and *ASCL2* [89], a crucial component of the Wnt signaling pathway that acts as a regulator of cell growth [90]. Moreover, one of the major characteristics of this subtype of tumor is that it neither contains estrogen receptors nor progesterone receptors, which distinguishes it from other subtypes. Metastatic HER2-positive breast cancer can be treated with various targeted therapies that include neratinib (Nerlynx), a tyrosine kinase inhibitor, trastuzumab (Herceptin), a monoclonal antibody that blocks Her2's transduction pathway by binding to it, ado-trastuzumab emtansine (Kadcyla), which is a combination of monoclonal antibody trastuzumab and cytotoxic DM1, and pertuzumab (Perjeta), a monoclonal antibody that binds to subdomain II of HER2. These targeted therapies can be combined with chemotherapy, hormone therapy, radiation therapy, and/or mastectomy surgery for better results [91]. The aggressive nature of HER2 positive breast cancer is attributed to its propensity to accumulate genetic and epigenetic alterations that contribute to aberrant cell growth and proliferation, ultimately leading to increased metastatic potential [92].

1.9. Metastasis

Metastasis comprises the process in which malignant cells progress from the primary tumor to another organ establishing a secondary neoplasm at distant anatomical sites [93]. Despite significant progress in scientific research and therapeutic interventions, metastasis continues to be the foremost contributor to cancer-related mortality on a global scale [94]. In order for cancer cells to metastasize, a complex and multi-step process is required, with only a small proportion of cancer cells that enter circulation colonizing distant organs, while the majority of them are eliminated by the immune system or other host defenses [95–97]. Given that our research in this project focuses on elucidating the mechanisms of tumor dissemination during different stages of tumor development, particular attention is directed toward investigating the pivotal role of local invasion as the initial and distinguishing step in this process. The most studied cell migration process is called epithelial-to-mesenchymal transition (EMT). Epithelial-derived cancer cells, which are generally tightly connected to each other, gain molecular and cellular changes to acquire a more mesenchymal phenotype. This is achieved by a complex interplay of signaling pathways, including TGF-β, Wnt, and Notch activating transcription factors (TFs), like Slug and Snail [98]. These TFs control the expression of genes repressing mesenchymal phenotype like E-cadherin or activate the expression of pro-mesenchymal proteins like N-cadherin and fibronectin. These changes enable the transformed cancer cells to breach the extracellular matrix (ECM) and invade the stroma. The second step is termed intravasation, whereby carcinoma cells that are locally invasive, penetrate the lumina of lymphatic or blood vessels. Metastatic carcinoma cells are mainly spread through the hematogenous circulation, but nevertheless, the spread through lymphatic vessels is often a marker for disease progression in human tumors [99]. After cancer cells have entered the bloodstream, they have to survive within the circulation, marking the third step of the metastatic process. Here circulating tumor cells (CTCs) function as metastatic intermediates losing the anchor to the ECM and widely spreading through veins and arterial vessels. CTCs remain relatively

short within the blood system as their diameter (20–30 um) is decently larger than the diameter of the capillaries, eventually clocking themselves in the bloodstream [100]. This observation marks also one hypothesis of why cancer cells arrest at specific distant organs to establish metastasis as a passive process dictated by the layout of the vasculature and size restrictions. An alternative hypothesis suggests that CTCs are prone to settle in specific tissues due to genetically programmed ligand-receptor interactions [101]. Once tumor cells have reached distant organs through the microvasculature, they initiate intraluminal growth and form microcolonies. These colonies can eventually rupture the cell walls of blood vessels, then allowing direct contact between metastatic cells and the parenchyma of surrounding tissue [102]. Carcinoma cells can also cross capillaries by infiltrating the endothelial cells and pericyte layers that divide the vessel lumina from the stroma. This process marks the next step of the metastatic cascade and is known as extravasation. After the tumor cells have successfully extravasated into the distant microenvironment, they have to adapt and survive in the foreign tissue. The microenvironment of the metastatic site typically exhibits distinct differences from the primary tumor site, encompassing variations in stromal cells, extracellular matrix (ECM) composition, and overall tissue architecture. Some studies propose that tumor cells can address this challenge by establishing a "pre-metastatic niche", where primary tumors release signals that induce changes in the local microenvironment of future metastatic sites [103]. These changes lead to the mobilization of hematopoietic progenitor cells originated from the bone marrow, modifying the metastatic site's local microenvironment. Importantly, these changes occur before the arrival of carcinoma cells at the metastatic sites, making them more hospitable for disseminated tumor cells to settle and form micrometastases. Moreover, disseminated cancer cells need to adapt diverse cellular functions to adjust to the requirements of foreign tissues. For instance, the activation of Src tyrosine kinase signaling contributes to the survival of breast carcinoma cells in bone but not in the lung, highlighting organ-specific variations in adaptive mechanisms [104]. Disseminated carcinoma cells, which have migrated from the original neoplasm to distant sites in the organism, may face obstacles in their ability

to proliferate and form metastases. These cells might gradually decrease or remain dormant as microcolonies. The dormancy of disseminated tumor cells could result from incompatibilities with the foreign microenvironments they face, demanding the activation of cell-non-autonomous mechanisms to exit dormancy and begin proliferating. On the other hand, some disseminated tumor cells may proliferate continuously but fail to undergo a net increase in overall cell number due to high apoptotic rates [105]. The "seed-and-soil" theory, proposed by Stephen Paget over a century ago, indicates that certain organ microenvironments can be more or less favorable for the growth and persistence of disseminated tumor cells [106]. Recent studies have discovered genes that promote organ-specific metastatic colonization by addressing the incompatibilities between tumor cells and foreign tissue microenvironments [107–110]. For example, IL-11 has been shown to facilitate the formation of bone metastases by breast cancer cells by perturbing the normal physiological processes in the bone microenvironment [107]. Taken together, carcinoma cells carry out complex, step-by-step biological processes that alter genetic and epigenetic mechanisms to develop clinically relevant, large growths at distant organ sites.

1.10. Early versus late dissemination of cancer cells

In 1976 Nowell first proposed the theory of monoclonal cancer origin, which is widely accepted in science to date [111]. However, there is an ongoing debate between two prominent theoretical models that seek to explain the timeline of genomic and epigenetic changes that enable the process of metastasis, as well as the degree of molecular similarity between the primary neoplasm and their corresponding metastatic lesions. Figure 1.5 shows an overview of the two general cancer progression models. The linear progression model hypothesizes that cancerous cells undergo several cycles of genetic transitions and selective pressures within the microenvironment of the primary tumor prior to successfully spreading and establishing a metastatic site. In this scenario,

metastases arise from the most malignant and progressive carcinoma clone that also prevails at the primary tumor site [112]. For instance, these investigations into the intricate mechanisms underlying metastasis have often utilized cell lines from advanced tumors or metastatic sites.

Nevertheless, several studies have shown that metastatic lesions are comprised of similar gene expression profiles also found in a majority of bulk cancer cells within the primary tumor [113, 114]. These findings strengthen the hypothesis that metastatic cells disseminate at a later stage from an already solid tumor. Furthermore, it was postulated that primary tumor size often correlates with successful metastasis. For example, the tumor suppressor gene *TP53* is often mutated in the late stages of breast cancer, where the primary tumor already has a size larger than 5 cm [115].

In contrast, the parallel progression model posits that metastatic tumors advance autonomously and concurrently from primary tumor tissue at an early stage, leading to increased divergence in the genetic and epigenetic profile between the primary tumor and the metastatic lesions. This model highlights the autonomous acquisition of genetic and epigenetic changes in metastases as they are subjected to distinct selective pressures in different anatomical sites that differ from the microenvironment at the primary tumor site. Early dissemination may establish distant metastases before diagnosing the primary tumor, making early detection and treatment more challenging. Evidence suggests that cancer cells can spread early in the disease before the primary tumor becomes clinically detectable [116]. Moreover, Schmidt-Kittler et al. have analysed early disseminated tumor cells (DCCs) from patients with stage M0 tumors and revealed a significantly reduced occurrence of chromosomal aberrations compared to both primary tumors and DCCs obtained from patients at the M1 stage where the tumor already has a detectable size [117]. Despite these advances, the precise mechanisms and timing of cancer cell dissemination in breast cancer remain poorly understood. Several questions are still unanswered, such as whether early and late disseminated cells have distinct molecular and genetic profiles and respond differently to therapy. Therefore, there is a need for

further research to improve our understanding of the mechanisms and timing of cancer cell dissemination in breast cancer.

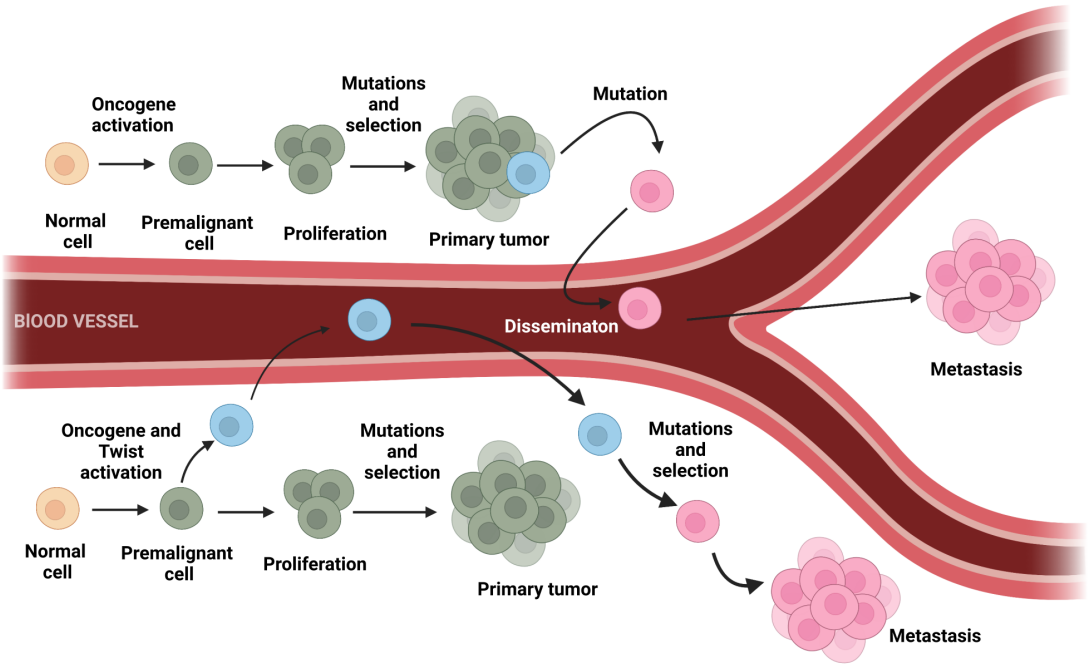


Figure 1.5.: Dissemination Models. Malignant cells which disseminate late show genetic features like the primary tumor (top). Cells that disseminate early therefore accumulate mutations at the distant anatomical site (bottom). Figure adapted from [118].

1.11. The Balb-NeuT mouse model

To investigate the dynamics and molecular processes defining breast cancer cell dissemination, a transgene mouse model was used in the work for this thesis. The BalbNeu-T mouse model is described as a very aggressive type of cancer model due to its *Her2* over-expression. The HER2 receptor is normally expressed in mammary tissue when the *Erbb2* oncogene is activated. The importance of tumor development of the rodent

counterpart of the *Erbb2* gene, *Neu*, was first explicated in a study from 1984 that employed a chemically-induced neuroblastoma model, thereby constituting a seminal milestone in the elucidation of the genetic promotion of oncogenic signaling pathways [119]. As a result of this research, it was possible to develop targeted therapies for breast cancer and better understand the molecular basis of the disease. Breast cancer patients with HER2 overexpression who received trastuzumab (Herceptin), a monoclonal antibody that targets ERBB2, experienced superior disease-free and prolonged survival outcomes compared to those with HER2 amplification [120].

The aforementioned observations have initiated the creation of transgenic animal models as a means to examine the involvement of HER2 in distinct breast cancer sub-types. For this purpose, the approach from Sinn et al. was utilized to regulate the transgene using 16 long terminal repeats of the mouse mammary tumor virus (MMTV-LTR) to facilitate transgene expression at high levels exclusively within the mammary epithelium [121]. Nevertheless, obstacles had to be overcome as the expression of the HER2 receptor in transgenic mice showed a variety of phenotypes based on the inserted oncogene (wild-type or mutated), the recipient host strain, and the genomic location and frequency of the insertion [122]. The integration of the wild-type NEU oncogene led to very slow tumor development with a tumor remission of one year. Therefore a new activated mutant NEU (NeuT) oncogene was developed and investigated in the FVB mouse strain. However, this model generated tumors in organs other than the mammary gland [123]. Lucchini et al. reported heterogeneous tumor onset and fertility issues with a redesigned NeuT model [124]. The Balb-NeuT murine model was then established by strategically crossing transgenic CD1 male mice with BALB/c females of random-bred ancestry [125]. The utilization of this genetic manipulation facilitated the expression of a modified variant of the rat HER-2/neu oncogene, regulated by the murine mammary tumor virus (MMTV) promoter, in juvenile BALB/c mice, commencing at week 3. This experimental model exhibits a rapid and localized tumor development phenotype, limited to the mammary gland, which offers a distinctive and effective interplay

between tumorous signals and repressive immune responses, owing to the continuous emergence of neoplastic cells expressing rat *p185neu* within the Balb-NeuT genome.

1.12. Aim of work

Several studies have supported the notion that cancer cells can disseminate at an early stage; however, the relationship between primary tumor progression and metastasis remains a topic of ongoing debate [100]. Empirical findings obtained from in vitro studies on cell lines indicate that the process of metastasis is associated with the accumulation of assertive single-cell clones within the tumor, the transition from epithelial to mesenchymal phenotype, and further genetic and epigenetic instability of the genome. It still remains elusive how metastatic cancer cells are capable of mastering critical metastatic barriers in the early dissemination cancer model, as they were not able to accumulate genetic and epigenetic abnormalities from late-stage primary tumor cells. [100].

According to Patel and Vanharanta [126], no mutations or epigenetic changes have been identified as specific drivers of metastatic progression across various tumor types. Consequently, the mechanisms underlying cancer cell dissemination, immune evasion, resistance to chemotherapy, and successful growth in a specific niche remain poorly understood [127].

Furthermore, the number and nature of progression pathways to metastasis and the relative contribution of epigenetic versus genetic alterations to this process are yet to be determined. The timing of methylation changes during systemic progression and their interplay also require further investigation. Given these gaps in knowledge, the critical question of whether DNA methylation promotes systemic progression and the mechanisms involved remain unanswered. To address some of these questions, the consortium Forschergruppe 2127 (FOR2127) has employed the Balb-NeuT mouse model,

which does not exhibit genetic evidence for metastatic driver mutations but displays significant epigenetic differences between primary tumors and metastases. This model is, therefore, an appropriate tool for investigating the methylation progression and shedding light on the critical issues mentioned above [128]. In a previous study of FOR2127, the aggregation of epigenetic alterations in metastatic tumors was detected that involved various recurring modes of promoter methylation and demethylation. However, it remains unclear whether these epigenetic changes occur adaptively or if they emerge as early epigenetic lesions in individual cells of primary tumor lesions that undergo subsequent selection. To investigate the selection hypothesis, we aimed to identify sub-clones within primary tumors that exhibit early epigenetic progression patterns. Proving selection requires the identification of existing small sub-clones in primary tumors. This is not feasible with the broad but shallow sequencing approach used in the first funding period. Therefore, a marker panel of 9 representative regions of 1 Kb on average was selected, that covers various patterns of methylation progression. Ultra-deep sequencing of 9 marker regions allows us to study DNA methylation on a single allele level which will recover rare subclones in primary tumors and early lesions or strengthens the hypothesis that increased DNA methylation is the effect of progressive cell proliferation and cell aging.

2. Material

2.1. Equipment

Device	Supplier/Manufacturer
Agilent 2200 Tape Station	Agilent Technologies, Santa Clara, USA
Autoclave	Technomara, Fernwald, Germany
Biofuge Fresco	Heraeus, Hanau, Germany
Biofuge Pico	Heraeus, Hanau, Germany
Centrifuge 5804 R	Eppendorf, Hamburg, Germany
CFX Connect Real-Time PCR Detection System	Bio-Rad Laboratories, California, USA
Electrophoresis equipment	Biometra, Goettingen, Germany
Eppendorf Research (plus) (2.5 μl, 10 μl, , 20 μl, 100 μl, , 200 μl, 1000 μl)	Eppendorf, Hamburg, Germany
EVOS M5000	Life Technologies, Carlsbad, USA
Fusion Pulse 6	Vilber Lourmat, Collegien, France
Heat sealer	Eppendorf, Hamburg, Germany
Hemocytometer	Marienfeld, Lauda-Koenigshofen, Germany
Herasafe Laminar air flow cabinet	Heraeus, Hanau, Germany
IKA MS 3 Vortexer	IKA-Werke, Staufen Germany
Incubator BBD 6220	Heraeus, Hanau, Germany
Intelli-Mixer RM-2L	Elmi-Tech, Riga, Latvia

Mastercycler ep realplex	Eppendorf, Hamburg, Germany
Megafuge 3.0R	Heraeus, Hanau, Germany
Microscope Axiolab	Zeiss, Jena, Germany
Microscope Axiovert	Zeiss, Jena, Germany
Microwave Privileg 8017	Privileg, Stuttgart, Germany
Mini-PROTEAN Tetra Handcast System	Bio-Rad Laboratories, California, USA
Mini-PROTEAN Tetra Vertical Electrophoresis Cell	Bio-Rad Laboratories, California, USA
Multifuge 3S-R	Heraeus, Osterode Germany
Multipipette plus	Eppendorf, Hamburg, Germany
NanoDrop 1000	PeqLab, Erlangen, Germany
Neon Transfection System	Thermo Fisher Scientific, Waltham, USA
PCR Mastercycler nexus X2	Eppendorf, Hamburg, Germany
PCR Thermocycler PTC-200	MJ-Research, Biometra, Goettingen, Germany
pH-Meter	Knick, Berlin, Germany
Pipetboy	Integra Biosciences, Zizers, Switzerland
PIPETMAN Classic (10 μl, 20 μl, 200 μl, 1000 μl)	Gilson, Midleton, USA
Power supplies	Biometra, Goettingen, Germany
PowerPacTM Basic Power Supply	Bio-Rad Laboratories, California, USA
Qubit 2.0 Fluorometer	Life Technologies, Carlsbad, USA
Refrigerated SIGMA 2-Sartorius Centrifuge	Sartorius, Goettingen, Germany
Rotilabo-mini-centrifuge	Carl Roth, Karlsruhe, Germany
SPROUT MINI CENTRIFUGE 12V	Heathrow Scientific, IL, USA
Tecan Spark multimode reader	Tecan Trading AG, Switzerland
Thermo Cell	Hangzhou Bioer Co. Ltd., Hangzhou, China
Thermomixer	Eppendorf, Hamburg, Germany

Typhoon 9200	Molecular Dynamics, Krefeld, Germany
UVT-20 M Transilluminator	Herolab, Wiesloch, Germany
Vacuum manifold QIAvac 24 Plus	Qiagen, Hilden, Germany
Vortex-Genie 2	Scientific Industries, New York, USA
Water purification system	Milipore, Eschborn, Germany
Waterbath	Julabo, Seelstadt, Germany
ZX4 Advanced IR Vortex Mixer	VELP Scientifica, Usmate, Italy
MiSeq	Illumina, Inc., San Diego, CA, USA
Eppendorf nexus	Eppendorf, Hamburg, Germany
PCR kabinet	Jena Biosciences, Jena, Deutschland
Rainin mulitchannel pipet 8 channel	Mettler-Toledo, LLC, Columbus, OH, USA
magnetic rack	in-house production
10x magnetic rack	10x Genomics, Pleasanton, CA, USA

Table 2.1.: Technical devices and equipment

2.2. Consumables

Product	**Supplier/Manufacturer**
AMPure XP Magnetic beads	Beckman Coulter, Brea, USA
Broad tips	Sigma-Aldrich, St. Louis, USA
Cell culture flasks and dishes	Eppendorf, Hamburg, Germany
Cell culture pipettes	Costar, Cambridge, USA nerbe plus, Winsen (Luhe), Germany
Cell culture plates (6-well)	Costar, Cambridge, USA
Cell Scraper 25 cm	Sarstedt, Nuemrecht, Germany
Combitips advanced (0.5 ml, 2.5 ml, 5.0 ml)	Eppendorf, Hamburg, Germany

Cryo tubes	Nunc, Wiesbaden, Germany
Disposable scalpel	Feather, Osaka, Japan
DNA LoBind Tubes, DNA LoBind, 1.5 ml	Eppendorf, Hamburg, Germany
Electroporation cuvettes (0.4 cm)	PeqLab, Erlangen, Germany
Eppendorf Safe-Lock Tubes, 0.5 ml	Eppendorf, Hamburg, Germany
Eppendorf Tubes 5.0 ml	Eppendorf, Hamburg, Germany
epT.I.P.S. LoRetention Reloads, Eppendorf QualityTM, 2 – 200 μl / 50 – 1000 μL	Eppendorf, Hamburg, Germany
Falcon tubes (200 ml / 50 ml / 15 ml)	Corning, New York, USA
Falcon 100 mm x 15 mm TC-Treated Bacteriological Petri Dish, 20/Pack, 500/Case, Sterile	Corning, New York, USA
Falcon 5mL Round Bottom Polystyrene Test Tube	Corning, New York, USA
Grade 3MM Chr Blotting Paper, sheet, 46 x 57 cm	GE Healthcare, Little Chalfont, UK
Heatsealing Film	Eppendorf, Hamburg, Germany
High Sensitivity D1000 ScreenTape	Agilent Technologies, Santa Clara, USA
Immobilon-P Membran	Merck Millipore, Burlington, USA
Incidin OxyWipe S	Ecolab Deutschland, Monheim am Rhein, Germany
Micro tube 1.5 ml/ 0.5 ml DNA LowBind	Sarstedt, Nuemrecht, Germany
Neon Transfection Kit 10 μl	Thermo Fisher Scientific, Waltham, USA
NeoTouch Premium Disposable Neoprene Gloves	Ansell, New Jersey, USA
nProtein A Sepharose 4 Fast Flow	GE Healthcare, Little Chalfont, UK
Optical cap 8x strip	Agilent Technologies, Santa Clara, USA

Optical tube 8x strip	Agilent Technologies, Santa Clara, USA
Parafilm	Pechiney Plastic Packaging, Chicago, USA
PCR Plate 96, semi-skirted	Eppendorf, Hamburg, Germany
Pipet tips 10 μl	Sarstedt, Nuemrecht, Germany
Protein G Sepharose 4 Fast Flow	GE Healthcare, Little Chalfont, UK
Safe Seal Tube 2.0 ml/ 1.5 ml	Sarstedt, Nuemrecht, Germany
SafeSeal SurPhob filter tips (10 μl/ 100 μl/ 1250 μl)	Biozym Scientific, Hessisch Oldendorf
TouchNTuff Disposable Gloves	Ansell, New Jersey, USA
VersaTouch Disposable Nitrile Gloves	Ansell, New Jersey, USA
Rainin pipette Tips	Mettler-Toledo, LLC, Columbus, OH, USA

Table 2.2.: Consumables

2.3. Chemicals and reagents

Product	**Supplier/Manufacturer**
Aqua (1000 ml)	B. Braun Melsungen, Melsungen, Germany
Ammoniumpersulfat	Merck KGaA, Darmstadt, Deutschland
Biozym LE Agarose	Biozym Scientific, Hessisch Oldendorf, Germany
Collagenase/hyaluronidase (10×)	STEMCELL Technologies, Vancouver, Canada
Deoxycholate (10%)	Roche, Basel, Switzerland
DMEM	Thermo Fisher Scientific, Waltham, USA
DSG	Thermo Fisher Scientific, Waltham,

	USA
Ethanol > 99,5%, Ph.Eur., reinst.	Carl Roth, Karlsruhe, Germany
Gel Loading Dye, Purple (6 ×)	New England Biolabs, Frankfurt, Germany
GelPilot DNA Loading Dye (5 ×)	Qiagen, Hilden, Germany
GibcoTM Distilled Water	Thermo Fisher Scientific, Waltham, USA
Glycin	Carl Roth, Karlsruhe, Germany
Glycogen (5,0 mg/ml)	Thermo Fisher Scientific, Waltham, USA
Heparin Solution	STEMCELL Technologies, Vancouver, Canada
Human Recombinant bFGF	STEMCELL Technologies, Vancouver, Canada
Human Recombinant EGF	STEMCELL Technologies, Vancouver, Canada
Matrigel	BD Biosciences, Franklin Lakes, USA
Natriumdodecylsulfat	Merck KGaA, Darmstadt, Deutschland
Polyhema	Merck KGaA, Darmstadt, Deutschland
RIPA	Merck KGaA, Darmstadt, Deutschland
Rotiphorese Gel 30 (37,5:1)	Karl Roth, Karlsruhe, Germany
TE Buffer, 1 ×, Molecular Biology Grade	Promega, Madison, USA
Water (RNase/DNase free)	Thermo Fisher Scientific, Waltham, USA

Table 2.3.: Chemicals and reagents

2.4. Enzymes and kits

Product	Supplier/Manufacturer
Agencourt® AMPure XP- Kit	Beckman Coulter, Brea USA
Alt-R S.p. Cas9 Nuclease V3	Integrated DNA Technologies, Loewen, Belgien
Alt-R S.p. Cas9 D10A Nickase V3	Integrated DNA Technologies, Loewen, Belgien
Qubit dsDNA HS Assay Kit	Thermo Fisher Scientific, Waltham, USA
ECL Prime Western Blotting System	Sigma-Aldrich, Taufkirchen, Germany
Blood and Tissue Culture Kit	Qiagen, Hilden, Germany
dNTP Mix	Agena Bioscience, San Diego, USA
dNTPs	Thermo Fisher Scientific, Waltham, USA
EZ Lightning DNA Methylation Kit	Zymo Research, Irvine, USA
EpiMark Hot Start Taq DNA Polymerase	NEB, Frankfurt, Germany
MiSeq Reagent Kit v2	Illumina, Inc., San Diego, California, USA
MiSeq Reagent Kit v2 nano	Illumina, Inc., San Diego, California, USA
Monarch DNA Gel Extraction Kit	NEB, Frankfurt, Germany
Monarch PCR & DNA Cleanup Kit	NEB, Frankfurt, Germany
MycoAlert Mycoplasma Detection Kit	Lonza, Basel, Switzerland
NEXTflex-16S-V1-V3-Amplicon-Seq-Kit	Bio Scientific Cooperation, Austin, USA
50x Proteaseinhibitor Cocktail	Thermo Fisher Scientific, Waltham, USA
Phusion High-Fidelity DNA Polymerase	Thermo Fisher Scientific, Waltham, USA
Proteinase K	Thermo Fisher Scientific, Waltham,

	USA
Reverse Transcriptase SuperScript II	Promega, Madison, USA
HS Taq DNA Polymerase	Thermo Fisher Scientific, Waltham, USA

Table 2.4.: Enzymes and kits

2.5. Western blot antibodies

Antibody	**Supplier/Manufacturer**
Dnmt3a Antibody (C-12): sc-365769	Santa Cruz Biotechnology, Dallas, USA
Goat-anti-rabbit IgG, HRP-conjugated Goat-anti-mouse IgG, HRP-conjugated	Agilent Technologies, Böblingen, Germany

Table 2.5.: Antibodies for Western Blot

2.6. Custom oligonucleotides

2.6.1. CRISPR

Primername	**Sequence 5′- 3′**
D3Anick1_forward	ACTTCTGCCTGCTGCTAGAC
D3Anick1_reverse	AGCAATCCCATCAAAGAGAGAC
D3Anick2_forward	GACTCCAGATGTTCTTTGCC
D3Anick2_reverse	CATGCTTGCTTCTTTCTTTCC
D3Anick3_forward	AGTAGGGTGGAACAGAAATGG
D3Anick3_reverse	TGAAGGATGACAGAGGCAAG

Table 2.6.: Primer sequences CRISPR/Cas9 gene knockdown

Primername	Sequence 5′- 3′
D3A.CatDomain.qPCR.Fwd1	GGCGTTAGTGACAAGAGGGA
D3A.CatDomain.qPCR.Rev1	GGTCCTCACTTTGCTGAACTTG
D3B.CatDomain.qPCR.Fwd1	CCATCAAGGTGTCTGCTGCT
D3B.CatDomain.qPCR.Rev1	TGTCTGCACTTTCTTTAACTTTGCT

Table 2.7.: Primer sequences used for qPCR

guideRNA	Sequence 5′- 3′
DNMT3Anick1.1	CGCACATGTAGCAGTTCCAGGGG
DNMT3Anick1.2	CTTTGCCAATAACCATGACCAGG
DNMT3Anick2.1	CAGGTGGGTAAACCTTTGGGGGG
DNMT3Anick2.2	ATGGGATTGCTACAGGTGAGTGG
DNMT3Anick3.1	TTCAGGCTCCACCCACATGTCGG
DNMT3Anick3.2	ATCATTGATGAGCGCACAAGGG

Table 2.8.: Sequences CRISPR guideRNAs with PAMs

2.6.2. BBA-seq

Primername	Sequence 5′- 3′
TargetBFwd1	CTCTTTCCCTACACGACGCTCTTCCGATCTTTAGAGTTGTAGTGGGGTTG
TargetBRev1	CTGGAGTTCAGACGTGTGCTCTTCCGATCTTAAATACCTCCTACTCTCCTTC
TargetCFwd1	CTCTTTCCCTACACGACGCTCTTCCGATCTGTGTAGGTGAATTAGAAAGTGGGTAT
TargetCRev1	CTGGAGTTCAGACGTGTGCTCTTCCGATCTTAAATCTACTTCTCCTTCCAAAA
TargetDFwd1	CTCTTTCCCTACACGACGCTCTTCCGATCTTTTGGATTTTTTTTGTTGTTAATTG
TargetDRev1	CTGGAGTTCAGACGTGTGCTCTTCCGATCTTCCAACACTTTAACCCTCAAACTAC
TargetEFwd1	CTCTTTCCCTACACGACGCTCTTCCGATCTAGGGAGAGGATGATTTTATTT
TargetERev1	CTGGAGTTCAGACGTGTGCTCTTCCGATCTCAATAACCCAACACCAAAACTCTA
TargetFFwd1	CTCTTTCCCTACACGACGCTCTTCCGATCTTAAGTGGTTGGAGGAGGTTGAT

TargetFRev1	CTGGAGTTCAGACGTGTGCTCTTCCGATCTCCTCCTTTTCCAACTATAAACTA
TargetGFwd1	CTCTTTCCCTACACGACGCTCTTCCGATCTTATAGAGAGTGGTTTTTGGGGA
TargetGRev1	CTGGAGTTCAGACGTGTGCTCTTCCGATCTTCCCAAATACTAAAATAAAATATACA
TargetJFwd1	CTCTTTCCCTACACGACGCTCTTCCGATCTAGAGGGTAAGGTGGTTTTGGTT
TargetJRev1	CTGGAGTTCAGACGTGTGCTCTTCCGATCTCACAAATTAAAATTACCTCCAAAAA
TargetKFwd1	CTCTTTCCCTACACGACGCTCTTCCGATCTATGTGGAAATGGTGGGTTAGTT
TargetKRev1	CTGGAGTTCAGACGTGTGCTCTTCCGATCTACAAACTTTATCTACAAACAAAATCTTCAA
TargetLFwd1	CTCTTTCCCTACACGACGCTCTTCCGATCTGTGGGAAAAGGGATTGTATTGGAGTT
TargetLFwd2	CTGGAGTTCAGACGTGTGCTCTTCCGATCTCTCCATAACTAAAACCATATTTATCACAA

Table 2.9.: Primer sequences BBA-seq first PCR

Primername	**Sequence 5′- 3′**
NEXTflex Fwd	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
NEXTflex Rev	CAAGCAGAAGACGGCATACGAGATXXXXXXXXXXXXGTGACTGGAGTTCAGACGTGTG CTCTTCCGATCT

Table 2.10.: Primer sequences BBA-seq second PCR

3. Methods

3.1. General cell culture methods

In general for washing and harvesting, primary cells and cell lines were centrifuged using the centrifugation program: 8 min, 300 $\times$g, 4 °C.

3.1.1. Cell culture conditions and passaging

If not otherwise indicated, cells were cultured in high glucose DMEM (Gibco, Thermo Fisher Scientific, Waltham, USA) supplemented with 10% heat inactivated FCS, sodium pyruvate (1 mM), L-glutamine (2 mM), antibiotics (50 U/ml penicillin and 50 μg / ml streptomycin), 2 ml vitamins, non-essential amino acids and 50 μM β-mercaptoethanol. Media supplements were purchased from Gibco (Waltham, USA) and Merck Millipore (Darmstadt, Germany; L-glutamine). Each batch of FCS was tested and heat inactivated for 20 min at 56 °C prior to use. Cells were cultured at 37 °C, 5% CO_2 and with 95% relative humidity in an incubator. Cell cultures were split 1:2 to 1:4 in fresh medium every 2-3 days. Adherent cells were washed once with PBS and detached with 1 $\times$ Trypsin-EDTA solution incubated for 10 minutes and deactivated with 10 % heat inactivated FCS.

3.1.2. Assessing cell number and viability

Cell viability and concentration were assessed by means of Trypan blue exclusion. Cell suspensions were prepared and mixed with an equal volume of Trypan blue solution. The mixture was immediately transferred onto a Neubauer haemocytometer and the number of unstained (viable) and stained (non-viable) cells were counted under a light microscope. The concentration of viable cells was calculated using the following equation:

$$\text{Concentration of Viable Cells (cells/ml)} = \frac{\text{Number of Viable Cells} \times \text{Dilution Factor}}{\text{Volume of one square (in ml)}}$$

3.1.3. Freezing and thawing cells

Cells were collected before reaching confluence and then suspended in 1 ml of ice-cold freezing medium at a concentration of 5 - 10 $\times$ 10^6 cells/ml. Subsequently, the cells were placed into cryo tubes. To enable gradual freezing at a rate of 1 °C/min, the cryo tubes were put in isopropanol-filled cryo containers (Nalgene, Sigma-Aldrich, Taufkirchen, Germany) and maintained at -80 °C for 24 hours. Afterward, the tubes were moved to liquid nitrogen storage at -196 °C for long-term preservation.

3.1.4. Primary cultures

Mammary glands were obtained from either Balb-NeuT or BALB/c mice and placed in a 50 ml Falcon tube containing phosphate-buffered saline (PBS). The tissue was then minced into small pieces using surgical blades and digested in digestion medium (consisting of Basal Medium + Proliferation Supplement + 10 ng/ml EGF + 10 ng/ml bFGF + 4 μg/ml heparin, and 0.5 $\times$ penicillin/streptomycin supplemented with 200 U/ml collagenase, and 200 U/ml hyaluronidase). The tissue was incubated for 1.5 - 2 hours at 37 °C, with intermittent mixing every hour. The digested tissue was then

transferred to a fresh 50 ml tube. The organoids were isolated by centrifugation at 80 ×g for 1 minute, and the supernatant was collected into another 50 ml tube. The mammary epithelial cells were obtained by centrifuging of the supernatant at 350 × g for 5 minutes. Following this, the cell pellet was incubated in a combination of 1 unit cold HBSS containing 10 mM HEPES (without Phenol red) and 2% FBS, along with 4 units Ammonium Chloride solution. This mixture was then centrifuged at 350 ×g for 5 minutes to eliminate erythrocytes. The cell suspension was then diluted with an additional 10 ml of cold HF and filtered through a 37 μm reversible cell strainer into a new 50 ml Falcon tube. Then cells were centrifuged at 450 × g for 5 minutes and the supernatant discarded. The cell pellet was then either resuspended in Buffer R for subsequent electroporation or in EpiCult-B Medium for sphere generation.

3.1.5. Creating poly-HEMA and low-attachment plates

The coating solution was prepared by dissolving 2.4 g of poly-hydroxy-ethyl-methyl-acrylate (poly-HEMA) in 20 ml of 95% ethanol, which was then placed on a shaker at 65 °C for about eight hours. Once the poly-HEMA had fully dissolved, 80 ml of 95% ethanol was added to the solution and mixed thoroughly. This solution can be stored indefinitely at 4 °C.

For cell culture dishes, the poly-HEMA solution was added at a final density of 0.8 mg/cm (w/v). The plates were left to dry overnight in a sterile PCR hood. The poly-HEMA-coated plates were UV irradiated for 15 minutes and could be stored at 4 °C indefinitely until use.

3.1.6. Sphere culture

Isolated epithelial cells were subsequently re-suspended in EpiCult-B Medium (Basal Medium + Proliferation Supplement + 10 ng/ml EGF + 10 ng/ml bFGF + 4 μg/ml

heparin) and cultured at a density of 5×10^5 cells/ml on ultra-low adherent plates coated with 1.2% poly-HEMA. The sphere cultures were incubated at 37 °C with 5% CO_2 and 7% O_2 for 10 days, after which the size and number of mammospheres were evaluated under a light microscope and measured using Invitrogen Evos Cell Imaging Systems. A comparable sphere generation process was carried out for the TUBO cell line. Cultured cells underwent trypsinization and were dissociated into a single-cell suspension. Subsequently, the cells were re-suspended in EpiCult-B Medium and plated in poly-HEMA coated plates at suitable densities. After three to five days, the size and number of spheres were assessed using a light microscope and measured with the Invitrogen Evos Cell Imaging System.

3.1.7. Counting cell spheres in suspension culture

The quantification of spheres was carried out using a time-dependent approach, whereby spheres derived from primary cells were examined after 10 - 14 days, and spheres from TUBO cell line were evaluated after 3 - 5 days. Only spheres with a diameter larger than 50 μm were considered for analysis. The enumeration of spheres in cell lines was accomplished by visually screening the entire plate under 10 $\times$ magnification of a microscope. Spheres from primary mammary tissue were harvested and centrifuged at low speed, 200 $\times$g for 5 minutes. Following aspiration of the supernatant, the cells were resuspended in EpiCult-B medium to a final volume of 500 μl. A 96-well plate was used, and one well was marked with a plus sign on its underside using a pencil. This helped to create four quadrants in the well, which were used to keep track of the counted spheres. 50 μl of sphere suspension was added to the well, and the spheres were observed under a microscope at 10 $\times$ magnification. Spheres were counted in each quadrant, ensuring at least 50 spheres were counted for accuracy. The sphere concentration was calculated by dividing the count by the volume (50 μl), and total count was determined by multiplying concentration by the total volume (e.g., 500 μl).

3.1.8. Mycoplasma assay

The cell lines underwent regular screening for the presence of mycoplasma contamination, utilizing the MycoAlert Mycoplasma Detection Kit in accordance with the instructions provided by the manufacturer.

3.1.9. Transient transfection of mammalian cells

For CRISPR RNP transfection via electroporation into primary cells and TUBO cell line, the Neon Transfection System from Life Technologies Inc. was used following the manufacturer's instructions with 10 μl electroporation tips and two electroporation pulses of 1400 V for 20 ms.

3.2. Animal experiments

3.2.1. Maintenance

The Balb-NeuT transgenic mice used in this study were obtained in collaboration with AG Klein and were maintained in accordance with the European Union guidelines at our facilities. All animal experiments were conducted in compliance with EU and national institutional regulations, as stipulated by the Government of Upper Palatinate (55.2-2532.1-27/14).

3.2.2. Surgery

In this study, BALB/c mice served as models. Mice were anesthetized with Midazolam (5 mg/kg), Fentanyl (0.05 mg/kg), and Medetomidin (0.5 mg/kg) intraperitoneally before transplantation experiments. The thorax and abdomen were shaved, and a

midline incision was made to expose the surgical site. For sphere injection, 50 spheres (section 3.1.6) were mixed with matrigel at 40% concentration and injected into the fourth right mammary gland of each mouse. Skin closure was done using polygelatin sutures. To counteract anesthesia effects, Flumazenil (0.5 mg/kg), Atipamezol (2.5 mg/kg), and Naloxone (1.2 mg/kg) were administered subcutaneously. Postoperative pain was managed with Buprenorphine (0.1 mg/kg). If the tumor diameter was 5-10 mm, curative surgery or sacrifice was performed.

3.2.3. Mice dissection

Euthanasia of mice was performed by either CO_2 asphyxiation or cervical dislocation. The organs of interest, including mammary glands, tumors and lungs, were either snap-frozen or directly used for the generation of Dnmt3a knockout spheres. AG Klein oversaw the handling of this process.

3.3. Methods in molecular biology

3.3.1. DNA preparation and analysis

Genomic DNA (gDNA) was isolated using the Thermo Scientific Genomic DNA Purification Kit. Genomic DNA concentration was determined with the NanoDrop spectrophotometer or Qubit fluorometer and quality was observed by agarose gel electrophoresis.

3.3.2. Agarose gel electrophoresis

Agarose gels for electrophoresis were prepared following a standardized procedure. Initially, the appropriate amount of agarose was calculated based on Table 3.1 and

combined with the corresponding volume of 1 × TAE. The mixture was then heated in a microwave until the agarose fully dissolved. After cooling the solution to 50 – 60 °C, SYBR Safe DNA gel stain was introduced. Next, the gel was placed in the electrophoresis tank and covered with 1 × TAE. For the DNA samples, a 5:1 dilution with DNA loading dye (6 ×) was prepared, combined thoroughly, and loaded into the gel slots. The gels were operated at a voltage between 90 and 120 volts for a period of 45 minutes to 1.5 hours, depending on the size of the DNA fragments and the resolution required.

Efficient range of separation (kb)	**Percent of agarose in gel**
0.1 – 2	2.0
0.2 – 3	1.5
0.4 – 6	1.2
0.5 – 7	0.9
0.8 – 10	0.7
genomic DNA	0.5

Table 3.1.: Agarose concentration for different separation ranges.

TAE (50 ×)	253.3 g	(2 M)	Tris
	20.5 g	(250 mM)	NaOAc / HOAc, pH 7.8
	18.5 g	(50 mM)	EDTA
	Add ddH_2O to 1 l		
EDTA (0.5 m)	18.6 g	(50 mM)	EDTA / NaOH, pH 8.0
	Add H_2O to 100 ml		
1.0% Agarose	1 g	(1%)	Agarose (Biozym)

Table 3.2.: Buffers for gel electrophoresis

3.3.3. Quantification of DNA

In order to determine the precise DNA concentration, two methods were employed: the Qubit fluorometer (Thermo Scientific) and the NanoDrop spectrophotometer.

3.3.4. Polymerase Chain Reaction (PCR)

PCR is a method of amplifying DNA through sequence-specific polymerization of nucleotide triphosphates, which is catalyzed by DNA polymerase in an in vitro environment. The technique allows for the synthesis of large quantities of DNA with high specificity and accuracy. [129]. This approach involves "priming" the polymerization process with short oligonucleotides that bind to the template DNA strand via base pairing, thereby conferring specificity by defining the limits of the DNA fragment to be amplified. The sequences used for primer generation were extracted from the UCSC genome browser, unless otherwise stated. Primer design was carried out using the `SnapGene` Software, and PCR and BLAT functions of the UCSC genome browser were utilized to ensure primer quality. PCR amplifications were carried out using a Mastercycler Nexus M2 (Eppendorf, Hamburg, Germany), with 0.2 ml PCR strips and a reaction volume ranging from 10 - 50 μl. To prevent condensation, the lid of the PCR device was heated to 105 °C. The nucleotide sequences of the primers used in the PCR reactions are provided in Table 2.6. The primer annealing temperatures were set between 55 and 72 °C. For standard amplifications, the Phusion^{TM} Hot Start High-Fidelity DNA Polymerase (NEB, Massachusetts, USA) was utilized, using the standard reaction conditions listed in Table 3.3. Table 3.4 summarizes the general parameter settings for analytical PCR.

Component	Volume	Final concentration
5 × Phusion HF buffer	10 μl	1 ×
10 mM dNTPs	1 μl	200 μM each
Primer S	1 μl	0.2 μM
Primer AS	1 μl	0.2 μM
Template DNA	X μl	
Phusion HF Polymerase (2 U/ml)	0.5 μl	0.02 U/μl
H_2O	Add 50 μl	

Table 3.3.: Standard PCR reaction.

Step	Temperature	Time
Initial Denaturation	98 °C	30 seconds
25 - 35 Cycles	98 °C	5 - 10 seconds
	45 - 72 °C	10 - 30 seconds
	72 °C	15 - 30 seconds per kb
Final Extension	72 °C	5-10 minutes
Hold	4 - 10 °C	

Table 3.4.: Thermocycling conditions for a routine PCR.

3.3.5. RNA isolation

The extraction of total cellular RNA was carried out using the Qiagen RNeasy Mini (Hilden, Germany) in accordance with the manufacturer's instructions. A DNase treatment was included in all RNA isolations. The concentration of the purified total RNA was ascertained using the qubit fluorometer (Invitrogen), while the quality of the RNA was assessed in line with the manufacturer's instructions using the Agilent TapeStation (Boeblingen, Germany).

3.3.6. Reverse transcription PCR (RT-PCR)

In order to measure mRNA transcripts and gene expression, reverse transcription of isolated RNA was performed. This was accomplished with MMLV reverse transcriptase (Promega, Germany) alongside random decamers (Ambion, Germany) in a total reaction volume of 20 μl (Table 3.5). Subsequently, the produced cDNA was quantified using specific primers through real-time PCR.

1 μg	Total RNA
1 μl	Randon decamers
1 μl	dNTPs (10 pmol/ml)
Add 20 μl H_2O	
Incubate for 5 min at 65 °C, cool on ice and centrifuge	
4 μl	M-MLV Buffer (5x)
Mix and incubate for 2 min at 42 °C	
1 μl	M-MLV Reverse transcriptase
Incubate for 50 min at 42 °C followed by 15 min at 70 °C	

Table 3.5.: Reaction setup for reverse transcription.

3.3.7. Quantitative real time PCR (qPCR)

qPCR technique was employed to detect and quantify a specific DNA sequence in a given sample. This method enables absolute quantification of the target sequence, as well as relative quantification after normalization to DNA input or normalizing genes. qPCR follows the fundamental principle of polymerase chain reaction, and its unique attribute is the real-time quantification of amplified DNA as it accumulates after each amplification cycle. The method was employed for quantification of cDNA after

reverse transcription for expression kinetics and to quantitatively analyze PCR products after BBA-seq library generation. PCR reactions were conducted using the Quantifast SYBR Green Kit from Qiagen or the KAPA library quantification Kit (F. Hoffmann-La Roche Ltd, Basel, Switzerland) in a 96-well format on the CFX Connect Real-Time PCR Detection System (Bio-Rad Laboratories, California, USA). The SYBR green dye, which intercalates in double-stranded DNA, emits light that is used to determine the relative amount of amplified DNA. In order to determine the amplification efficiency for each primer pair, a standard curve was created by amplifying four different dilutions (1:10, 1:50, 1:100, 1:1000). Realplex software was utilized to calculate the DNA amounts based on the generated slope and intercept. To confirm specific amplification, melting-curve analysis was performed. Data was then imported and processed using a spreadsheet. To analyze mRNA expression, all samples were measured in triplicates and normalized to the β-actin housekeeping gene. For the BBA-seq library samples, triplicates were quantified using a DNA quantity standard obtained from F. Hoffmann-La Roche Ltd., Basel, Switzerland.

3.4. Protein biochemical methods

3.4.1. Preparation of whole cell lysates

Cells were harvested and washed twice with ice-cold PBS. Following the washes, the supernatant was completely removed, and the cell pellet was resuspended in 50 μl of RIPA buffer. The samples were incubated on ice for 20 minutes before being centrifuged at full speed for 15 minutes at 4 °C. The supernatant containing the protein was then transferred to a new tube, and concentration was determined using a BCA assay as per the manufacturer's instructions. Lysates were stored at -20 °C for later SDS-PAGE and Western Blot analysis. Adherent cells were rinsed with cold PBS right on the plate. Then, they were scraped into RIPA buffer and transferred to tubes.

3.4.2. SDS-PAGE (sodium dodecyl sulfate polyacrylamide gel electrophoresis)

Protein samples were separated using a discontinuous gel system consisting of stacking and separating gel layers differing in salt and acrylamide (AA) concentrations. The day before electrophoresis, the separating gel was prepared and covered with water-saturated isobutanol until polymerization occurred. Next, the stacking gel was poured on top of the separating gel, and a comb was quickly inserted. Once the gel polymerized, it was placed in the electrophoresis tank, filled with 1 $\times$ Laemmli buffer. The protein samples were subsequently added, and the gel ran at 80 V until the sample buffer bands reached the stacking gel's surface. The voltage was then raised to 110 V, allowing the gel to run for 1 to 3 hours. The proteins were separated in the gel according to their respective sizes.

3.4.3. Analysis through Western blot and immunostaining

After separation using SDS-PAGE, proteins were transferred to a polyvinylidene fluoride (PVDF) membrane via a three-buffer semi-dry method [130]. Specific antibodies and the ECL Prime Western Blotting System were used for immunostaining to visualize the proteins. The membrane was trimmed to the size of the gel, pre-wetted with isopropanol and buffer B, and carefully stacked on top of three Whatman 3MM filter papers soaked in buffer A at the bottom, adjacent to the anode. Another layer of three Whatman 3MM filter papers soaked in buffer B was then added. The SDS-PAGE gel was submerged in buffer B and placed on the membrane. To ensure efficient transfer, three additional Whatman 3MM filter papers soaked in buffer C were added over the gel, followed by connecting the cathode. Attention was given to prevent air bubbles between the layers. The protein transfer was carried out for 60 minutes at a current density of 0.8 mA/cm^2 of the gel's surface area. The blotted membranes underwent a thorough wash, starting with water for 2-3 minutes, followed by blocking in 3% non-fat milk in

TBS for 30 minutes at room temperature. After that, the membranes were washed once for 10 minutes with TBS and incubated with the primary antibody for 30 minutes at room temperature. Following this, they were washed again for 10 minutes with TBS and incubated with a horseradish-peroxidase (HRP)-coupled secondary antibody that detects the isotype of the primary antibody for another 30 minutes at room temperature. Eight washing cycles with TBST were carried out to visualize the bound antibodies, totaling 20 minutes, before using the ECL Prime Western Blotting System. The bound antibody was visualized using the Fusion Pulse imaging system from Vilber Lourmat, with imaging times ranging from 5 seconds to 10 minutes based on signal intensity. For re-blotting, the membranes were washed once with TBS, followed by application of the ReBlot Plus Mild Antibody Stripping Solution, which was incubated with the blots for 15 minutes at room temperature. The stripped blots were then washed once with TBS, and the blocking and staining steps were repeated as described prior.

3.4.4. Delivery of ribonucleoprotein complexes into Balb-NeuT mammary gland cells

For the geneartion of CRISPR RNP complexes the protocol from IDT was used following the manufacturer's instructions. In brief, guide and tracer RNA were resuspended to a final concentration of 200 μM in IDTE buffer. Equimolar concentrations of guide and tracer RNAs were mixed to a final concentration of 44 μM. The combined RNAs were heated for 5 minutes in a thermo cycler and cooled to room temperature. Afterwards the duplex RNA was diluted 1:5 in Neon Resuspension Buffer R and mixed with a prediluted (36 μM) Cas9 nickase D10A in equimolar ratio. After a 10 minute incubation the CRISPR RNP complex was ready to use for subsequent electropoartion.

3.4.5. Analysis of Cas9 gene editing with T7 endonuclease assay

First, the PCR reaction was set up using ≈ 100 ng of targeted and negative control gDNA, c.f.3.4. Afterwards the PCR product was cleaned up with NEB Monarch PCR & DNA Clean Up Kit following manufacturer's instructions. Next, 200 ng of the PCR product was digested with T7 endonuclease. After 15 min the reaction was stopped with 1.5 μl 0.25 M EDTA and cleaned-up again. The product was then analyzed on a gel imager. Gene modification was calculated with the following formula according to [131]:

$$r_{\text{mod}}[\%] = 100 \times (1 - \sqrt{1 - r_{\text{cleaved}}})$$

3.5. Analysis of DNA methylation

3.5.1. Bisulfite conversion

Genomic DNA was subjected to bisulfite treatment using the EZ DNA Methylation-Lightning Kit from Zymo Research Corporation (California, USA). This kit utilizes sodium bisulfite to convert cytosine to uracil. The manufacturer's protocol was followed to carry out the conversion reaction, whith very low input of genomic DNA.

3.5.2. Primer desing for BBA-seq approach

The genomic DNA sequences were obtained from the University of California, Santa Cruz genome browser (http://www.genome.ucsc.edu). To increase the coverage of the target region, both the forward and reverse strands were included in the amplicon design process. The selected genomic sequence was exported to `MethPrimer` (http://www.urogene.org/methprimer), a primer design software, for further analysis. An in silico bisulfite conversion was conducted within `MethPrimer` to facilitate primer

design. The PCR primer design criteria consisted of the following specifications: optimal primer melting temperature of 50 °C, primer length ranging from 18 - 25 nucleotides (excluding adaptors), and amplicon length varying from 100 - 400 bp, with a desired length of 300 bp. After successfull primer design, handle sequences for the subsequent barcoding step were added to the primer sequences.

3.5.3. PCR amplification

Target sequences with candidate CpG sites were amplified using EpiMark Hot Start Taq DNA Polymerase, ideal for use on bisulfite-converted DNA, in a first PCR step. PCR amplifications were carried out using a Mastercycler Nexus M2 (Eppendorf, Hamburg, Germany), with 0.2 ml PCR strips and a reaction volume of 25 μl. To prevent condensation, the lid of the PCR device was heated to 105 °C. The nucleotide sequences of the primers used in the PCR reactions are provided in Table 3.6. PCR conditions and components of the first DNA are depicted in Table 3.6 and 3.7. The amplicons of each

Step	**Temperature**	**Time**
Initial Denaturation	95 °C	15 minutes
45 Cycles	94 °C	30 seconds
	55 °C	30 seconds
	72 °C	30 seconds
Final Extension	72 °C	10 minutes
Hold	4-10 °C	

Table 3.6.: Thermocycling conditions for first BBA-seq PCR.

sample were pooled at equal concentrations, quantified with Qubit (Invitrogen), and cleaned up with paramagnetic beads from Agencourt AMPure PCR Purification system in a ratio between sample and beads of 1:0.7 according to manufacturer's instructions

Component	Volume	Final concentration
5 × EpiMark® Hot Start Taq Reaction Buffer	5 μl	1 ×
10 mM dNTPs	0.5 μl	200 μM each
10 μM Forward Primer	0.5 μl	0.2 μM
10 μM Reverse Primer	0.5 μl	0.2 μM
Template DNA	× μl	
EpiMark® Hot Start Taq DNA Polymerase	0.125 μl	0.6 units/25 μl PCR
H_2O	Add 25 μl	

Table 3.7.: Reaction setup for first BBA-seq PCR.

(Beckman Coulter). 36 μl of PCR products were subsequently added to 12 μl NEXTflex PCR Master Mix and 2 μl containing 10 pmol of barcoded primers (NEXTflex 16S V1-V3 Amplicon Seq Kit, Bioo Scientific, Austin, USA) for a second round of PCR (3.8). PCR products were quantified by Qubit (Invitrogen), pooled in equimolar ratios, and

Step	Temperature	Time
Initial Denaturation	98 °C	4 minutes
12 Cycles	98 °C	30 seconds
	60 °C	30 seconds
	72 °C	30 seconds
Final Extension	72 °C	4 minutes
Hold	4-10 °C	

Table 3.8.: Thermocycling conditions for subsequent BBA-seq PCR.

cleaned up with magnetic beads from Agencourt AMPure PCR Purification system in a ratio between sample and beads of 1:1 according to manufacturer's instructions (Beckman Coulter). Ten picomolar DNA library was then spiked-in with 15% PhiX control and subjected to 250 bp paired-end sequencing on a MiSeq lane (Illumina, CA,

USA) using MiSeq reagent V2 kit (Illumina, CA, USA).

3.5.4. Measurement of mean methylation values

Mean methylation values were calculated with the `getMethylationStats` function from **R**-package `methylKit` for each sample and respective region [132]. The function takes `*.bam` files of each sample per region into account and reprints the specific methylation statistics per base, namely "Minimum", "1st Quantil", "Median", "Mean", "3rd Quantil" and the "Maximum". The values for "Mean" were inserted into a table, and a heatmap was generated out of this matrix.

3.5.5. Measurement of bisulfite conversion rate

Bisulfite conversion is a widely used technique for detecting and quantifying DNA methylation at single-base resolution. The method relies on the ability of sodium bisulfite to selectively deaminate unmethylated cytosines to uracils while methylated (or hydroxymethylated) cytosines are not converted. This process enables the distinction between methylated and unmethylated cytosines in genomic DNA. The process involves the initial denaturation of the DNA, followed by treatment with sodium bisulfite, which can significantly alter the chemical and physical properties of the DNA [133].

Bisulfite conversion efficiency must be carefully evaluated to ensure accurate analysis and false positive errors must be excluded. This can be achieved by using a lambda phage unmethylated DNA spike-in control or by calculating the conversion rate as the number of non-methylated cytosines divided by the coverage for each non-CpG cytosine. A coverage of at least ten times and a phred score smaller than 20 are required for a cytosine to be included in the calculation.

In this study, the average conversion rate was calculated using the second approach, and an average conversion rate score (ACR) was determined using the `processBismarkAln` function from the **R**-package `methylKit`[132].

3.5.6. From Bismark Methylation extractor to aligned `*.fasta` file

The BBA-seq workflow entails several crucial steps to analyze every read processed in an aligned `*.fasta` file at single base cytosine resolution. In order to compare reads from different samples, it is necessary to ensure that they have identical lengths, an appropriate distribution, and adequate sequencing depth to cover all alleles of the bulk samples from which they were derived. After alignment with the relevant reference genome, the `bismark_methylation_extractor` script is used to extract the methylation call for every C analyzed (3.1). The position of each cytosine is then written out to a new output, `*.txt`-file, depending on its context (e.g., CpG). Methylated Cs are labeled as forward reads (+) in the methylation state column or upper 'Z' in the methylation call column, while non-methylated Cs are labeled as reverse reads (-) in the methylation state column or lower 'z' in the methylation call column.

The methylation extractor output files (File Exerpt 3.1) contained information on every methylated cytosine from every processed read at single base resolution and are subsequently used to construct an aligned `*.fasta` file for comparative analysis. The **R**-package `tidyverse`, particularly its core-package `dplyr`, was utilized to transform the existing dataset into a `*.fasta` format best suited for phyloepigenetic analysis [134]. Code snippet A.1 shows the **R**-script. In particular, the `*.txt`-file is first loaded with the `read_tsv` function, and various other functions are then applied to generate a `tempStr`-element, which retrieves the methylation information from the methylation call column as well as the read length from the seqID's start- and end-positions information for each read. A `fastaStr`-element is also created, which contains the greater-than character (>) followed by a description of the sequence.

The file is then further processed with the package's piping (%>%), whereby a chain of processing functions is applied to obtain the tab-separated information of each read into a one-line sequence that is found in standard `*.fasta` files. The file is first grouped by its readID, and the methylation information is mutated to a single line. These two pieces of information are then selected to create a new file with the ending `*.fasta` (File Excerpt 3.2).

In a subsequent **R**-script (Code snippet A.2), the newly generated `*.fasta` file is further processed to meet the demands required for analysis. Specifically, the methylation-state information at cytosines is isolated, and each non-cytosine base (" - ") is excluded. Additionally, the 'Z' indicating a positive methylated cytosine is replaced by a 'C', while every 'z' denoting an unmethylated cytosine is replaced by a 'T'. Furthermore, the seqID information is modified to the form `>sample_region_sequence`, and two functions generate the final two output `*.fasta` files (File Excerpts 3.3 and 3.4). The first file contains cytosine information for each individual read, which is necessary for sequence logo generation, consensus sequence analysis, and the phyloepigenetic trees for regions B, D, E, F, and G. Subtracting low-methylated reads in these regions resulted in a reduction of total reads, allowing each individual read to be represented. The second file merges identical reads to facilitate the generation of phyloepigenetic trees for regions C, J, K, and L for the reason of condensation.

```
Bismark methylation extractor version v0.23.1
M00563:201:000000000-KMMB6:1:2103:18989:13865_1:N:0:13   +  chr1  3670901  Z
M00563:201:000000000-KMMB6:1:2103:18989:13865_1:N:0:13   -  chr1  3670930  z
M00563:201:000000000-KMMB6:1:2103:18989:13865_1:N:0:13   -  chr1  3670939  z
M00563:201:000000000-KMMB6:1:2103:18989:13865_1:N:0:13   -  chr1  3670956  z
```

File Excerpt 3.1: Bismark Methylation Extractor `*.txt` output file. The Bismark Methylation Extractor output file displays lines in a tabular format with five columns. The first column presents the read information, while the second and fifth columns provide the methylation information. The third and fourth columns indicate the location of CpG's in the sequence.

```
>M00563:201:000000000-KMMB6:1:1101:15277:5347_1:N:0:13 startpos=3671821 chromosome=chr1
z--z----------z-------------z-------------z-------------------------z--z------
------------------------z-z--------------z------z------------------z---------z
------z----------------z--------------------------z-z--z-z
>M00563:201:000000000-KMMB6:1:1101:14622:5820_1:N:0:13 startpos=3671821 chromosome=chr1
z--z----------z-------------z-------------z-------------------------z--z------
------------------------z-z--------------z------z------------------z---------z
------z----------------Z--------------------------z-z--z-z
```

File Excerpt 3.2: Intermediate `*.fasta` file. Example lines from the intermediate `*.fasta` file produced by the initial conversion script exhibits the usual characteristics of a standard `*.fasta` file format.

```
>PT3145_regionE_CCTTTTTTTTCTCCTTTTT-1
CCTTTTTTTTCTCCTTTTT
>PT3145_regionE_CCTTTTTTTTCTCCTTTTT-2
CCTTTTTTTTCTCCTTTTT
>PT3145_regionE_CCTTTTTTTTCTTTTTTTT-1
CCTTTTTTTTCTTTTTTTT
>PT3145_regionE_CCTTTTTTTTTTCTTTTTT-1
```

File Excerpt 3.3: Final `*.fasta` all file. The second conversion script generates one `*.fasta` file that presents all sequencing reads files for conducting consensus sequence logo analysis and generating phyloepigenetic trees for regions B, D, E, F, and G.

```
>PT3145_regionE_CCTTTTTTTTCTCCTTTTT-2
CCTTTTTTTTCTCCTTTTT
>PT3145_regionE_CCTTTTTTTTCTTTTTTTT-1
CCTTTTTTTTCTTTTTTTT
>PT3145_regionE_CCTTTTTTTTTTCTTTTTT-1
CCTTTTTTTTTTCTTTTTT
>PT3145_regionE_CTTTTTTTTCCCCTTTTCC-4
CTTTTTTTTCCCCTTTTCC
```

File Excerpt 3.4: Final `*.fasta` count file. The second conversion script creates a second `*.fasta` file by merging identical sequencing reads and annotating them with the corresponding occurrence count. These files were used to generate phyloepigenetic trees for samples in regions C, J, K, and L.

4. Results I: Phyloepigenetic analysis of metastatic breast tumor progression

4.1. Target regions determination

During the previous funding period of the FOR2127 project, the DNA methylation status of primary tumors and their corresponding metastases was investigated using methyl-CpG immunoprecipitation (MCIp) combined with subsequent sequencing (-seq) data from Balb-NeuT mammary gland tissue transplanted into Balb/C mice. The results revealed a notable increase in overall DNA methylation levels as the tumor progressed from its initial primary site to fully developed lung metastatic lesions, representing diverse methylation patterns [135]. However, the underlying mechanisms behind this observation and whether those changes serve as an advantage for early dissemination are not fully understood yet. One hypothesis suggests that the DNA methylation changes may result from tumor cell adaptation to varying selective pressures at the metastatic site. Alternatively, it may arise from the selection of sub-clones with higher fitness during the various stages of systemic progression or as a consequence of cell aging and increased proliferation. To address these possibilities and to elucidate the associated methylation patterns resulting from different mechanisms, a barcoded bisulfite amplicon sequencing (BBA-seq) approach was implemented to analyze cytosine methylation at single base resolution.

The main research questions and dataset for this thesis originate from the application of MCIp-seq to affinity-enriched methylated mouse DNA. These laboratory procedures were conducted by Sophie Gaertner [135]. MCIp is a technique employed to selectively isolate and enrich DNA segments harboring methylated cytosine residues. It utilizes the fusion of the human Ig1 Fc portion to the protein of the CpG-binding domain of MBD2, which facilitates the capture of CpG methylation sites within double-stranded DNA. Following immunoprecipitation, the enriched fragments of methylated DNA were eluted and fractionated based on varying degrees of methylation using increasing salt concentrations in the elution buffer. Subsequently, these fractions can be further analyzed using diverse downstream methodologies [136]. In order to acquire fully methylated reference samples, healthy mammary gland tissue DNA underwent treatment with the procaryotic DNA(cytosine-5)methyltransferase M.SssI before DNA extraction. Additionally, there were two control samples each for breast and lung tissue, serving as reference points for the biological context surrounding primary tumors and metastases. The dataset comprised 80 primary tumor and metastasis samples, some of which were paired. The MCIp-seq data was aligned to the GRCm38 (mm10) mouse reference genome and preprocessed using `Baymeth` from the **R**-package `Repitools` [137]. As a result, each sample and 100 bp window within the genome provided a relative methylation estimate, represented by values ranging from 0 to 1. Moreover, upper and lower credible intervals and variance information were provided, reflecting the precision of the estimate. This enabled the comprehensive examination of DNA methylation in terms of counts per region and in a highly localized manner. The analysis has unveiled prominent recurring motifs, intriguing methylation patterns, and genomic regions examined in this thesis. Figure 4.1 shows specific basic patterns that occurred repeatedly across the whole genome. However, MCIp possesses a lower resolution compared to targeted bisulfite sequencing methods. It can identify methylated regions but may not identify the exact methylated cytosines within those regions. To overcome this disadvantage and to characterize methylated CpGs at base resolution in an allele-specific manner, to analyze the exact methylation behavior of primary tumor

reads, and to identify possibly existing small high methylated sub-clones, two targeted bisulfite sequencing approaches were utilized in this thesis.

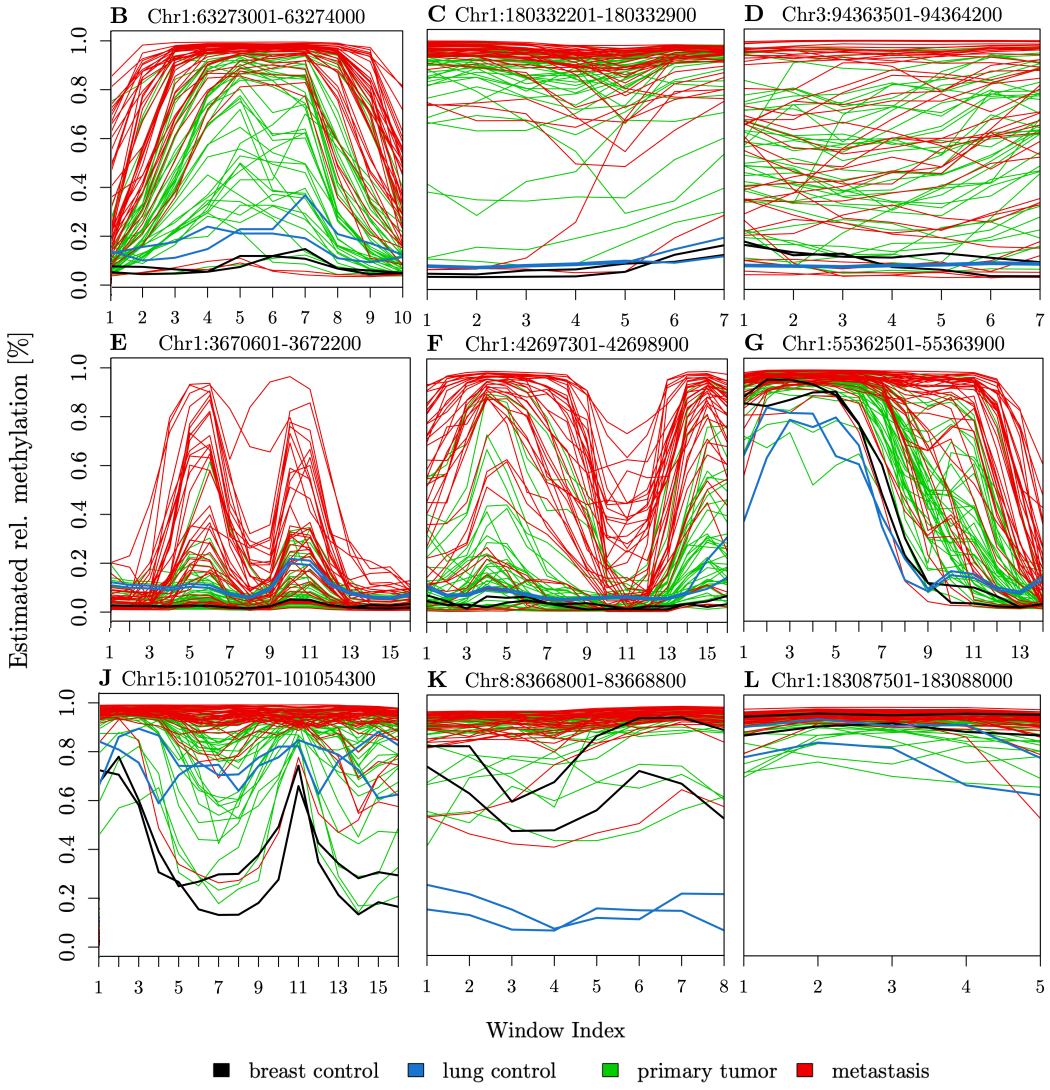


Figure 4.1.: Regions with exemplary methylation types and patterns that served as targets for the BBA-seq approach implemented in this thesis. Data was generated using MCIp and binned into windows using Baymeth implementation. Figure adapted from [138].

4.2. Identification of DNA methylation changes with bisulfite padlock probes

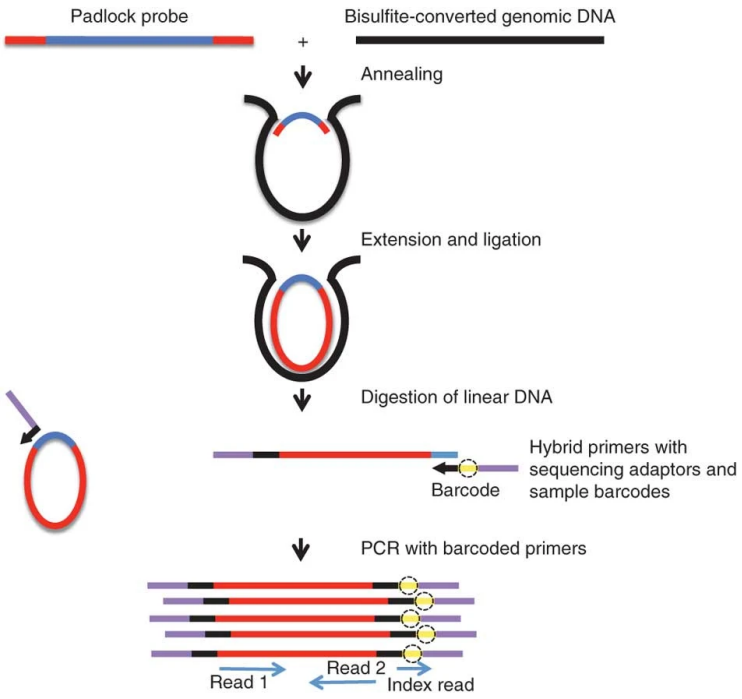


Figure 4.2.: Schematic overview of the targeted, library-free bisulfite padlock probes approach. Figure taken from [139].

A significant challenge that had to be overcome during the experiments conducted for this thesis was the limited quantity of initial DNA material available from tumor and metastasis samples from the first funding period. As the comprehensive methodology relied on the results of the MCIp analysis, the subsequent experiments required the use of the same sample sets analyzed in funding period 1. Additionally, the animal experiments were complex and laborious, so there was a need to use this DNA. Hence, there was a requirement for a technique that could target specific genomic regions, with the added ability to capture multiple targets simultaneously while also yielding efficient results even with a small amount of input DNA. The method of choice was the capture of genomic targets for single-molecule bisulfite sequencing with bisulfite

padlock probes. In 2009, Deng et al. published a method that utilizes sets of forward and reverse primers connected by a shared linker oligonucleotide sequence of 58 bases (also known as "padlocks") [139]. Figure 4.2 shows the main steps characterizing this technique. The padlock probe consists of a common linker sequence among all probes and two target-specific capturing arms (colored in red) that bind to bisulfite-converted genomic DNA. Once the ligation process is completed, the Stoffel fragment from *Thermus aquaticus* polymerase extends the DNA locus from the 3′ to 5′ end and fills the gaps. This crucial step is responsible for enclosing the circular DNA. Once linear DNA has been removed with exonuclease I and II, the captured targets that have been circularized are amplified using barcoded primers and then sequenced directly using an Illumina MiSeq sequencer. This method has some advantages over PCR-based targeted techniques, which are highly specific but cannot be used for genome-wide assays. In addition, the padlock probes have a strong ability to capture specific genetic regions, allowing for single-tube reactions that are appropriate for processing small amounts of DNA input. The padlock probes were generated using a specialized primer design algorithm known as `ppDesigner` [140]. This software creates the best possible group of padlock probes that will cover any given set of unique genomic targets. When selecting candidate probes, the algorithm takes into account various sequence features such as melting temperature, size, and word statistics (which refer to the distribution of 12-mers in the bisulfite-converted genome) of the capturing arms and gap sizes. These features are used to determine the weight of each probe. However, designing padlock probes for targets located in CpG-rich regions proved to be challenging. The algorithm used to generate the capturing arms avoided CpG sites since bisulfite conversion causes methylated cytosines to become thymines, leading to improper primer generation. This problem was solved by making certain adjustments to the `ppDesigner` parameters, which changed the length of the final PCR product and reduced the default value range of the melting temperature. To ensure comprehensive coverage of the identified regions, 91 padlock probes were finally generated. The approach was carefully designed to guarantee maximum efficacy and accuracy.

Nevertheless, it was not possible to create appropriate bisulfite padlock probe libraries that could be sequenced after successful primer generation. By making significant changes to the protocol, such as minimizing the temperature intervals during the ligation process of the padlock probes to the DNA template, adding an oil droplet to prevent evaporation, performing a PA-gel for library clean-up, and increasing the number of PCR cycles to generate more product, it became feasible to produce libraries with the appropriate size (Figure 4.3A). However, the sequencing results from these libraries were unsatisfactory. It was discovered that the libraries had a significant quantity of adapter sequences, even though they were properly cleaned up. The `FastQC` [141] analysis showed a noticeable decrease in the phred score quality after 100 bases, indicating poor quality of the padlock libraries (Figure 4.3B). Additionally, the libraries were enriched and contaminated with sequences that were blasted on the genome of *Spisula solida*, which is a type of common surf clam (Figure 4.3C). In conclusion, a new approach was needed to analyze abnormal methylation in the development and progression of metastasis in breast cancer.

4.3. Identification of DNA methylation changes with BBA-seq

BBA-seq was chosen to investigate allele-specific changes in DNA methylation from primary tumor to metastatic lesions, as an alternative method to the unsatisfactory capturing of target regions with bisulfite padlock probes. BBA-seq enables the simultaneous processing of multiple DNA sequences on a single lane, accommodating relatively long amplicons of up to 500 bases, and providing extensive coverage of over 1000-fold [142]. Having a high coverage was crucial for achieving single-base resolution in the samples. To amplify the genomic locus of interest and add Illumina sequencing adapters, a two-step PCR approach was used. In the first step, locus-specific, barcoded primers were employed, and then handle sequences were added in the second PCR. Following bead clean-up and pooling (Figure 4.4A), the library was sequenced on an Illumina

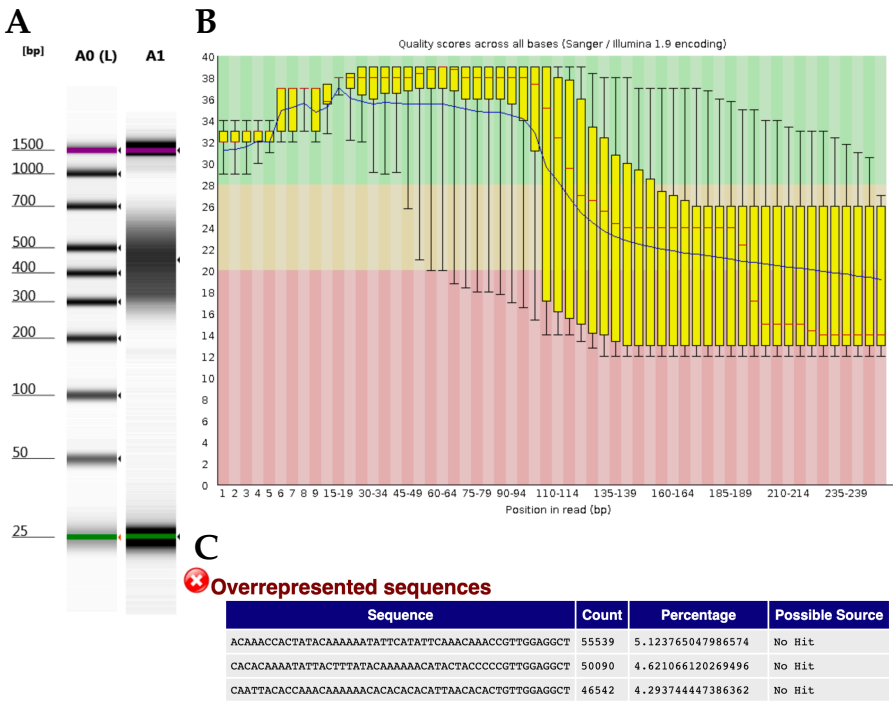


Figure 4.3.: Bisulfite padlock probes approach achieved unsatisfying results. (A) Tape station lane A1 shows padlock library with the correct size. (B) `FastQC` analysis shows drop of phred quality score of sequenced library after 100 bases. (C) Processed library contained a high amount of overrepresented sequences blasted onto genome of common surf clam, *Spisula solida*.

MiSeq sequencer (Detailed in 3.5). In comparison to the sequenced library following the padlock probes protocol, the BBA-seq libraries exhibited much higher quality. The phred score remained consistently above 28 throughout the reads until >300 bp (Figure 4.4B). In summary, the BBA-seq approach was found to be the most suitable method for analyzing the provided samples at single-base resolution. A total of 7 primary tumors, 11 matching metastases, and 8 control samples (5 breast and 3 lung) were analyzed at 9 different loci (Figure 4.4C). Table 4.1 summarizes the locations of the PCR products and the location within the corresponding CpG island and provides further information about CpG coverage, PCR product length, and the genomic context. Unfortunately, the PCR product covering region K contained only 4 CpGs that could be analyzed, so this region was excluded from detailed phyloepigenetic analysis.

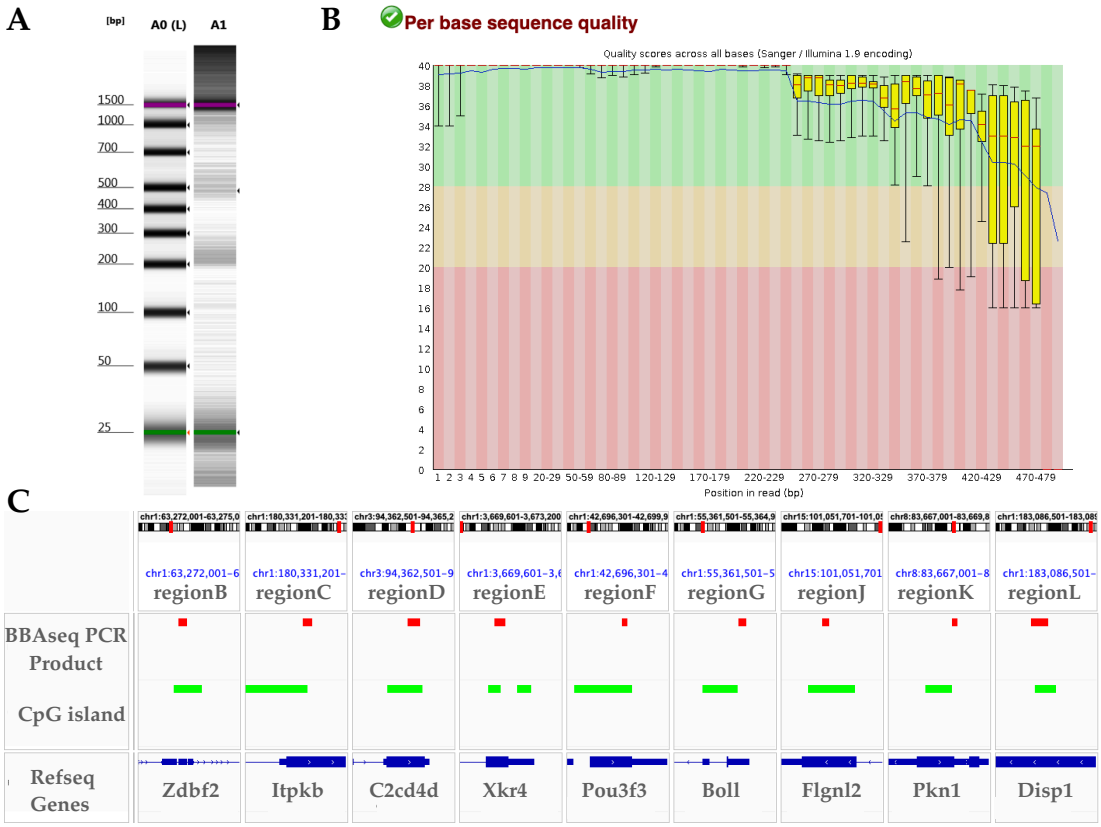


Figure 4.4.: DNA methylation analysis with BBA-seq. (A) Tape station lane A1 shows pooled BBA-seq library with the correct size. (B) `FastQC` analysis represents high-quality BBA-seq library with phred scores above 28. (C) Overview of BBA-seq PCR product locations.

4.4. Data processing

One of the main objectives of this thesis was to accurately map DNA methylation at the level of individual bases. This was done to detect any possibly existing rare sub-groups with high levels of methylation within primary tumor samples and to conduct a detailed analysis of abnormal DNA methylation patterns in the progression of metastatic breast cancer. For this purpose, a new and original process was developed which entails the analysis of DNA methylation at single-base cytosine resolution. Figure 4.5 represents a schematic representation of the BBA-seq approach together with the phyloepigenetic visualization and analysis. First, the PCR amplicons were sequenced

	CpG island location	Gene	Genetic feature	Total CpGs	PCR product location	Length (bp)	CpGs captured
Region B	Chr1:63273001-63274000	Zdbf2	Promoter	90	Chr1:63273215-63273466	253	40
Region C	Chr1:180332201-180332900	Itpkb	Exon	157	Chr1:180332746-180332994	250	13
Region D	Chr3:94363501-94364200	C2cd4d	Exon	112	Chr3:94363980-94364323	345	37
Region E	Chr1:3670601-3672200	Xkr4	Promoter	79	Chr1:3670863-3671256	282	19
Region F	Chr1:42697301-42698900	Pou3f3	Exon	158	Chr1:42698279-42698481	203	13
Region G	Chr1:55362501-55363900	Boll	Promoter	115	Chr1:55363697-55363946	251	12
Region J	Chr15:101052701-101054300	Fignl2	Exon	196	Chr15:101053177-101053430	254	24
Region K	Chr8:83668001-83668800	Pkn1	Exon	74	Chr8:83668779-83668937	159	4
Region L	Chr1:183087501-183088000	Disp1	Exon	39	Chr1:183087393-183087817	378	25

Table 4.1.: Overview of the locations of the CpG islands and characteristics (Name of gene, feature at location, total CpGs in CpG island, location of PCR product, length of the PCR product, and amount of CpGs caputered in PCR product) of PCR products generated to capture specific regions using the BBA-seq method for resolving DNA methylation at base resolution.

using an Illumina MiSeq next-generation sequencer. The raw sequence data coming from the high throughput sequencer were obtained as `*.fastq` files and quality checked with `FastQC` software [141]. After successful inspection, Illumina adapters were trimmed from the sequencing reads using `Trim Galore!`, which is a wrapper script for read filtering and further quality control [143]. To this end, the analysis contained just reads with an acceptable phred score for further investigations. As the forward and reverse reads from the paired-end sequencing approach would form two distinct clusters in the afterward phyloepigenetic analysis, both reads were combined into a single read using `PEAR` software, which merges raw Illumina paired-end reads from target fragments by overlapping the forward with its corresponding reverse read [144]. The quality-checked, trimmed, and paired reads were then mapped to the corresponding mm10 mouse genome using the `Bismark` alignment software [145]. The acquired `*.bam` files were further processed using `SAMtools` [146]. In brief, the `*.bam` files were sorted and indexed according to their genomic location, and the reads for each genomic region were obtained using `SAMtools sort`, `view`, and `index` commands. The analysis included a total of 48,859,019 methylated C's in the CpG context, which required the alignment output files to be filtered on the number of bisulfite conversion-related non-bisulfite mismatches. This was accomplished using the `filter_non_conversion` option in the

`Bismark` alignment software, which examined the methylation calls of reads and filtered out reads that exceeded a certain threshold (cut-off) [145]. The filtering resulted in the removal of 14,141 sequencing reads, representing an average of 1.02% filtered reads from each sample. Afterward, the files were handled as described in 3.5.6 and 4.7.

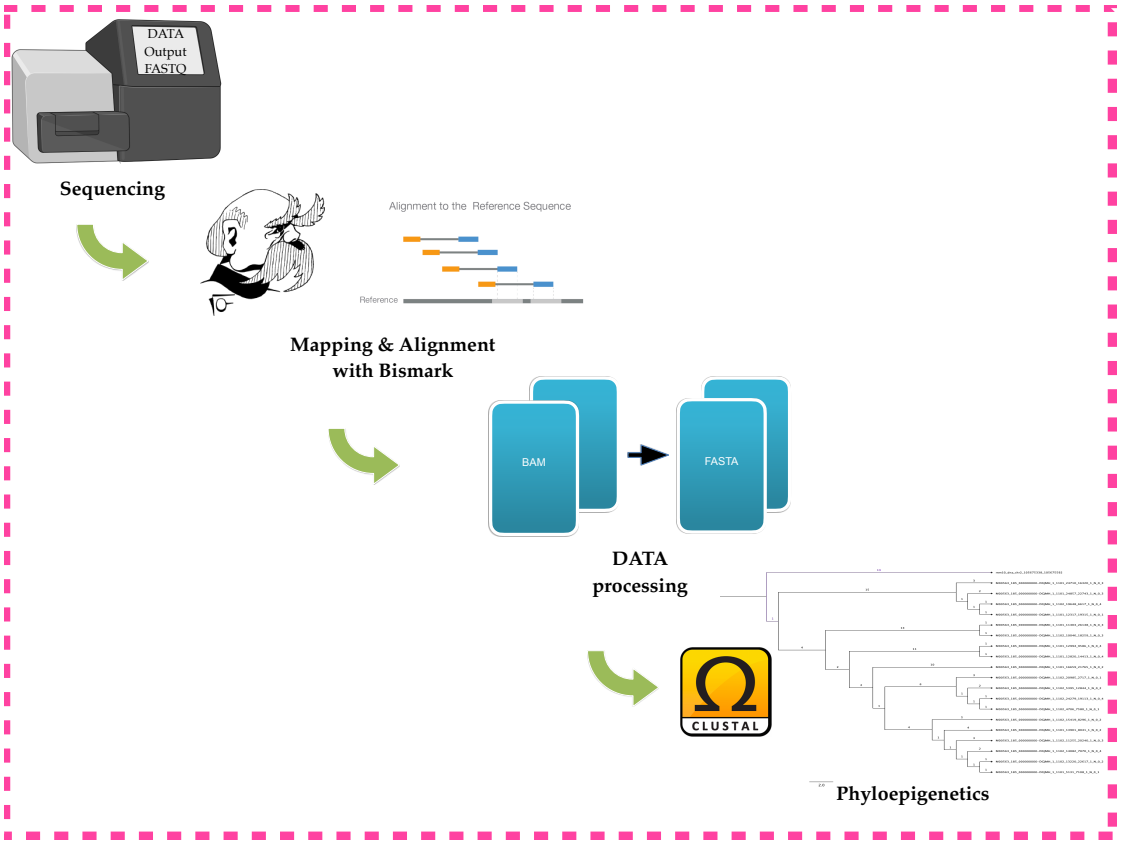


Figure 4.5.: Schematic overview of the BBA-seq and phyloepigenetic workflow implemented to identify, possibly existing, highly methylated sub-clones from bulk DNA primary tumor samples.

4.5. Isolating tumor-specific DNA methylation signatures by eliminating non-tumor cell contamination

Stroma, normal non-tumorous, and lymphoid/myeloid cells contribute to the cellular heterogeneity within tumors, making it difficult to distinguish between tumor-specific methylation aberrations and those related to infiltrated normal cells in bulk bisulfite-sequencing [147]. This leads to diluted epigenetic signals of the tumor cells, reducing the sensitivity of variation and abnormality detection. Mixed populations of cells require bioinformatic tools to dissect the diversity and deconvolute sequencing data [148]. This has been done with transcriptomic data from RNA sequencing, where the gene expression patterns of infiltrated cells in the tumor were analyzed and successfully subtracted from the subsequent analyses of the tumor. Despite the absence of gene expression data, a successful approach was used to estimate and subtract infiltrated non-tumorous cells. This involved analyzing normal BalbC breast and lung tissue and generating distribution plots to identify regions lacking increased DNA methylation in those control tissues. For these regions, it was hypothesized that the increase in methylation truly comes from tumorous cells. However, for regions with a high DNA methylation signature in the control tissue, a subsequent subtraction of sequenced reads arising from non-tumorous cells was not feasible. The fasta files generated by BBA-seq were combined per region for breast tissue control (n = 5), lung tissue control (n = 3), primary tumors (n = 7), and metastases (n = 11). The CpGs for each read were counted and then arranged in ascending order. The distribution and relative frequency (read count / total reads) were then plotted in histograms (Figure 4.6). Cut-offs containing most control tissue reads were used to define the primary tumor and metastases reads for subsequent analysis. Five of the nine analyzed regions showed a predominantly unmethylated signature in breast and lung control tissues, namely regions B, D, E, F, and G. Regions C, J, K, and L showed already high methylation in control tissues. For this reason, subtracting non-tumorous sequenced reads for those regions was impossible. Table 4.2 shows the cut-off values, total discarded, and kept

single-end records and percentages. Due to an unusually high level of unmethylated characteristics, it was determined that sample 5423 contained a substantial amount of non-tumorous tissue, and consequently, it was excluded from subsequent analysis. For the rest of the samples, analysis of aberrant methylation progression from the primary tumor to metastatic sites was performed using the subtracted samples for the specific regions.

	Region B		Region D		Region E		Region F		Region G	
	Low methylated	Methylated	Low methylated	Methylated	Low methylated	Methylated	Low methylated	Methylated	Low methylated	Methylated
PT3145	1384 (22.26%)	4834 (77.74%)	893 (69.33%)	395 (30.67%)	985 (81.00%)	231 (19.00%)	2002 (51.78%)	1864 (48.22%)	2517 (89.60%)	292 (10.4%)
PT3524	127 (35.88%)	227 (64.12%)	353 (29.42%)	847 (70.58%)	1402 (96.89%)	45 (3.11%)	682 (72.40%)	260 (27.60%)	3941 (98.21%)	72 (1.79%)
PT3525	369 (31.73%)	794 (68.27%)	1503 (72.12%)	581 (27.88%)	417 (91.45%)	39 (8.55%)	4724 (84.22%)	885 (15.78%)	3948 (93.22%)	287 (6.78%)
PT3569	1543 (23.73%)	4958 (76.27%)	2469 (31.34%)	5408 (68.66%)	2471 (86.46%)	387 (13.54%)	13559 (88.19%)	1816 (11.81%)	6531 (98.36%)	109 (1.64%)
PT3620	741 (25.26%)	2193 (74.74%)	831 (67.07%)	408 (32.93%)	403 (98.77%)	5 (1.23%)	1411 (77.19%)	417 (22.81%)	409 (84.50%)	75 (15.50%)
PT3631	360 (32.35%)	753 (67.65%)	5653 (35.02%)	10491 (64.98%)	3369 (80.06%)	839 (19.94%)	7066 (79.21%)	1855 (20.79%)	1761 (96.65%)	61 (3.35%)
PT5423	14425 (97.06%)	437 (2.94%)	320 (96.97%)	10 (3.03%)	219 (100.00%)	0 (0.00%)	4185 (99.76%)	10 (0.24%)	543 (99.45%)	3 (0.55%)
MET3145_1	8705 (56.25%)	6759 (43.68%)	834 (76.23%)	260 (23.77%)	1425 (51.78%)	1327 (48.22%)	4119 (41.67%)	5692 (57.58%)	2810 (59.41%)	1920 (40.59%)
MET3145_2	799 (7.43%)	9958 (92.57%)	2100 (59.09%)	1454 (40.91%)	510 (54.96%)	418 (45.04%)	383 (13.74%)	2404 (86.26%)	769 (61.42%)	483 (38.58%)
MET3145_3	5416 (81.90%)	1197 (18.10%)	3503 (88.28%)	465 (11.72%)	1741 (91.73%)	157 (8.27%)	4050 (93.53%)	280 (6.47%)	642 (87.95%)	88 (12.05%)
MET3524	2711 (41.16%)	3876 (58.84%)	1234 (39.35%)	1902 (60.65%)	2757 (90.42%)	292 (9.58%)	10439 (57.32%)	7772 (42.68%)	6830 (98.60%)	97 (1.40%)
MET3525	808 (54.12%)	685 (45.88%)	785 (70.34%)	331 (29.66%)	720 (92.66%)	57 (7.34%)	1129 (61.03%)	721 (38.97%)	714 (92.01%)	62 (7.99%)
MET3569	14134 (41.52%)	19907 (58.48%)	2535 (41.01%)	3647 (58.99%)	3603 (83.04%)	736 (16.96%)	19077 (59.86%)	12791 (40.14%)	7468 (73.52%)	2690 (26.48%)
MET3620_1	1061 (31.26%)	2326 (68.53%)	494 (58.32%)	353 (41.68%)	310 (95.98%)	13 (4.02%)	1204 (92.69%)	95 (7.31%)	1185 (98.83%)	11 (1.17%)
MET3620_2	1243 (30.56%)	2825 (69.44%)	1442 (65.81%)	749 (34.19%)	340 (99.71%)	1 (0.29%)	2230 (68.98%)	1003 (31.02%)	671 (93.85%)	44 (6.15%)
MET3620_3	1471 (53.16%)	1296 (46.84%)	2841 (100.00%)	0 (0.00%)	1083 (88.77%)	137 (11.23%)	2800 (96.39%)	105 (3.61%)	915 (96.01%)	38 (3.99%)
MET3631	183 (5.93%)	2905 (94.07%)	705 (10.34%)	6114 (89.66%)	672 (24.65%)	2054 (75.35%)	6278 (43.59%)	8124 (56.41%)	1817 (56.24%)	1410 (43.76%)
MET5423	12115 (95.56%)	563 (4.44%)	2163 (98.63%)	30 (1.37%)	923 (99.89%)	1 (0.11%)	10863 (99.92%)	9 (0.08%)	2138 (99.35%)	14 (0.65%)

Table 4.2.: Overview of Cut-offs, total and percentage of subtracted and kept reads in regions B, D, E, F, and G. The "Low methylated" columns show the subtracted reads, whereas the "Methylated" column shows the reads that were kept for subsequent analysis. Cut-offs were >= 7 CpGs for region B, >= 9 CpGs for region D, >= 5 CpGs for region E, >= 5 CpGs for region F, and >= 3 CpGs for region G.

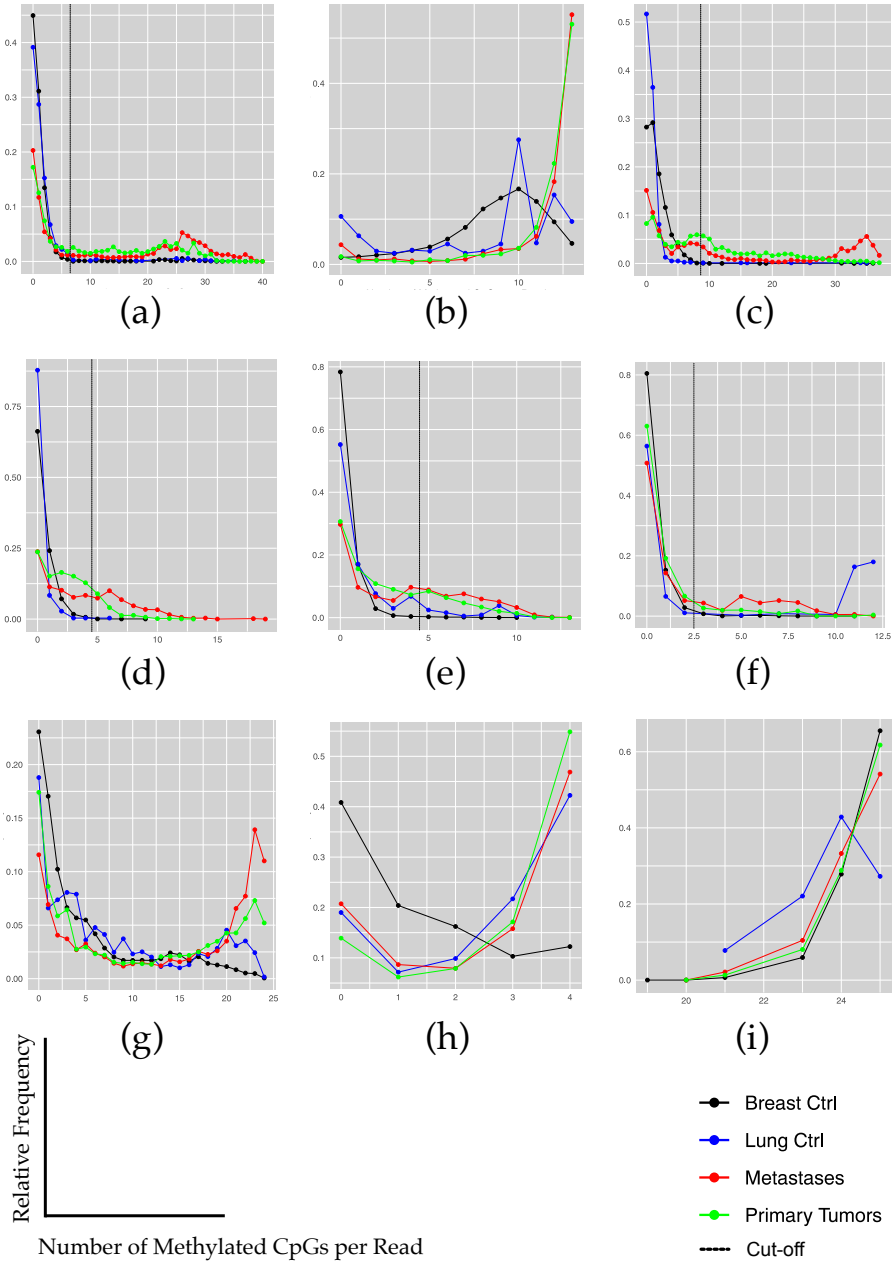


Figure 4.6.: Histograms showing the relative frequency and distribution of methylated CpGs in each region. (a) Region B. (b) Region C. (c) Region D. (d) Region E. (e) Region F. (f) Region G. (g) Region J. (h) Region K. (i) Region L.

	Sequences analyzed	Mapping efficiency (%)	Number of C's analyzed	Methylated C's in CpG context	Unmethylated C's in CpG context	Average conversion rate	Mean methylation
Breast Control 1	381516	17.4	5310148	395735	862455	99.4	28
Breast Control 2	66890	2.6	125092	3179	27523	99.4	10.4
Breast Control 3	72206	3.3	179783	6114	41139	99.4	12.9
Breast Control 4	163592	6.4	725735	25994	129698	99.3	16.7
Breast Control 5	183682	9.0	981624	30662	110524	99.3	21.7
Lung Control 1	354350	1.9	442312	26957	63833	99.2	29.7
Lung Control 2	533607	0.9	252661	7557	64360	99.2	10.5
Lung Control 3	2491	48.6	97234	5116	16106	99.3	24.1
PT3145	955126	23.6	2540527	345077	293954	99.5	54.9
PT3524	296100	11.1	1768642	248823	69364	99.4	78.2
PT3525	470513	19.8	5584906	539610	417195	99.5	56.4
PT3569	422063	17.9	3531136	466109	299258	99.4	60.9
PT3620	389913	17.9	1015146	150515	96231	99.4	61.0
PT3631	642701	15.9	3899441	568317	308715	99.4	64.8
MET3145 1	122223	42.8	2496082	473016	184864	99.4	71.9
MET3145 2	141905	19.5	2001220	443965	189366	99.5	70.1
MET3145 3	70765	34.6	711659	90624	81352	99.5	52.7
MET3524	439663	16.4	3148140	408885	210637	99.4	66.0
MET3525	28242	46.8	635773	80655	49657	99.4	61.9
MET3569	572736	23.3	6243833	983388	560392	99.5	63.7
MET3620 1	128169	10.6	660094	99306	77395	99.3	56.2
MET3620 2	123170	21.8	1685894	276430	131284	99.5	67.8
MET3620 3	125254	16.2	837123	97738	92045	99.4	51.5
MET3631	479371	18.8	3984814	727233	193315	99.3	79.0

Table 4.3.: Overview of diverse quality metrics for all samples analyzed, including total sequences analyzed, mapping efficiency in percentage, total number of cytosines analyzed, total methylated cytosines in CpG context, total unmethylated cytosines in CpG context, average conversion rate (ACR), mean methylation in percentage.

4.6. General methylation analysis

Methylation in the context of CpG dinucleotides serves as an epigenetic silencing mark in vertebrate genomes. Furthermore, recent findings show that DNA methylation functions differentially context-wise throughout different regions in the genome. To investigate CpG methylation in the context of metastatic breast cancer progression, DNA from six BalbNeu-T primary tumors, ten matching lung metastasis, and eight control samples were bisulfite converted and analyzed and single-base resolution.

Generally, the BBA-seq approach provided reasonable and reliable results. It generated a total of 7166248 sequences that could be aligned to the corresponding mm10 mouse genome with an average mapping efficiency of 18.6%. Nevertheless, some samples exhibited low mapping efficiency (e.g., Lung control 1 with 1.9%). The bisulfite treatment leads to low mapping efficiency due to the potential loss of entropy [149]. Additionally, the results showed an average conversion rate of 99.4%, which is considered very high

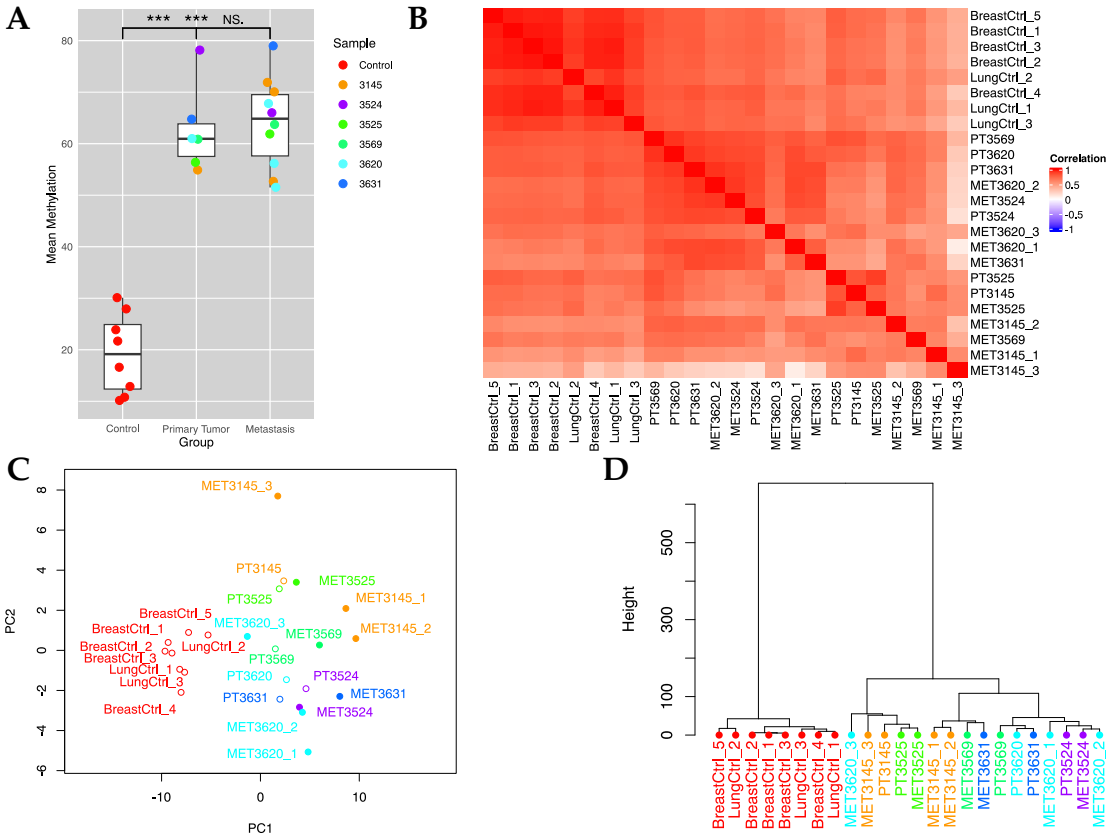


Figure 4.7.: General methylation analysis covering all regions. Low-methylation subtracted reads were applied to regions B, D, E, F, and G for primary and metastatic samples. No reads were subtracted for the other regions nor for the control tissue samples. (A) Boxplots showing mean methylation values for control, primary tumor and metastatic tissues. Mean methylation showed significant increase in Ctrl vs. PT and Ctrl vs. MET (p-values = <.0001 & <.0001, linear mixed model, followed by tukey method for comparing a family of 3 estimates) and none significant change from PT vs. MET (p-value = 0.9453, linear mixed model, followed by tukey method for comparing a family of 3 estimates). (B) Heatmap displaying the overall correlation between each sample. Samples were hierarchically clustered based on their correlation coefficient (C) PCA showing overall relationship between all samples analyzed. (D) Hierachical clustering representing relationship between samples analyzed (distance measure= "manhattan", agglomeration method= "ward").

compared to published data, with conversion effectiveness ranging from 96 to 100% among different kits [150]. These results demonstrated a high average conversion rate,

indicating successful bisulfite conversion. After filtering false positive CpGs according to the ACR, the analyzed reads are highly accurate, providing a reliable foundation for further DNA methylation analysis. Table 4.3 summarizes the general quality control of all samples analyzed. Mean methylation, expressed as a percentage, was calculated as follows:

$$r_{\text{mean meth}}[\%] = 100 \times \frac{n_{\text{meth}}}{n_{\text{meth}} + n_{\text{unmeth}}}$$

Overall, mean methylation values covering all regions were low for the control breast and lung tissues (average = 19.2%). They were significantly higher in the analyzed primary tumors and metastases, with mean methylation values of 62.7% and 64%, respectively (Table 4.3, Figure 4.7A). Nevertheless, the mean methylation change was insignificant between the malign conditions (Figure 4.7A, p-value = 0.9453, linear mixed model, followed by tukey method for comparing a family of 3 estimates). **R**-package `methylKit` provides the function `getCorrelation`, which calculates correlation coefficients between each sample. Using this functionality to calculate the Pearson correlation coefficients with subsequent hierarchical clustering revealed two distinct clusters as illustrated in figure 4.7B using a heat map. As the control tissue samples showed the lowest overall methylation, they clustered on the top left of the heatmap, indicating a close relation. The primary tumor and metastases samples were clustered on the bottom right of the heatmap. PT3569, PT3620, PT3631, PT3524, MET3620 2, and MET3524 formed a sub-cluster with a high correlation. Samples PT3525, PT3145, and MET3525 also showed high correlation. Interestingly, all the most distinct samples from the control tissue were metastatic samples.

Principle component analysis (PCA) revealed similar results, as illustrated in figure 4.7C. The control tissue samples were clustered nearby. However, it was noticed that the metastatic samples were predominantly positioned on the right side of the PCA plot. This may be due to the slight increase in DNA methylation. Most metastatic samples were located close to their respective primary tumors (3524, 3525, 3569, 3631). This could be explained by the fact that despite the increase in methylation, the metastatic samples

still exhibit the distinctive DNA methylation patterns characteristic of their primary tumors. Interestingly, samples displaying multiple metastases (3145, 3620) showed less similarity within each other and a higher variety in distribution, concluding a higher distinction in DNA methylation patterns among the metastases samples (Figure 4.7C). In 2011, Clifford et al. compared different clustering methods to analyze DNA methylation data and found that the distance method "manhattan" in combination with the agglomeration method "ward" outperformed other approaches and provided the most consistent results [151]. Therefore, these two metrics were used to build a hierarchical clustering of the samples, shown in figure 4.7D. Again, the control tissues showed high similarity as they built their tree branch. Furthermore, the clustering method reflected the results of the PCA analysis, with some primary tumors being close to their metastatic samples. Samples 3145 and 3620 exhibited a wider distribution.

Taken together, DNA methylation levels increased significantly from non-malignant tissue to both, the primary tumor and further to the metastatic tissues at the specific loci analyzed (Figure 4.7A). Correlation, principal component analysis, and hierarchical clustering indicated that methylation changes were more individual-specific than linked to ongoing metastasis progression, as most metastatic sites were closely related to their individual primary tumor. Further, the findings indicate varying methylation progression patterns across the different samples, with some showing increases and others showing decreases in methylation as they progress from the primary tumors to metastases. For example, in sample 3524, a reduction in methylation from primary tumor to metastasis could be observed, whereas sample 3631 exhibited increased DNA methylation, cf Table 4.3. Therefore, primary tumors with higher methylation did not necessarily have higher methylated metastases. This could suggest that methylation changes are more sample-specific and potentially related to unique biological factors in each case. However, the methylation pattern could not be analyzed in detail by calculating mean methylation values. This is because it does not indicate whether the values result from a mix of low and highly methylated cells in the primary tumors or

if most of the primary tumors have an equal level of DNA methylation. For this, the regions were analyzed separately to investigate the methylation pattern and progression in primary tumors at single-base resolution and to determine whether single highly methylated clones already existed in the primary tumor sites.

4.7. Phyloepigenetic analysis of BBA-seq-derived data and tree generation

In this study, potentially existing rare, highly methylated primary tumor sub-clones from bulk samples were to be identified by resolving DNA methylation at single-base allele resolution. Furthermore, the alleles were to be grouped and ordered according to their methylation content. To accomplish this, first proper `*.fasta` files for each sample were produced, and those files that should be compared later were combined, such as primary tumor I + metastasis I (II-III) + controls. The combined files were aligned using the multiple-sequence-alignment (MSA) software `Clustal Omega` [152]. Five methods were essayed to choose the suitable tree algorithm for sorting the sequenced reads from unmethylated to methylated: neighbor-joining (NJ), minimum evolution (ME), maximum parsimony (MP), maximum likelihood (ML), and unweighted pair group method with arithmetic mean (UPGMA). The NJ algorithm is derived from the ME method. Both algorithms start in a star-like structure and connect the closest related taxa, in this case, the sequenced reads, iteratively. The advantages of NJ are its speed and the ability to calculate distances on large datasets [153]. MP creates a tree structure that tries to explain the data with the fewest evolutionary changes. Nevertheless, in evolutionary models, the algorithm tends to produce "long-branch attraction"-artifacts that falsely group taxa with high mutation rates. ML uses the log-likelihood function to construct trees with optimized branch lengths. The ML method is widely considered the best algorithm for constructing phylogenetic trees, as it helps identify the most optimal arrangement of genetic sequences to explain assumed evolutionary development [154].

The UPGMA clustering method assumes a constant rate of evolution. It constructs ultrametric trees where the distance from the root to any branch is the same. Like NJ, it begins with the closest pair of taxa, merges them, and combines clusters based on average linkage until a single tree is formed [155]. All those algorithms were implemented to reconstruct evolutionary relationships and are rarely used to cluster sequenced reads derived from bisulfite-converted DNA.

For this reason, all methods were used to construct trees from the BBA-seq data to determine the most suitable algorithm. The method of choice had to cluster the reads according to their DNA methylation signature in diverse conditions, namely in regions with a low or high methylation profile and everything in between. Figure 4.8 illustrates sample 3524 within region B, representing a gradient from low to high methylated reads. The UPGMA algorithm (Figure 4.8A) and the ME algorithm (Figure 4.8B) proved the best-suited methods to cluster the reads. Region C of sample 3631 is exemplary for regions containing mostly methylated reads and the results of the clustering with different algorithms is shown in figure 4.9. As one can see, here all algorithms had problems clustering the samples in the correct order from low to high methylated. Nevertheless, UPGMA (Figure 4.9A) showed the best resolution and clustering. The same was true for regions containing mostly un-methylated reads. Figure 4.10 shows sample 3569 within region D, where most methods showed clustering of the reads according to their increased DNA methylation, with UPGMA (Figure 4.10A) and ME (Figure 4.10B) having the best resolution from low to high methylated.

In conclusion, the UPGMA algorithm seemed to be the best choice for arranging the sequencing reads from low to highly methylated and subsequent clustering. Furthermore, the UPGMA method could be executed within `Clustal Omega` through the command line, making it more user-friendly and faster. On the other hand, the ME algorithm was run in `MEGA` software, which is significantly slower. As a result, the UPGMA algorithm was used for tree generation and analysis. Figure 4.11 shows and explains an example of such a tree of sample 3620 within region J. Among all other methods, the result-

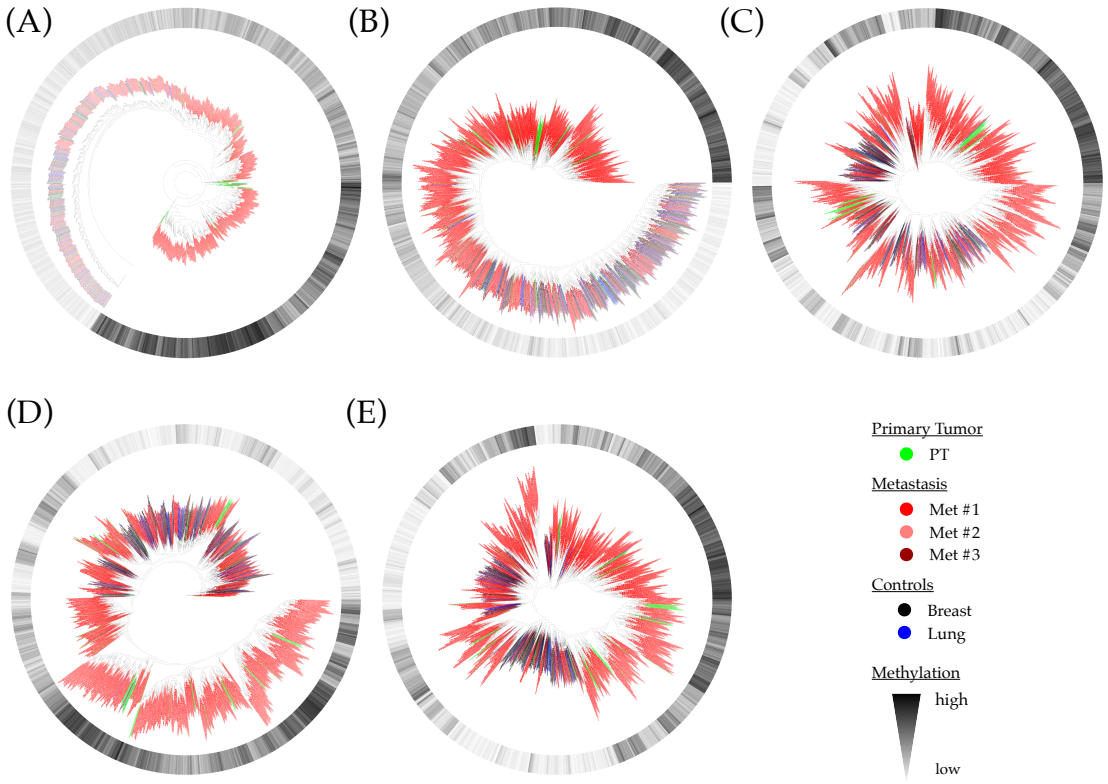


Figure 4.8.: A visual representation of phyloepigenetic trees of sample 3524 within region B, covering all phylogenetic algorithms that had been essayed to resolve and cluster sequenced reads from BBA-seq data. (A) UPGMA. (B) ME. (C) ML. (D) MP. (E) NJ.

ing phyloepigenetic tree allows the visualization of the close relationships between different samples, controls, and gradients of methylation levels. The tips of the tree comprise the first line of each read from the `*.fasta` file containing the information `>sample_region_sequence` (File Excerpt 3.4). A specific color code was employed to facilitate the interpretation of the phyloepigenetic tree. Primary tumors that developed metastasis were highlighted in green, metastasis samples in red, breast control in black, and lung control reads in blue. If there was more than one sample per type, a gradient in the same color was used to distinguish between the samples, e.g. different metastases from one sample. To further aid the interpretation of the phyloepigenetic tree, a ring surrounding the tree that indicates the methylation level of each indivdiual read from

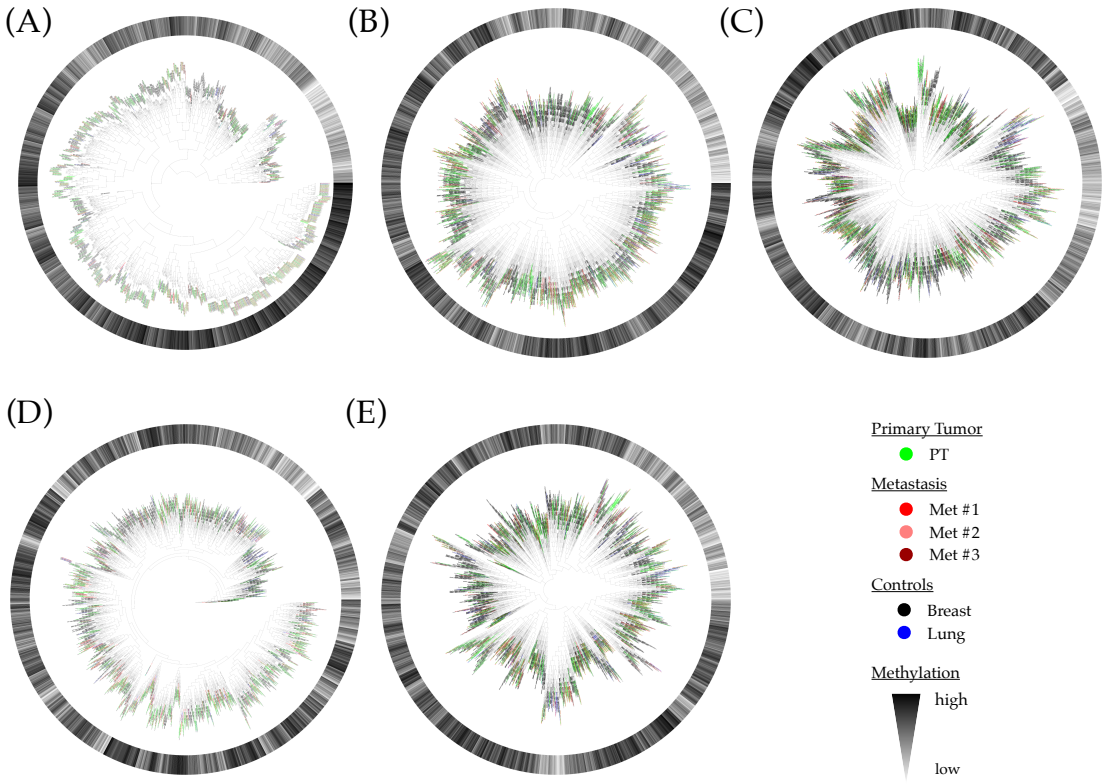


Figure 4.9.: A visual representation of phyloepigenetic trees of sample 3631 within region C, covering all phylogenetic algorithms that had been essayed to resolve and cluster sequenced reads from BBA-seq data. (A) UPGMA. (B) ME. (C) ML. (D) MP. (E) NJ.

zero to full methylated was included. Additionally, the tree was divided into clusters comprising differentially methylated reads. The clusters were denoted with Roman numerals and used to generate sequence logos to further identify rare high-methylated reads from primary tumor samples. Next to the tree, the legend comprises the sample, controls, and the gradient of the surrounding ring. Trees were bootstrapped 1000 times with `Clustal Omega` to improve the reliability of the obtained results. This method generates phyloepigenetic trees using a resampling technique called Bootstrap, which randomly samples the data and then repeats the analysis multiple times. The goal is to estimate the reliability of the phyloepigenetic tree topology and branch lengths [156]. Together, these features provide a comprehensive and detailed visualization of

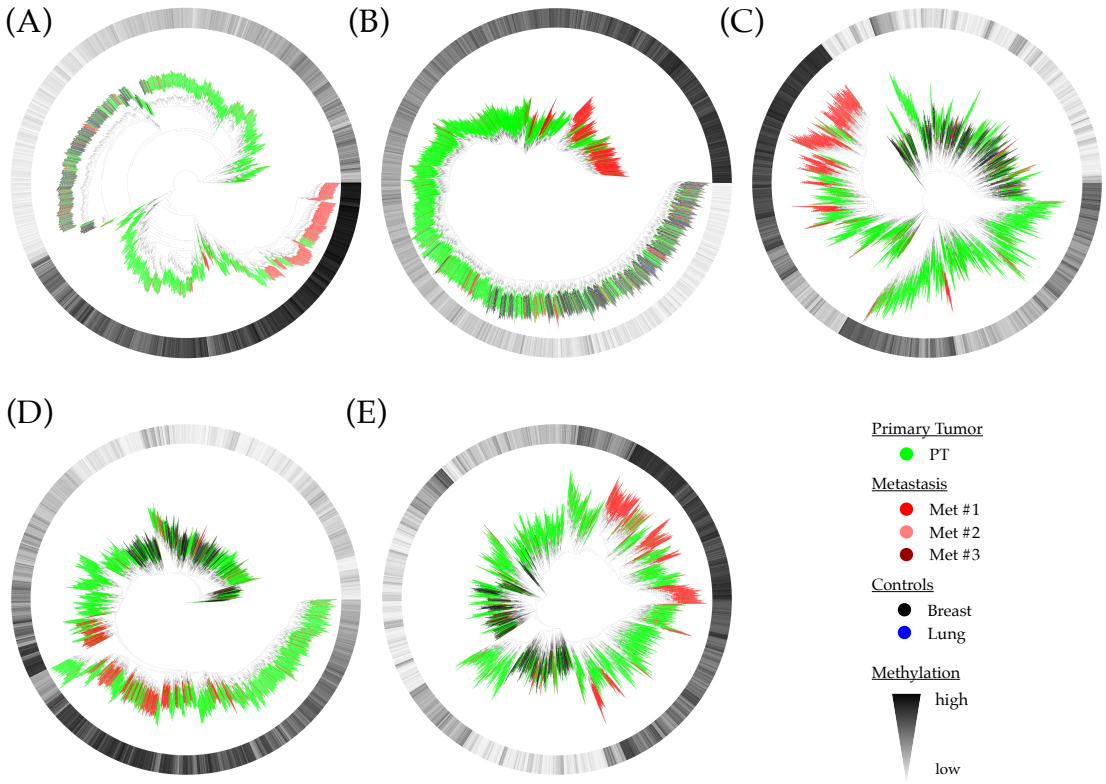


Figure 4.10.: A visual representation of phyloepigenetic trees of sample 3569 within region D, covering all phylogenetic algorithms that had been essayed to resolve and cluster sequenced reads from BBA-seq data. (A) UPGMA. (B) ME. (C) ML. (D) MP. (E) NJ.

the relationships between the different samples and controls and the methylation level at each stage of the progression process.

4.8. Methylation analysis in region B

Region B covers 40 CpG sites at a CpG island in the promoter region of gene *Zdbf2* on Chr1:63273001-63274000. This region, located in a promoter, shows promise as a candidate for further analysis. Promoter regions are known for their dynamic methylation patterns, particularly regarding differentiation and proliferation. In this study, six

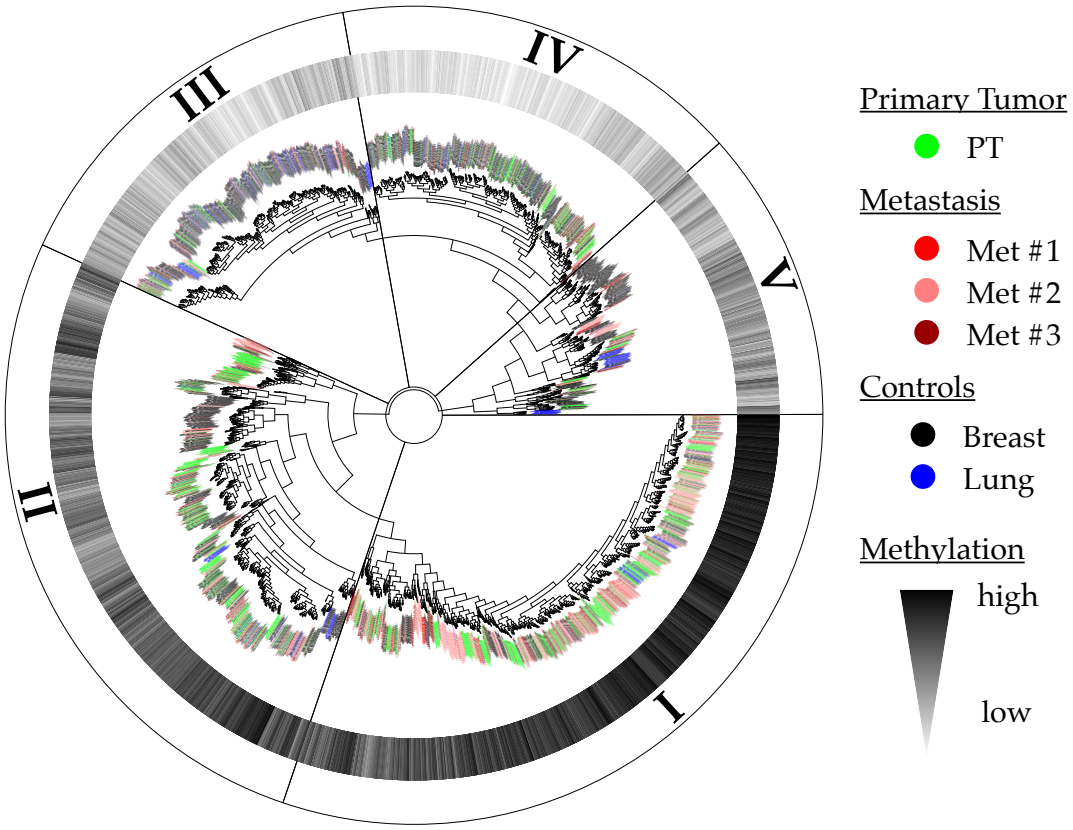


Figure 4.11.: A visual representation of a phyloepigenetic tree has been created using the BBA-seq and phyloepigenetic workflow. The tree has been sorted based on an increase in DNA methylation within the sequencing reads of the respective samples. A legend is provided on the right-hand side of the illustration.

primary tumors with matched metastases were analyzed. Two tumors (3145 and 3620) exhibited three metastatic sites, while the remaining showed only one. As the control tissues showed no increase in methylation, it was assumed that the infiltrated stroma cells also did not show an increase in methylation. As a result, the non-methylated reads were excluded from the analysis in all samples for this specific region to achieve the most accurate results at the highest resolution. The unmethylated cluster contained $\approx$ 99% of the breast and lung control reads in all samples. For a better comparison all phyloepigenetic trees before subtraction of the non-methylated reads from primary tumors, and metastases can be observed in figure B.2. An example for this procedure

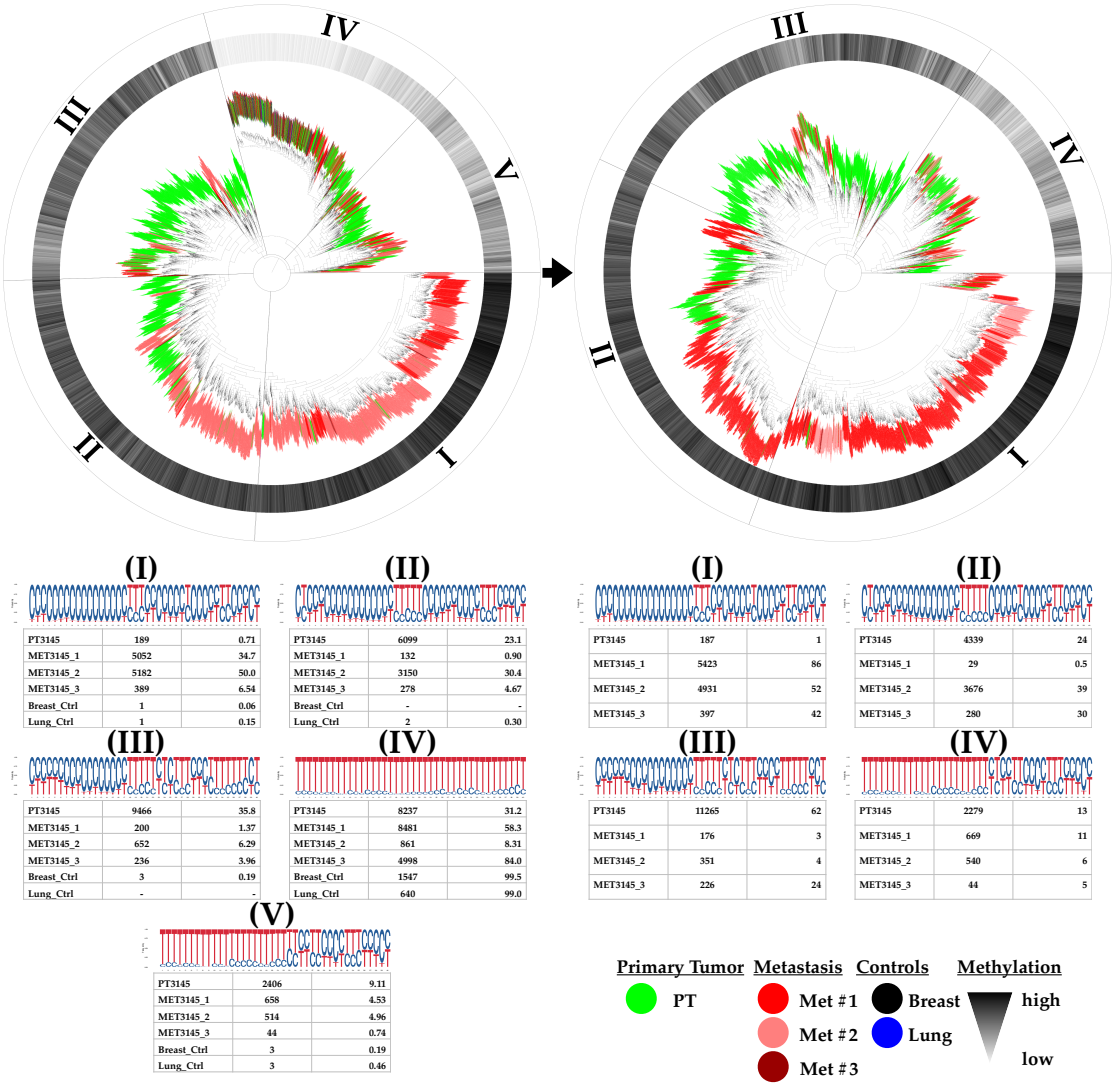


Figure 4.12.: Example on sample 3145 illustrates the subtraction of non-methylated reads due to non-tumorous stroma cell infiltration. The sequence logos at the bottom summarize the base composition of each cluster. The tables display the sample ID (left column), the number of reads within each cluster per sample (middle column), and the overall percentage of reads per sample (right column). The legend is located at the bottom, far right.

is illustrated in figure 4.12 for sample 3145. The left tree displays all reads before the unmethylated cluster IV was removed, and the right tree displays the tree where the unmethylated reads were excluded.

When comparing the primary tumors with their corresponding metastases, there was no significant change in mean methylation. The primary tumors had a mean methylation value of 43.3% (31.3% before subtraction), whereas the metastatic samples showed an increase in mean methylation with an average value of 52.9% (31.3% before subtraction). However, this change was insignificant (p-value = 0.2716, linear mixed model, followed by tukey method for comparing a family of 3 estimates) (Figure 4.13A). Calculating correlation coefficients with subsequent hierarchical clustering showed that most samples were not highly correlated. However, PT3620, PT3569, PT3631, MET3620 1, and MET3620 3 showed a cluster on the top left of the plot. These samples also exhibited low mean methylation values. Sample pair 3525 showed a high correlation as clustering in a cluster in the middle of the plot (Figure 4.13B).

Additionally, the principle component analysis showed no defined clustering of the primary tumors or the metastases (Figure 4.13C). Most of the variance could be explained through component 1 (Figure B.1B). Primary tumor samples and their matched metastasis samples clustered more closely together than primary tumors and metastases, respectively. These observations support the hypothesis that DNA methylation is more animal or sample-specific and does not randomly increase over time and proliferation. Primary tumors with already high methylation values also exhibited a high methylation signature in metastasis (e.g. PT3145= 52.9%, MET3145 1= 75.2%, MET3145 2= 70.5%, MET3145 3= 63%). In contrast, primary tumors that exhibited low methylation had corresponding metastases with low methylation values as well (e.g., PT3620= 38.4%, MET3620 1= 31.5%, MET3620 2= 40.3%, MET3620 3= 30.8%). Clustering the methylation data not via correlation but with “manhattan” distance and “ward” agglomeration once again showed a close relationship between primary tumors and metastases from individual samples, rather than between primary tumors or metastases from different samples (3525, 3620) (Figure 4.13D). However, for a more detailed view at a single base resolution and in order to to analyze the actual distribution of CpG methylation, the phyloepigenetic approach was used to identify possibly existing

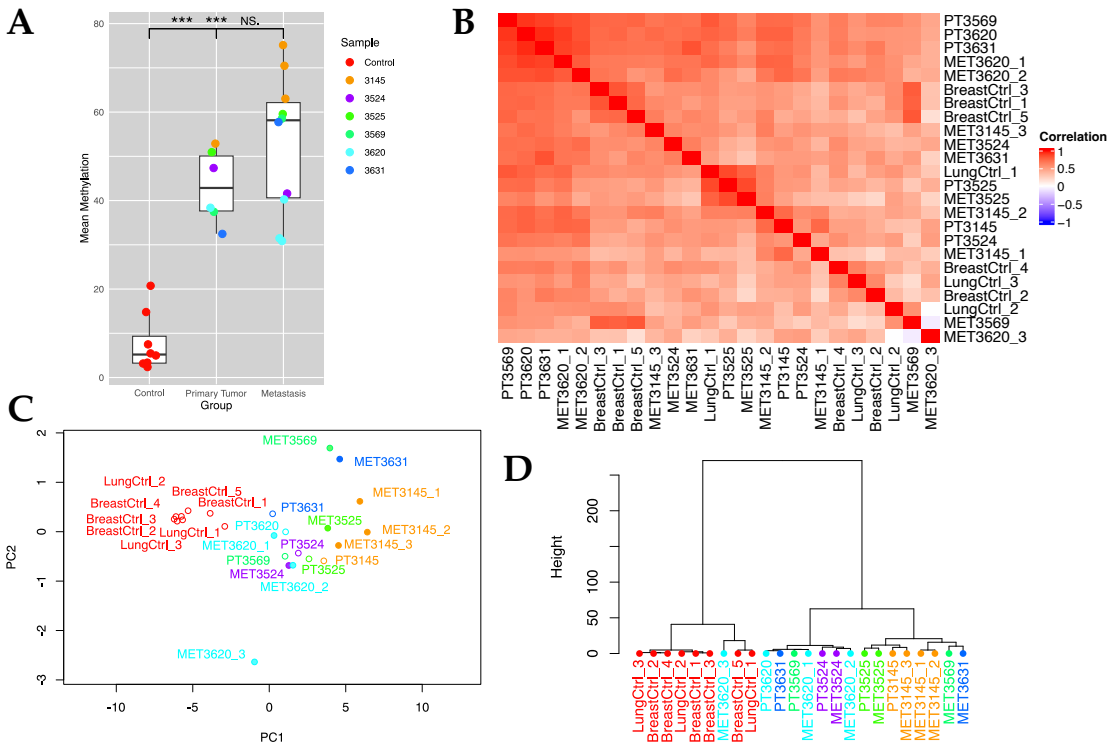


Figure 4.13.: General analysis of methylation in region B. (A) Boxplots showing mean methylation values for control, primary tumor and metastatic tissues in region B. Mean methylation showed significant increase in Ctrl vs. PT and Ctrl vs. MET (p-values = <.0001 & <.0001, linear mixed model, followed by tukey method for comparing a family of 3 estimates) and no significant change from PT vs. MET (p-value = 0.2716, linear mixed model, followed by tukey method for comparing a family of 3 estimates). (B) Heatmap displaying the correlation between each sample. (C) PCA analysis indicates no significant clustering of samples. Neither primary tumors nor metastasis show signs of unity. Matched primary tumors and metastases are uniquely colored. Controls and primary tumors are represented by unfilled points, while metastasis samples are represented by filled points. (D) Clustering of CpG methylation does not indicate a distinct separation between primary tumors and metastasis samples (distance measure= "manhattan", agglomeration method= "ward").

highly methylated reads from primary tumor samples (Figure 4.14). In all samples, over 99% of the control reads from breast and lung tissues were found in the cluster with the lowest methylation levels. This observation supported the assumption that primary tumor reads with a high methylation profile originated from primary tumor

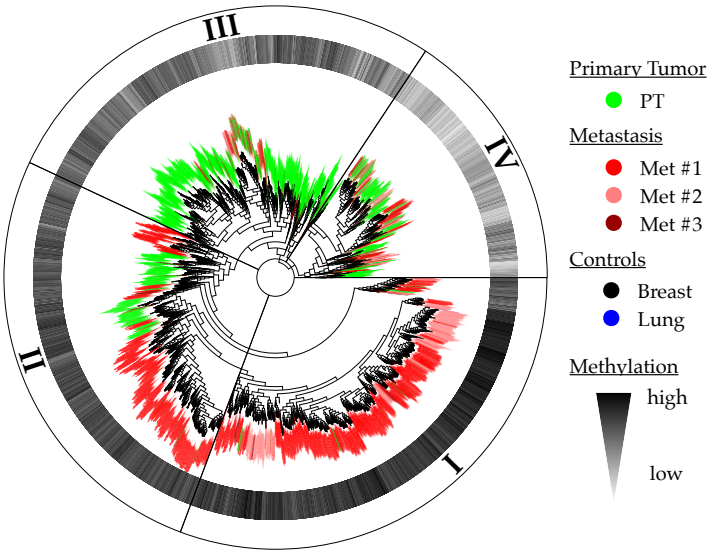
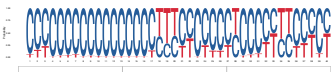
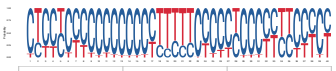


Figure 4.14.: DNA methylation analysis in region B at single base resolution. The phyloepigenetic tree for sample 3145 is shown as an example for region B. The DNA methylation pattern observed in primary tumors exhibits variations in methylation progression within region B.



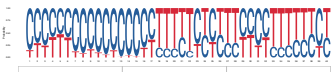
PT3145	187	1
MET3145_1	5423	86
MET3145_2	4931	52
MET3145_3	397	42

(a) 3145 Cluster I



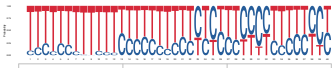
PT3145	4339	24
MET3145_1	29	0.5
MET3145_2	3676	39
MET3145_3	280	30

(b) 3145 Cluster II



PT3145	11265	62
MET3145_1	176	3
MET3145_2	351	4
MET3145_3	226	24

(c) 3145 Cluster III



PT3145	2279	13
MET3145_1	669	10.5
MET3145_2	540	5
MET3145_3	44	4

(d) 3145 Cluster IV

Figure 4.15.: The sequence logos summarize the base composition of each cluster for sample 3145 in region B. The tables display the sample ID (left column), the number of reads within each cluster per sample (middle column), and the overall percentage of reads per sample (right column).

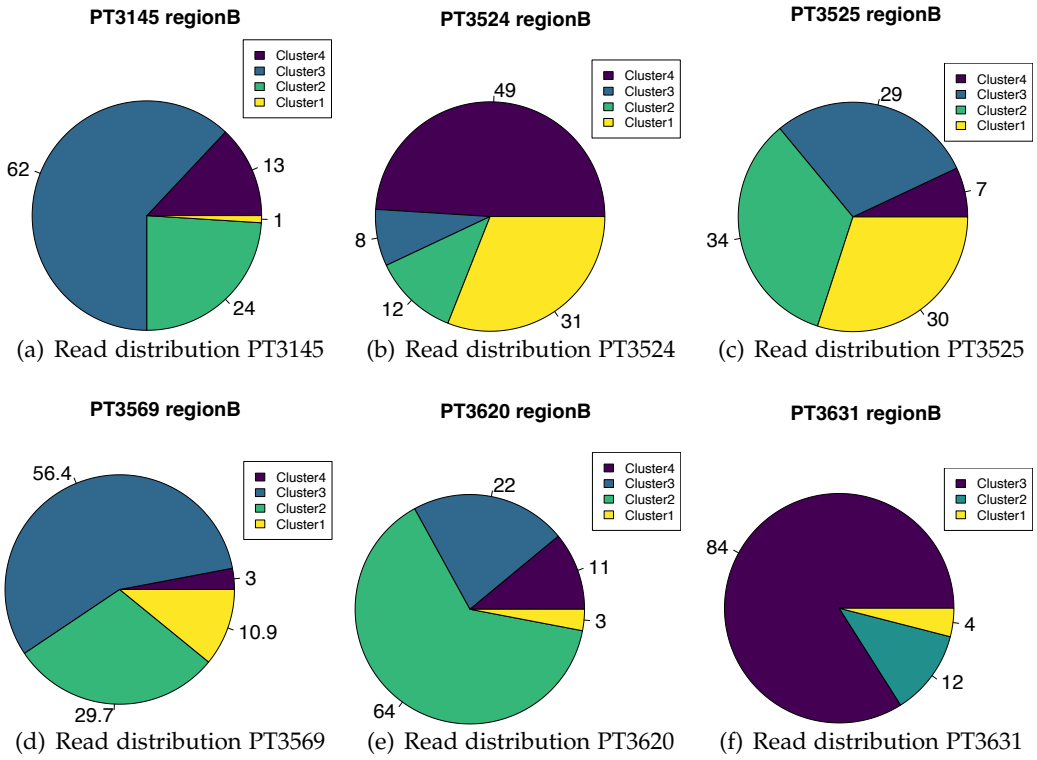
cells, not infiltrated stroma cells. For sample 3145, cluster analysis revealed that only 1% of primary tumor reads were located in the cluster with the highest methylation

(Figures 4.15 (a) and 4.16 (a)). In contrast, this cluster contained the highest number of reads derived from metastatic samples (86%, 52%, and 42% of reads for MET3145 1, 2, and 3, respectively) (4.15 (a)). Further, highly methylated sequencing reads were identified within the tumor sample. More than 80% of the primary tumor sample reads were found in clusters II and III. It's worth noting that these clusters have a high methylation profile (Figure 4.15). In cluster II, the primary tumor reads were located at the beginning of the cluster, indicating that they have the lowest methylation within this cluster. This observation could suggest that the upcoming metastatic cells are derived from those primary tumor cells. Further, the rare, highly methylated reads from the primary tumor found in cluster I strengthen the hypothesis of a selection process. Compared to the phyloepigenetic tree and the cluster analysis without subtracting the non-methylated reads from the samples, a comparable amount of sequencing reads of the primary tumor sample could be detected in the cluster with the highest methylated reads. More than 30% of the primary tumor reads were excluded following this analysis because they possessed no methylation. More than 80% of the reads from the metastatic sample 3 were removed due to a non-methylated read profile, indicating that these reads were from non-metastatic tissue. This shifted the distribution toward the highest methylated cluster within that sample. Nevertheless, even in the non-excluded tree, rare highly methylated reads could be identified in the cluster comprising the highest methylation (Figure 4.12). In sample 3524, 31% of the primary tumor reads were located in the cluster with the highest methylation (Figures 4.16 (b) and B.3A). Compared to the tree where no reads were removed, no shift in distribution was observed (29% to 31%). 35% of reads from the primary tumor were removed from this analysis as they showed no methylation (Figure B.3A). Subsequently, 50% of the primary tumor reads could have been found in cluster IV. Still, the distribution of sequencing reads from low to high methylation was nearly the same. In each cluster, single methylated reads from the primary tumor were observed surrounded by reads belonging to the metastatic sample. 37% of the reads from the metastasis showed no methylation and were therefore removed from this analysis. Unfortunately, for this sample in this region,

only approx. 150 reads from the primary tumor could be analyzed. Compared to 1800 reads from the corresponding metastasis, a meaningful comparison between these two samples was difficult. The identified rare high methylated reads from the primary tumor in the high methylated clusters could come from the small sample read size or the selective process where the high methylated primary tumor samples evolve into metastatic cells. In sample 3525, 30% of the primary tumor reads were located in the high methylation cluster I, compared to 43% for the metastatic sample (Figure 4.16 (c)). An equal distribution could also be observed in the tree and the cluster analysis where no reads were removed. Furthermore, both primary tumor and corresponding metastasis samples showed an equal distribution throughout the whole tree. No rare high methylated reads could be identified in the primary tumor sample (Figure B.3B). After no-methylated read removal, more than 700 reads from each sample could be analyzed, making the tree and the cluster analysis reliable. Within sample 3525, an even distribution of primary tumor reads could be observed comparable to its corresponding metastasis. Here, no significant differences in methylation and progression between the two samples could be identified. Like 3145, sample 3569 is an example where rare, highly methylated reads could be found in the primary tumor sample. The tumor displayed only 11% of reads in the cluster with the highest methylation (Figures B.3C and 4.16 (d)). This cluster also contained 90% of the metastasis reads. Gain of methylation from the primary tumor to the metastasis occurred as in all other samples at the 3' end. This CpG location seems to be of further importance. Most primary tumor reads were located in clusters II and III, ancient to cluster I. This leads to the conclusion that primary tumor cells have a characteristic methylation fingerprint, with only a few exceptions, in particular those becoming metastatic cells. Also, in the tree where no reads have been removed, rare high methylated reads in the tumor samples could be identified, with only 4% supporting this observation. An almost equal amount of reads had been removed for this analysis, with 29% for the primary tumor and 37% for the metastasis sample, respectively. Sample 3620 had an overall low level of methylation across all samples. Tree and cluster analysis showed an even distribution

of all sample reads throughout the methylation profile (Figures B.3D and 4.16 (e)). No rare, high methylated reads could be observed for the primary tumor sample. Sample 3620 showed no clear difference between the primary tumor and its corresponding metastasis regarding their methylation pattern. Sample 3631 again showed signatures of a selective process, as only 4% of the primary tumor reads were located in cluster I (Figures B.3D and 4.16 (f)). In contrast, almost 50% of the reads from the metastasis were assigned to that cluster. Therefore, rare high methylated reads from the primary tumor could be observed. The CpG methylation was increased from the 5′ to 3′ prime end within the primary tumor sample. Most of its reads (84%) were located in cluster III. This observation was further strengthened when looking at the tree and the cluster analysis, where no reads were removed, as 62% of the primary tumor reads showed a non-methylated pattern. Additionally, in regions where low methylated reads may have been excluded, the distribution of primary and metastatic reads was plotted based on their CpG count rather than the UPGMA algorithm. This was possible because removing the unmethylated reads resulted in a much smaller sample size. It allowed the distribution plots to validate the phyloepigenetic trees and further simplified the identification of specific DNA methylation progression patterns. In detail, the methylated cytosines were counted for each sequenced read and sorted in ascending order. Similar to the phyloepigenetic tree plots, a ring surrounding the distribution plots indicates the DNA methylation proportion from low to high methylation. In summary, these plots confirm the observations deduced from the phyloepigenetic trees (Figure 4.17). Samples 3145, 3569, and 3631 showed small groups of highly methylated cells compared to the rest of their sequenced reads. Within samples 3524, 3525, and 3620, the primary tumor reads were equally distributed throughout the plots, showing signs of a constant increase of DNA methylation (Figure 4.17). In conclusion, within region B, heterogeneous DNA methylation patterns were observed in the primary tumor samples (Figure 4.17). Three out of six tumor samples showed single high methylated reads within clusters, which were highly methylated and predominately harbored by metastatic samples. In these samples, the hypothesis is that those single high-methylated primary

tumor cells gained the capability of becoming metastatic cells due to their higher DNA methylation pattern. Additionally, some metastatic sites showed lower methylation profiles than their corresponding primary tumor, and most of the highly methylated reads from primary tumors were located at the beginning of the subsequent cluster. All these observations supported the hypothesis of a selection process taking place. The PCA analysis showed no defined separation between these groups, indicating that the metastasis samples are similar to the primary tumor samples they derived from. Nevertheless, three of the six samples showed a continuation in the gain of DNA methylation. These observations strengthen the hypothesis that the gain of methylation is a byproduct of time and proliferation rather than providing a selective advantage.



(a) Read distribution PT3145 (b) Read distribution PT3524 (c) Read distribution PT3525

(d) Read distribution PT3569 (e) Read distribution PT3620 (f) Read distribution PT3631

Figure 4.16.: Pie charts show how sequenced reads are distributed across all primary tumor samples from region B in each individual cluster. These charts clarify the read distribution that reflects the phyloepigenetic trees and consensus sequence logo cluster analysis.

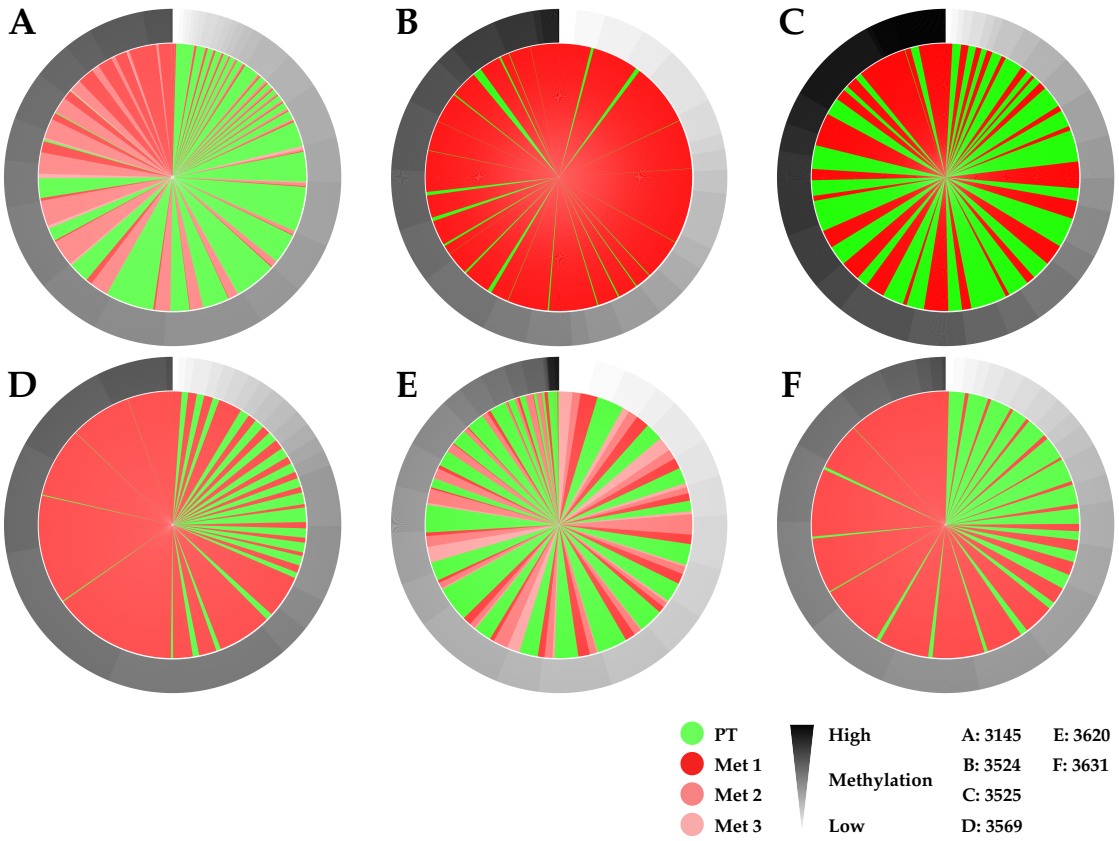


Figure 4.17.: The arrangement of reads in region B is based on their mCpG count. The reads are plotted according to their mCpG count in increasing order and displayed in a circular format. A gradient ring surrounding the circle indicates the methylation status.

4.9. Methylation analysis in region C

Region C is located within exon two of eight of the *Itpkb* gene, and in the samples analyzed, it exhibted a generally hypermethylated state. However, the dataset revealed a range of methylation patterns, varying from fully methylated to unmethylated, with most sequenced reads displaying high methylation levels. The region spans approximately 700 bases and contains 13 CpGs subjected to thorough analysis. Most CpG sites within intragenic features tend to be methylated [157]. This was also true for this locus, as most samples showed an overall high CpG methylation profile with no significant change in mean methylation between primary tumor and metastatic

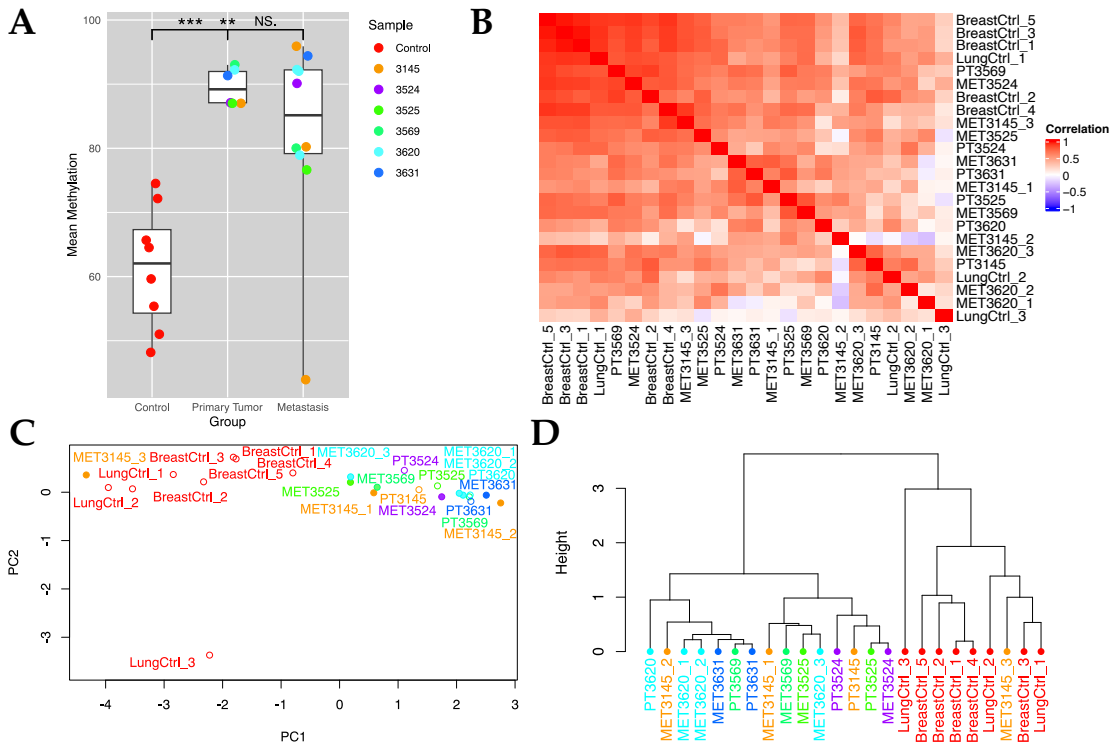
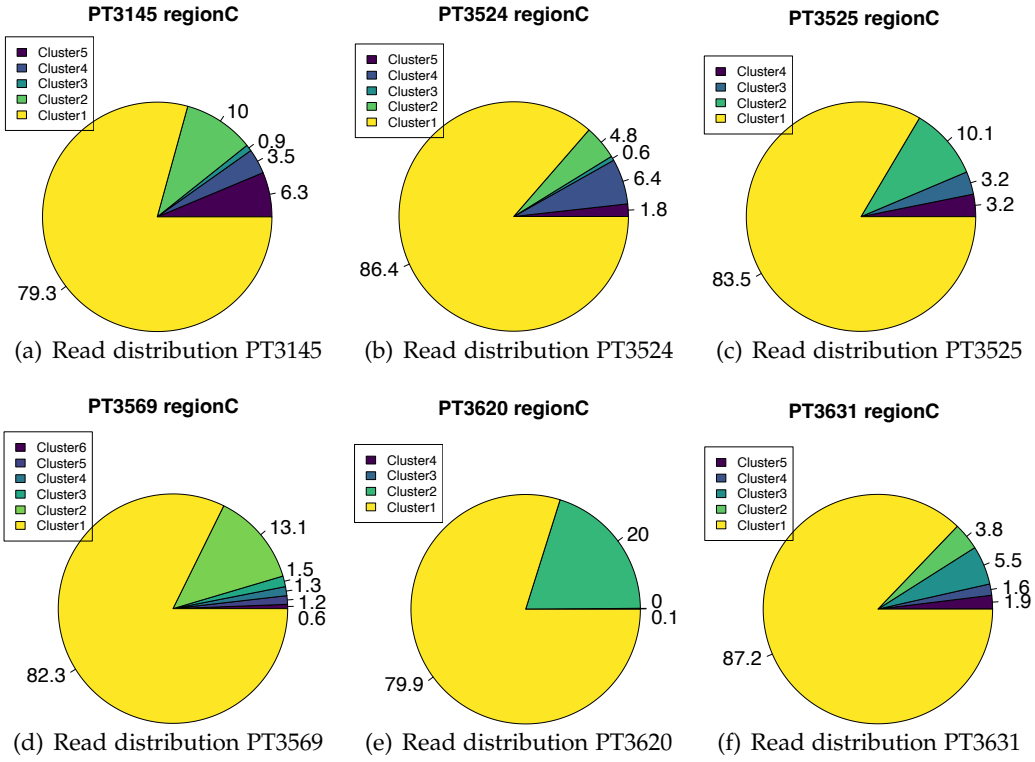


Figure 4.18.: General analysis of methylation in region C. (A) Boxplots showing mean methylation values for control, primary tumor and metastatic tissues in region C. Mean methylation showed significant change in Ctrl vs. PT, Ctrl vs. MET (p-values = 0.0005, 0.0026 & 0.2528, linear mixed model, followed by tukey method for comparing a family of 3 estimates) and no significant change in PT vs. MET (p-value = 0.4624, linear mixed model, followed by tukey method for comparing a family of 3 estimates). (B) Heatmap displaying the correlation between each sample. (C) PCA analysis indicates no significant similarity of samples. Neither primary tumors nor metastasis show signs of unity. Matched primary tumors and metastases are uniquely colored. Controls and primary tumors are represented by unfilled points, while metastasis samples are represented by filled points. (D) Clustering of CpG methylation does not indicate a distinct separation between primary tumors and metastasis samples (distance measure= “manhattan”, agglomeration method= “ward”).

samples (Figure 4.18A). The controls exhibited an average mean methylation value of 61.38%, primary tumor samples 89.61%, and metastases 82.43%. Interestingly, most of the metastatic samples in this region showed lower mean CpG methylation than the primary tumors from which they originated. No increase in methylation could be

observed over time, nor were specific CpG sites differently methylated. The correlation coefficients indicated that the control lung tissue sample 3 had the lowest correlation with the other samples. Furthermore, no division between primary tumor controls and metastatic samples could be observed. Also, besides sample 3631, no primary tumor/metastases pairs grouped (Figure 4.18B). This could also be observed in the PCA analysis, as most samples were close to each other without a clear division of sample pairs (Figure 4.18C). PC1 and PC2 covered most variances (Figure B.1C). Additionally, CpG methylation clustering showed clear separation of the most unmethylated samples, as the control tissues and MET3145 3 branching off at the earliest from the rest. MET3145 3 had the lowest mean methylation value (43.9%). The remaining samples clustered near each other, with no division between primary tumors and metastases or within sample pairs (Figure 4.18D). As the changes in methylation were minimal and the region was already highly methylated, no definitive statement could be made regarding methylation alteration and progression. In addition, the control samples displayed a significant high methylation pattern. This makes it challenging to distinguish between reads obtained from tumor samples and infiltrated stroma cells. The phyloepigenetic trees demonstrated that the samples are evenly distributed across the entire tree (Figure B.4). Nevertheless, the pie charts indicated that cluster I had the highest number of reads across all samples (Figure 4.19). In conclusion, within region C, it was impossible to identify rare high-methylated reads within tumor samples due to the high-methylated character of all reads and the significant high-methylation properties of the control samples.



(a) Read distribution PT3145 (b) Read distribution PT3524 (c) Read distribution PT3525

(d) Read distribution PT3569 (e) Read distribution PT3620 (f) Read distribution PT3631

Figure 4.19.: Pie charts show how sequenced reads are distributed across all primary tumor samples from region C in each individual cluster. These charts clarify the read distribution that reflects the phyloepigenetic trees and consensus sequence logo cluster analysis.

4.10. Methylation analysis in region D

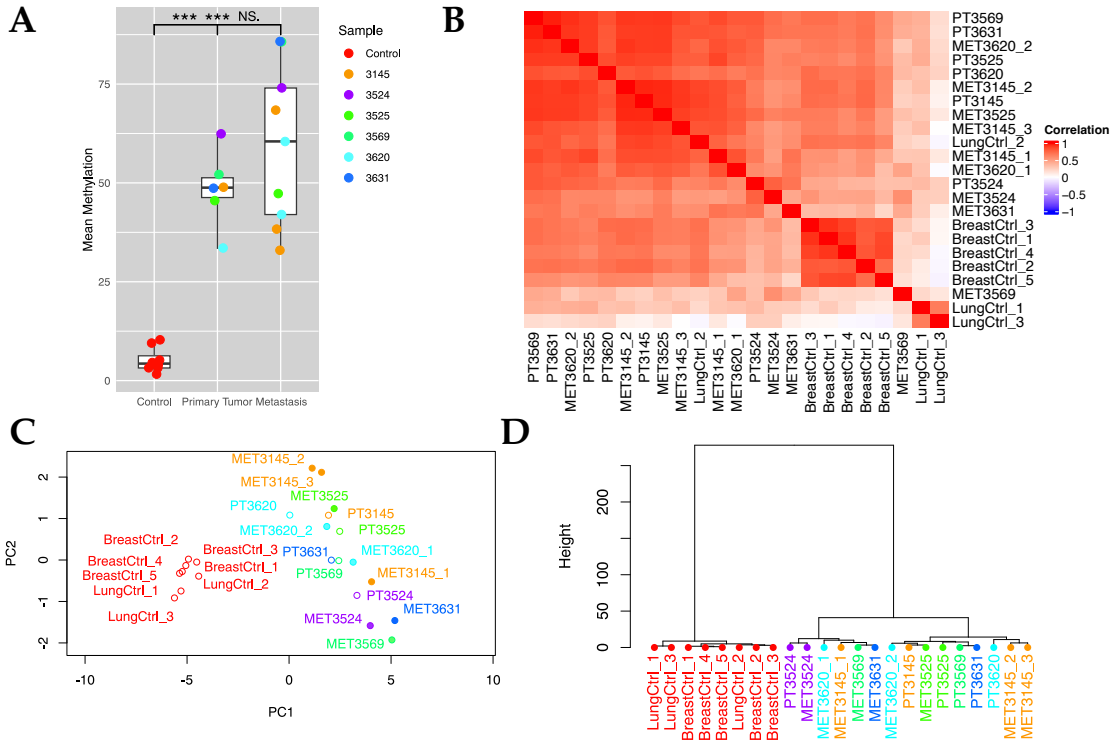


Figure 4.20.: General analysis of methylation in region D. (A) Boxplots showing mean methylation values for control, primary tumor and metastatic tissues in region D. Mean methylation showed significant increase in Ctrl vs. PT and Ctrl vs. MET (p-values = <.0001 & <.0001, linear mixed model with tukey method for comparing a family of 3 estimates) and no significant change from PT vs. MET (p-value = 0.3087, linear mixed model with tukey method for comparing a family of 3 estimates). (B) Heatmap displaying the correlation between each sample. (C) PCA analysis indicates no significant clustering of samples. Neither primary tumors nor metastasis show signs of unity. Matched primary tumors and metastases are uniquely colored. Controls and primary tumors are represented by unfilled points, while metastasis samples are represented by filled points. (D) Clustering of CpG methylation does not indicate a distinct separation between primary tumors and metastasis samples (distance measure= “manhattan”, agglomeration method= “ward”).

The amplicon analyzed in region D contained 37 CpGs and is located in exon 2 of the gene *C2cd4d*. Despite being situated in an exon, the overall methylation was low. Nevertheless, the low methylation status, especially in the control tissues (mean

methylation = 5.2%), allowed to the exclusion of low methylated reads within the primary and metastases samples (Figure B.5). After subtracting, the primary tumor and corresponding metastasis samples showed an average methylation value of 48.5% and 59.4%, with no significant change between these two sample types (Figure 4.20A). Correlation analysis revealed a distinct difference between two lung tissue controls and the rest of the samples. This could be explained by the fact that the malignant tissues originated from the mammary gland tissue of the breast and then spread to the lung. The breast tissue controls also clustered together and showed a high correlation. Most samples showed high correlation coefficients, except for PT3524, MET3524, MET3631, and MET3569. These samples had a higher or lower CpG methylation pattern, so they were clustered at the bottom right of the correlation heatmap (Figure 4.20B). Principal component analysis showed no significant separation between primary tumor and metastasis samples but an equal distribution throughout the two components, but most of the sample pairs clustered close together, indicating a close relationship (e.g. 3524, 3525) (Figure 4.20C). The two first components could explain most of the variances (Figure B.1D). Clustering the samples based on CpG methylation did not differentiate between primary tumors and metastases. Similar to PCA and correlation analysis, some animals clustered closely (3524, 3525), whereas others (3631) clustered further apart (Figure 4.20D). Overall, these results emphasize the heterogeneity of methylation in the context of breast tumor progression.

Nevertheless, the region contained minimal highly methylated reads in all samples. Therefore, those reads were extracted, and the phyloepigenetic approach was used to investigate the methylation patterns in detail. In sample 3145, the primary tumor sample showed higher methylated reads than two of the three corresponding metastatic samples. 37% of the primary tumor reads could be found in the cluster comprising the highest methylated reads, whereas only 16.3% and 11.6% of metastasis reads from samples MET3145 2 and MET3145 3 were located in that cluster. MET3145 1 showed a clear shift toward higher methylation compared to the primary tumor,

which might suggest that this metastasis has undergone epigenetic changes, favoring hypermethylation, with 75.3% of its reads located in cluster I (Figures 4.21 and 4.23 (a)). On the other hand, MET3145 2 and MET3145 3 showed a decrease in higher methylation (4.23 (a))) and an increase in lower methylation (Figures 4.23 (b) and 4.23 (c)). This suggests a different progression path where the DNA methylation pattern is more diverse. This could be explained as exon regions often exhibit a loss in genome-wide DNA methylation in cancer progression. No single highly methylated sequenced reads could be found in the primary tumor, as more than one-third of primary tumor reads could be found in the highly methylated cluster. The distribution plot of the reads, according to their mCpG count, further strengthened this observation (Figure 4.24A). Sample 3524 showed increased methylation from the primary tumor sample to the corresponding metastasis (62.4% to 74%). The methylation pattern visualized in the phyloepigenetic tree (Figure 4.21) revealed that the primary tumor reads were mainly clustered and located at the beginning of the clusters. MET3524 shows a strong shift toward higher methylation, with an increase from 15.1% in the primary tumor to 48.0% in the metastasis in cluster I (Figures B.6A and 4.23 (b)). The decrease in lower methylation in the metastasis further supported the trend toward increased methylation. This pattern suggests a clear progression toward higher methylation as the tumor metastasizes. Nevertheless, the sequenced reads from the primary tumor showed an equal distribution throughout the phyloepigenetic tree, with no identification of single highly methylated reads. This result was further emphasized with the arrangement of the reads according to their mCpG count (Figure 4.24B). In sample 3525, the metastasis showed a modest increase in higher methylation from the primary tumor to the metastasis, which indicated epigenetic changes favoring hypermethylation. 16% compared to 6.6% of primary tumor reads were located in cluster I (Figure B.6B). Interestingly, the metastasis showed lower methylated reads as its corresponding primary tumor, as 74.2% of its reads were located in cluster II comprising the lowest methylated reads. This suggests a divergence toward more extreme methylation states. Again, the phyloepigenetic tree and the distribution of the

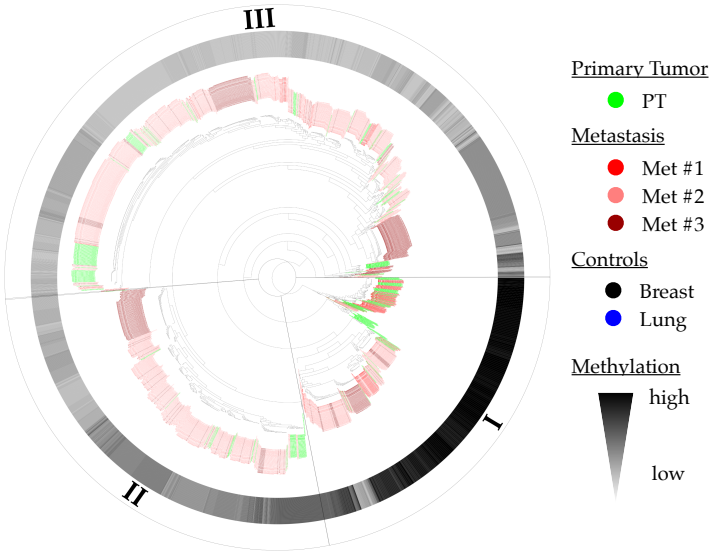
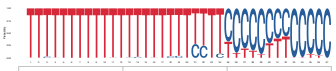


Figure 4.21.: DNA methylation analysis in region D at single base resolution. The phyloepigenetic tree for sample 3145 is shown as a representive for region D. The DNA methylation pattern observed in primary tumors exhibits variations in methylation progression within region D.



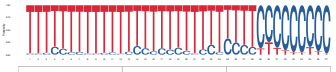
PT3145	116	37
MET3145 1	143	75.3
MET3145 2	219	16.3
MET3145 3	47	11.6

(a) 3145 Cluster I



PT3145	48	15.3
MET3145 1	1	0.5
MET3145 2	418	31.0
MET3145 3	108	26.7

(b) 3145 Cluster II



PT3145	149	47.6
MET3145 1	46	24.2
MET3145 2	710	52.7
MET3145 3	250	61.7

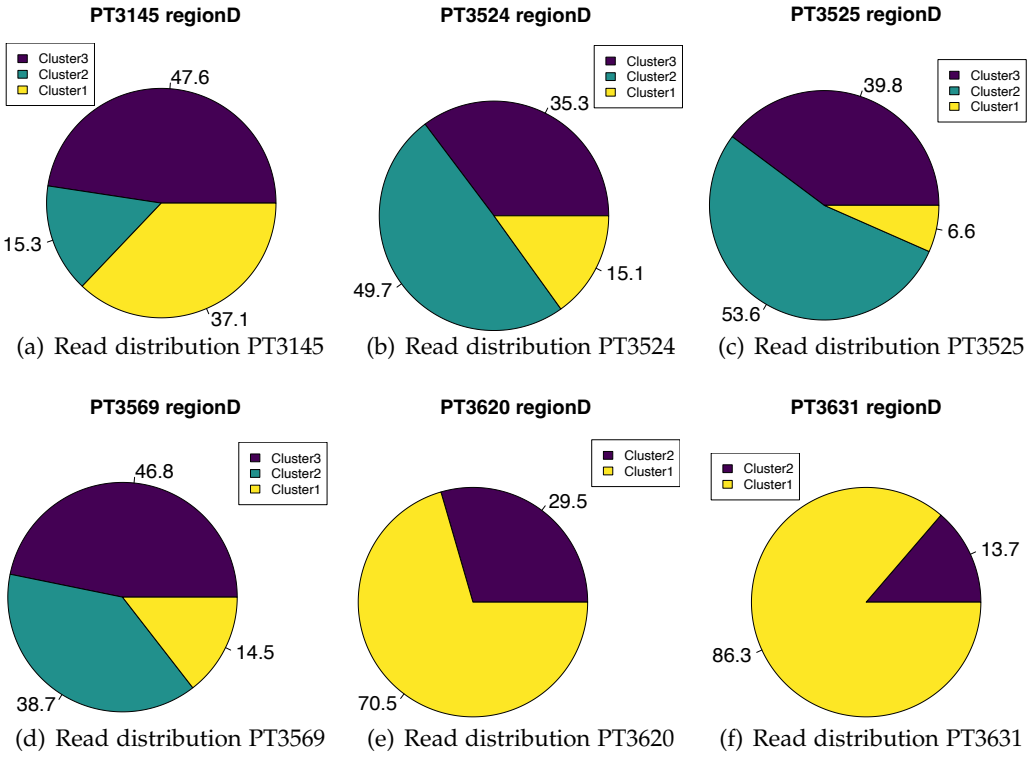
(c) 3145 Cluster III

Figure 4.22.: The sequence logos summarize the base composition of each cluster for sample 3145 in region D. The tables display the sample ID (left column), the number of reads within each cluster per sample (middle column), and the overall percentage of reads per sample (right column).

reads based on their mCpG count revealed an equal distribution of the primary tumor reads with no single highly methylated reads (Figures B.6B and 4.24C). Within sample 3569, an increase in methylation toward the metastasis could have been observed. MET3569 shifted toward hypermethylation compared to the primary tumor (89.5% vs. 14.5% reads in cluster I). This was coupled with a significant reduction of reads in both moderately methylated in cluster II and lower methylated in cluster III (Figure B.6C). This pattern reflected a clear progression toward higher DNA methylation in

the metastatic site, indicating that as the tumor advances, the epigenetic landscape becomes increasingly dominated by hypermethylation. The primary tumor did not have a single cluster dominated by highly methylated reads. Instead, the reads were more equally distributed, with a larger proportion falling into the moderate cluster II and lower cluster III methylation categories. However, cluster I (higher methylation) had the smallest share of primary tumor reads, indicating that hypermethylation was less prevalent in the primary tumor compared to the other clusters (Figure 4.23 (d)). The distribution suggested that the primary tumor had a more balanced or less extreme methylation profile, with a greater amount of reads in lower and moderate methylation levels (Figures B.6C and 4.24D). PT3620 showed a strong dominance of lower methylated reads in cluster I (70.5%) with proportions of higher methylation in cluster II (29.5%) (Figures B.6D and 4.23 (e)). MET3620 1 exhibited higher methylation, with 39.6% of its reads located in cluster II. It moved toward a more methylated state, with increased methylation. MET3620 2, on the other hand, was diversifying its methylation landscape with a more balanced distribution across the clusters (Figure B.6D). The phyloepigenetic tree, as well as the distribution plot based on increased mCpG count, showed an equal distribution of sequenced reads in the primary tumor (Figures B.6D and 4.24E). Sample 3631 showed an overall increase in methylation in the progression from primary tumor to metastasis. The mean methylation values were 48.7% for the primary tumor and 85.8% for the corresponding metastasis. The distribution and the phyloepigenetic tree revealed that only a small proportion (5.4%) of the reads from the primary tumor were located in the cluster containing those reads with the highest methylation (Figures B.6E and 4.23 (f)). The rest of the reads showed an equal distribution (Figure 4.24F). The metastasis showed a shift, with an increase in higher methylation (83.9%) and a significant reduction in lower methylation (16.1%) (Figure B.6E). The higher methylation reads in Cluster I of the primary tumor (PT3631) could be considered as single rare methylated clones within the overall distribution of reads. This is due to their relatively small proportion (5.4%) compared to the dominant lower methylation reads in cluster II (94.6%). These reads likely represent specific clonal

events or subpopulations within the primary tumor, reflecting a unique methylation profile that is not widespread across the entire tumor. These observations strengthen the hypothesis of selection and the contribution of high-level methylation to metastatic progression. In summary, region D showed a heterogeneous pattern in methylation progression throughout the primary tumors analyzed (Figure 4.23).



(a) Read distribution PT3145 (b) Read distribution PT3524 (c) Read distribution PT3525

(d) Read distribution PT3569 (e) Read distribution PT3620 (f) Read distribution PT3631

Figure 4.23.: Pie charts show how sequenced reads are distributed across all primary tumor samples from region D in each individual cluster. These charts clarify the read distribution that reflects the phyloepigenetic trees.

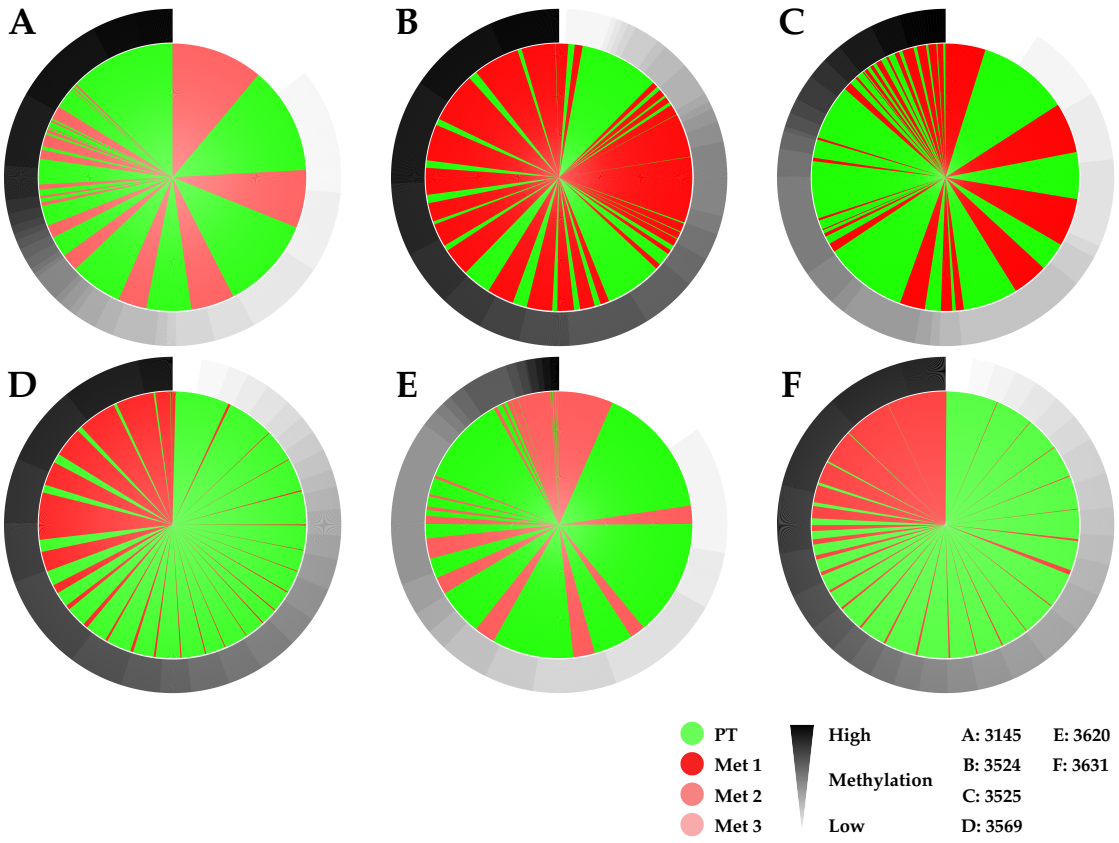


Figure 4.24.: Arrangement of reads in region D according to their mCpG count. The reads are plotted according to their mCpG count in increasing order and displayed in a circular format. A gradient ring surrounding the circle indicates the methylation status.

4.11. Methylation analysis in region E

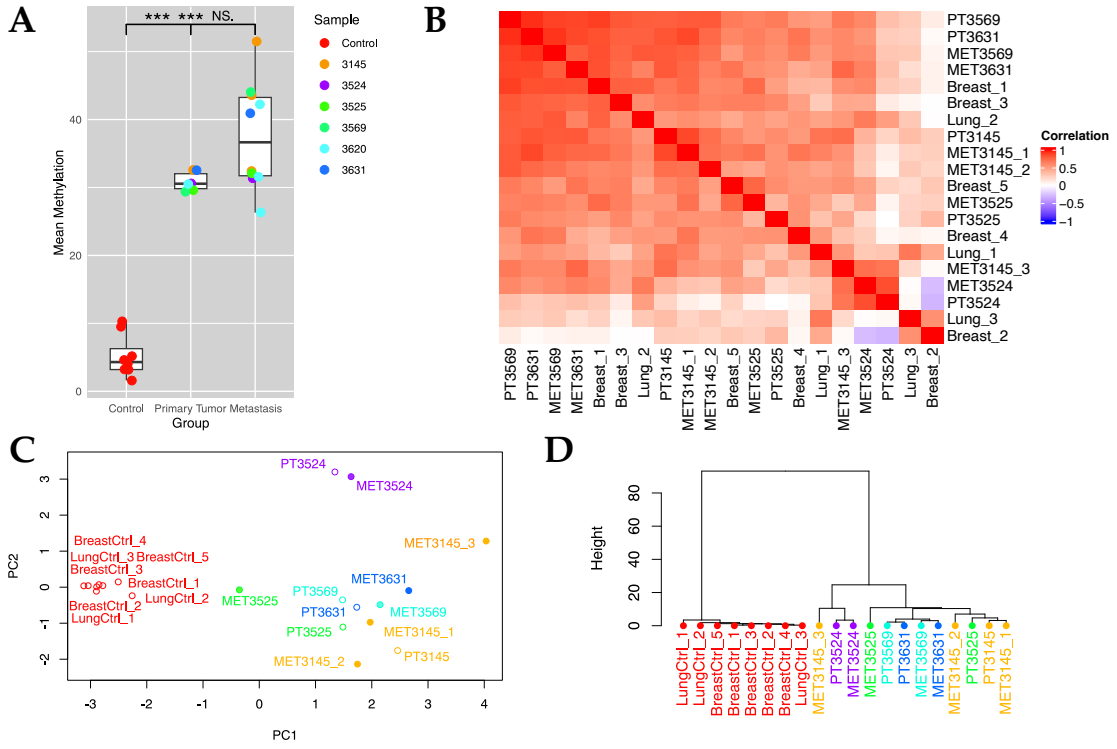


Figure 4.25.: General analysis of methylation in region E. (A) Boxplots showing mean methylation values for control, primary tumor and metastatic tissues in region E. Mean methylation showed significant increase in Ctrl vs. PT and Ctrl vs. MET (p-values = <.0001 & <.0001, linear mixed model with tukey method for comparing a family of 3 estimates) and no significant change from PT vs. MET (p-value = 0.0532, linear mixed model with tukey method for comparing a family of 3 estimates). (B) Heatmap displaying the correlation between each sample. (C) PCA analysis indicates no significant clustering of samples. Neither primary tumors nor metastasis show signs of unity. Matched primary tumors and metastases are uniquely colored. Controls and primary tumors are represented by unfilled points, while metastasis samples are represented by filled points. (D) Clustering of CpG methylation does not indicate a distinct separation between primary tumors and metastasis samples (distance measure= "manhattan", agglomeration method= "ward").

The promoter region of gene *Xfr4* was investigated in region E, and DNA methylation analyses revealed that this region showed a prominent pattern of unmethylated CpGs across the entire locus. Mean methylation values were low but significantly increased

between the controls and the two malign groups. Therefore, the low-methylated reads were subtracted from the primary tumor and metastases samples. Supplementary Figure B.7 displays the phyloepigenetic trees before the subtraction of low-methylated reads. With subtracting, the mean methylation of the primary tumor changed from 9.8% to 30.8 %, and from 12.8% to 37.8% in the metastatic samples. However, the methylation profile was similar between the primary tumor lesions and their metastatic sites, with no significant difference, but showed a wider distribution within metastases (Figure 4.25A). Pearson correlation with subsequent hierarchical clustering revealed that most of the sample pairs clustered close to each other, suggesting a close relationship between primary tumors and metastases for each sample. This leads to the assumption that the pattern in DNA methylation is more sample-specific rather than increasing over time and proliferation. The PCA analysis further strengthened this observation, as no clear distribution between the samples could be observed. Moreover, the metastatic samples clustered closer to their corresponding primary tumors as to each other (Figure 4.25C). Clustering the samples with distance method "manhattan", and agglomeration method "ward" showed the same result. Most of the sample pairs clustered nearby (Figure 4.25D).

Using again the phyloepigenetic approach to identify highly methylated clones within the primary tumor revealed that in sample 3145 most of the primary tumor reads were accumulated in clusters II and III, comprising reads with moderate to low methylation. None of the reads were found in cluster I, harboring the highest methylation reads. This cluster only contained reads from MET3145 3 (Figures 4.26 and 4.27). This sample was also the sample with the highest mean methylation in all samples analyzed (51.5%). The data indicated a mild progression in methylation patterns from the primary tumor to metastases, with an increasing presence of moderate methylation in the metastases and a potential emergence of high methylation in MET3145 3. Nevertheless, neither the phyloepigenetic tree nor the distribution plot based on mCpG counts identified single highly methylated reads within the primary tumor (Figures 4.26 and 4.29). The reads

were equally distributed throughout the plots. In sample 3524, the primary tumor and the metastasis were characterized by low methylation, harboring most of the reads in cluster II. The primary tumor showed no high methylation at all, while the metastasis exhibited a very minor increase in high methylation (Cluster I), though it remained low-methylated (Figures B.8A and 4.28 (b)). The minimal change in methylation patterns from the primary tumor to the metastasis suggests that this tumor's progression did not involve significant alterations in DNA methylation, with the methylation landscape remaining largely stable. PT3525 displayed a mixed methylation profile with a notable presence of higher and lower methylation, indicating some diversity in the methylation states within the primary tumor. MET3525, however, shifted toward a more homogenous lower methylation profile, reducing the higher methylation reads (Figures B.8B and 4.28 (c)). This could suggest a loss of the higher methylation clones during metastasis, leading to a more uniformly low-methylated state. Both the primary tumor, PT3569 and the corresponding metastasis, MET3569 exhibited 100% of their reads in cluster I, representing moderate methylation (Figures B.8C and 4.28 (d)). This suggested that the methylation pattern was consistent between the primary tumor and the metastasis, with no variation in methylation levels. The consistency in the methylation pattern indicates no significant epigenetic change as the tumor progresses from the primary site to the metastatic site. This uniformity suggested that the tumor maintains a stable, moderate methylation state throughout its development and spread. The primary tumor, PT3620 and all corresponding metastases (MET3620 1, MET3620 2, MET3620 3) were characterized by a consistent and uniform moderate methylation pattern, with all reads falling into cluster I (Figures B.8D and 4.28 (e)). The same applies for primary tumor PT3631 and its corresponding metastasis MET3631 (Figures B.8E and 4.28 (f)). In summary, region E did not serve as an appropriate candidate to investigate methylation progression as most reads exhibited low methylation levels in primary tumor samples (Figures 4.28 and 4.29). Furthermore, there was only a marginal change in methylation surrounding one specific CpG.

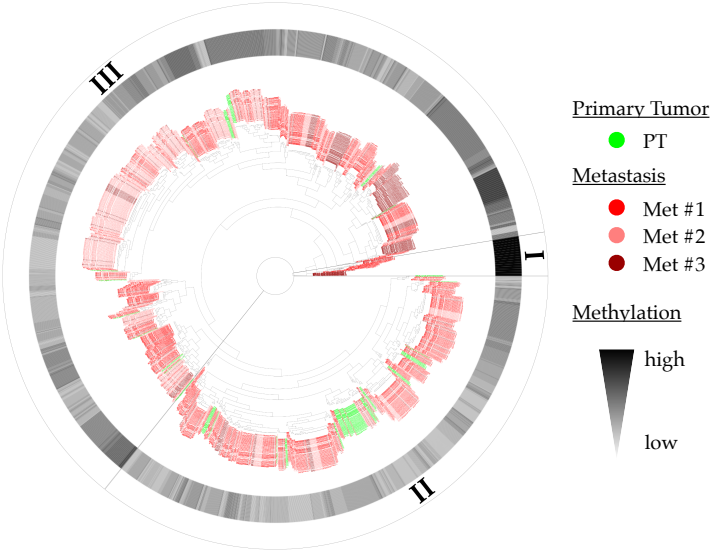


Figure 4.26.: DNA methylation analysis in region E at single base resolution. The phyloepigenetic tree for sample 3145 is shown as a representive for region E. The DNA methylation pattern observed in primary tumors exhibits variations in methylation progression within region E.



PT3145	-	0
MET3145_1	-	0
MET3145_2	-	0
MET3145_3	16	11.3

(a) 3145 Cluster I



PT3145	25	26.6
MET3145_1	482	52.1
MET3145_2	352	91.4
MET3145_3	116	81.7

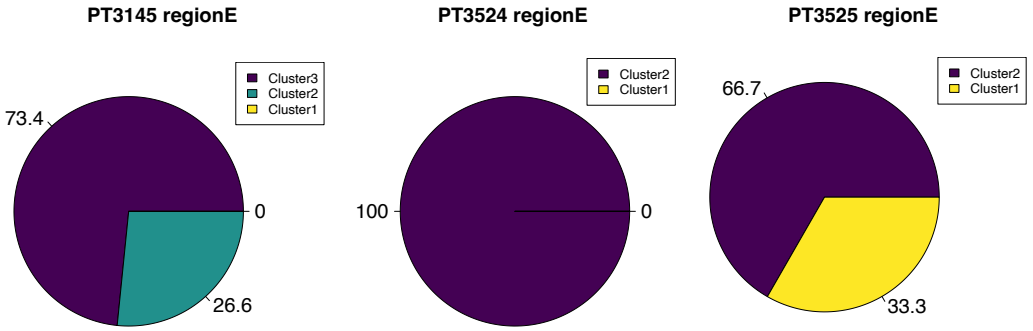
(b) 3145 Cluster II



PT3145	69	73.4
MET3145_1	443	47.9
MET3145_2	33	8.6
MET3145_3	10	7

(c) 3145 Cluster III

Figure 4.27.: The sequence logos summarize the base composition of each cluster for sample 3145. The tables display the sample ID (left column), the number of reads within each cluster per sample (middle column), and the overall percentage of reads per sample (right column).



(a) Read distribution PT3145 (b) Read distribution PT3524 (c) Read distribution PT3525

(d) Read distribution PT3569 (e) Read distribution PT3620 (f) Read distribution PT3631

Figure 4.28.: Pie charts show how sequenced reads are distributed across all primary tumor samples from region E in each individual cluster. These charts clarify the read distribution that reflects the phyloepigenetic trees. Within this region, all primary tumors exhibited an overall low methylation profile.

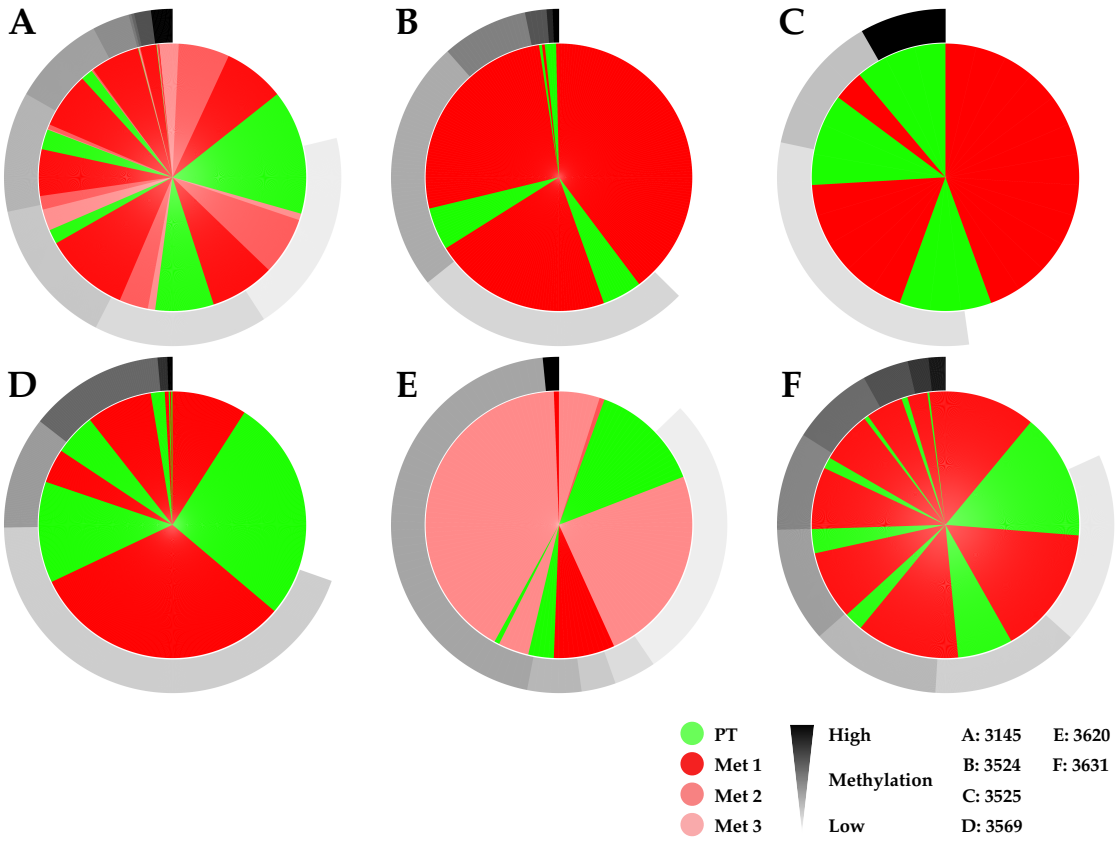


Figure 4.29.: Arrangement of reads in region E according to their mCpG count. The reads are plotted according to their mCpG count in increasing order and displayed in a circular format. A gradient ring surrounding the circle indicates the methylation status.

4.12. Methylation analysis in region F

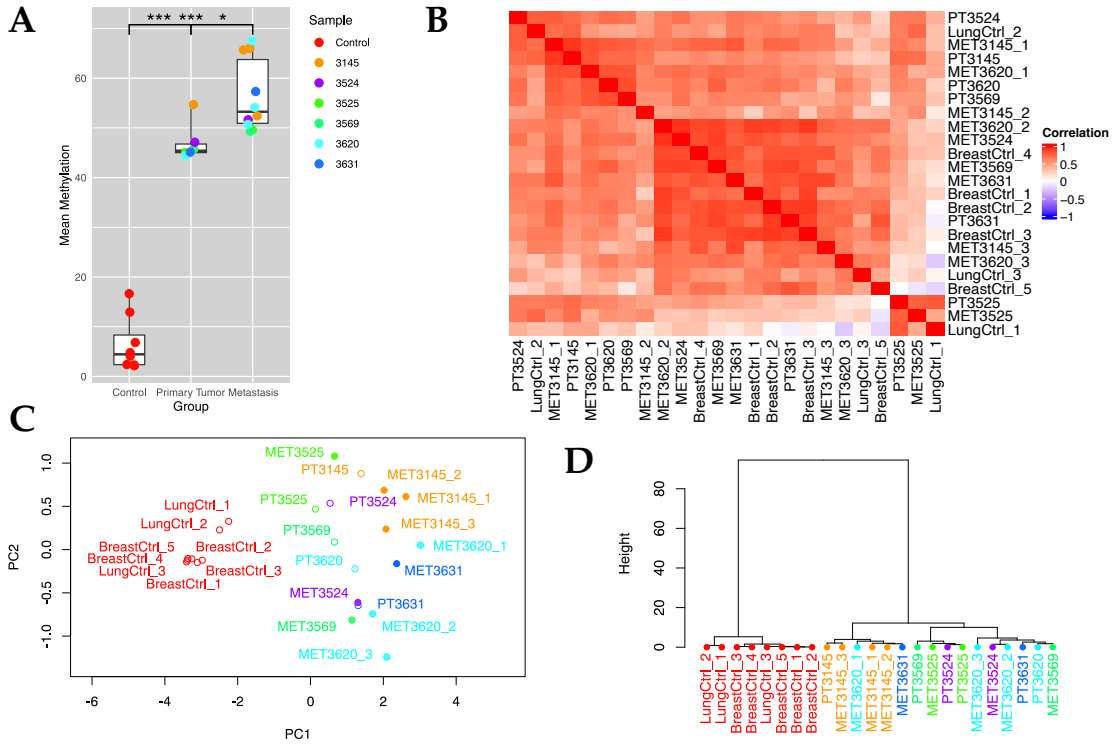


Figure 4.30.: General analysis of methylation in region F. (A) Boxplots showing mean methylation values for control, primary tumor and metastatic tissues in region F. Mean methylation showed significant difference between each sample type (Ctrl vs. PT, Ctrl vs. MET, PT vs. MET; p-values = <.0001 & <.0001 & 0.0165, linear mixed model, followed by tukey method for comparing a family of 3 estimates). (B) Heatmap displaying the correlation between each sample. (C) PCA analysis indicates clustering of metastases samples. Matched primary tumors and metastases are uniquely colored. Controls and primary tumors are represented by unfilled points, while metastasis samples are represented by filled points. (D) Clustering of CpG methylation does not indicate a distinct separation between primary tumors and metastasis samples (distance measure= "manhattan", agglomeration method= "ward").

Region F was generally unmethylated across all samples analyzed, including the control samples. Therefore, the proportion of low-methylated reads was subtracted from the primary tumor and metastases samples according to the methylation patterns of the breast and lung tissue controls (Figure B.9). The primary tumor and metastasis

samples showed a minor shift in mean methylation values, with a significant difference (Figure 4.30A). The metasatases samples tended toward hypermethylation, when being compared to their primary tumors (47% vs. 56.4% mean methylation). Calculating the correlation coefficients with subsequent hierarchical clustering revealed that primary tumors and metastases were evenly divided (Figure 4.30B). In general, the correlation between each sample was distinct within this region. Interestingly, PCA analysis revealed that the metastatic samples clustered more on the right side of the plot. In contrast, the primary tumors were more arranged within the middle of the two components plotted. This would suggest a progressive gain of DNA methylation toward the metastases, as the control tissue samples further clustered on the left side of the plot (Figure 4.30C). Most of the variance could be explained through components 1 & 2 (Figure B.1F). When clustering the CpG methylation based on "manhattan" distance, rather than correlation, it showed a clear distinction of the control tissues from the malign samples. Within the groups of primary tumors and metastases no separation of either group could be observed. Interestingly, within this region, no sample pair showed signs of unity (Figure 4.30D).

The phyloepigenetic approach and the arrangement based on mCpG count showed that all primary tumors displayed a minimal consistent increase in DNA methylation, with no noticeable single-cell subclones exhibiting an abnormally high methylation pattern. This could be further seen by the fact that it was not possible to determine any specific delimitation when finding clusters within the phyloepigenetic trees (Figure B.10). In summary, the findings prove that every primary tumor exhibited an evenly distributed minimal increase in DNA methylation in region F, with no distinct single-cell subclones displaying a unique high methylation pattern (Figure 4.31). A binomial test (uniform increase vs. single rare highly methylated reads) yielded a highly significant p-value of 0.007812, as all six tumors exhibited a consistent minor increase in DNA methylation.

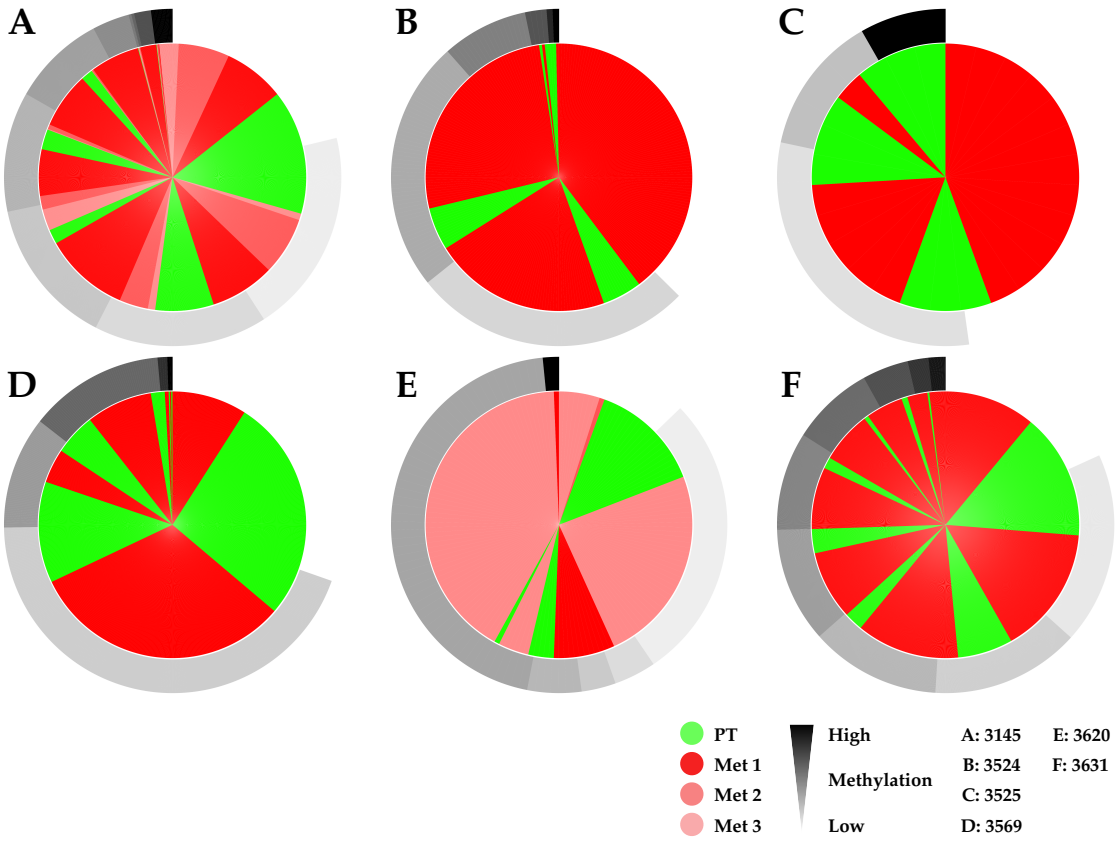


Figure 4.31.: Arrangement of reads in region F according to their mCpG count. The reads are plotted according to their mCpG count in increasing order and displayed in a circular format. A gradient ring surrounding the circle indicates the methylation status.

4.13. Methylation analysis in region G

The MCIp data generated during the initial funding period of FOR2127 has provided insights into the epigenetic landscape of gene *Boll* in metastatic breast cancer progression. The data revealed a gradual increase in CpG methylation proceeding from the promoter transcription start toward the gene body of gene *Boll*, particularly in samples derived from lung metastases. To investigate the methylation progression, the PCR product generated for analysis was set at the end of the CpG island, including a total of 12 CpGs for analysis. Unfortunately, a decrease to the 3′ end was observed in all samples. The PCR primer design algorithm placed the PCR product too close to the end of the

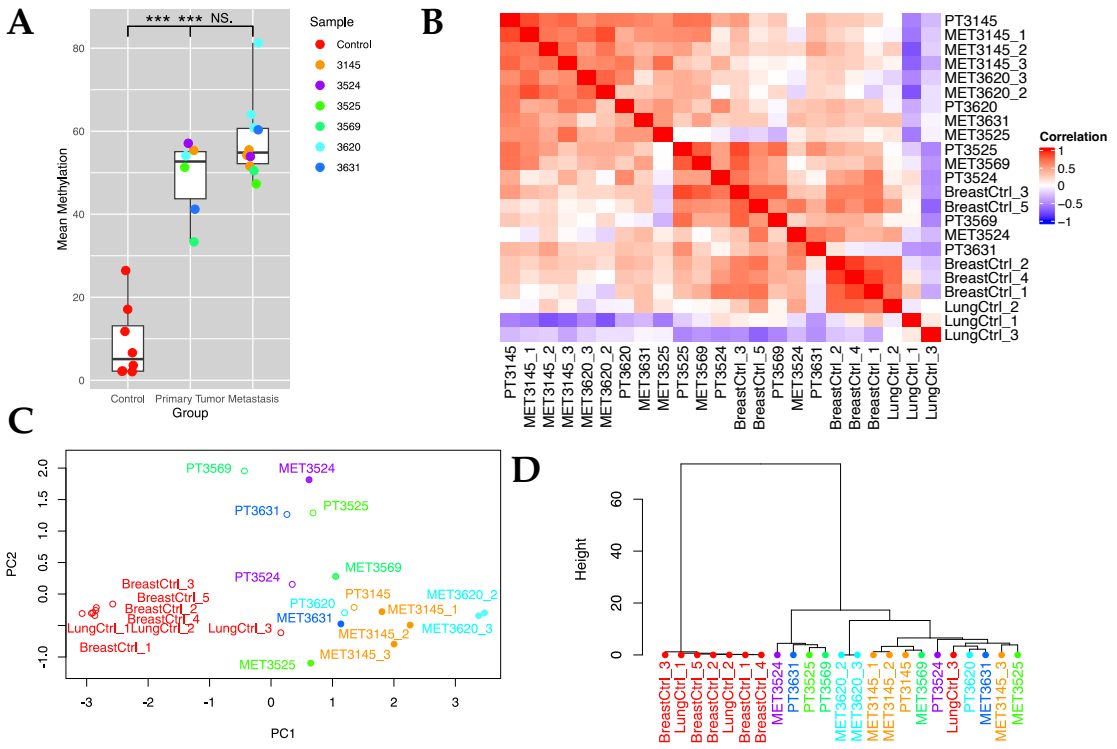


Figure 4.32.: General analysis of methylation in region G. (A) Boxplots showing mean methylation values for control, primary tumor and metastatic tissues in region G. Mean methylation showed significant difference between Ctrl vs. PT & Ctrl vs. MET (p-values = <.0001 & <.0001 & 0.9959, linear mixed model, followed by tukey method for comparing a family of 3 estimates), and no significance between PT & Ctrl (p-values = 0.1617, linear mixed model, followed by tukey method for comparing a family of 3 estimates). (B) Heatmap displaying the correlation between each sample. (C) PCA analysis indicates no significant clustering of samples. Neither primary tumors nor metastasis show signs of unity. Matched primary tumors and metastases are uniquely colored. Controls and primary tumors are represented by unfilled points, while metastasis samples are represented by filled points. (D) Clustering of CpG methylation does not indicate a distinct separation between primary tumors and metastasis samples (distance measure= “manhattan”, agglomeration method= “ward”).

CpG island. Unfortunately no better suitable PCR product could have been generated. The MCIp data also showed a decrease in methylation at the end of the analyzed locus. Nevertheless, as the breast tissue controls were generally low methylated, the reads harboring the same methylated pattern were subtracted from the primary tumor

and metastases samples (Figure B.11). However, mean methylation values showed a non-significant difference between primary tumor and metastasis samples (PT= 48.7% vs. MET= 57.9%) (Figure 4.32A). This suggests a high similarity between the sample pairs that were analyzed. The high similarity could also be observed in the correlation heatmap; many sample pairs were clustered together (e.g., 3145, 3620, 3525, 3524) (Figure 4.32B). PCA analysis showed that most metastatic samples were clustered on the bottom right of the plot. This could indicate a minor DNA methylation increase and close relationship between the metastasis samples (Figure 4.32C). Again, most of the variances could be explained through components 1 and 2 (Figure B.1G). When clustering the samples with "manhattan" distance and "ward" agglomeration, the samples were homogenously distributed throughout the tree, with no signs of clustering between sample pairs or primary tumor and metastatic entities (Figure 4.32D).

When analyzing the subtracted samples with the phyloepigenetic approach and the arrangement based on mCpG count, the CpGs chosen for the analysis were too homogeneous and still low-methylated (Figure B.12). None of the primary tumor samples showed a significantly higher methylation profile compared to the rest of the primary tumors or the metastasis samples (Figure B.12). Similar to region F the minor increase in DNA methylation within the primary tumors did not reveal any single highly methylated clones, however, showed more characteristics of equal distribution (Figure 4.33). In conclusion, despite being a promising candidate region to investigate methylation progression from primary tumor lesion to corresponding metastasis, the PCR product location was poorly chosen. It is worth noting that primer design and generating an appropriate PCR product was challenging within this region.

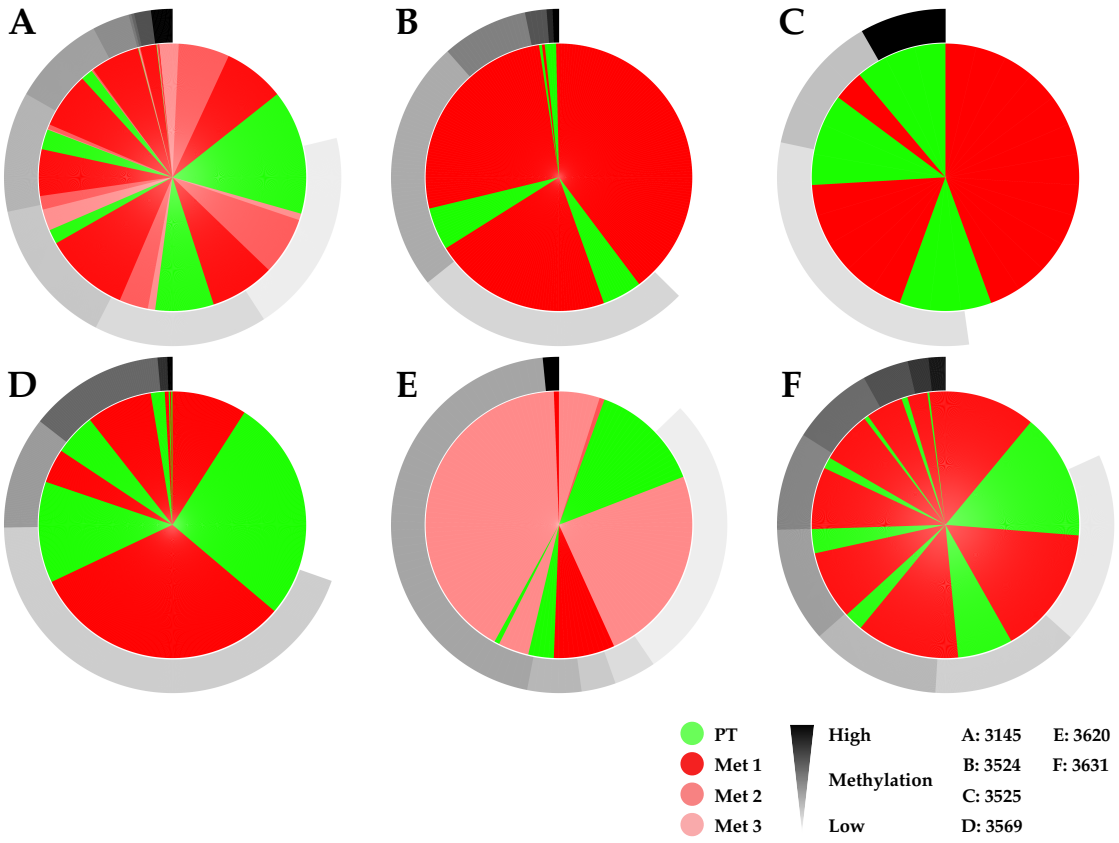


Figure 4.33.: Arrangement of reads in region G according to their mCpG count. The reads are plotted according to their mCpG count in increasing order and displayed in a circular format. A gradient ring surrounding the circle indicates the methylation status.

4.14. Methylation analysis in region J

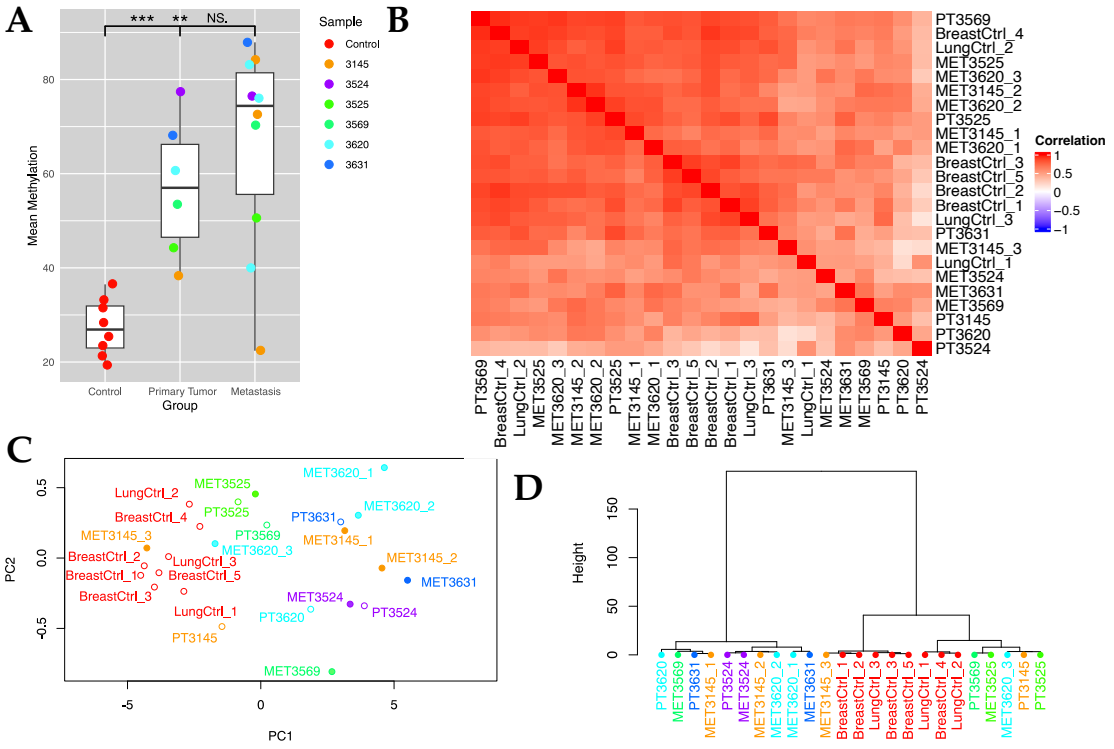
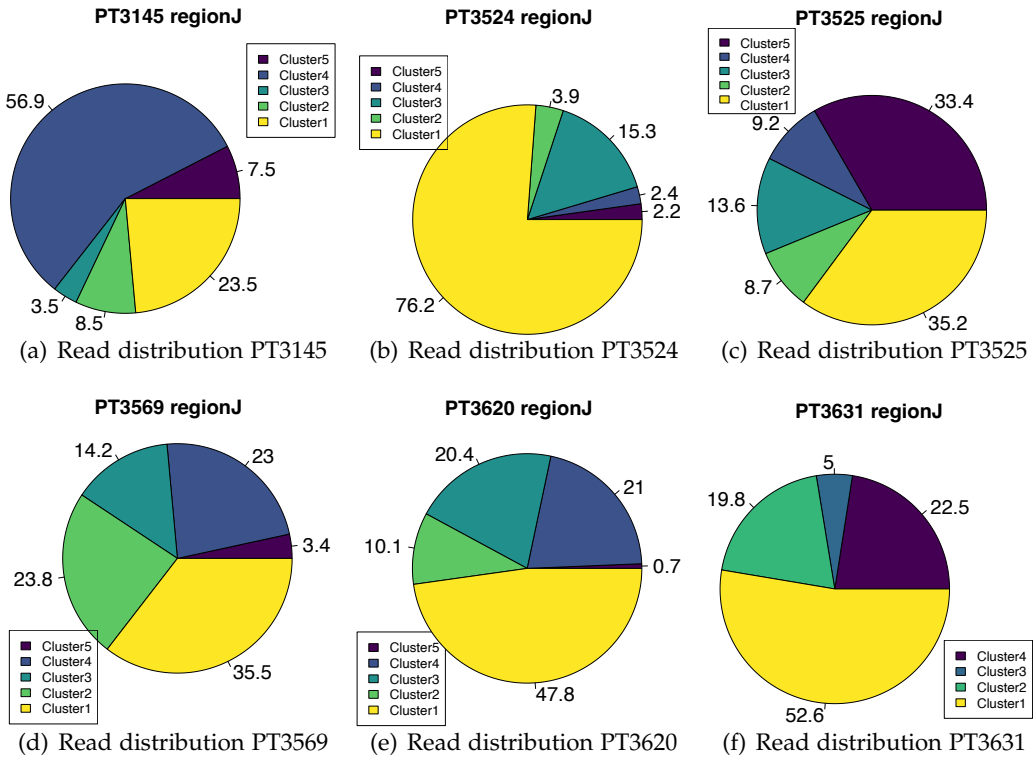


Figure 4.34.: General analysis of methylation in region J. (A) Boxplots showing mean methylation values for control, primary tumor and metastatic tissues in region J. Mean methylation showed significant differences in Ctrl vs. PT and Ctrl vs. MET (p-values = 0.0076 & 0.0001, linear mixed model, followed by tukey method for comparing a family of 3 estimates) and no significant change from PT vs. MET (p-value = 0.5138, linear mixed model, followed by tukey method for comparing a family of 3 estimates). (B) Heatmap displaying the correlation between each sample. (C) PCA analysis indicates no significant clustering of samples. Neither primary tumors nor metastasis show signs of unity. Matched primary tumors and metastases are uniquely colored. Controls and primary tumors are represented by unfilled points, while metastasis samples are represented by filled points. (D) Clustering of CpG methylation does not indicate a distinct separation between primary tumors and metastasis samples (distance measure= “manhattan”, agglomeration method= “ward”).

A total of 24 CpGs were analyzed within region J. The PCR product lay in the middle of the region, which spanned 1.6kb and contained 2 CpG islands. The MCIp data showed an overall methylated profile within metastases and a heterogeneous pattern in

methylation in control tissues and primary tumor samples (Figure 4.1J). This observation was further strengthened with the BBA-seq analysis. Nevertheless metastasis samples with a low methylation profile were identified. The average mean methylation value for the control samples was 27.4%. Primary tumors exhibited 57% and metastases 66.36%, respectively. Some primary tumors with low methylation levels also developed metastases with low methylation (for example, PT3145 38.3%, MET3145 3 22.4%), while primary tumors with high mean methylation levels also led to metastases with higher methylation values (PT3524 77.4%, MET3524 76.4%). Overall mean methylation showed significant differences between control tissues and malign sample types, whereas primary tumors and metastases didn't change significantly. This suggests that within this region, the primary tumors were similar to their corresponding metastases in CpG methylation (Figure 4.34A). Clustering the samples based on their correlation coefficients revealed that no cluster formed; the heatmap showed an even distribution of correlation between all samples (Figure 4.34B). PCA analysis revealed that all samples were evenly distributed throughout the plot (Figure 4.34C). Component 1 could explain the majority of variances (Figure B.1H). Creating a dendrogram with "manhattan" distance revealed no distinct separation between primary tumors and metastases; the sample types were evenly dispersed throughout the tree. This underlines again the heterogeneity in CpG methylation in metastatic breast cancer progression (Figure 4.34D).

While analyzing the read distribution using the phyloepigenetic approach and clustering the reads based on their methylation profile, it was observed that the methylation was already so high in primary tumors that no rare, highly methylated single reads could be detected (Figure B.13). All primary tumor samples had over 14% of their reads in the cluster with the highest methylated reads. Primary tumor 3524 even had more than 76% highly methylated. In summary, the primary tumor samples had a very high methylation profile, which made this region unsuitable for discovering rare high methylated reads (Figure 4.35).



(a) Read distribution PT3145 (b) Read distribution PT3524 (c) Read distribution PT3525

(d) Read distribution PT3569 (e) Read distribution PT3620 (f) Read distribution PT3631

Figure 4.35.: Pie charts show how sequenced reads are distributed across all primary tumor samples from region J in each individual cluster. These charts clarify the read distribution that reflects the phyloepigenetic trees.

4.15. Methylation analysis in region L

Region L served as the positive methylation control. The PCR product spanned 25 CpGs in the 10th exon of the gene *Disp1*. MCIp data showed an overall high methylation profile of $\approx$ 100% (Figure 4.1L). The same was true for the samples analyzed with the BBA-seq approach. Control tissue showed a mean methylation value of 96.25%, primary tumors 96.61%, and metastases 96.35%, with no significant difference between each sample type (Figure 4.36A). Despite being very high methylated, the samples were generally uncorrelated (Figure 4.36B). Nevertheless, PCA showed homogenous clustering of most samples, indicating high similarity in 2D space (Figure 4.36C). Most of the variance could be explained through components 1 and 2, but this region showed

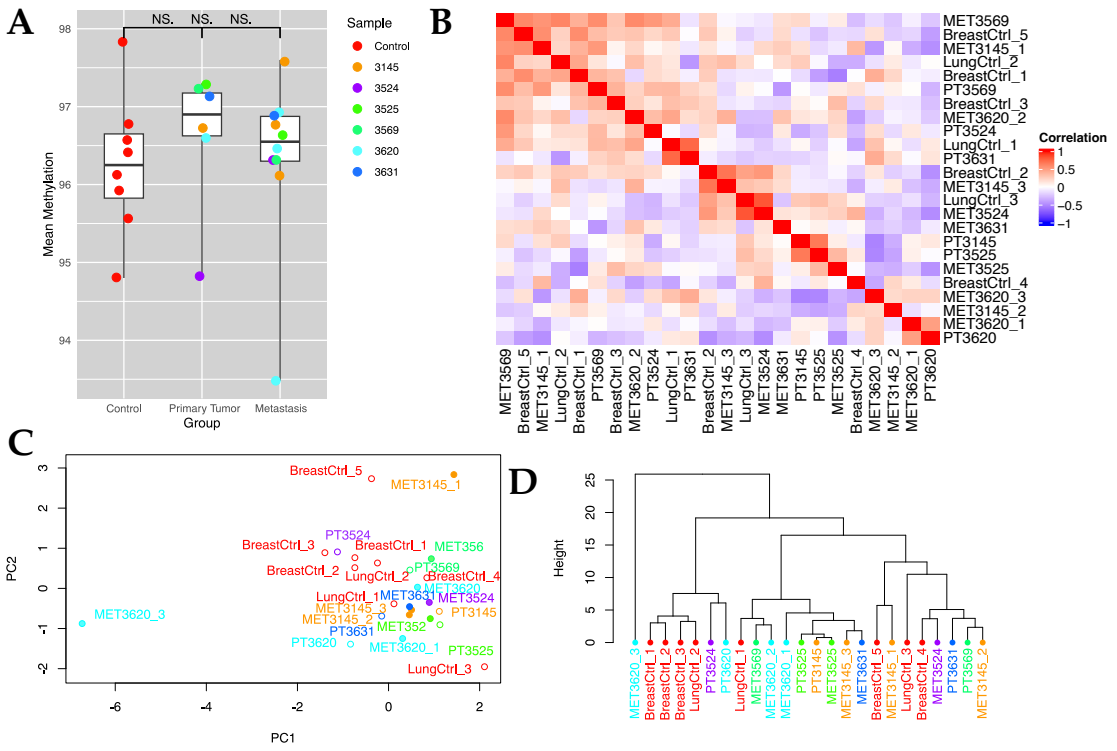
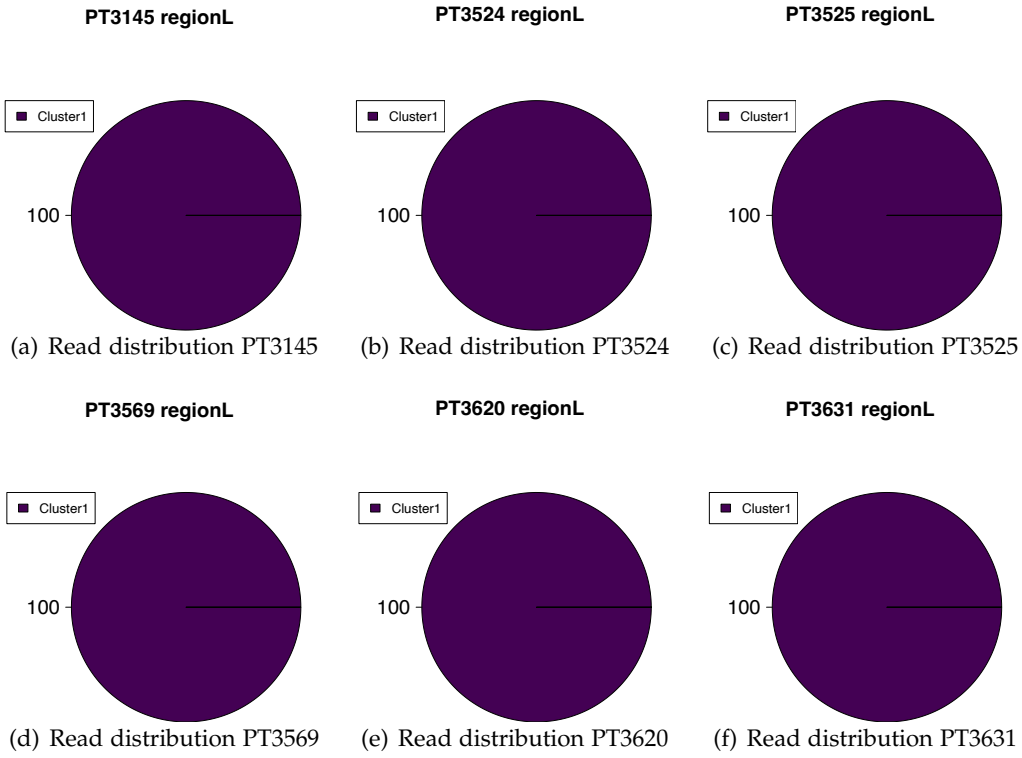


Figure 4.36.: General analysis of methylation in region L. (A) Boxplots showing mean methylation values for control, primary tumor and metastatic tissues in region L. Mean methylation showed no significant difference between all sample types analyzed (Ctrl vs. PT, Ctrl vs. MET, PT vs. MET; p-values = 0.7736, 0.9753 & 0.8613, linear mixed model, followed by tukey method for comparing a family of 3 estimates). (B) Heatmap displaying the correlation between each sample. (C) PCA analysis indicates homogenous clustering of samples. Matched primary tumors and metastases are uniquely colored. Controls and primary tumors are represented by unfilled points, while metastasis samples are represented by filled points. (D) Clustering of CpG methylation does not indicate a distinct separation between primary tumors and metastasis samples (distance measure= "manhattan", agglomeration method= "ward").

the highest variance in component 3. Analyzing component 3 in combination with 2 did not reveal any differences but instead showed the same result. The dendrogram tree clustering of the CpG methylation with the "manhattan" distance revealed neither a clustering of the primary tumors nor the metastases (Figure 4.36D). Although no discernible variation was detected between the analyzed samples due to the high CpG methylation feature exhibited by most of the reads, this region can be utilized as a

reliable methylation positive control. To summarize, region L served as an appropriate positive control region, as all samples displayed an even, highly methylated composition and a homogenous distribution of highly methylated reads (Figures 4.37 and B.14).



(a) Read distribution PT3145 (b) Read distribution PT3524 (c) Read distribution PT3525

(d) Read distribution PT3569 (e) Read distribution PT3620 (f) Read distribution PT3631

Figure 4.37.: Pie charts show how sequenced reads are distributed across all primary tumor samples from region L in each individual cluster. These charts clarify the read distribution that reflects the phyloepigenetic trees.

4.16. Identifying differentially methylated cytosines

The DNA methylation changes in cancer progression from the primary tumors toward their metastases showed a heterogeneous pattern. Capturing the variability within these changes helps to understand how stable or unstable the epigenetic changes are as the tumor progresses. Furthermore, identifying the most variable methylated CpGs in the different stages of the cancer could lead to the identification of specific biomarkers. Mitchell et al. used genome-wide methylation analysis and multiplexed bisulfite sequencing on colorectal cancer (CRC) tissue and normal tissue samples to identify DNA methylation biomarkers for early detection within CRC. They found 23 genes that showed elevated methylation in over 50% of CRC samples. They also postulated six genes that could serve as biomarkers for stool-based assays due to low methylation in normal tissue controls and eleven genes for blood-based diagnostic tests due to low methylation in peripheral blood DNA [158].

Identifying highly variable CpG sites between primary tumor and metastasis samples could also benefit this study. It could help to characterize the primary tumor better and make specific statements about the primary tumor stage. If a primary breast tumor biopsy or a blood sample with subsequent DNA methylation sequencing yields a high proportion of metastases-related, variable CpGs, it could be assumed that the tumor has already spread from the primary tumor site. Therefore, the differentially methylated cytosines could serve as potential biomarkers of metastasis. Additionally, in correlation with gene expression data, these cytosines could indicate a contribution to the downregulation of metastasis-related genes. Higher variability in the metastasis group could indicate epigenetic instability, which is often associated with cancer progression and the development of metastatic potential.

For this reason, regions with low methylation in control tissues and increased DNA methylation within the malign samples underwent differentially methylated cytosine analysis using the `calculateDiffMeth` function from the **R**-package `methylKit`[132].

The standard deviation was calculated and plotted to assess heterogeneity and variability. In detail, methylation percentages for each cytosine in every sample were assigned. Subsequently, the standard deviation of methylation percentages for each cytosine was calculated across the primary tumor and metastasis groups. Then, the standard deviation was visualized to compare the variability between the primary tumor and metastasis groups. In region B, eight cytosines could be identified that were hypermethylated by more than 25% in the metastasis group compared to the primary tumor group (Table 4.4). Within region D, there were 18 differentially methylated cytosines (Table 4.5). Within both sites, the variability in methylation was significantly higher in metastasis samples than in primary tumor samples (Figures 4.38 and 4.39). No differentially methylated cytosines could be detected within region E and in regions F and G, just one, so the analysis was not feasible. This further underlined the results obtained with the phyloepigenetic and general analysis as the DNA methylation pattern within these regions only changed in a marginal amount. In summary, highly variable and differentially methylated cytosines could be detected within regions B and D. These regions also reflected the most variable patterns in DNA methylation progression. Identifying those variable features could generate biomarker assays for breast cancer metastasis analysis.

CpG	chr	Position	Methylation difference in percentage
4	chr1	63273272	29.2976524158215
5	chr1	63273274	25.75129963382
21	chr1	63273352	39.0960903796944
22	chr1	63273355	40.9109569791542
23	chr1	63273361	25.6327688129689
24	chr1	63273366	30.5476777827499
25	chr1	63273371	28.2305821908981
35	chr1	63273423	30.9435859877868

Table 4.4.: Differentially methylated cytosines in region B between primary tumor samples and their corresponding metastases.

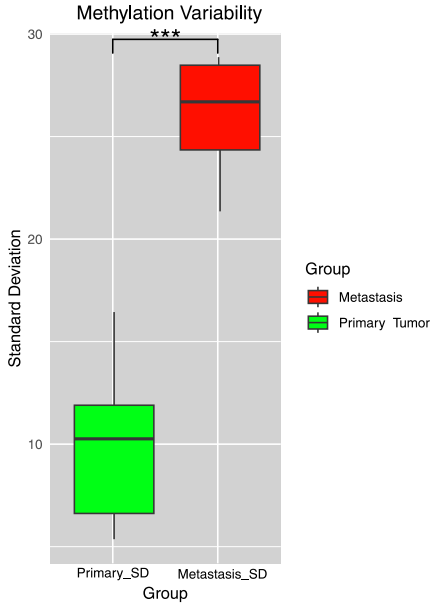


Figure 4.38.: Boxplots showing more variability within metastases compared to their corresponding primary tumors. The standard deviation change was highly significant between the two groups in region B (p-value = 5.606e-05, paired t-test).

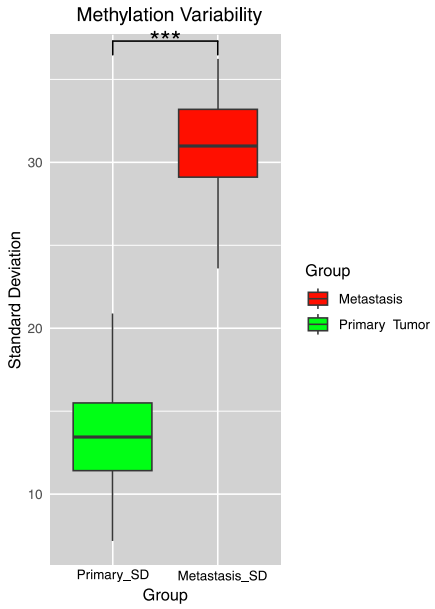


Figure 4.39.: Boxplots showing more variability within metastases compared to their corresponding primary tumors. The standard deviation change was highly significant between the two groups in region D (p-value = 2.298e-09, paired t-test).

CpG	chr	Position	Methylation difference in percentage
1	chr3	94364005	40.5638835853416
2	chr3	94364009	45.0713422692958
3	chr3	94364022	43.750515964842
4	chr3	94364029	35.2869267007377
5	chr3	94364041	31.1434518985756
7	chr3	94364046	40.2771712512192
8	chr3	94364056	25.7738383842187
9	chr3	94364067	27.6493990262301
11	chr3	94364085	44.7605748796839
14	chr3	94364112	26.1278709672435
15	chr3	94364115	29.7615508931156
16	chr3	94364128	27.4836203333044
17	chr3	94364132	33.2373032502528
18	chr3	94364134	27.1786039472699
20	chr3	94364145	38.1395985805914
21	chr3	94364151	33.1956700119599
24	chr3	94364169	27.0785792097575
30	chr3	94364218	26.2529160594741

Table 4.5.: Differentially methylated cytosines in region D between primary tumor samples and their corresponding metastases.

5. Results II: Altering de-novo DNA methylation capabilities

In mouse mammary epithelial cells, the de novo DNA methyltransferases DNMT3A/B play a critical role in the establishment of all de novo DNA methylation events, including the aberrant hypermethylation that was frequently observed during the progression of a metastatic tumor in the first funding period from FOR2127. The acquisition of such epigenetic changes is thought to possess a selective advantage that drives tumor progression, thus underscoring the importance of DNMT3A/B activity in this context. To investigate the potential impact of altered *Dnmt3a/b* expression in metastatic progression, a CRISPR/Cas9 RNP editing approach was employed in the work for this thesis to selectively inactivate one or both copies of *Dnmt3a/b* in mammary epithelial cells.

5.1. Expression kinetics of *Dnmt3a* and *Dnmt3b*

Recent studies have provided evidence supporting a significant relationship between the expression of *Dnmt3a* and *Dnmt3b* and the clinical outcomes of patients with breast cancer [159]. Specifically, it has been demonstrated that mammary tumors exhibit considerably higher levels of *Dnmt3a* expression in comparison to fibroadenomas, indicating a potential role for *Dnmt3a* in the development and progression of breast

cancer [160]. These findings suggest that *Dnmt3a* may serve as a valuable target for gene knockout within the tumor resulting in a decrease in metastatic potential.

The simultaneous disruption of all alleles of the two enzymes proved inefficient and impractical, necessitating a focus on the predominant family member that is active in Balb-NeuT mammary gland cells. To this end, mRNA was isolated, and quantitative real-time PCR was employed to evaluate the expression kinetics of the two different forms, thus revealing the most abundant enzyme that could alter the epigenetic landscape by adding methyl groups to cytosine residues in this specific cell type. The findings indicated that *Dnmt3a* exhibits a fold change four times higher than *Dnmt3b*, implying that this gene is expressed 16 times more in Balb-NeuT mammary gland cells. This observation is consistent with published data that suggests that *Dnmt3b* is a key player during the embryonic stages of development, while *Dnmt3a* gains greater prominence in more mature and complex cell types [161]. Consequently, subsequent experiments focused on DNMT3A as the primary enzyme responsible for de novo CpG methylation in Balb-NeuT mammary gland cells.

5.2. In silico analysis - *Dnmt3a*

The nucleotide sequence of *Dnmt3a* was retrieved from the National Library of Medicine (NCBI). The present study aims to investigate the molecular and genetic characteristics of the mouse gene encoding the enzyme DNMT3A, namely *Dnmt3a*, located on chromosome 12. The RefSeq accession number for the mouse *Dnmt3a* gene is NC_000078.7. According to the NCBI Reference Sequence database, the coding sequence (CDS) of the mouse *Dnmt3a* gene spans 4,095 base pairs. However, the entire gene's length, including intronic and untranslated regions (UTRs), is much longer. Specifically, the genomic sequence of the mouse *Dnmt3a* gene extends approximately 82 kilobases (kb), which comprises a 5' UTR of approximately 1.3 kb, a CDS of 4.1 kb, and a 3' UTR of approximately 76.6 kb. Such genetic and molecular properties are of great

significance in fully understanding the functional implications of DNMT3A in diverse biological processes and pathological conditions, including cancer development and progression.

5.3. Functional domain prediction and target site design

The UCSC genome browser was used to extract relevant protein domains and structural information associated with *Dnmt3a*, a critical epigenetic regulator [162]. A comprehensive examination revealed the existence of three fundamental structural domains within DNMT3A: An ATRX domain, a PWWP, a DNMT3 and ADD domain, and a methyltransferase (MTase) domain, which can be seen in Figure 5.1. These domains play pivotal roles in modulating gene expression and other essential cellular processes through DNMT3A's epigenetic modification activities [163]. The PWWP domain demon-

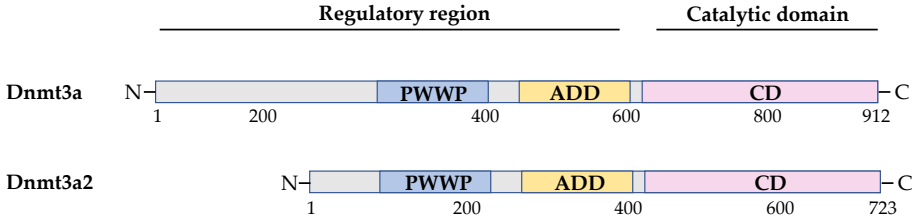


Figure 5.1.: Domain structure of DNMT3A. The two bars indicate the full-length MTase DNMT3A and the isoform DNMT3A2. ADD = ATRX-DNMT3-DNMT3L; PWWP = Pro-Trp-Trp-Pro.

strated remarkable specificity for recognizing and targeting nucleic acids. Its critical role in recruiting the enzyme to heterochromatin regions has also been previously documented [164]. This domain's versatile functions may be attributed to its ability to bind to different histone modifications and DNA sequences, thereby modulating chromatin structure and accessibility. In contrast, the ADD domain has been demonstrated to promote essential protein-protein interactions with diverse TFs, such as Myc and proteins HP1 and Suv39h1, which are responsible for chromatin packaging and histone deacetylases [165]. This domain was particularly significant, as adding

methyl groups should have been prevented. Furthermore, the findings identified six highly conserved Cytosine C5-DNA methyltransferase motifs within the MTase domain. These motifs facilitate the enzyme's ability to add methyl groups to Cytosine residues, modulating gene expression and regulating other critical cellular processes. The presence of these conserved motifs underscores the significance of DNMT3A in maintaining cellular homeostasis and highlights its essential role in regulating gene expression. This domain was of particular significance for this work as the addition of methyl groups should have been prevented. As postulated, it has been discerned that a truncated intragenic transcript, denoted as DNMT3A2, coexists with the full-length DNMT3A isoform in mice, wherein the latter lacks the N-terminal 219 amino acid residues [166]. Leveraging this insight, guide RNAs targeting gene knockout were deliberately situated in close proximity to the C-terminal catalytic domain that starts encoding from exon 16/23, illustrated in Figure 5.3. To ensure the utmost specificity and efficiency of the gene-editing approach, candidate guide RNAs were subjected to a comprehensive evaluation using the UCSC genome browser, and subsequently, the top-scoring candidates were cross-checked with IDT's CRISPR-Cas9 guide RNA design checker tool, thereby enabling the identification of guide RNAs with high on-target and low off-target scores. Initially, single-cut guide RNAs were employed for gene disruption. However, the process was found to have limitations regarding efficiency and reproducibility. Analyzing genome editing efficiency via T7 endonuclease assay revealed 30% for *Dnmt3a* and 7% for *Dnmt3b* gene modification with the single gRNA approach (Figure 5.2). As a result, a second set of experiments was implemented using the CRISPR/Cas9 nickase approach, aimed at reducing the likelihood of off-target editing and enhancing the efficacy of the knockout process. The CRISPR/Cas9 nickase approach involves using Cas9 nickase to cleave the target DNA's single strand rather than induce a double-strand break. When administered with two adjacent gRNAs in opposing directions, this creates a staggered cut. Specifically, the D10A mutation inactivates the RuvC domain of the Cas9 nickase, thereby enabling it to cleave only the target strand. This alteration in the Cas9 nuclease cutting mechanism has contributed to

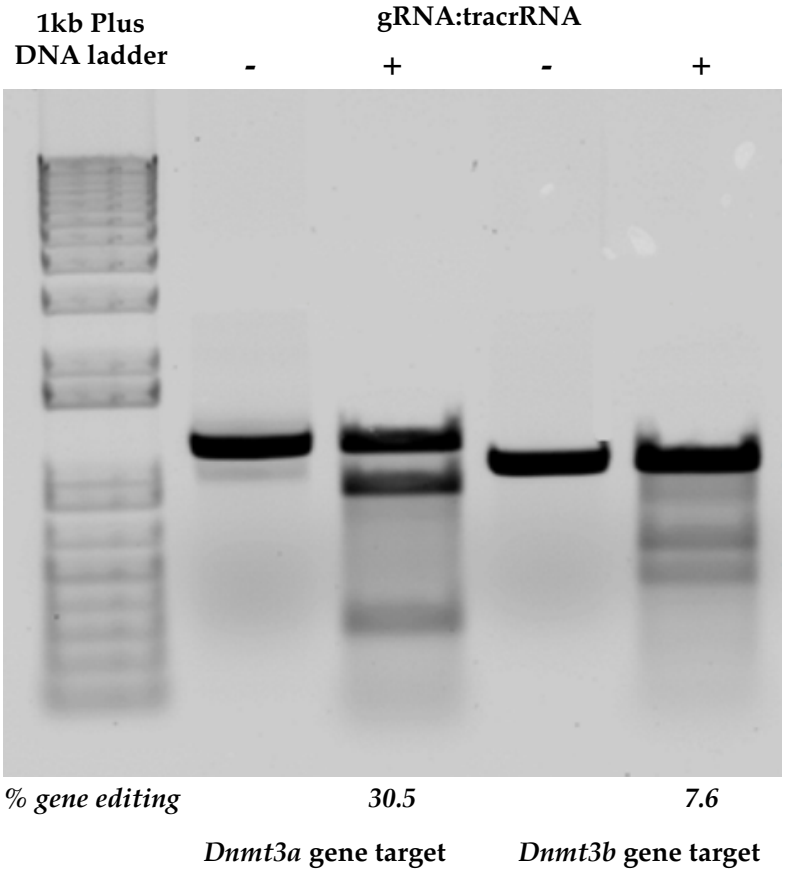


Figure 5.2.: T7 endonuclease assay of *Dnmt3a* & *Dnmt3b*. Gel image showing genome editing efficiencies of single-cut gRNAs Cas9 approach for genes *Dnmt3a* & *Dnmt3b*.

the improved precision and success rate of the gene knockout process while reducing undesired genetic alterations via off-target effects [167]. Overall, the findings of this study provide valuable insights into the use of the CRISPR/Cas9 system for gene editing purposes, particularly in the context of achieving an effective gene knockdown in mammalian cells.

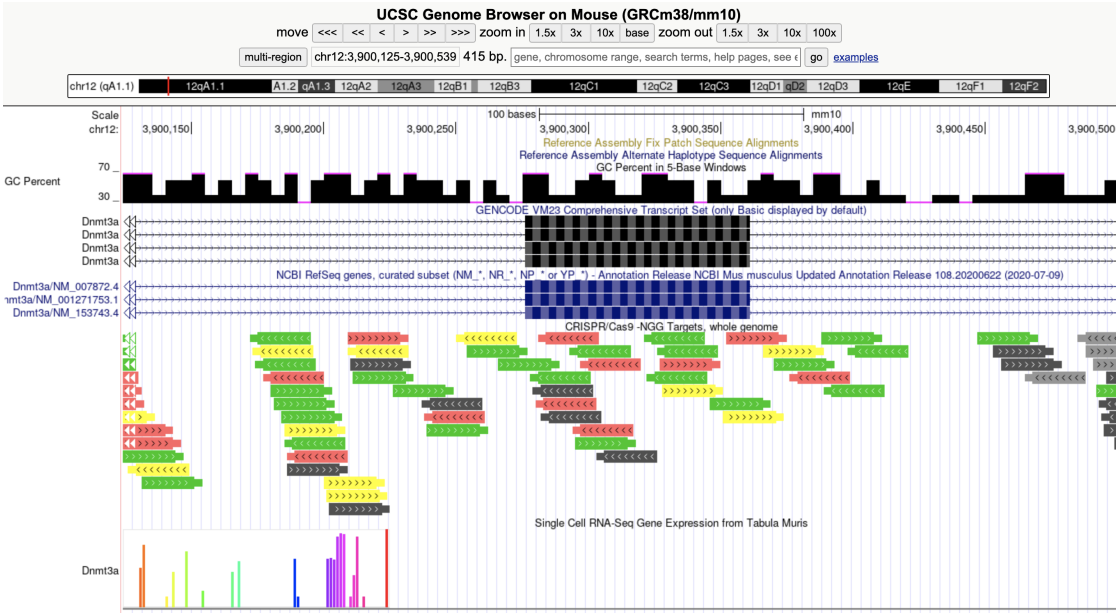


Figure 5.3.: Screenshot of UCSC genome browser at location exon 16 of *Dnmt3a*. The overview shows the position of guideRNAs relative to *Dnmt3a* exon 16. The green bars indicate the presence of highly effective guideRNAs in the corresponding region.

5.4. Determining Cas9 nickase knockdown efficiency

The objective of the experiments was to achieve a highly efficient knockdown of the *Dnmt3a* gene in primary mouse mammary gland cells using CRISPR/Cas9 genome editing technology. Three pairs of guide RNAs (gRNAs) were designed and tested before selecting the optimal candidates for use in distinct transplantation experiments. Genetically modified primary mammary spheres were isolated and transfected with RNP complexes of Cas9 D10A nickase and the selected gRNA pairs via electroporation. Proper sphere formation was checked after three days, and 500 spheres were picked and transferred to a new poly-Hema-coated cell culture plate after another six days. These 500 spheres were designated for transplantation into the mammary fat pad of healthy donor BALB/c mice the following day. In contrast, the remaining knockout spheres were harvested for further analysis to check knockdown efficiency. After DNA isolation, specific PCR-amplified target regions were sequenced using Sanger

sequencing with their respective forward and reverse primers. The inference of CRISPR Edits ICE analysis was used to evaluate the knockout efficiency of the experiments from the Sanger sequencing data, as this approach is a cost-effective means of measuring CRISPR edits. The ICE analysis revealed a successful gene knockdown of *Dnmt3a*, with a knockdown efficiency of 63% for the forward read and 65% for the reverse read. This observation was further validated by western blot analysis of the isolated proteins, where a knockdown efficiency of 65% was observed and calculated, as illustrated in Figure 5.4. After estimating the knockdown efficiency, the spheres containing with knockdown were mixed with wild-type balb-neut spheres in a ratio of 1:1 before transplantation, in order to observe which spheres had an advantage in establishing tumor lesions with further metastatic growth as the primary tumor evolves. Overall, these findings provide valuable insights into the efficacy of using CRISPR/Cas9 genome editing technology to knock down the *Dnmt3a* gene and its potential impact on tumor development and progression in mammary gland cells.

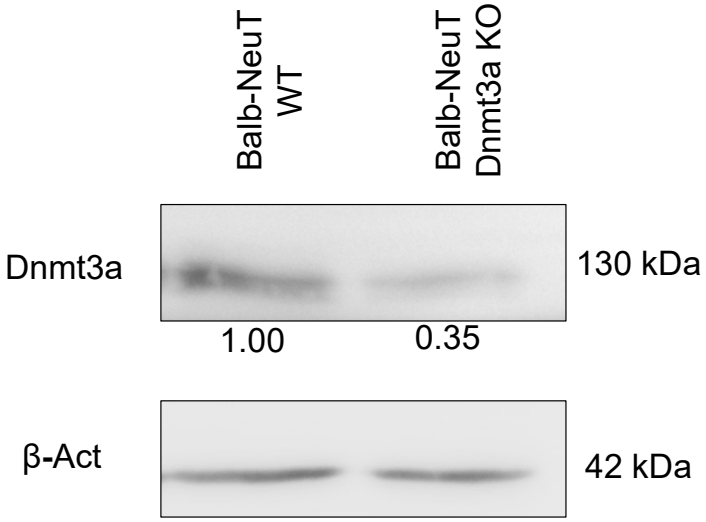


Figure 5.4.: Gene knockout of *Dnmt3a* in Balb-NeuT mice led to a reduction of DNMT3A protein levels compared to the empty vector (WT), as assessed by western blotting. 10 μg protein lysate from spheres derived from mammary gland tissue were used per lane. β-Actin was used as a loading control.

5.5. Generating spheres from Balb-NeuT mammary gland tissue

The experiments aimed to investigate the role of the *Dnmt3a* gene in mammary gland cell differentiation and in tumor growth and progression in mice carrying the NeuT transgene. To this end, mammary stem cells were isolated from Balb-NeuT mice at weeks 4-5, a time point before transgene activation, and cultured under mammosphere conditions. Additionally, short-term cultures were established for primary tumor cells. The ultimate goal was to alter the *Dnmt3a* expression of the mammary stem cells and retransplant them into healthy BALB/c recipient mice. However, given the harsh Cas9-RNP transfection into these cells by electroporation, it was necessary to improve the isolation protocol to increase the viability and quantity of stem cell spheres. To this end, changes were made to the protocol, including the use of a new culture medium, an additional centrifugation step to exclude fibroblast cells, and subsequent incubation with ACK buffer to eliminate erythroid blood cells during the isolation of the mammary gland stem cells. These alterations led to the formation of spheres with a higher mutation rate, suitable for retransplantation into recipient mice. A phase II brightfield microscopic image of a mammary stem cell sphere just before retransplantation is presented in Figure 5.5. Following successful gene knockdown, 50 spheres were re-transplanted into five wild-type BALB/c mice each. The experiments were conducted within the last six months of this thesis, and tumor growth and progression are still ongoing, awaiting evaluation. The findings from this study have the potential to contribute significantly to the understanding of mammary breast cancer progression in the context of aberrant DNA methylation and may provide insights into the development of targeted therapies for breast cancer patients.

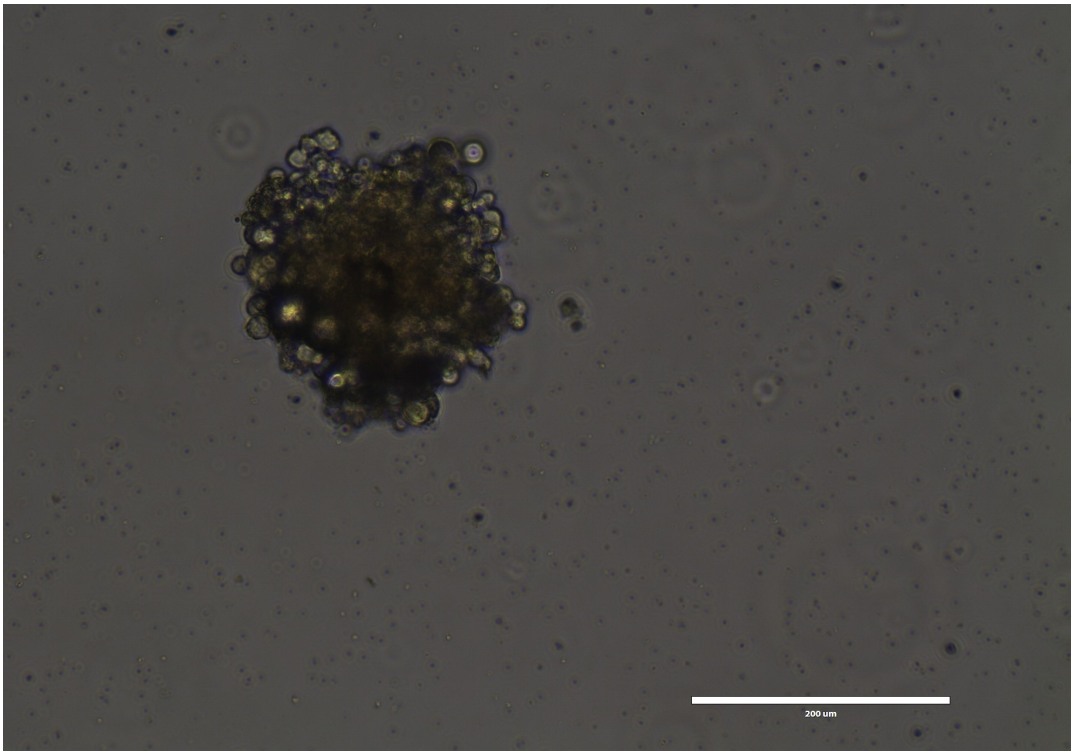


Figure 5.5.: Successful generation of Balb-NeuT mammary spheres. Brightfield microscope image shows that mouse mammary gland cells have been successfully isolated and used to generate spheres.

6. Discussion and Outlook

6.1. Discussion

Breast cancer metastasis is associated with active transcription changes and chromatin architecture alterations. The mechanisms of these changes have yet to be fully discovered; therefore, extensive research in this area is required. Aberrant DNA methylation in malignant breast neoplasms is a main driver of various metastatic processes [168]. While different hypotheses exist concerning the dissemination of cancerous cells from primary breast tumors and their transformation into metastatic cells to our knowledge, no scientific publication has yet analyzed to what extent an increase in DNA methylation contributes to the early dissemination and metastatic breast cancer spread. To address this gap in knowledge, the FOR2127 project aimed to elucidate the molecular mechanisms underlying the primary tumor dissemination and metastasis formation process. In the first funding period, MCIp-derived data revealed an overall increase in DNA methylation from normal mammary gland tissues toward the primary tumor and further to the metastatic lesion in the lung. However, the broad but shallow MCIp sequencing approach could only provide a general indication of overall methylation level increase and was therefore not useful in identifying individual cells with unique high DNA methylation patterns that could contribute to the metastatic process.

To overcome this limitation, a new, in-depth approach to characterize methylome remodeling during all stages of primary tumor progression was developed and evaluated

in this thesis. Specifically, barcoded bisulfite amplicon sequencing was used to identify and characterize DNA methylation on a single allelic level within breast cancer primary tumors in the HER2-driven Balb-NeuT mouse model. This sequencing approach allowed the identification of DNA methylation changes at single base resolution and enabled the clustering of the obtained sequencing reads based on their methylation increase. This approach was applied in order to to investigate whether a small subset of primary tumor cells exhibit an aberrant high, increased DNA methylation leading to metastasis or whether all the primary tumor cells exhibit equal levels of DNA methylation with a constant, steady increase toward the metastatic lesion. The resulting data were then aligned, clustered with the UPGMA algorithm, and visualized using phyloepigenetic trees to identify single cells with an unique high DNA methylation pattern.

Using this new analysis workflow, the DNA methylation landscape of 7 primary breast tumor samples and their corresponding lung metastases within 9 DNA regions representing various methylation patterns was profiled. These regions included high, low, and intermediate conditions of DNA methylation. Although the regions were thoroughly chosen, it turned out that they might not have been the best choice for the above outlined aim of this study. In particular, in most PCR products analyzed, the overall methylation did not change between the primary tumor and its metastatic sides. This observation made it difficult to check for rare high-methylated subpopulations within the primary tumor lesions, as the overall methylation pattern between the primary tumors and their corresponding metastasis was too similar. Additionally, within most regions analyzed, it was not feasible to distinguish primary tumor reads from reads arising from infiltrating stroma cells. These stroma cells have the same DNA methylation patterns as the normal tissue control samples, which made a distinction impossible. Only the PCR product for region B and D could serve as appropriate candidates for investigating this process, as all other regions showed an already too-high or too-low methylation content. Within region B, a mixture of DNA methylation was observed, where primary tumors 3145, 3569, and 3631 showed raw high methylated

reads, whereas the other primary tumor samples were either not analyzable (3524) or showed a constant increase in DNA methylation. It was impossible to identify raw high methylated reads within tumor samples in region C due to the high methylation properties of both the control and tumor samples. The primary tumors analyzed in region D exhibited a heterogeneous pattern in methylation progression. Samples 3145, 3524, and 3525 showed an equal distribution of sequenced reads, with no raw highly methylated reads in the primary tumor sample. Single higher methylated reads were detected in samples 3569, 3620 and 3631 of the primary tumor. The methylation progression investigation was unsuitable in region E as the primary tumor samples exhibited low methylation levels in most reads. Although single higher methylated sequenced reads were observed in the primary tumors, these reads only displayed a marginal increase in CpG methylation. Furthermore, no significant increase in methylation could have been observed from the primary tumor to the metastatic lesion. The analysis in region F showed a minor, constant increase of DNA methylation within all primary tumors. Due to the lack of evidence of aberrant, high methylated single cells in the primary tumors, the hypothesis is strenghtend that the increase in DNA methylation towards metastasis is driven by ongoing proliferation and cell division rather than by a selective process. Despite being a promising candidate for investigating methylation progression, region G could not serve for appropriate analysis as the sequenced PCR product was placed too close to the end of the CpG island, showing an already decreased methylation pattern. Therefore, the samples were homogenous unmethylated, and no single highly methylated reads could have been identified in the primary tumor samples. During the analysis of the read distribution in region J, a phyloepigenetic approach was taken where the reads were grouped together based on their methylation profile. The results showed that the methylation levels were already high in the primary tumors. As a result, no raw, highly methylated single reads were detectable. The same was observed in the PCR product of region K. Despite only having four CpGs in the PCR product analyzed, it was not a suitable candidate to study methylation progression from primary tumor to metastasis due to its overall methylated character. Region L served as

a good positive control region for the study, with all samples displaying a consistent and highly methylated composition. The even distribution of highly methylated reads was a clear indication of the efficacy of the methodology. The results indicate that there are diverse DNA methylation patterns during the progression of primary breast cancer tumors. Some tumors displayed small subsets of highly methylated reads in specific regions, suggesting that these cells may have an advantage in becoming metastatic (e.g. PT3145 in region D). The metastases analyzed tended to show hypermethylation when compared to their corresponding primary tumor. Additionally, there was an even distribution of reads gaining DNA methylation in a steady and constant manner within primary tumor samples (e.g. PT3145 in region F). This observation supports the hypothesis that increased DNA methylation accumulation is closely linked to the constant high proliferation rate and ongoing cell aging seen in breast cancer [169]. These findings underscore the complex and diverse nature of primary tumor samples, comprised of cells with a wide range of gene expression patterns and epigenetic modifications. Typically, an increase in DNA methylation in CpG islands within promoters goes along with a widespread reduction in methylation across most of the genome [170]. Carmona et al. showed that an increase in DNA methylation in these promoter structures goes along with silencing tumor suppressor genes. A comprehensive characterization of matched cell lines derived from primary tumors and metastatic sites was conducted, and the analysis identified *CDH11* as a candidate gene targeted for silencing via DNA methylation. This analysis subsequently provided evidence supporting the involvement of *CDH11* in the downregulation of metastasis [171]. Importantly, results indicate that aberrant DNA methylation can activate oncogenes, promoting uncontrolled cell proliferation and tumor development. These findings underscore the significance of DNA methylation as a crucial epigenetic mechanism that drives breast cancer metastasis and progression. Previous studies have demonstrated the heterogeneity in gene expression and DNA methylation patterns among breast cancer subtypes. Győrffy et al. showed that DNA methylation significantly influences gene expression, ultimately influencing the prognosis of breast cancer patients based on the common breast cancer subtypes

[172]. By performing large-scale methylome analyses, they identified numerous genes with consistently aberrant methylation patterns associated with prognosis. These genes may be promising biomarkers for predicting clinical outcomes and stratifying patients with distinct breast cancer subtypes. A comprehensive analysis of 880 human breast carcinomas conducted by Elsheikh et al. revealed that global histone modifications are closely associated with tumor phenotypes, prognostic factors, and patient outcome [173]. This study provided compelling evidence that alterations in histone acetylation and DNA methylation patterns may indicate an initial indication of breast cancer, with the potential to serve as valuable diagnostic markers. The study by Elsheikh et al. on histone modifications and methylation changes has shed light on the critical role of epigenetic regulation in breast cancer pathogenesis and may pave the way for more effective early detection and personalized treatment strategies.

Rodenhiser et al. explored the epigenetic alterations occurring in highly metastatic breast cancer cells, specifically in the context of gene methylation status [174]. The researchers compared the methylation patterns of the upstream promoter regions of genes belonging to various pathways in the highly metastatic MDA-MB-468LN cell line relative to the less metastatic MDA-MB-468GFB cell line, using whole-genome promoter tiling microarrays. They observed extensive alterations in gene methylation patterns likely to contribute to carcinogenesis and metastasis. Intriguingly, approximately 20-30% of genes exhibited differential methylation when the cell lines were compared. The *SWI/SNF* gene, encoding a chromatin remodeler, was among the genes displaying altered methylation levels. These changes were linked to the silencing of the TGF-β pathway and overexpression of MYC, both of which are involved in carcinogenesis and metastasis.

Additionally, the study revealed that hypomethylation of genes such as *ZEB2* and *SNAIL2* is linked to their overexpression, leading to repression of the E-cadherin (*CDH1*) gene, which is part of the epithelial-mesenchymal transition (EMT). The researchers also observed differential methylation of genes belonging to the integrin signaling

pathway, with some genes being hypermethylated while others were hypomethylated. Integrin receptors play a crucial role in cell-cell attachment and motility, which must be carefully regulated during metastasis. Taken together, these findings underscore the complexity of metastasis, suggesting that distinct patterns of gene silencing or re-expression may contribute to the more metastatic forms of breast cancer. It should be noted, however, that the study utilized different breast cancer models, making direct comparisons of the results unfeasible. This emphasizes once more the heterogeneous nature of primary breast cancer tumors and their various subtypes.

In the year 2018, Ming et al. conducted a study that highlighted the complex genetic and epigenetic mechanisms that underlie the progression of breast cancer [175]. Specifically, their work reveled the role of promoter hypermethylation in the regulation of several key genes, including *RARb2*, *DAPK*, *hMLH1*, *p14*, and *p15*, in the context of breast cancer progression. Their study employed a meta-analysis approach, which allowed for the comprehensive evaluation of existing literature on the topic. Through this analysis, they were able to demonstrate significant associations between promoter hypermethylation of *RARb2* and lymph node metastasis, as well as TNM-stage breast cancer patients. These findings have important clinical implications, as they suggest that *RARb2* promoter hypermethylation may serve as a potential biomarker for disease progression and prognosis in breast cancer patients. In addition to the association between *RARb2* hypermethylation and disease progression, the study revealed significant associations between promoter hypermethylation of *DAPK*, *hMLH1*, and p14 genes and breast cancer susceptibility. These findings suggest that epigenetic alterations in these genes may play a key role in the development and onset of breast cancer. Overall, the study by Ming et al. represents a significant contribution to breast cancer research, providing important insights into the molecular mechanisms that drive disease progression. The findings further have important implications for developing new diagnostic and therapeutic strategies for breast cancer patients and underscore the importance of ongoing research efforts in this area.

The present study utilized a comprehensive BBA-seq approach to investigate the role of DNA methylation in the progression of metastatic HER2-driven breast cancer. The results obtained from this study are consistent with current literature, suggesting that DNA methylation is a heterogeneous process that contributes to cancer progression. However, the precise mechanisms involved in this process are not yet fully understood. The obtained results suggest that breast cancer cells change DNA methylation patterns as they progress from primary tumors to metastatic disease. These changes can result in silencing tumor-suppressor genes and activating oncogenes, thereby facilitating cancer progression and metastasis. Furthermore, the results suggest that some breast cancer cells may acquire high levels of methylation in primary tumors, providing them with a selective advantage in becoming metastatic cells. Likely, the relationship between DNA methylation and metastatic HER2-driven breast cancer progression is multifaceted and complex, and more research is needed to comprehend the underlying mechanisms fully.

In conclusion, this thesis provides new insights into the role of DNA methylation in the progression of metastatic HER2-driven breast cancer. The findings presented indicate the need for further investigation into the mechanisms involved in DNA methylation-mediated cancer progression and metastasis, which may ultimately lead to the development of more effective therapeutic strategies for patients with metastatic HER2-driven breast cancer.

6.1.1. Stroma cells in tumor microenvironment

As already described, cancer is a complex disease involving both tumor cells and surrounding stromal cells. Stroma cells consist of tumor-associated macrophages (TAMs), lymphocytes like CD4+, CD8+, and regulatory T-cells, Endothelial cells, and Fibroblasts [176]. In the context of breast cancer, it has been shown that cancer-associated fibroblasts

(CAFs) contribute to tumor onset, angiogenesis, progression, and metastasis development. Those CAFs share common similarities with fibroblasts found in inflammatory tissue [177]. They get activated and secrete extracellular matrix (ECM) components due to the release of TGF-beta and CXCL12/SDF-1 from cancer cells [178]. Furthermore, Caveolin-1 (Cav-1), a tumor suppressor gene that interacts with and downregulates the Ras-p42/44 MAP kinase, was found to be less expressed in CAFs than in non-cancer-associated fibroblasts (NAFs) in the breast cancer microenvironment. Through the downregulation of these genes, tumor-associated fibroblasts proliferate more and faster, which leads to tumor growth and metastasis progression [179]. Al-Ansari et al. have shown that downregulating the tumor-suppressor gene p16INK4A in CAFs increased the EMT in breast cancer cells, leading to higher levels of carcinogenesis [180]. Particularly in breast cancer progression, CAFs impact Estradiol (E2) levels. By synthesizing estrogen via aromatase activity, they stimulate the growth of pre-cancerous breast epithelial cells while inhibiting their differentiation [181]. Apart from affecting E2 levels, CAFs contribute to breast cancer progression by increasing the secretion of growth factors. High levels of IL-6 secreted by CAFs promote migration and induce EMT in breast cancer cells, promoting an invasive phenotype. Co-inoculation of CAFs with breast cancer cells in mouse models results in accelerated tumor development [182]. CAFs not only play an essential role in breast cancer initiation and invasion, they further contribute to the development of metastasis. They induce EMT in breast epithelial cells and promote angiogenesis through the secretion of CXCL12/SDF-1 [183]. Afterward, cancer cells secrete factors like CCL2 to activate CAFs and recruit macrophages, which results in intravasation and promotes the formation of breast cancer stem cells (CSCs) [184]. After metastasis cells arrive at distant sites like the lung, CAFs promote metastasis. The secretion of CCL5 leads to the activation of regulatory T-cells. Those infiltrating lymphocytes then produce RANKL, which activates HER2 over-expressing metastatic breast cancer cells by binding to RANK and activating RANKL–RANK signaling [185]. Additionally to fibroblasts, within the tumor microenvironment TAMs, which are predominantly of the M2 phenotype, are known for secreting growth factors

that facilitate various aspects of tumor progression, including angiogenesis, invasion, migration, metastasis, and, importantly, immunosuppression [186]. In breast cancer, high TAM infiltration correlates with poor prognostic features and higher tumor grade. Studies have highlighted the role of chemokine ligand 18 (CCL18) in promoting invasion and metastasis by enhancing cancer cell adhesion to the extracellular matrix through interaction with its receptor PITPNM3 (chen2011ccl18). Further, epigenetic changes, such as miR92a expression and IL-4 activation, influence TAM function and breast cancer invasiveness [187]. In conclusion, it has been shown that the carcinogenic cell alone is not responsible for tumor initiation and metastatic progression; rather, the tumor microenvironment, with its variety of tumor-infiltrating cells, plays an important role, too.

One major weakness within the analysis applied in this thesis of aberrant DNA methylation progression was the contamination by infiltrating stroma cells in the primary tumor and metastasis tissues. The lysis and isolation procedure did not allow for the enrichment of actual tumor cells or discrimination of infiltrating stroma cells. The stroma cells generally have the same methylation profile as the surrounding tissue, that is, mammary epithelial and lung tissue. For this reason, the low methylated reads were subtracted from the primary tumor and metastasis samples in regions where the control tissues showed no signs of high methylation. This led to improved comparability between the two malign groups. This was the case in regions B,D,E,F, and G as the other regions did show high DNA methylation reads in the controls. To circumvent this problem in future DNA methylation analysis to investigate metastatic progression, one could employ a strategy to fully isolate cells from the primary tumor and their corresponding metastatic sides. In 2021, Rodriguez de la Fuente et al. published a protocol for isolating viable single cells from mouse breast tumor tissue, followed by isolation and sorting via flow cytometry [188]. After tumor tissue dissection and enzymatic digestion to obtain a single cell suspension, the cells were labeled with specific antibody combinations to exclude unwanted and enrich desired cells - that

is, primary tumor and metastasis cells. CD45 was used to exclude immune-infiltrated cells, as this cytokine is exclusively expressed in immune cells. Staining with anti-Her2 could be conducted to enrich tumor cells, as this receptor is expressed on tumor cells. Finally, double-negative (CD45-, Her2-) selection could be conducted to exclude the infiltrated stroma cells, especially fibroblasts. This process guarantees an enrichment for tumor cells applicable for downstream processes, like the investigation of the DNA methylome.

6.1.2. Strengths and potential limitations of the BBA-seq approach

With the BBA-seq approach implemented for this thesis, the DNA methylome of 7 primary breast cancer samples with matched metastatic sites could be analyzed at a single-base resolution. One of the strengths of this study is that accurate results were obtained despite using a small amount of DNA in the samples. This is because the DNA had already been utilized in the MCIp experiments conducted during the initial funding period. A multiplex approach was first implemented by covering the genomic loci with bisulfite padlock probes to overcome the low amount of input DNA. Unfortunately, producing high-quality libraries was impossible, which led to poor sequencing results. A second approach, BBA-seq, was employed to analyze the progression of methylation from primary tumor to corresponding metastasis. Although a single reaction was used for each PCR product, the method yielded analyzable results as multiplexing was impossible. This was achieved through extensive test experiments to improve bisulfite conversion and library generation with low DNA concentrations. Nevertheless, retrieving data from all tumor samples from the first funding period was impossible as the amount of DNA was too low or completely gone. Furthermore, a very high bisulfite conversion rate could be achieved, leading to a high read coverage after filtering out false positive sequencing reads. Potential limitations lie in the generation of the tree files. In the first tests, it was attempted to illustrate every sequencing read, as this would represent the right read distributions.

Unfortunately, because of size and capacity restrictions, these trees could not be loaded in `Figtree` software. This issue was solved by combining equal reads in the MSA files, leading to more condensed trees. Nevertheless, when covering a decent amount of CpGs within the sequencing read, the variety of probable CpG compositions increases, leading to an increase in tree file size with limitations in illustration. Here, a condensing step should be developed by clustering reads with a certain similarity threshold. In future studies, more regions should be investigated to develop more insights into the complex behavior of DNA methylation in the context of metastatic cancer progression. Further, BBA-seq and progression models should be combined with RNA-seq gene expression experiments to elucidate functional regions and genes and their contribution to the metastatic process. Most recently published studies and publications combine methylome analysis with transcriptional analysis. In 2022, Li et al. showed that the tumor suppressor gene *PRGMA* is down-regulated in progressive breast cancer due to its promoter hypermethylation [189]. They further showed that the absence of *PRGMA* correlates with poor patient outcomes. Suppressing *DNMT1* and *PDNMT3A* expression via the drug 5-Aza-CdR could result in the regained expression of *PRGMA*, which inhibits tumor growth. In conclusion, identifying RGMA as a pivotal molecular marker offers novel insights into potential therapeutic targets and prognostic markers in breast cancer progression. Yeon Park et al. investigated the progression of DNA methylation throughout the different stages of breast cancer, namely from pre-invasive lesions, atypical ductal hyperplasia, and ductal carcinoma in situ to invasive ductal carcinoma [190]. They saw a significant increase of methylation in the early stages of breast cancer but showed similar CpG island methylation between ductal carcinoma in situ and invasive ductal carcinoma. They further suggested that aberrant DNA methylation in breast cancer occurs early in disease onset. The analysis in the thesis also supported this observation, as most of the primary tumors showed comparable methylation patterns to their corresponding metastasis. Some studies have reported a shift from methylation to unmethylation events from the primary tumor to the metastatic lesion, which stands in contrast with the increase in methylation observed in the MCIp data obtained during

the first funding period. Urrutia et al. reported that the methylation pattern, particularly the methylation index (MI), did not change significantly between the primary tumor and corresponding lymph node metastasis. Nevertheless, they identified one CpG site in the promoter region of the gene PAX6 that showed a decreased methylation level toward the metastatic lesion. They further demonstrated a correlation between *PAX6* expression level and methylation status of promoter region using qPCR. This research highlights the effectiveness of using DNA methylation analysis along with gene expression data. Another publication conducted by Reyngold et al. investigated the progression of methylation toward the metastatic lesion in different breast cancer subtypes [168]. They found that most metastases retain their DNA methylation pattern in CpG islands and promoter regions when compared with their origin primary tumor. Nevertheless, they saw a gain in methylation toward the metastasis in CpG island poor areas. These observations might explain why the MCIp data from the first funding period show an overall gain in methylation and the data obtained within this thesis showed similar methylation patterns between primary tumor and metastasis matches. Additionally to the methylation analysis, they conducted gene expression analysis and found a narrow set of genes (= 8) that are downregulated in metastasis when compared to the primary tumor lesion. This observation in combination with the dramatic change in methylome composition from normal breast tissue to primary tumor lesion let them conclude that the methylation remodeling starts early in tumor onset and does not change significantly toward metastasis. In summary, combining DNA methylation and gene expression analysis underscored even more the complexity and heterogeneous character of cancer. Most studies see an overall increase in DNA methylation from the primary breast tumor toward its corresponding metastasis. Nevertheless, loss of methylation with subsequent gene activation could have been also observed. Additionally, most publications did not see an overall change in methylation and progression when specific loci were compared between primary tumors and metastasis. All those observations are in accord with the data analyzed in this thesis, as most of the metastasis samples showed similar methylation patterns when compared to their primary tumors. This leads to the

conclusion that aberrant DNA methylation occurs early in the onset of breast cancer evolution, and the increase in DNA methylation toward the metastatic lesion occurs primarily in genomic regions, which are not responsible for gene expression and are generally a by-product of extensive cell division and proliferation, rather than providing additional advantages for metastasis. Nevertheless, this increase in methylation could lead to genomic instability and aberrant genomic organization within the nucleus.

6.1.3. DNA methylation in other cancers

Aberrant DNA methylation in the context of human cancer progression could not only be observed in breast cancer but also serves as a hallmark of malignancy and poor patient prognosis in other cancer types. In general, it is shown that DNA methylation in cancer exhibits genome-wide hypomethylation but shows hypermethylation in CpG-rich promoter regions, especially in tumor suppressor gene-associated transcription start sites [191, 192]. So far, no genetic mutation that facilitates the spread of primary tumor cells to distant organs has been identified. Instead, epigenetic changes in the genome are promising therapeutic targets because they can be reversed and have dynamic properties that could alter the metastasis process. In 2020, Ili et al. focused on understanding the DNA methylation landscape in colorectal cancer (CRC) and its role in tumor progression from primary tumors to lymph node metastasis [193]. They conducted genome-wide methylation sequencing on paired normal tissue samples, primary tumors, and corresponding lymph node metastasis samples. They focused on identifying specific methylation changes that are associated with CRC progression. In their study, primary tumors exhibited global DNA hypomethylation and showed significant hypermethylation of CpG islands, particularly in gene promoter regions, compared to the normal tissue controls. Similar results were observed in this thesis. In most CpG islands, where the normal tissue exhibited a low methylation profile, DNA methylation increased from normal tissue to primary tumor sites. In contrast, lymph node metastasis displayed high global methylation in non-CpG island regions

but lower CpG island methylation compared to their matched primary tumors. The metastatic sites retained methylation patterns from primary tumors, indicating a milder but persistent tumor signature. They explain these observations with the fact that metastasis cells might represent a subpopulation of Cancer-Initiating Cells (CICs) that have the capabilities of self-renewal, differentiation, and migration. Those CICs make up less than 5% of the solid tumor mass and might have separated from the primary tumor before most of the aberrant DNA methylation changes occurred. Additionally, they further postulate the problem of immune cell contamination that could result in misleading methylation levels, especially in the lymph node metastasis samples. Unfortunately, no bioinformatic tool is currently available to deconvolute lymphocyte methylation patterns as there is for RNA sequencing [194]. Taken together, the study showed similar methylation progression from normal tissue towards the primary tumor lesion but observed striking differences in the methylation pattern of the metastatic samples. Here, they found a decrease in CpG island methylation, whereas the data analyses in this thesis showed an increase in CpG island methylation in metastasis. This further illustrated the heterogeneity of cancer progression throughout the different tumor types. Zheng et al. investigated the role of DNA methylation in renal cell carcinoma (RCC) metastasis, specifically with its association with the TGF-β/RUNX3 pathway [195]. They conducted a targeted gene array to compare methylation-regulated gene expression metastatic versus primary RCC tissue. They found that the metastatic tissue displayed higher CpG methylation in RUNX3 and TGF-β in their respective promoter regions compared to primary tumors, leading to decreased expression of these tumor suppressor genes. Additionally, murine models showed that tumors from metastatic cells grew larger, and the mice had an overall shorter survival time. Treatment with a methylation inhibitor restored RUNX3 and TGF-β expression, reduced the tumor size, and prolonged the survival of the treated mice. In summary, they showed that methylation is critical in regulating RCC metastasis by inhibiting the TGF-β/RUNX3 pathway. In comparison to the data from this thesis, a similar increase in CpG island methylation could be observed. Unfortunately, the authors did not make any statement

or analysis about when a metastasis cell could disseminate from the primary tumor lesion or the DNA methylation progression.

6.2. Outlook

Finally, phyloepigenetic analysis, together with the BBA-seq approach implemented for this thesis, represents a promising, powerful tool for the identification of aberrant DNA methylation patterns during metastatic breast cancer progression. The approach provides a comprehensive and high-resolution view of DNA methylation at single base resolution, allowing for the identification of novel methylated regions and the characterization of methylation dynamics across different samples from bulk cell datasets. One of the major strengths of BBA-seq is its ability to overcome the limitations of other DNA methylation profiling methods, such as MCIp, by providing a very deep sequencing approach as only a small fraction of the genome is covered. Moreover, BBA-seq enables the identification of conserved and lineage-specific DNA methylation patterns, which can shed light on the process and progression of DNA methylation and its functional significance. Looking ahead, the BBA-seq approach combined with new models such as machine learning and single-cell RNA-seq might have a significant impact on DNA methylation analysis, particularly in the areas of comparative and evolutionary multiomics. Additionally, the deep sequencing approach of BBA-seq should be combined with other cancer progression models like the cross-sectional model conjunctive bayesian networks [196], the new capri algorithm [197] or using mutual hazard networks [198]. The algorithms aim to establish functional relations between alterations, where the DNA methylation at a specific position selects for a later methylation at another site. The methylation in the former makes the methylation in the latter more beneficial for the tumor, and hence, the methylation in the first site predominantly occurs before the methylation at the second site. With these approaches a more distinct timeframe of methylation progression should be feasible to obtain. In

conclusion, BBA-seq and phyloepigenetics, in combination with specific progression models, have the potential to uncover new insights into the diversity and development of epigenetic regulation across different species and tumors.

Appendix

Appendix A.

Code snippets

```
args = commandArgs(trailingOnly = TRUE)
library(tidyverse)
input <- args[1]
output <- gsub("txt","fasta", input)
paste(input,output)

inputDF <- read_tsv(input, skip=1,
                    col_names = c("readID","methylation_state","chromosome",
                                  "position","methylation_call"))

comb <- function(a, b) {
if(nchar(a) != nchar(b)) {
 stop("unequal length")
}
res = a
for(i in 1:nchar(a)){
   res = paste0(substring(res, 0, i-1),
                ifelse(substring(a, i, i) != "-", substring(a, i, i), substring(b, i, i)),
                substring(res, i+1, nchar(a)))
}
return(res)
}

tempStr <- function(position,startpos,length,call){
return(paste0(strrep("-",position-startpos),
              call, strrep("-",length-(position-startpos)-1)))
}
```

```
astaString <- function(ID,startpos,chrom,methylationString){
return(paste0(">",ID," startpos=",startpos," chromosome=",
              chrom,"\n",methylationString))
}

inputDF %>% group_by(readID) %>%
mutate(startpos = min(position)) %>% mutate(length=max(position)-min(position)+1) %>%
mutate(tempStr=tempStr(position,startpos,length,methylation_call)) %>%
select(-methylation_state,-position,-methylation_call) %>%
mutate(methylationString = reduce(tempStr,comb)) %>% select(-tempStr) %>% unique %>%
mutate(fastaString=fastaString(readID,startpos,chromosome,methylationString)) %>%
ungroup() %>% select(fastaString) %>%

write_delim(file = output, col_names = FALSE, quote = "none")
```

File Excerpt A.1: `convertFastaI.R` is a **R**-script which converts the output `*.txt` file, cf. File Excerpt 3.1 generated by `bismark_methylation_extractor` to a preliminary `*.fasta` file, cf. File Excerpt 3.2 for subsequent use by `convertFastaII.R`

```
args = commandArgs(trailingOnly = TRUE)
library(tidyverse)
input <- args[1]
sampleID <- gsub("^([^\\.]*)\\..*","\\1",basename(input), perl=TRUE)
region <- gsub(".*(region[^_]+)_.*","\\1",basename(input), perl=TRUE)
all_data = readLines(input)
formatted <- c()

for(i in seq(1, length(all_data), 2)) {
   lineOne <- gsub(" $","", all_data[i])
   formatted <- c(formatted, paste(lineOne,all_data[i+1]))
}

shortSequence <- function(seq){
   tmp <- gsub("-","",seq)
   tmp <- gsub("Z","C",tmp)
   tmp <- gsub("z","T",tmp)
   return(tmp)
}
```

```
fasta <- function(id,sequence){
   return(paste0(">",id,"\n",sequence))
}

df <- read_delim(file = I(formatted), delim = " ",
                 col_names = c("id","startpos","chr","sequence")) %>%
      mutate(sequence = shortSequence(sequence)) %>%
      mutate(id = paste0(sampleID,"_",region,"_",sequence)) %>%
      select(-startpos,-chr) %>%
      count(id, sequence)

maxN <- max(df %>% pull(n))

targetLength <- str_length(df %>% filter(n==maxN) %>% head (n=1)  %>% pull(sequence))

df <- df %>% filter(str_length(sequence)==targetLength)

df %>% mutate(id = paste0(id,"_",n) ) %>% select(-n) %>%
   mutate(fasta = fasta(id,sequence)) %>%
   select(fasta) %>%
   write_delim(file = paste0(dirname(input),"/", sampleID,"_",region,"_count.fasta"),
                            col_names = FALSE, quote = "none")

outdf <- tibble()

for(irow in 1:dim(df)[1]) {
   row <- df[irow,]
      for(i in 1:row[["n"]]) {
         outdf <- bind_rows(outdf, tibble(id = paste0(row[["id"]], "-", i),
                                                  sequence = row[["sequence"]]))
   }
}

outdf %>% mutate(fasta = fasta(id,sequence)) %>%
   select(fasta) %>%
   write_delim(file = paste0(dirname(input),"/", sampleID,"_",region,"_all.fasta"),
                            col_names = FALSE, quote = "none")
```

File Excerpt A.2: `convertFastaII.R` is a **R**-script which converts the output `*.fasta` file produced by `convertFastaI.R` to two final `*.fasta` files, examples of which are shown in File Excerpts 3.3 and 3.4.

Appendix B.

Supplementary figures

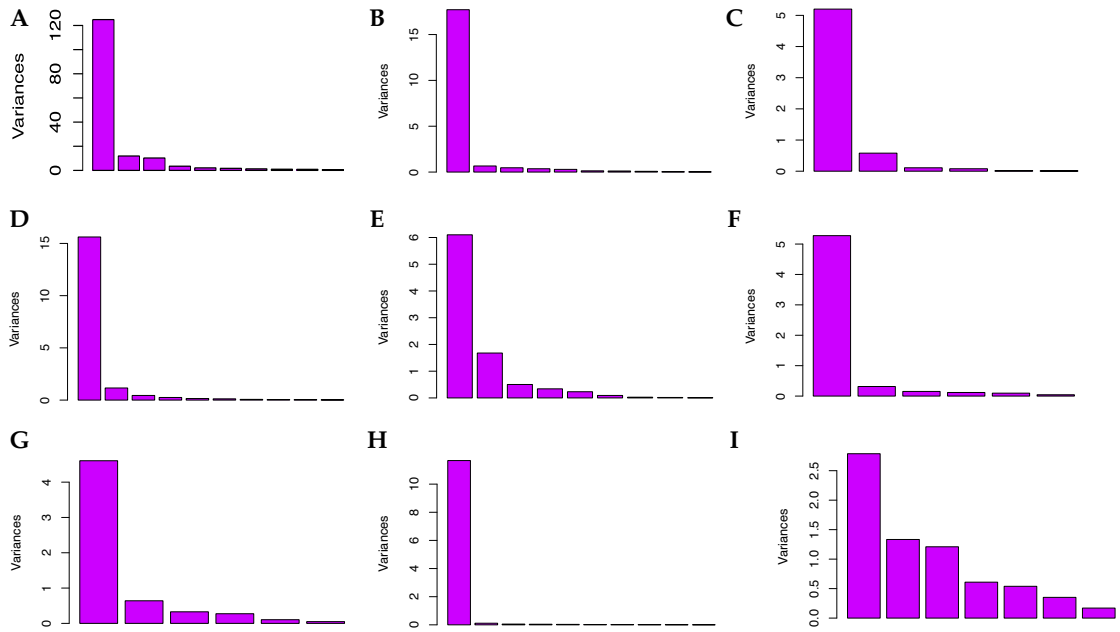


Figure B.1.: Screeplots showing variances for principal component analysis. (A) All regions combined (B) Region B (C) Region C (D) Region D (E) Region E (F) Region F (G) Region G (H) Region J (I) Region L.

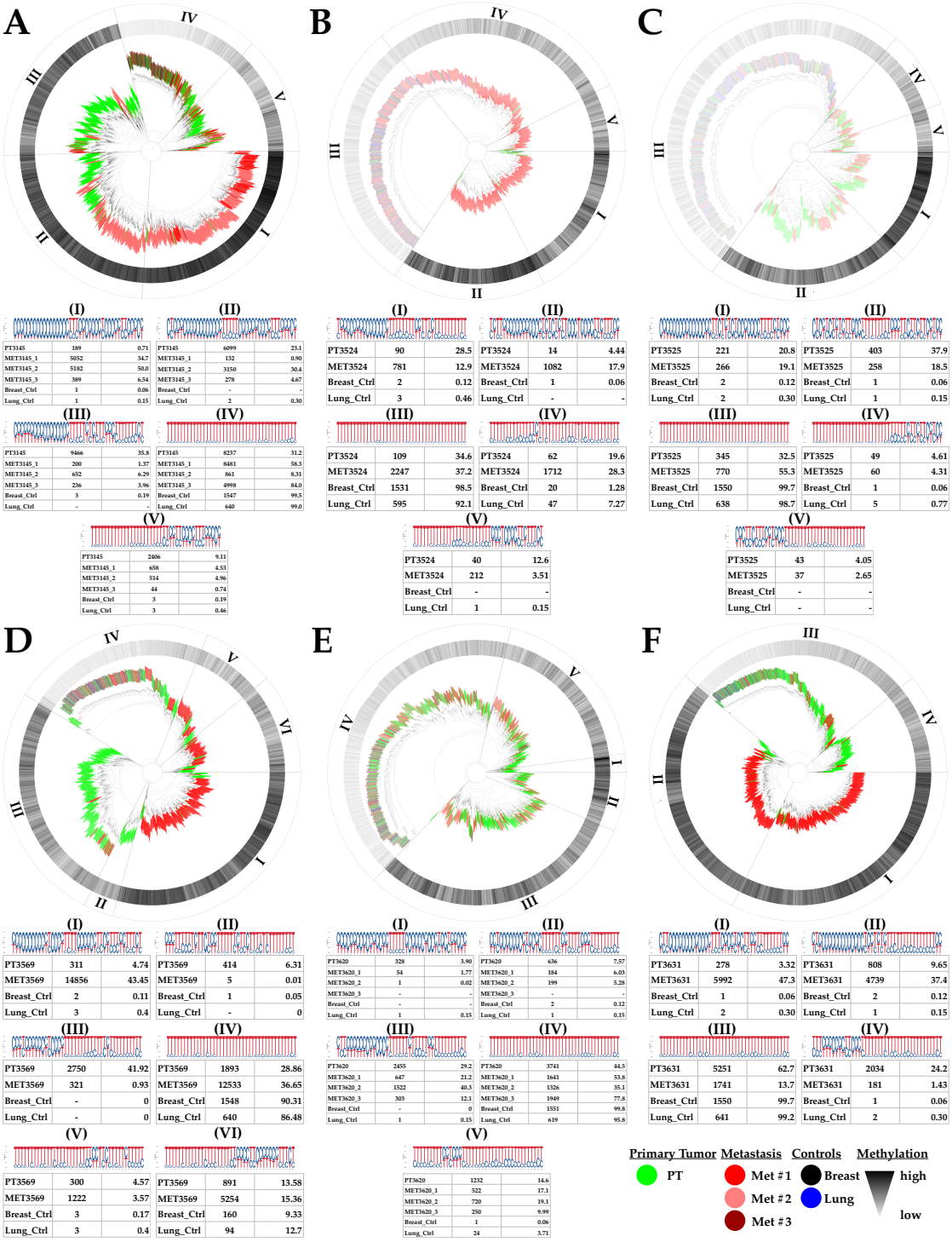


Figure B.2.: Phyloepignetic tree files and sequence logos for region B before subtracting low-methylated reads from primary tumor and metastasis samples. (A) 3145 (B) 3524 (C) 3525 (D) 3569 (E) 3620 (F) 3631.

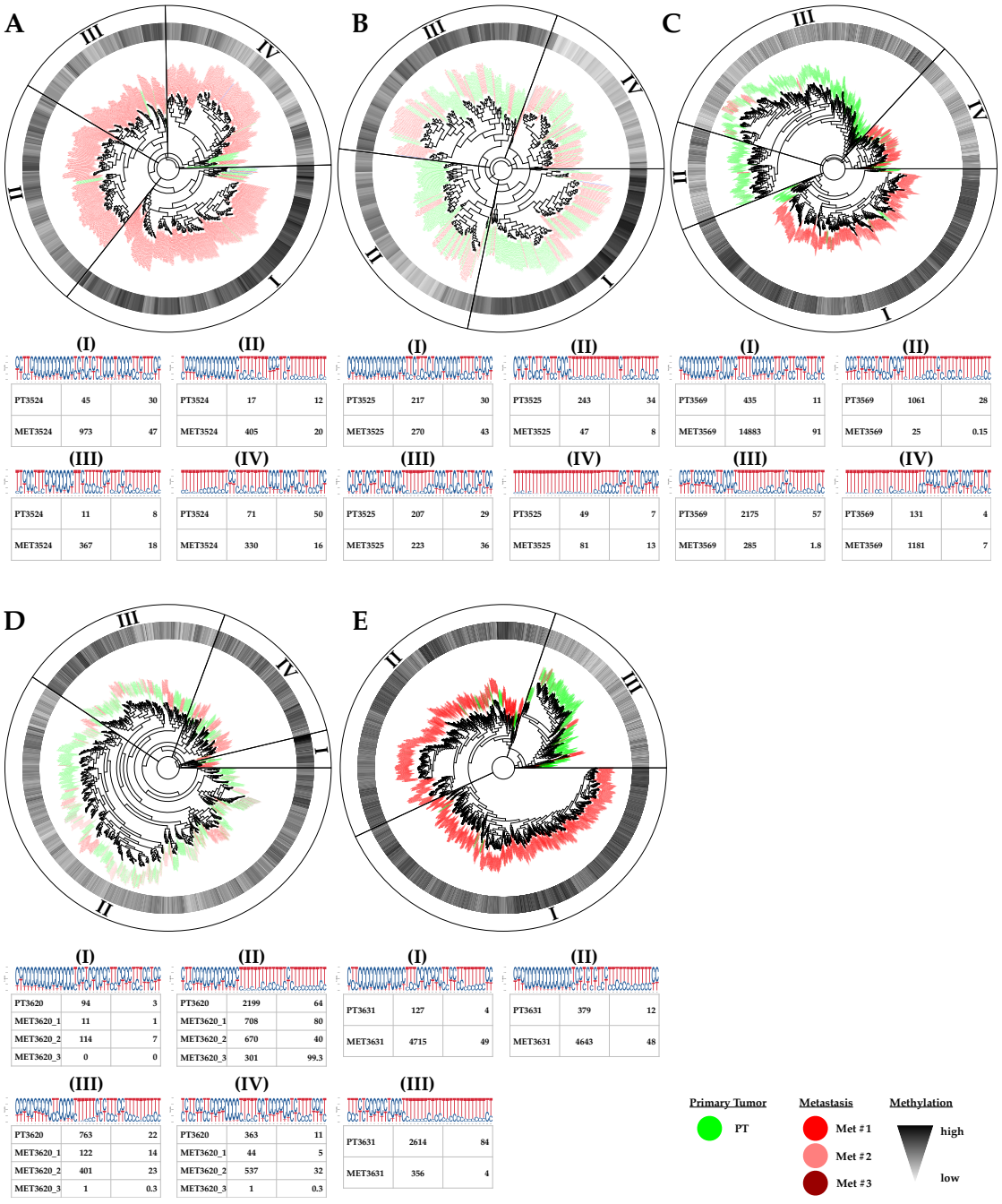


Figure B.3.: Phyloepignetic tree files & sequence logos for region B. (A) 3524 (B) 3525 (C) 3569 (D) 3620 (E) 3631.

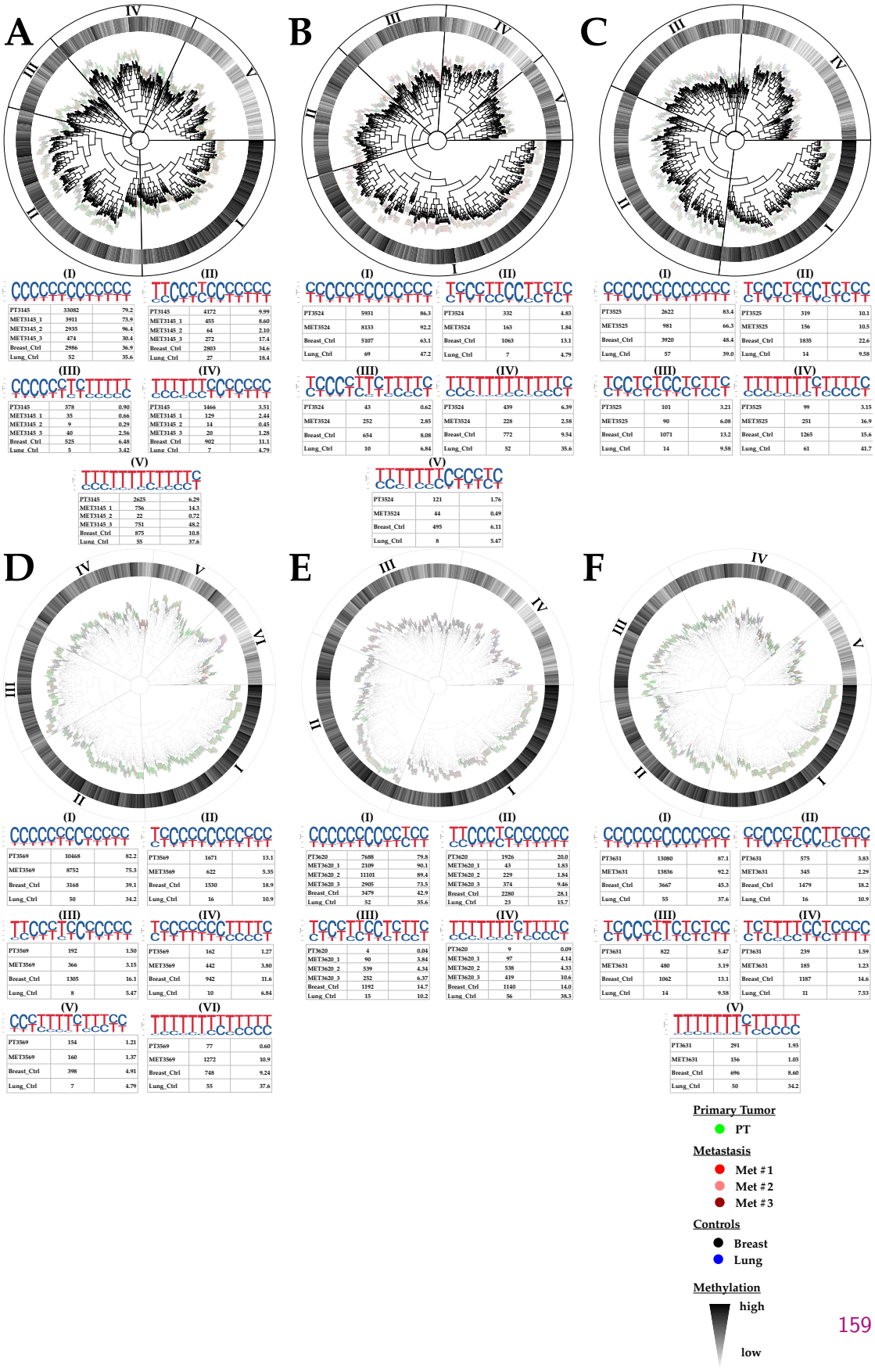


Figure B.4.: Phyloepignetic tree files & sequence logos for region C. (A) 3145 (B) 3524 (C) 3525 (D) 3569 (E) 3620 (F) 3631.

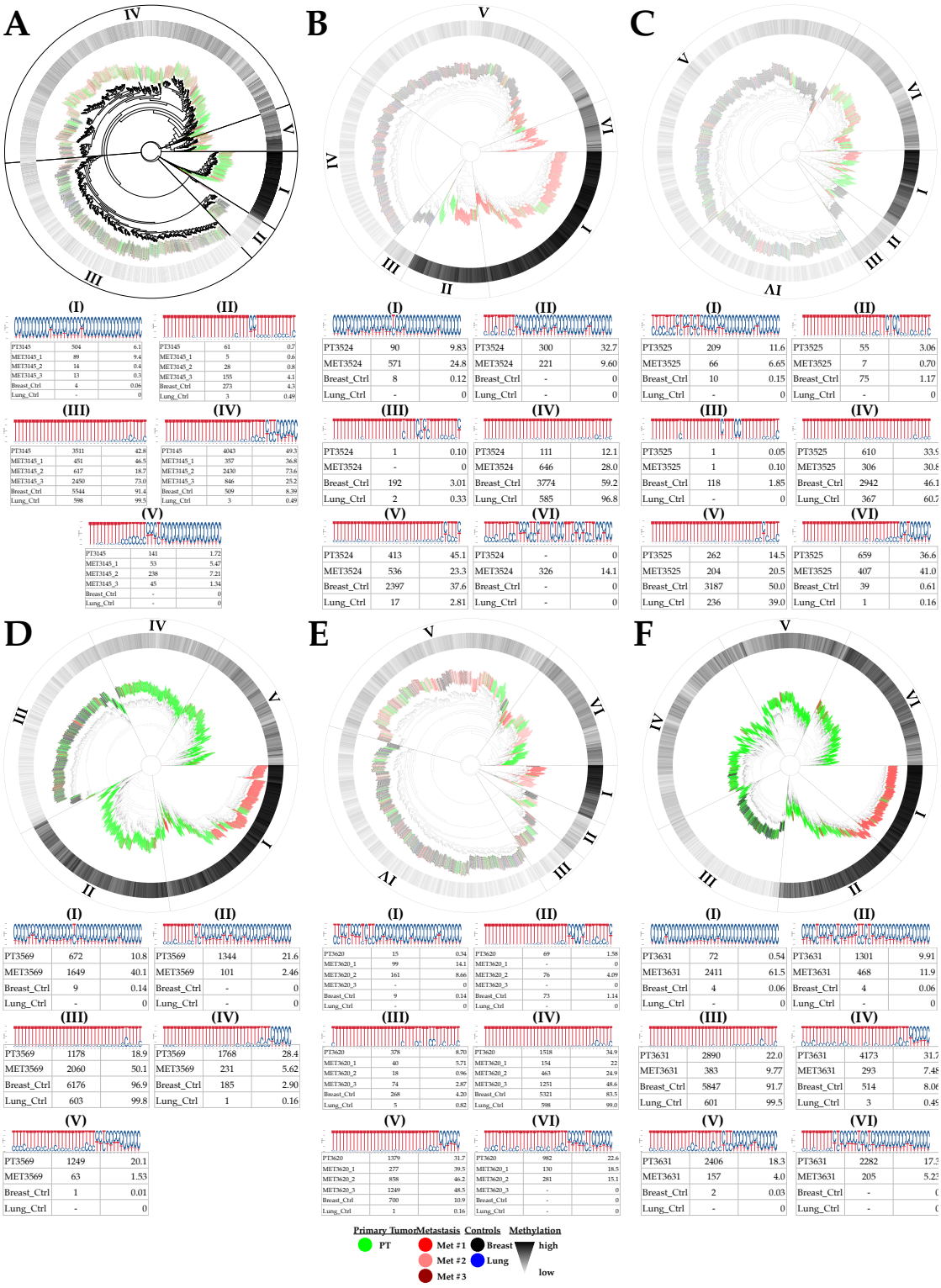


Figure B.5.: Phyloepignetic tree files and sequence logos for region D before subtracting low-methylated reads from primary tumor and metastasis samples. (A) 3145 (B) 3524 (C) 3525 (D) 3569 (E) 3620 (F) 3631.

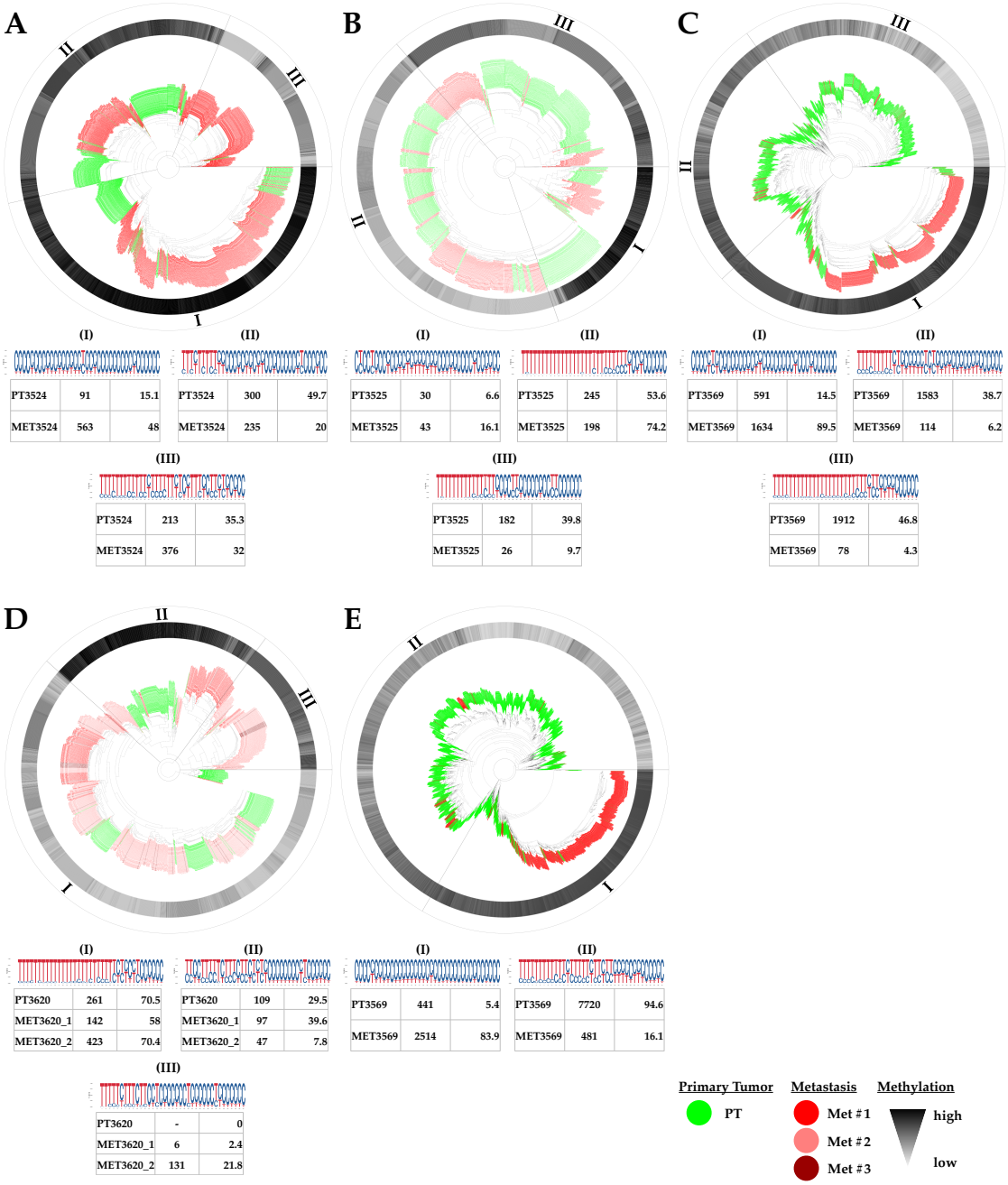


Figure B.6.: Phyloepignetic tree files & sequence logos for region D. (A) 3524 (B) 3525 (C) 3569 (D) 3620 (E) 3631.

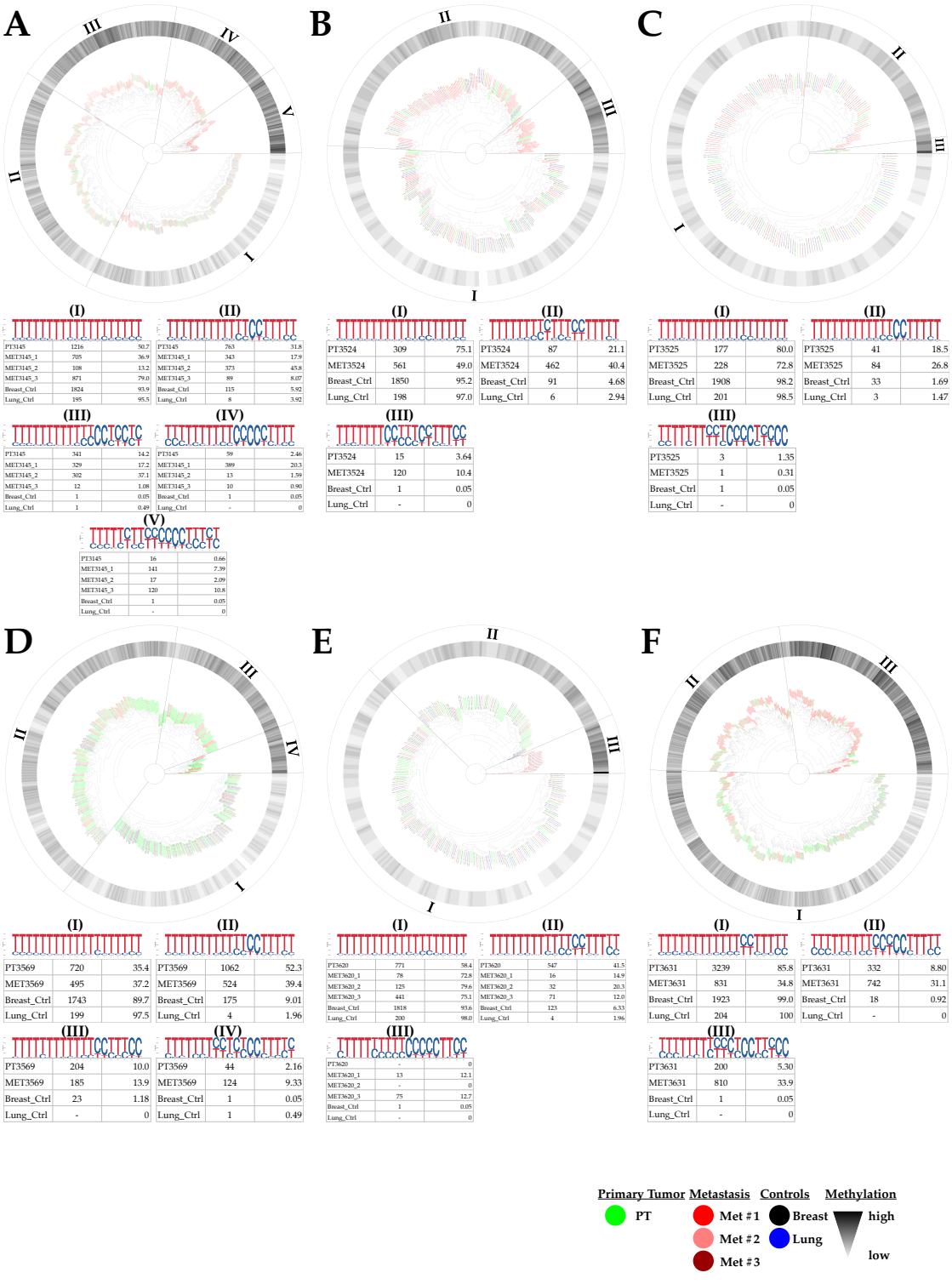


Figure B.7.: Phyloepignetic tree files and sequence logos for region E before subtracting low-methylated reads from primary tumor and metastasis samples. (A) 3145 (B) 3524 (C) 3525 (D) 3569 (E) 3620 (F) 3631.

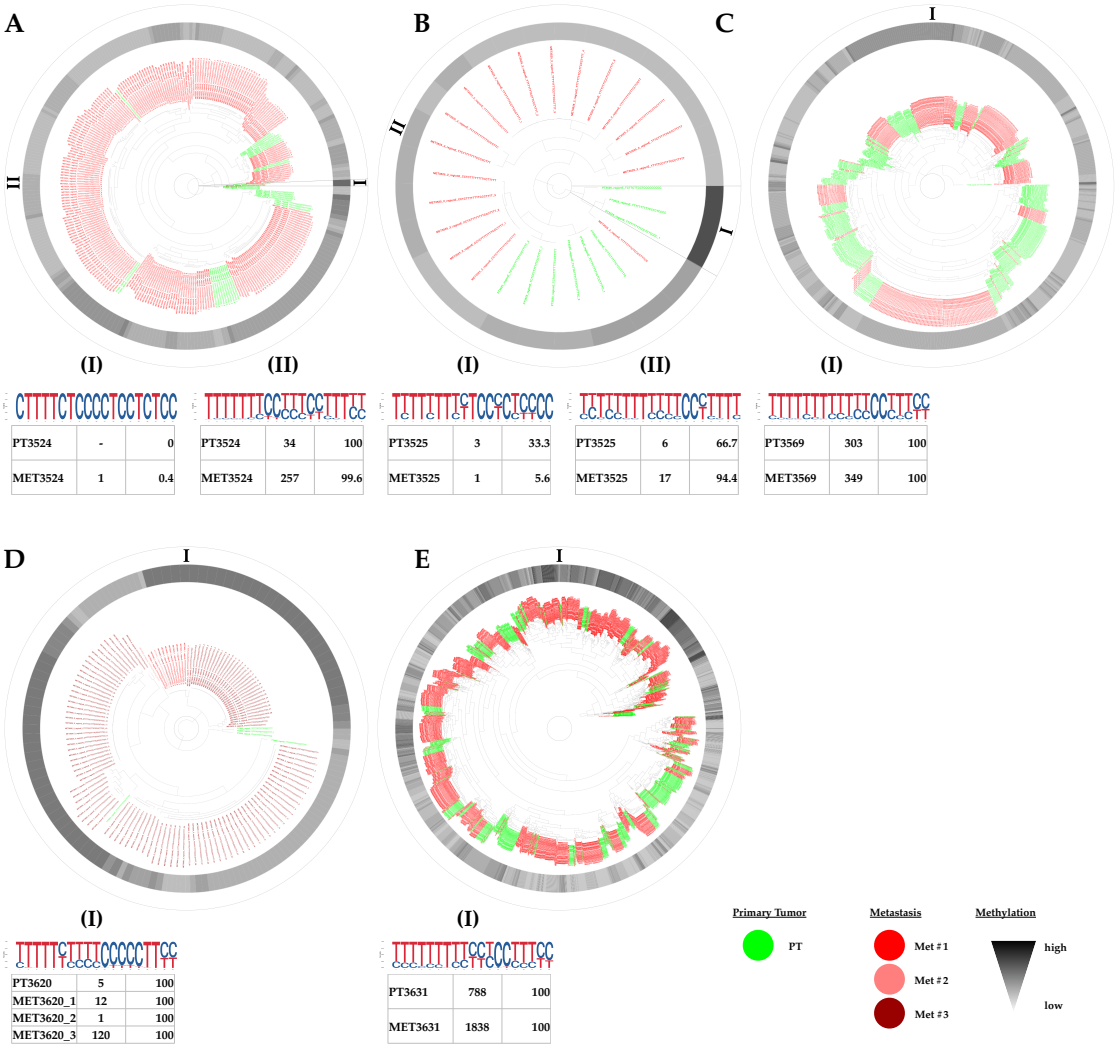


Figure B.8.: Phyloepignetic tree files & sequence logos for region E. (A) 3524 (B) 3525 (C) 3569 (D) 3620 (E) 3631.

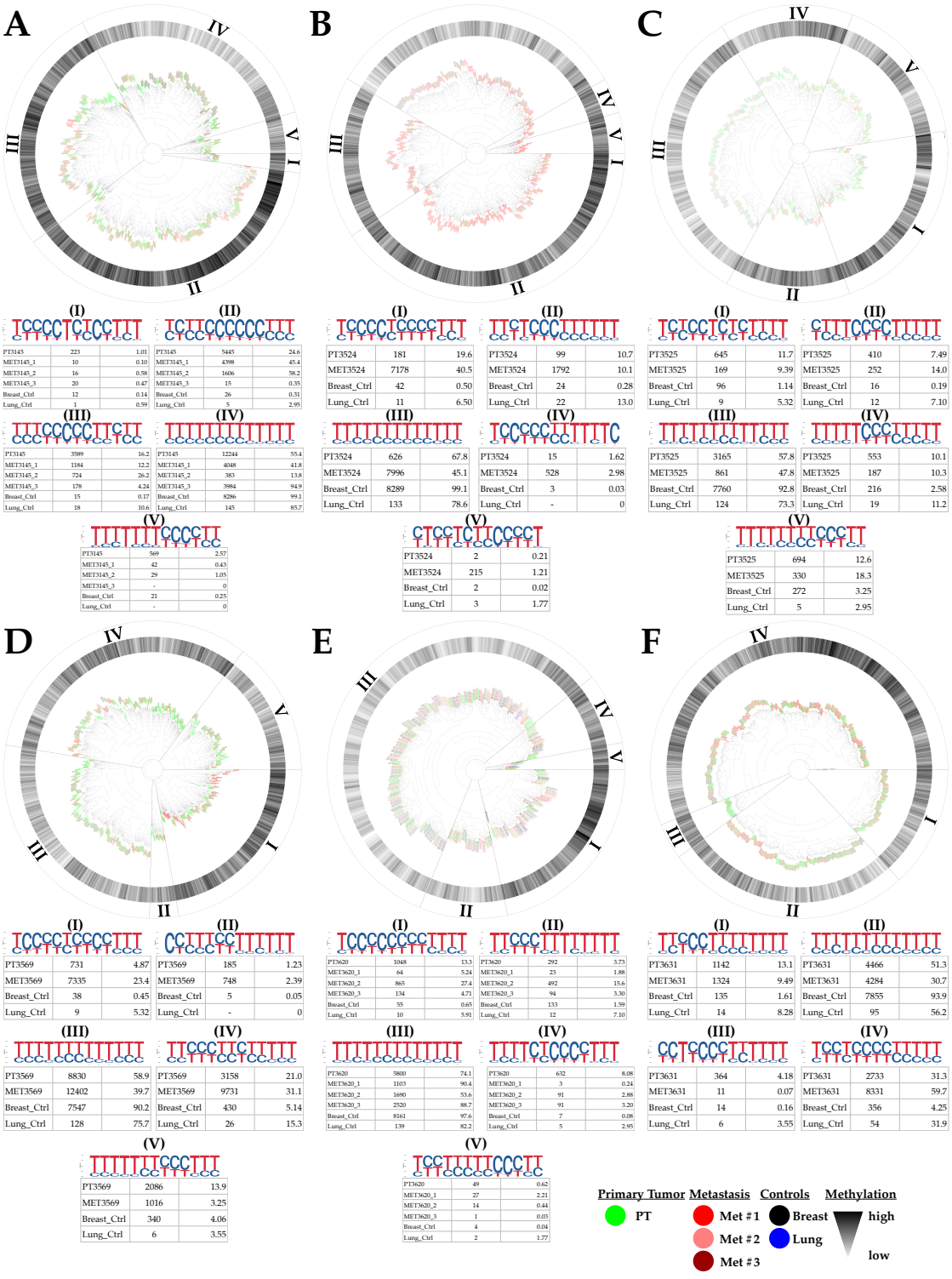


Figure B.9.: Phyloepignetic tree files and sequence logos for region F before subtracting low-methylated reads from primary tumor and metastasis samples. (A) 3145 (B) 3524 (C) 3525 (D) 3569 (E) 3620 (F) 3631.

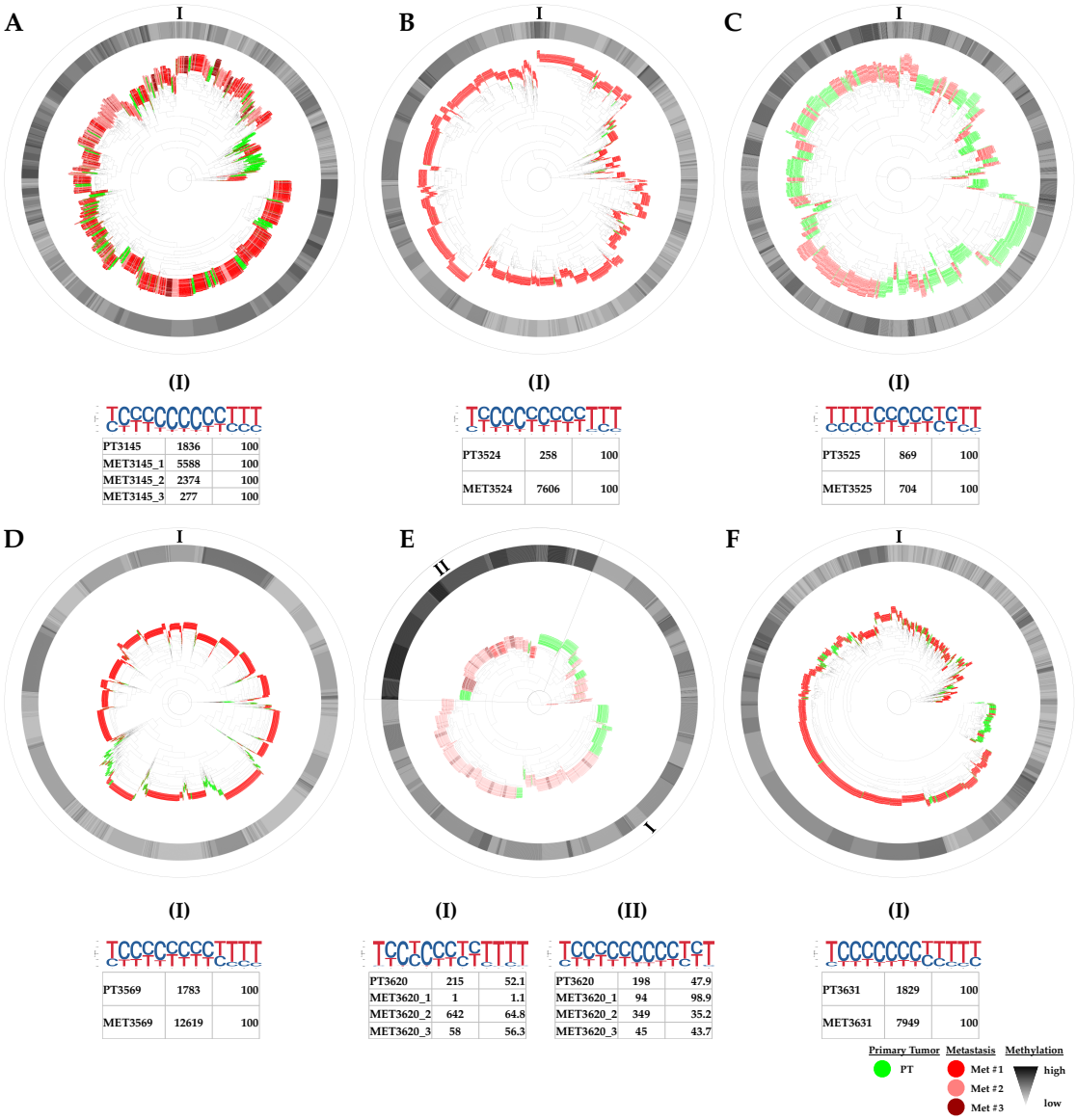


Figure B.10.: Phyloepignetic tree files & sequence logos for region F. (A) 3145 (B) 3524 (C) 3525 (D) 3569 (E) 3620 (F) 3631.

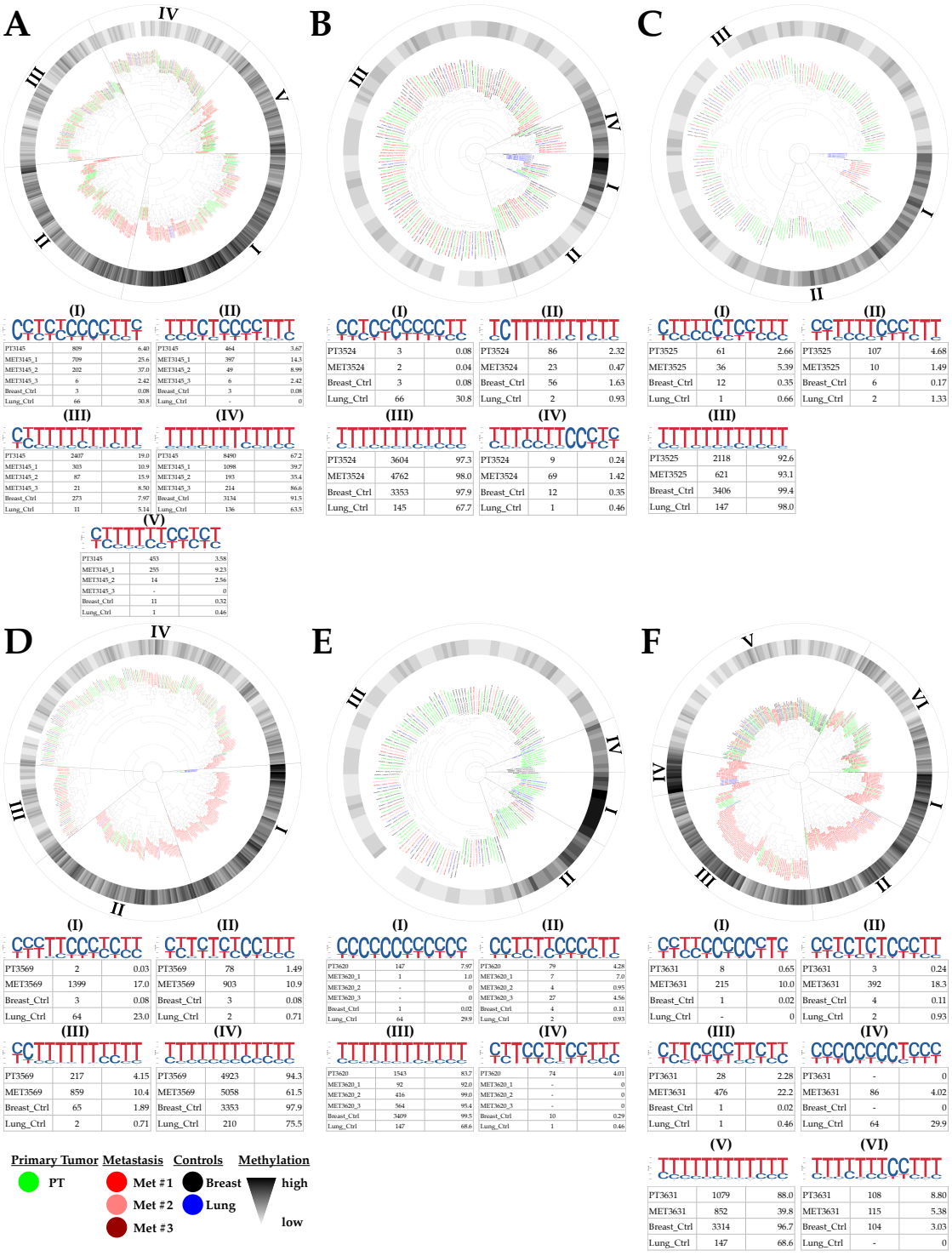


Figure B.11.: Phyloepignetic tree files and sequence logos for region G before subtracting low-methylated reads from primary tumor and metastasis samples. (A) 3145 (B) 3524 (C) 3525 (D) 3569 (E) 3620 (F) 3631.

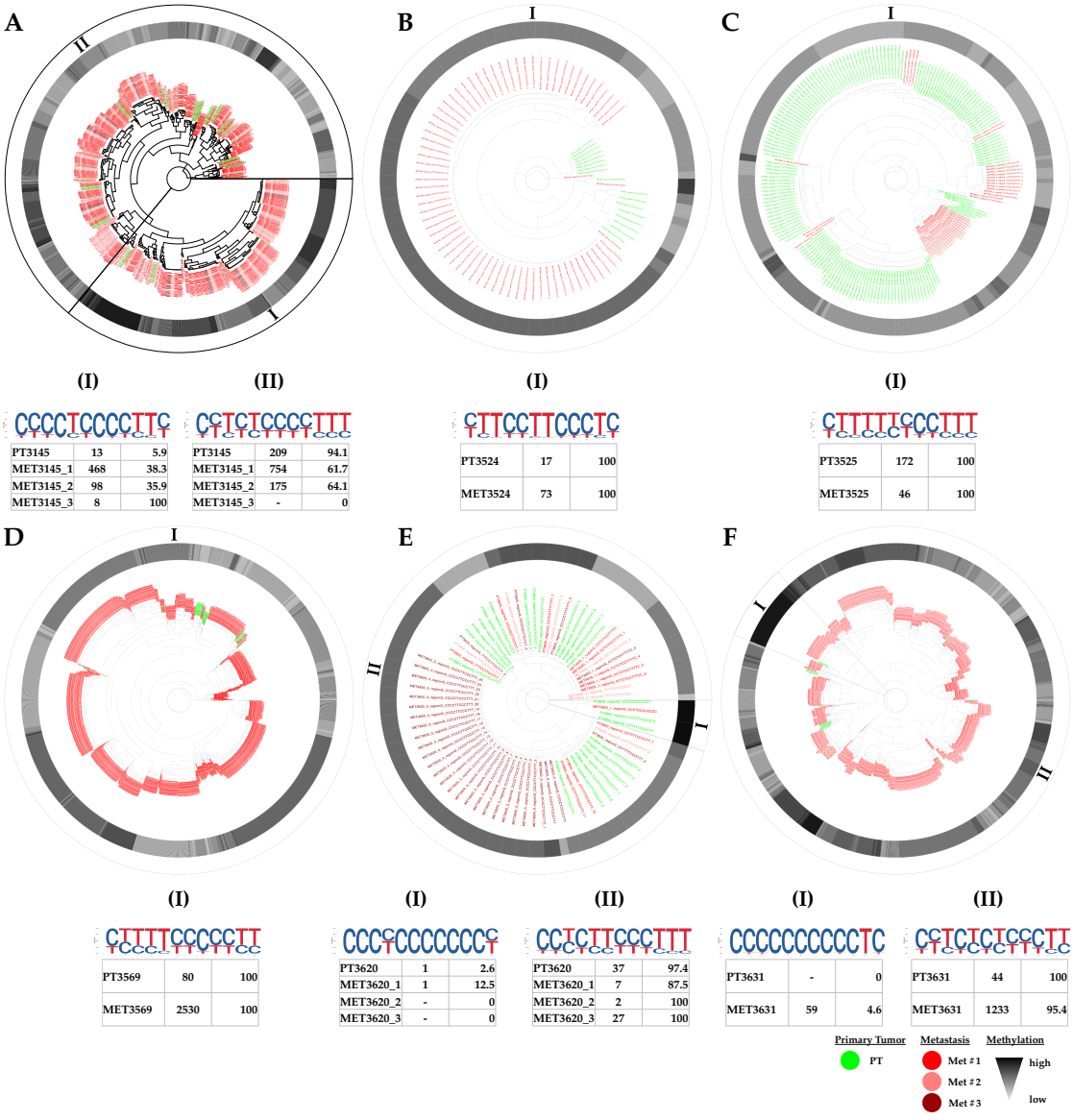


Figure B.12.: Phyloepignetic tree files & sequence logos for region G. (A) 3524 (B) 3525 (C) 3569 (D) 3620 (E) 3631.

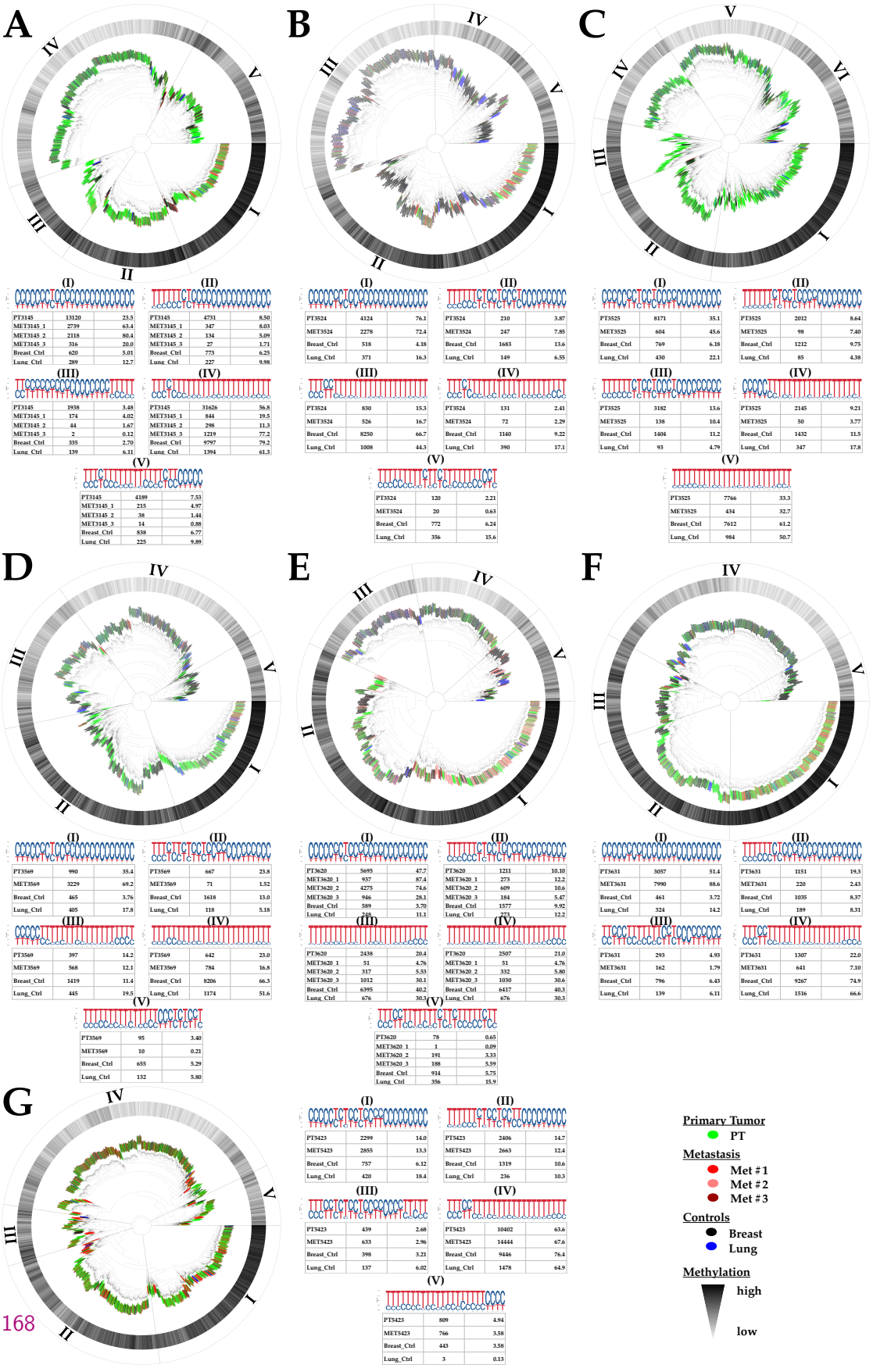


Figure B.13.: Phyloepignetic tree files & sequence logos for region J. (A) 3145 (B) 3524 (C) 3525 (D) 3569 (E) 3620 (F) 3631.

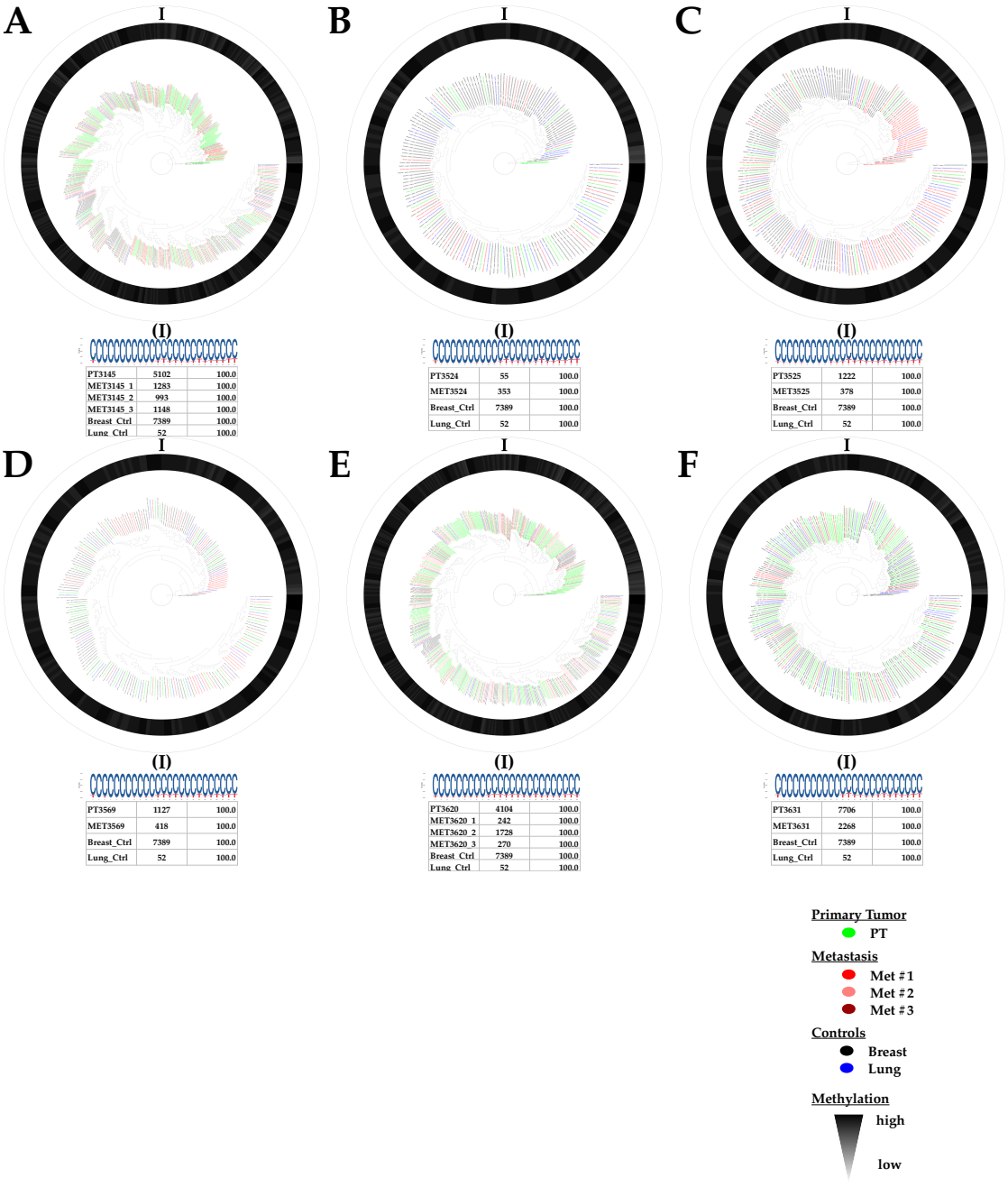


Figure B.14.: Phyloepignetic tree files & sequence logos for region L. (A) 3145 (B) 3524 (C) 3525 (D) 3569 (E) 3620 (F) 3631.

Bibliography

[1] J. P. Hamilton. "Epigenetics: principles and practice". In: *Digestive diseases* 29.2 (2011), pp. 130–135 (cit. on p. 1).

[2] G. Zhang and S. Pradhan. "Mammalian epigenetic mechanisms". In: *IUBMB life* 66.4 (2014), pp. 240–256 (cit. on p. 1).

[3] C. H. Waddington. "Canalization of development and the inheritance of acquired characters". In: *Nature* 150.3811 (1942), pp. 563–565 (cit. on pp. 1, 2).

[4] X. G. Pappalardo and V. Barra. "Losing DNA methylation at repetitive elements and breaking bad". In: *Epigenetics & chromatin* 14.1 (2021), pp. 1–21 (cit. on pp. 2, 3).

[5] A. Bird et al. "A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA". In: *Cell* 40.1 (1985), pp. 91–99 (cit. on p. 2).

[6] E. Li, C. Beard, and R. Jaenisch. "Role for DNA methylation in genomic imprinting". In: *Nature* 366.6453 (1993), pp. 362–365 (cit. on p. 2).

[7] B. Panning and R. Jaenisch. "DNA hypomethylation can activate Xist expression and silence X-linked genes." In: *Genes & development* 10.16 (1996), pp. 1991–2002 (cit. on p. 3).

[8] D. Jähner and R. Jaenisch. "Retrovirus-induced de novo methylation of flanking host sequences correlates with gene inactivity". In: *Nature* 315.6020 (1985), pp. 594–597 (cit. on p. 3).

[9] C. P. Walsh, J. R. Chaillet, and T. H. Bestor. “Transcription of IAP endogenous retroviruses is constrained by cytosine methylation”. In: *Nature genetics* 20.2 (1998), pp. 116–117 (cit. on p. 3).

[10] V. Colot, L. Maloisel, and J.-L. Rossignol. “Interchromosomal transfer of epigenetic states in Ascobolus: transfer of DNA methylation is mechanistically related to homologous recombination”. In: *Cell* 86.6 (1996), pp. 855–864 (cit. on p. 3).

[11] R. Z. Chen et al. “DNA hypomethylation leads to elevated mutation rates”. In: *Nature* 395.6697 (1998), pp. 89–93 (cit. on p. 3).

[12] T. Hashimshony et al. “The role of DNA methylation in setting up chromatin structure during development”. In: *Nature genetics* 34.2 (2003), pp. 187–192 (cit. on p. 3).

[13] D. N. Cooper and H. Youssoufian. “The CpG dinucleotide and human genetic disease”. In: *Human genetics* 78 (1988), pp. 151–155 (cit. on p. 3).

[14] P. W. Laird and R. Jaenisch. “DNA methylation and cancer”. In: *Human molecular genetics* 3.suppl_1 (1994), pp. 1487–1495 (cit. on p. 3).

[15] G. Pfeifer et al. “Polymerase chain reaction-aided genomic sequencing of an X chromosome-linked CpG island: methylation patterns suggest clonal inheritance, CpG site autonomy, and an explanation of activity state stability.” In: *Proceedings of the National Academy of Sciences* 87.21 (1990), pp. 8252–8256 (cit. on p. 3).

[16] M. Ehrlich et al. “Amount and distribution of 5-methylcytosine in human DNA from different types of tissues or cells”. In: *Nucleic acids research* 10.8 (1982), pp. 2709–2721 (cit. on p. 4).

[17] W. Schulz, C. Steinhoff, and A. Florl. “Methylation of endogenous human retroelements in health and disease”. In: *DNA Methylation: Development, Genetic Disease and Cancer* (2006), pp. 211–250 (cit. on p. 4).

[18] B. Gwynn et al. “Intracisternal A-particle element transposition into the murine β-glucuronidase gene correlates with loss of enzyme activity: a new model for β-glucuronidase deficiency in the C3H mouse”. In: *Molecular and cellular biology* 18.11 (1998), pp. 6474–6481 (cit. on p. 4).

[19] L. K. Hutnick et al. “Repression of retrotransposal elements in mouse embryonic stem cells is primarily mediated by a DNA methylation-independent mechanism”. In: *Journal of Biological Chemistry* 285.27 (2010), pp. 21082–21091 (cit. on p. 4).

[20] A. P. Bird. “Use of restriction enzymes to study eukaryotic DNA methylation: II. The symmetry of methylated sites supports semi-conservative copying of the methylation pattern”. In: *Journal of molecular biology* 118.1 (1978), pp. 49–60 (cit. on p. 4).

[21] R. S. Illingworth et al. “Orphan CpG islands identify numerous conserved promoters in the mammalian genome”. In: *PLoS genetics* 6.9 (2010), e1001134 (cit. on p. 4).

[22] T. S. Mikkelsen et al. “Genome-wide maps of chromatin state in pluripotent and lineage-committed cells”. In: *Nature* 448.7153 (2007), pp. 553–560 (cit. on p. 4).

[23] P. Carninci et al. “Genome-wide analysis of mammalian promoter architecture and evolution”. In: *Nature genetics* 38.6 (2006), pp. 626–635 (cit. on p. 5).

[24] J. D. Choi et al. “A novel variant of Inpp5f is imprinted in brain, and its expression is correlated with differential methylation of an internal CpG island”. In: *Molecular and cellular biology* 25.13 (2005), pp. 5514–5522 (cit. on p. 5).

[25] R. Zwart et al. “Bidirectional action of the Igf2r imprint control element on upstream and downstream imprinted genes”. In: *Genes & development* 15.18 (2001), pp. 2361–2366 (cit. on p. 5).

[26] F. Neri et al. “Intragenic DNA methylation prevents spurious transcription initiation”. In: *Nature* 543.7643 (2017), pp. 72–77 (cit. on p. 5).

[27] S. Naftelberg et al. “Regulation of alternative splicing through coupling with transcription and chromatin structure”. In: *Annual review of biochemistry* 84 (2015), pp. 165–198 (cit. on p. 6).

[28] S. K. Biswas et al. “The mammary gland: basic structure and molecular signaling during development”. In: *International Journal of Molecular Sciences* 23.7 (2022), p. 3883 (cit. on p. 6).

[29] A. Hermann, R. Goyal, and A. Jeltsch. “The Dnmt1 DNA-(cytosine-C5)-methyltransferase methylates DNA processively with high preference for hemimethylated target sites”. In: *Journal of Biological Chemistry* 279.46 (2004), pp. 48350–48359 (cit. on p. 6).

[30] M. Okano et al. “DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development”. In: *Cell* 99.3 (1999), pp. 247–257 (cit. on p. 6).

[31] S. Xie et al. “Cloning, expression and chromosome locations of the human DNMT3 gene family”. In: *Gene* 236.1 (1999), pp. 87–95 (cit. on p. 6).

[32] K. V. Morris et al. “Small interfering RNA-induced transcriptional gene silencing in human cells”. In: *Science* 305.5688 (2004), pp. 1289–1292 (cit. on p. 7).

[33] C. Brenner et al. “Myc represses transcription through recruitment of DNA methyltransferase corepressor”. In: *The EMBO journal* 24.2 (2005), pp. 336–346 (cit. on p. 7).

[34] C. Gebhard et al. “General Transcription Factor Binding at CpG Islands in Normal Cells Correlates with Resistance to De novo DNA Methylation in Cancer CellsTF Binding Excludes De novo DNA Methylation”. In: *Cancer research* 70.4 (2010), pp. 1398–1407 (cit. on p. 7).

[35] J. Otani et al. “Structural basis for recognition of H3K4 methylation status by the DNA methyltransferase 3A ATRX–DNMT3–DNMT3L domain”. In: *EMBO reports* 10.11 (2009), pp. 1235–1241 (cit. on p. 7).

[36] X. Nan, R. R. Meehan, and A. Bird. "Dissection of the methyl-CpG binding domain from the chromosomal protein MeCP2". In: *Nucleic acids research* 21.21 (1993), pp. 4886–4892 (cit. on p. 7).

[37] H. Kimura and K. Shiota. "Methyl-CpG-binding protein, MeCP2, is a target molecule for maintenance DNA methyltransferase, Dnmt1". In: *Journal of Biological Chemistry* 278.7 (2003), pp. 4806–4812 (cit. on p. 7).

[38] J. Sharif et al. "The SRA protein Np95 mediates epigenetic inheritance by recruiting Dnmt1 to methylated DNA". In: *Nature* 450.7171 (2007), pp. 908–912 (cit. on p. 8).

[39] J. M. Daniel et al. "The p120 ctn-binding partner Kaiso is a bi-modal DNA-binding protein that recognizes both a sequence-specific consensus and methylated CpG dinucleotides". In: *Nucleic acids research* 30.13 (2002), pp. 2911–2919 (cit. on p. 8).

[40] P. Blecua, L. Martinez-Verbo, and M. Esteller. "The DNA methylation landscape of hematological malignancies: An update". In: *Molecular Oncology* 14.8 (2020), pp. 1616–1639 (cit. on p. 8).

[41] S. Cortellino et al. "Thymine DNA glycosylase is essential for active DNA demethylation by linked deamination-base excision repair". In: *Cell* 146.1 (2011), pp. 67–79 (cit. on p. 8).

[42] S. Ito et al. "Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine". In: *Science* 333.6047 (2011), pp. 1300–1303 (cit. on p. 9).

[43] R. D. Kornberg. "Chromatin Structure: A Repeating Unit of Histones and DNA: Chromatin structure is based on a repeating unit of eight histone molecules and about 200 DNA base pairs." In: *Science* 184.4139 (1974), pp. 868–871 (cit. on p. 10).

[44] T. Kouzarides. "Chromatin modifications and their function". In: *Cell* 128.4 (2007), pp. 693–705 (cit. on p. 10).

[45] A. Dobrovic and L. S. Kristensen. “DNA methylation, epimutations and cancer predisposition”. In: *The international journal of biochemistry & cell biology* 41.1 (2009), pp. 34–39 (cit. on p. 10).

[46] A. Barski et al. “High-resolution profiling of histone methylations in the human genome”. In: *Cell* 129.4 (2007), pp. 823–837 (cit. on p. 10).

[47] F. Fuks et al. “The DNA methyltransferases associate with HP1 and the SUV39H1 histone methyltransferase”. In: *Nucleic acids research* 31.9 (2003), pp. 2305–2312 (cit. on p. 11).

[48] F. Fuks et al. “DNA methyltransferase Dnmt1 associates with histone deacetylase activity”. In: *Nature genetics* 24.1 (2000), pp. 88–91 (cit. on p. 11).

[49] P. Karagianni et al. “ICBP90, a novel methyl K9 H3 binding protein linking protein ubiquitination with heterochromatin formation”. In: *Molecular and cellular biology* 28.2 (2008), pp. 705–717 (cit. on p. 11).

[50] F. Fuks et al. “The methyl-CpG-binding protein MeCP2 links DNA methylation to histone methylation”. In: *Journal of Biological Chemistry* 278.6 (2003), pp. 4035–4040 (cit. on p. 11).

[51] E. Berezikov. “Evolution of microRNA diversity and regulation in animals”. In: *Nature Reviews Genetics* 12.12 (2011), pp. 846–860 (cit. on p. 11).

[52] L. Sinkkonen et al. “MicroRNAs control de novo DNA methylation through regulation of transcriptional repressors in mouse embryonic stem cells”. In: *Nature structural & molecular biology* 15.3 (2008), pp. 259–267 (cit. on p. 11).

[53] G. Cooper and K. Adams. *The cell: a molecular approach*. Oxford University Press, 2022 (cit. on p. 12).

[54] S. B. Baylin and P. A. Jones. “Epigenetic determinants of cancer”. In: *Cold Spring Harbor perspectives in biology* 8.9 (2016), a019505 (cit. on p. 12).

[55] R. Lakshminarasimhan and G. Liang. "The role of DNA methylation in cancer". In: *DNA Methyltransferases-Role and Function* (2016), pp. 151–172 (cit. on pp. 12, 13).

[56] A. P. Feinberg and B. Vogelstein. "Hypomethylation distinguishes genes of some human cancers from their normal counterparts". In: *Nature* 301.5895 (1983), pp. 89–92 (cit. on p. 12).

[57] M. Ehrlich and M. Lacey. "DNA hypomethylation and hemimethylation in cancer". In: *Epigenetic alterations in oncogenesis* (2012), pp. 31–56 (cit. on p. 12).

[58] D. B. Joseph, D. W. Strand, and C. M. Vezina. "DNA methylation in development and disease: an overview for prostate researchers". In: *American journal of clinical and experimental urology* 6.6 (2018), p. 197 (cit. on p. 13).

[59] S. G. Zhao et al. "The DNA methylation landscape of advanced prostate cancer". In: *Nature genetics* 52.8 (2020), pp. 778–789 (cit. on p. 13).

[60] K. Ohshima et al. "Whole-genome and epigenomic landscapes of malignant gastrointestinal stromal tumors harboring KIT exon 11 557–558 deletion mutations". In: *Cancer Research Communications* 3.4 (2023), pp. 684–696 (cit. on p. 13).

[61] R. Singal and G. D. Ginder. "DNA methylation". In: *Blood, The Journal of the American Society of Hematology* 93.12 (1999), pp. 4059–4070 (cit. on p. 13).

[62] N. Parsa. "Environmental factors inducing human cancers". In: *Iranian journal of public health* 41.11 (2012), p. 1 (cit. on p. 13).

[63] P. A. Jones. "Functions of DNA methylation: islands, start sites, gene bodies and beyond". In: *Nature Reviews Genetics* 13.7 (2012), pp. 484–492 (cit. on p. 14).

[64] S. B. Baylin et al. "DNA methylation patterns of the calcitonin gene in human lung cancers and lymphomas". In: *Cancer research* 46.6 (1986), pp. 2917–2922 (cit. on p. 14).

[65] T. Hinoue et al. "Genome-scale analysis of aberrant DNA methylation in colorectal cancer". In: *Genome research* 22.2 (2012), pp. 271–282 (cit. on p. 14).

[66] Q. Yi et al. “TRIM29 hypermethylation drives esophageal cancer progression via suppression of ZNF750”. In: *Cell Death Discovery* 9.1 (2023), p. 191 (cit. on p. 14).

[67] E. B. Kartaloglu et al. “Identification of ALX3 Gene Promoter Hypermethylation as a Potential Biomarker for Lung Cancer”. In: *Anticancer Research* 43.7 (2023), pp. 3029–3036 (cit. on p. 14).

[68] G. P. Pfeifer. “Defining driver DNA methylation changes in human cancer”. In: *International journal of molecular sciences* 19.4 (2018), p. 1166 (cit. on p. 15).

[69] P. J. Stephens et al. “The landscape of cancer genes and mutational processes in breast cancer”. In: *Nature* 486.7403 (2012), pp. 400–404 (cit. on p. 14).

[70] C. Curtis et al. “The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups”. In: *Nature* 486.7403 (2012), pp. 346–352 (cit. on p. 15).

[71] B. W̃. H. H̃. M. S. C. L. 9. 1. P. P. J. 1. K. R. 13, G. data analysis: Baylor College of Medicine Creighton Chad J. 22 23 Donehower Lawrence A. 22 23 24 25, I. for Systems Biology Reynolds Sheila 31 Kreisberg Richard B. 31 Bernard Brady 31 Bressler Ryan 31 Erkkila Timo 32 Lin Jake 31 Thorsson Vesteinn 31 Zhang Wei 33 Shmulevich Ilya 31, et al. “Comprehensive molecular portraits of human breast tumours”. In: *Nature* 490.7418 (2012), pp. 61–70 (cit. on p. 16).

[72] F. Fang et al. “Breast cancer methylomes establish an epigenomic foundation for metastasis”. In: *Science translational medicine* 3.75 (2011), 75ra25–75ra25 (cit. on p. 16).

[73] S. Pandya and R. G. Moore. “Breast development and anatomy”. In: *Clinical obstetrics and gynecology* 54.1 (2011), pp. 91–95 (cit. on p. 17).

[74] J. R. Hens and J. J. Wysolmerski. “Key stages of mammary gland development: molecular mechanisms involved in the formation of the embryonic mammary gland”. In: *Breast Cancer Research* 7 (2005), pp. 1–5 (cit. on p. 18).

[75] T. Sakakura, Y. Sakagami, and Y. Nishizuka. "Dual origin of mesenchymal tissues participating in mouse mammary gland embryogenesis". In: *Developmental biology* 91.1 (1982), pp. 202–207 (cit. on p. 18).

[76] D. L. Page et al. "Lobular neoplasia of the breast: higher risk for subsequent invasive cancer predicted by more extensive disease". In: *Human pathology* 22.12 (1991), pp. 1232–1239 (cit. on p. 18).

[77] W. Lyons. "Hormonal synergism in mammary growth". In: *Proceedings of the Royal Society of London. Series B-Biological Sciences* 149.936 (1958), pp. 303–325 (cit. on p. 18).

[78] S. Nandi. "Endocrine control of mammary-gland development and function in the C3H/He Crgl mouse". In: *Journal of the National Cancer Institute* 21.6 (1958), pp. 1039–1063 (cit. on p. 18).

[79] J. D. Figueroa et al. "Terminal duct lobular unit involution of the normal breast: implications for breast cancer etiology". In: *JNCI: Journal of the National Cancer Institute* 106.10 (2014) (cit. on p. 18).

[80] C. E. DeSantis et al. "Breast cancer statistics, 2015: Convergence of incidence rates between black and white women". In: *CA: a cancer journal for clinicians* 66.1 (2016), pp. 31–42 (cit. on p. 18).

[81] G. M. Crisi et al. "Invasive mammary carcinoma after immediate and short-term follow-up for lobular neoplasia on core biopsy". In: *The American journal of surgical pathology* 27.3 (2003), pp. 325–333 (cit. on p. 19).

[82] I. Diel et al. "Detection of tumor cells in bone marrow of patients with primary breast cancer: a prognostic factor for distant metastasis." In: *Journal of clinical oncology* 10.10 (1992), pp. 1534–1539 (cit. on p. 19).

[83] L. Zhang et al. "The identification and characterization of breast cancer CTCs competent for brain metastasis". In: *Science translational medicine* 5.180 (2013), 180ra48–180ra48 (cit. on p. 19).

[84] T. Sørlie et al. “Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications”. In: *Proceedings of the National Academy of Sciences* 98.19 (2001), pp. 10869–10874 (cit. on p. 19).

[85] C. S. Vallejos et al. “Breast cancer classification according to immunohistochemistry markers: subtypes and association with clinicopathologic variables in a peruvian hospital database”. In: *Clinical breast cancer* 10.4 (2010), pp. 294–300 (cit. on p. 19).

[86] G. Valabrega, F. Montemurro, and M. Aglietta. “Trastuzumab: mechanism of action, resistance and future perspectives in HER2-overexpressing breast cancer”. In: *Annals of oncology* 18.6 (2007), pp. 977–984 (cit. on p. 19).

[87] D. J. Slamon et al. “Studies of the HER-2/neu proto-oncogene in human breast and ovarian cancer”. In: *Science* 244.4905 (1989), pp. 707–712 (cit. on p. 19).

[88] S. Pollok et al. “Human Cdc45 is a proliferation-associated antigen”. In: *The FEBS journal* 274.14 (2007), pp. 3669–3684 (cit. on p. 19).

[89] A. Jubb et al. “Achaete-scute like 2 (ascl2) is a target of Wnt signalling and is upregulated in intestinal neoplasia”. In: *Oncogene* 25.24 (2006), pp. 3445–3457 (cit. on p. 19).

[90] D. J. Slamon et al. “Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2”. In: *New England journal of medicine* 344.11 (2001), pp. 783–792 (cit. on p. 19).

[91] J. Singh, K. Jhaveri, and F. Esteva. “HER2-positive advanced breast cancer: optimizing patient outcomes and opportunities for drug development”. In: *British journal of cancer* 111.10 (2014), pp. 1888–1898 (cit. on p. 19).

[92] H. Kennecke et al. “Metastatic behavior of breast cancer subtypes”. In: *Journal of clinical oncology* 28.20 (2010), pp. 3271–3277 (cit. on p. 19).

[93] C. L. Chaffer and R. A. Weinberg. “A perspective on cancer cell metastasis”. In: *science* 331.6024 (2011), pp. 1559–1564 (cit. on p. 20).

[94] I. J. Fidler. "The pathogenesis of cancer metastasis: the'seed and soil'hypothesis revisited". In: *Nature reviews cancer* 3.6 (2003), pp. 453–458 (cit. on p. 20).

[95] I. Baccelli et al. "Identification of a population of blood circulating tumor cells from breast cancer patients that initiates metastasis in a xenograft assay". In: *Nature biotechnology* 31.6 (2013), pp. 539–544 (cit. on p. 20).

[96] Y. Kienast et al. "Real-time imaging reveals the single steps of brain metastasis formation". In: *Nature medicine* 16.1 (2010), pp. 116–122 (cit. on p. 20).

[97] K. J. Luzzi et al. "Multistep nature of metastatic inefficiency: dormancy of solitary cells after successful extravasation and limited survival of early micrometastases". In: *The American journal of pathology* 153.3 (1998), pp. 865–873 (cit. on p. 20).

[98] J. P. Thiery et al. "Epithelial-mesenchymal transitions in development and disease". In: *cell* 139.5 (2009), pp. 871–890 (cit. on p. 20).

[99] G. P. Gupta and J. Massagué. "Cancer metastasis: building a framework". In: *Cell* 127.4 (2006), pp. 679–695 (cit. on p. 20).

[100] S. Valastyan and R. A. Weinberg. "Tumor metastasis: molecular insights and evolving paradigms". In: *Cell* 147.2 (2011), pp. 275–292 (cit. on pp. 21, 26).

[101] D. M. Brown and E. Ruoslahti. "Metadherin, a cell surface protein in breast tumors that mediates lung metastasis". In: *Cancer cell* 5.4 (2004), pp. 365–374 (cit. on p. 21).

[102] A. Al-Mehdi et al. "Intravascular origin of metastasis from the proliferation of endothelium-attached tumor cells: a new model for metastasis". In: *Nature medicine* 6.1 (2000), pp. 100–102 (cit. on p. 21).

[103] B. Psaila and D. Lyden. "The metastatic niche: adapting the foreign soil". In: *Nature Reviews Cancer* 9.4 (2009), pp. 285–293 (cit. on p. 21).

[104] X. H.-F. Zhang et al. "Latent bone metastasis in breast cancer tied to Src-dependent survival signals". In: *Cancer cell* 16.1 (2009), pp. 67–78 (cit. on p. 21).

[105] A. F. Chambers, A. C. Groom, and I. C. MacDonald. “Dissemination and growth of cancer cells in metastatic sites”. In: *Nature Reviews Cancer* 2.8 (2002), pp. 563–572 (cit. on p. 22).

[106] S. Paget. “The distribution of secondary growths in cancer of the breast.” In: *The Lancet* 133.3421 (1889), pp. 571–573 (cit. on p. 22).

[107] Y. Kang et al. “A multigenic program mediating breast cancer metastasis to bone”. In: *Cancer cell* 3.6 (2003), pp. 537–549 (cit. on p. 22).

[108] A. J. Minn et al. “Genes that mediate breast cancer metastasis to lung”. In: *Nature* 436.7050 (2005), pp. 518–524 (cit. on p. 22).

[109] P. D. Bos et al. “Genes that mediate breast cancer metastasis to the brain”. In: *Nature* 459.7249 (2009), pp. 1005–1009 (cit. on p. 22).

[110] S. Tabaries et al. “Claudin-2 is selectively enriched in and promotes the formation of breast cancer liver metastases through engagement of integrin complexes”. In: *Oncogene* 30.11 (2011), pp. 1318–1328 (cit. on p. 22).

[111] P. C. Nowell. “The Clonal Evolution of Tumor Cell Populations: Acquired genetic lability permits stepwise selection of variant sublines and underlies tumor progression.” In: *Science* 194.4260 (1976), pp. 23–28 (cit. on p. 22).

[112] E. R. Fearon and B. Vogelstein. “A genetic model for colorectal tumorigenesis”. In: *cell* 61.5 (1990), pp. 759–767 (cit. on p. 23).

[113] S. Ramaswamy et al. “A molecular signature of metastasis in primary solid tumors”. In: *Nature genetics* 33.1 (2003), pp. 49–54 (cit. on p. 23).

[114] B. Weigelt et al. “Molecular portraits and 70-gene prognosis signature are preserved throughout the metastatic process of breast cancer”. In: *Cancer research* 65.20 (2005), pp. 9155–9158 (cit. on p. 23).

[115] M. Olivier et al. “The clinical value of somatic TP53 gene mutations in 1,794 patients with breast cancer.” In: *Clinical cancer research: an official journal of the American Association for Cancer Research* 12.4 (2006), pp. 1157–1167 (cit. on p. 23).

[116] C. A. Klein and D. Hölzel. "Systemic cancer progression and tumor dormancy: mathematical models meet single cell genomics". In: *Cell cycle* 5.16 (2006), pp. 1788–1798 (cit. on p. 23).

[117] O. Schmidt-Kittler et al. "From latent disseminated cells to overt metastasis: genetic analysis of systemic breast cancer progression". In: *Proceedings of the National Academy of Sciences* 100.13 (2003), pp. 7737–7742 (cit. on p. 23).

[118] C. A. Klein. "The metastasis cascade". In: *Science* 321.5897 (2008), pp. 1785–1787 (cit. on p. 24).

[119] A. L. Schechter et al. "The neu oncogene: an erb-B-related gene encoding a 185,000-M r tumour antigen". In: *Nature* 312 (1984), pp. 513–516 (cit. on p. 25).

[120] N. Robert et al. "Randomized phase III study of trastuzumab, paclitaxel, and carboplatin compared with trastuzumab and paclitaxel in women with HER-2–overexpressing metastatic breast cancer". In: *Journal of Clinical Oncology* 24.18 (2006), pp. 2786–2792 (cit. on p. 25).

[121] E. Sinn et al. "Coexpression of MMTV/v-Ha-ras and MMTV/c-myc genes in transgenic mice: synergistic action of oncogenes in vivo". In: *Cell* 49.4 (1987), pp. 465–475 (cit. on p. 25).

[122] J. Ursini-Siegel et al. "Insights from transgenic mouse models of ERBB2-induced breast cancer". In: *Nature Reviews Cancer* 7.5 (2007), pp. 389–397 (cit. on p. 25).

[123] W. J. Muller et al. "Single-step induction of mammary adenocarcinoma in transgenic mice bearing the activated c-neu oncogene". In: *Cell* 54.1 (1988), pp. 105–115 (cit. on p. 25).

[124] F. Lucchini et al. "Early and multifocal tumors in breast, salivary, harderian and epididymal tissues developed in MMTY-Neu transgenic mice". In: *Cancer letters* 64.3 (1992), pp. 203–209 (cit. on p. 25).

[125] K. Boggio et al. “Interleukin 12–mediated prevention of spontaneous mammary adenocarcinomas in two lines of Her-2/neu transgenic mice”. In: *The Journal of experimental medicine* 188.3 (1998), pp. 589–596 (cit. on p. 25).

[126] S. A. Patel and S. Vanharanta. “Epigenetic determinants of metastasis”. In: *Molecular oncology* 11.1 (2017), pp. 79–96 (cit. on p. 26).

[127] N. McGranahan and C. Swanton. “Clonal heterogeneity and tumor evolution: past, present, and the future”. In: *Cell* 168.4 (2017), pp. 613–628 (cit. on p. 26).

[128] H. Hosseini et al. “Early dissemination seeds metastasis in breast cancer”. In: *Nature* 540.7634 (2016), pp. 552–558 (cit. on p. 27).

[129] R. K. Saiki et al. “Analysis of enzymatically amplified β-globin and HLA-DQα DNA with allele-specific oligonucleotide probes”. In: *Nature* 324.6093 (1986), pp. 163–166 (cit. on p. 46).

[130] H. Towbin, T. Staehelin, and J. Gordon. “Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: procedure and some applications.” In: *Proceedings of the national academy of sciences* 76.9 (1979), pp. 4350–4354 (cit. on p. 50).

[131] D. Y. Guschin et al. “A rapid and general assay for monitoring endogenous gene modification”. In: *Engineered Zinc Finger Proteins: Methods and Protocols* (2010), pp. 247–256 (cit. on p. 52).

[132] A. Akalin et al. “methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles”. In: *Genome biology* 13.10 (2012), pp. 1–9 (cit. on pp. 55, 56, 118).

[133] M. Frommer et al. “A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands.” In: *Proceedings of the National Academy of Sciences* 89.5 (1992), pp. 1827–1831 (cit. on p. 55).

[134] H. Wickham et al. “Welcome to the Tidyverse”. In: *Journal of open source software* 4.43 (2019), p. 1686 (cit. on p. 56).

[135] S. Gaertner. *Evolution epigenetischer Aberrationen bei der malignen Progression von Brustkrebs*. Jan. 2024. URL: https://epub.uni-regensburg.de/51419/ (cit. on pp. 59, 60).

[136] M. Sonnet et al. "Enrichment of methylated DNA by methyl-CpG immunoprecipitation". In: *Lymphoma: Methods and Protocols* (2013), pp. 201–212 (cit. on p. 60).

[137] A. Riebler et al. "BayMeth: improved DNA methylation quantification for affinity capture sequencing data using a flexible Bayesian approach". In: *Genome biology* 15.2 (2014), pp. 1–20 (cit. on p. 60).

[138] A. Holler. "Methylation Patterns in the BALB-neuT Breast Cancer Mouse Model". In: (Mar. 2018) (cit. on p. 61).

[139] J. Deng et al. "Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming". In: *Nature biotechnology* 27.4 (2009), pp. 353–360 (cit. on pp. 62, 63).

[140] D. Diep et al. "Library-free methylation sequencing with bisulfite padlock probes". In: *Nature methods* 9.3 (2012), pp. 270–272 (cit. on p. 63).

[141] S. Andrews et al. *FastQC: a quality control tool for high throughput sequence data*. 2010 (cit. on pp. 64, 67).

[142] Y. Han et al. "New targeted approaches for epigenetic age predictions". In: *BMC biology* 18 (2020), pp. 1–15 (cit. on p. 64).

[143] F. Krueger. "Trim Galore!: A wrapper around Cutadapt and FastQC to consistently apply adapter and quality trimming to FastQ files, with extra functionality for RRBS data". In: *Babraham Institute* (2015) (cit. on p. 67).

[144] J. Zhang et al. "PEAR: a fast and accurate Illumina Paired-End reAd mergeR". In: *Bioinformatics* 30.5 (2014), pp. 614–620 (cit. on p. 67).

[145] F. Krueger and S. R. Andrews. "Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications". In: *bioinformatics* 27.11 (2011), pp. 1571–1572 (cit. on pp. 67, 68).

[146] H. Li et al. "The sequence alignment/map format and SAMtools". In: *bioinformatics* 25.16 (2009), pp. 2078–2079 (cit. on p. 67).

[147] Y. Zhang and Z. Zhang. "The history and advances in cancer immunotherapy: understanding the characteristics of tumor-infiltrating immune cells and their therapeutic implications". In: *Cellular & molecular immunology* 17.8 (2020), pp. 807–821 (cit. on p. 69).

[148] E. Becht et al. "Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression". In: *Genome biology* 17 (2016), pp. 1–20 (cit. on p. 69).

[149] J. Porter et al. "Improving bisulfite short-read mapping efficiency with hairpin-bisulfite data". In: *2014 IEEE 4th International Conference on Computational Advances in Bio and Medical Sciences (ICCABS)*. IEEE. 2014, pp. 1–2 (cit. on p. 72).

[150] M.-B. Worm Ørntoft et al. "Comparative analysis of 12 different kits for bisulfite conversion of circulating cell-free DNA". In: *Epigenetics* 12.8 (2017), pp. 626–636 (cit. on p. 73).

[151] H. Clifford et al. "Comparison of clustering methods for investigation of genome-wide methylation array data". In: *Frontiers in genetics* 2 (2011), p. 88 (cit. on p. 75).

[152] F. Sievers et al. "Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega". In: *Molecular systems biology* 7.1 (2011), p. 539 (cit. on p. 76).

[153] N. Saitou and T. Imanishi. "Relative efficiencies of the Fitch-Margoliash, maximum-parsimony, maximum-likelihood, minimum-evolution, and neighbor-joining methods of phylogenetic tree construction in obtaining the correct tree". In: (1989) (cit. on p. 76).

[154] B. Kolaczkowski and J. W. Thornton. "Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous". In: *Nature* 431.7011 (2004), pp. 980–984 (cit. on p. 76).

[155] I. Gronau and S. Moran. "Optimal implementations of UPGMA and other common clustering algorithms". In: *Information processing letters* 104.6 (2007), pp. 205–210 (cit. on p. 77).

[156] B. Efron, E. Halloran, and S. Holmes. "Bootstrap confidence levels for phylogenetic trees". In: *Proceedings of the National Academy of Sciences* 93.23 (1996), pp. 13429–13429 (cit. on p. 79).

[157] E. Hodges et al. "High definition profiling of mammalian DNA methylation by array capture and single molecule bisulfite sequencing". In: *Genome research* 19.9 (2009), pp. 1593–1605 (cit. on p. 90).

[158] S. M. Mitchell et al. "A panel of genes methylated with high frequency in colorectal cancer". In: *BMC cancer* 14 (2014), pp. 1–15 (cit. on p. 118).

[159] X. Man et al. "DNMT3A and DNMT3B in Breast Tumorigenesis and Potential Therapy". In: *Frontiers in Cell and Developmental Biology* (2022), p. 1020 (cit. on p. 123).

[160] Z. Yu et al. "DNA methyltransferase 1/3a overexpression in sporadic breast cancer is associated with reduced expression of estrogen receptor-alpha/breast cancer susceptibility gene 1 and poor prognosis". In: *Molecular carcinogenesis* 54.9 (2015), pp. 707–719 (cit. on p. 124).

[161] D. Watanabe et al. "Stage-and cell-specific expression of Dnmt3a and Dnmt3b during embryogenesis". In: *Mechanisms of development* 118.1-2 (2002), pp. 187–190 (cit. on p. 124).

[162] D. Karolchik et al. "The UCSC genome browser database". In: *Nucleic acids research* 31.1 (2003), pp. 51–54 (cit. on p. 125).

[163] S. Tajima et al. “Domain structure of the Dnmt1, Dnmt3a, and Dnmt3b DNA methyltransferases”. In: *DNA Methyltransferases-Role and Function* (2016), pp. 63–86 (cit. on p. 125).

[164] K. Kibe et al. “The DNMT3A PWWP domain is essential for the normal DNA methylation landscape in mouse somatic cells and oocytes”. In: *PLoS genetics* 17.5 (2021), e1009570 (cit. on p. 125).

[165] S. Tajima et al. “Domain structure of the Dnmt1, Dnmt3a, and Dnmt3b DNA methyltransferases”. In: *DNA Methyltransferases-Role and Function* (2022), pp. 45–68 (cit. on p. 125).

[166] T. Chen et al. “Establishment and maintenance of genomic methylation patterns in mouse embryonic stem cells by Dnmt3a and Dnmt3b”. In: *Molecular and cellular biology* 23.16 (2003), pp. 5594–5605 (cit. on p. 126).

[167] B. Shen et al. “Efficient genome modification by CRISPR-Cas9 nickase with minimal off-target effects”. In: *Nature methods* 11.4 (2014), pp. 399–402 (cit. on p. 127).

[168] M. Reyngold et al. “Remodeling of the methylation landscape in breast cancer metastasis”. In: *PloS one* 9.8 (2014), e103896 (cit. on pp. 133, 144).

[169] S. Veronese et al. “Proliferation index as a prognostic marker in breast cancer”. In: *Cancer* 71.12 (1993), pp. 3926–3931 (cit. on p. 136).

[170] P. A. Jones and S. B. Baylin. “The epigenomics of cancer”. In: *Cell* 128.4 (2007), pp. 683–692 (cit. on p. 136).

[171] F. J. Carmona et al. “Epigenetic disruption of cadherin-11 in human cancer metastasis”. In: *The Journal of pathology* 228.2 (2012), pp. 230–240 (cit. on p. 136).

[172] B. Győrffy et al. “Aberrant DNA methylation impacts gene expression and prognosis in breast cancer subtypes”. In: *International journal of cancer* 138.1 (2016), pp. 87–97 (cit. on p. 137).

[173] S. E. Elsheikh et al. “Global histone modifications in breast cancer correlate with tumor phenotypes, prognostic factors, and patient outcome”. In: *Cancer research* 69.9 (2009), pp. 3802–3809 (cit. on p. 137).

[174] D. I. Rodenhiser et al. “Epigenetic mapping and functional analysis in a breast cancer metastasis model using whole-genome promoter tiling microarrays”. In: *Breast cancer research* 10.4 (2008), pp. 1–15 (cit. on p. 137).

[175] M. Qi and X. Xiong. “Promoter hypermethylation of RARβ2, DAPK, hMLH1, p14, and p15 is associated with progression of breast cancer: A PRISMA-compliant meta-analysis”. In: *Medicine* 97.51 (2018) (cit. on p. 138).

[176] Y. Mao et al. “Stromal cells in tumor microenvironment and breast cancer”. In: *Cancer and Metastasis Reviews* 32 (2013), pp. 303–315 (cit. on p. 139).

[177] H. F. Dvorak. “Tumors: wounds that do not heal—redux”. In: *Cancer immunology research* 3.1 (2015), pp. 1–11 (cit. on p. 140).

[178] Y. Kojima et al. “Autocrine TGF-β and stromal cell-derived factor-1 (SDF-1) signaling drives the evolution of tumor-promoting mammary stromal myofibroblasts”. In: *Proceedings of the National Academy of Sciences* 107.46 (2010), pp. 20009–20014 (cit. on p. 140).

[179] U. E. Martinez-Outschoorn et al. “Autophagy in cancer associated fibroblasts promotes tumor cell survival: Role of hypoxia, HIF1 induction and NFκB activation in the tumor stromal microenvironment”. In: *Cell cycle* 9.17 (2010), pp. 3515–3533 (cit. on p. 140).

[180] H. H. Al-Khalaf and A. Aboussekhra. “p16INK4A positively regulates p21WAF1 expression by suppressing AUF1-dependent mRNA decay”. In: *PloS one* 8.7 (2013), e70133 (cit. on p. 140).

[181] A. Sadlonova et al. “Breast fibroblasts modulate epithelial cell proliferation in three-dimensional in vitro co-culture”. In: *Breast Cancer Research* 7 (2004), pp. 1–14 (cit. on p. 140).

[182] H. J. Hugo et al. “Contribution of fibroblast and mast cell (afferent) and tumor (efferent) IL-6 effects within the tumor microenvironment”. In: *Cancer microenvironment* 5 (2012), pp. 83–93 (cit. on p. 140).

[183] A. Orimo et al. “Stromal fibroblasts present in invasive human breast carcinomas promote tumor growth and angiogenesis through elevated SDF-1/CXCL12 secretion”. In: *Cell* 121.3 (2005), pp. 335–348 (cit. on p. 140).

[184] B.-Z. Qian et al. “CCL2 recruits inflammatory monocytes to facilitate breast-tumour metastasis”. In: *Nature* 475.7355 (2011), pp. 222–225 (cit. on p. 140).

[185] W. Tan et al. “Tumour-infiltrating regulatory T cells stimulate mammary cancer metastasis through RANKL–RANK signalling”. In: *Nature* 470.7335 (2011), pp. 548–553 (cit. on p. 140).

[186] R. D. Leek et al. “Association of macrophage infiltration with angiogenesis and prognosis in invasive breast carcinoma”. In: *Cancer research* 56.20 (1996), pp. 4625–4629 (cit. on p. 141).

[187] S. Nilsson et al. “Downregulation of miR-92a is associated with aggressive breast cancer features and increased tumour macrophage infiltration”. In: *PloS one* 7.4 (2012), e36051 (cit. on p. 141).

[188] L. R. de la Fuente et al. “Tumor dissociation of highly viable cell suspensions for single-cell omic analyses in mouse models of breast cancer”. In: *STAR protocols* 2.4 (2021), p. 100841 (cit. on p. 141).

[189] Y. Li et al. “Aberrant promoter hypermethylation inhibits RGMA expression and contributes to tumor progression in breast cancer”. In: *Oncogene* 41.3 (2022), pp. 361–371 (cit. on p. 143).

[190] S. Y. Park et al. “Promoter CpG island hypermethylation during breast cancer progression”. In: *Virchows Archiv* 458 (2011), pp. 73–84 (cit. on p. 143).

[191] B. P. Berman et al. “Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina–associated domains”. In: *Nature genetics* 44.1 (2012), pp. 40–46 (cit. on p. 145).

[192] K. D. Hansen et al. “Increased methylation variation in epigenetic domains across cancer types”. In: *Nature genetics* 43.8 (2011), pp. 768–775 (cit. on p. 145).

[193] C. Ili et al. “Landscape of genome-wide DNA methylation of colorectal cancer metastasis”. In: *Cancers* 12.9 (2020), p. 2710 (cit. on p. 145).

[194] E. A. Fernández et al. “MIXTURE: an improved algorithm for immune tumor microenvironment estimation based on gene expression data”. In: *BioRxiv* (2019), p. 726562 (cit. on p. 146).

[195] J. Zheng et al. “DNA methylation affects metastasis of renal cancer and is associated with TGF-β/RUNX3 inhibition”. In: *Cancer Cell International* 18 (2018), pp. 1–11 (cit. on p. 146).

[196] M. Gerstung et al. “Quantifying cancer progression with conjunctive Bayesian networks”. In: *Bioinformatics* 25.21 (2009), pp. 2809–2815 (cit. on p. 147).

[197] D. Ramazzotti et al. “CAPRI: efficient inference of cancer progression models from cross-sectional data”. In: *Bioinformatics* 31.18 (2015), pp. 3016–3026 (cit. on p. 147).

[198] R. Schill et al. “Modelling cancer progression using mutual hazard networks”. In: *Bioinformatics* 36.1 (2020), pp. 241–249 (cit. on p. 147).