SILESIAN UNIVERSITY OF TECHNOLOGY

Faculty of Automatic Control, Electronics and Computer Science

Mathematical modelling in comparative analysis of methylation profiles of *de novo* and therapy-related AML

Doctoral Dissertation

by

Agnieszka Cecotka

Supervisors

Professor Joanna Polańska, PhD, DSc

Christophe Badie, PhD

2022

Gliwice, POLAND

I would like to thank prof. Joanna Polańska
for supervising this work and giving direction for each step.
Also, I wish to thank dr. Christophe Badie
for his priceless remarks and concepts.

Therefore, thanks to my beloved husband Adam,
for supporting and believing in me
and to my son Michał, who is my greatest motivation.
Many thanks to my whole family, who made it all possible.

Contents

List of abbreviations

AML	acute myeloid leukaemia
APL	acute promyelocytic leukaemia
BH	Benjamini-Hochberg
BIC	Bayesian Information Criterion
BM	bone marrow
BMIQ	Beta MIxture Quantile dilation
CDF	cumulative distribution function
ChAMP	Chip Analysis Methylation Pipeline
CHD	coronary heart disease
chemo-AML	chemotherapy-related acute myeloid leukaemia
CI	confidence interval
CLL	chronic lymphocytic leukaemia
ComBat	Combining Batches of expression data
CpG	5'—Cytosine—phosphate—Guanine—3'
DMP	differentially methylated position
DMR	differentially methylated regions
DNMT	DNA methyltransferase
EM	expectation-maximization
FAB	French-American-British classification of AMLs
FDR	false discovery rate
GEO	Gene Expression Omnibus
GMM	Gaussian mixture model
GO	Gene Ontology
hiPathia	High-throughput Pathway Analysis
HL	Hodges–Lehmann statistics
HP1	heterochromatin protein 1
HPCE	high-performance capillary electrophoresis
HPLC	high-performance liquid chromatography
HSC	hematopoietic stem cell
IMA	Illumina Methylation Analyzer
KEGG	Kyoto Encyclopedia of Genes and Genomes

lncRNA	long non-coding RNA
MAD	median absolute deviation
MBP	methyl-binding protein
MDS	myelodysplastic syndrome
MECP	methyl-CpG binding protein
miRNA	microRNA
MPN	myeloproliferative neoplasms
MS-DGGE	methylation-specific denaturing gradient gel electrophoresis
MS-DHPLC	methylation-specific denaturing high-performance liquid chromatography
MSP	methylation-specific polymerase chain reaction
MSRE	methylation-sensitive restriction endonuclease
Ms-SnuPE	methylation-sensitive single-nucleotide primer extension
MS-SSCA	methylation-specific single-strand conformation analysis
ncRNA	non-coding RNA
PCA	Principal Component Analysis
PCR	polymerase chain reaction
PDF	probability density function
piRNA	piwi-interacting RNA
qPCR	quantitative polymerase chain reaction
radio-AML	radiotherapy-related acute myeloid leukaemia
RLGS	restriction landmark genomic scanning
RRBS	reduced representation bisulfite sequencing
RS	regulatory sequence
SD	standard deviation
siRNA	small interfering RNA
SMART	simple modular architecture research tool
SNP	single-nucleotide-polymorphism
STRING	search tool for recurring instances of neighbouring genes
SWAN	Subset-quantile Within Array Normalization
t-AML	therapy-related acute myeloid leukaemia
TCGA	The Cancer Genome Atlas
TLC	thin-layer chromatography
t-MDS	therapy-related myelodysplastic syndrome
TSS	transcription start site
WHO	World Health Organization

Abstract

The expansion of high throughput experimental techniques leads to collecting a massive amount of data that must be processed. These data's characteristics are the huge number of features and the nontypical statistical distribution of the obtained signal. Hence, there is a need to develop new algorithms and pipelines to process them. One of phenomena measured with high throughput techniques is DNA methylation - epigenetic modification crucial for gene expression control and cancer development. Aberrations of DNA methylation in Acute Myeloid Leukaemia (AML) might be a reason for differences in treatment response and survival time between patients with *de novo* and therapy-related AML (AML being a side effect of previous malignancy treatment). Therefore, this thesis aims to develop pipelines for processing the DNA methylation data and investigate the DNA methylation profile in AML patients. The greatest approaches were developed, selected, and described from data preprocessing, statistical analysis, mathematical modelling, and functional analysis of detected features to validate the results with different experimental platforms.

Initially, the original algorithm for finding DNA methylation profile of AML is presented and implemented for data obtained with methylation microarrays. It is a composition of mathematical modelling and statistical approaches to conclude about low, medium, high and extremely high hyper- or hypomethylation of genome sites or regions. Subsequently, the detection of differences between genders in DNA methylation levels and survival factors in specific genomic regions in AML patients is investigated. It uses the integration of results obtained in a comparative and survival analyses. Moreover, the pipeline for detecting aberrations in DNA methylation in *de novo* and chemo- or radiotherapy-related AML is described. It is drawn upon supervised and unsupervised feature selection in epigenetics and functional analysis domains. The result is the detection of several biomarkers of therapy-related AML, confirmed in an independent, pyrosequencing experiment.

Streszczenie

Rozwój wysokoprzepustowych technik eksperymentalnych prowadzi do wytwarzania ogromnej ilości danych. Charakteryzuje je duża liczba mierzonych cech oraz nietypowy rozkład statystyczny otrzymywanego sygnału. Stąd potrzeba opracowania nowych algorytmów i sekwencji metod do ich przetwarzania. Jednym ze zjawisk mierzonych za pomocą technik wysokoprzepustowych jest metylacja DNA - modyfikacja epigenetyczna, kluczowa dla kontroli ekspresji genów i rozwoju raka. Zmiany metylacji DNA w ostrej białaczce szpikowej (AML) mogą być przyczyną różnic w odpowiedzi na leczenie i czasie przeżycia między pacjentami z samoistną białaczką i z białaczką będącą skutkiem ubocznym terapii innego nowotworu. Dlatego celem niniejszej pracy jest opracowanie schematów do przetwarzania danych dotyczących metylacji DNA i zbadanie profilu metylacji DNA u pacjentów z AML. Opracowano, wybrano i opisano najlepsze podejścia od wstępnego przetwarzania danych, poprzez analizę statystyczną, modelowanie matematyczne, do analizy funkcjonalnej wykrytych cech i walidacji wyników uzyskanych z wykorzystaniem różnych platform eksperymentalnych.

Na początku pracy przedstawiono i zaimplementowano oryginalny algorytm poszukiwania profilu metylacji DNA w AML, dla danych uzyskanych za pomocą mikromacierzy metylacyjnych. To połączenie modelowania matematycznego i metod statystycznych, pozwala na wnioskowanie o niskiej, średniej, wysokiej i bardzo wysokiej hiper- lub hipometylacji miejsc i regionów genomu. Następnie badano wykrywanie różnic między płciami w poziomach metylacji DNA i czynnikach przeżycia u pacjentów z ostrą białaczką szpikową. W tym podejściu wykorzystano integrację wyników uzyskanych w analizie porównawczej i analizie przeżycia. Ponadto opisano metodę wykrywania aberracji metylacji DNA u pacjentów z białaczką samoistną lub związaną z chemio- lub radioterapią. Wykorzystuje ona nadzorowaną i nienadzorowaną selekcję cech w przestrzeni epigenetycznej i analizy funkcjonalnej. Wykryto kilka biomarkerów ostrej białaczki szpikowej związanej z terapią, potwierdzonych w niezależnym eksperymencie, za pomocą pirosekwencjonowania.

1. Introduction

1.1. Motivation

The development of high throughput experimental techniques has progressed in the last years. The possibility to measure thousands of features in the same experiment leads to the collection of a massive amount of data that must be processed. However, huge number of features often do not correspond to similar number of observations. This difficulty and also nontypical distributions of measured data are challenges when seeking the algorithms for these data analysis. Hence, there is a demand for new data analysis and processing methods dedicated to a specific experimental platform.

Cancer diseases are commonly examined with the use of high throughput techniques. In such malignancies, aberrations occur in each stage of gene expression. They can be caused by mutations and epigenetic modifications, which impact gene expression level and, indirectly, cell proteome's qualitative and quantitative composition. All of these modifications influence cell functionality. One of the crucial phenomena is DNA methylation. It is one of the most important epigenetic processes. Methylation of gene promoters, especially those rich in CpG sites, is essential for gene expression control. The development of proper statistical methods would lead to detecting alterations in DNA methylation level during cancer. Such information is crucial for understanding the mechanism of malignancy and impacts the way of treatment.

One of the cancer diseases in which DNA plays an essential role is acute myeloid leukaemia (AML). It can initiate and develop by itself, and it is called *de novo* AML in this case. It can also be a long-term side effect of the previous malignancy treatment and then it is called therapy-related AML (t-AML). It can be distinguished into radiotherapy-related AML and chemotherapy-related AML. Patients with t-AML are characterised by worse treatment response and survival rate than *de novo* AML patients.

The epigenetic mechanism behind these differences needs to be investigated. A proper composition of different statistical approaches and mathematical modelling methods will enable the detection and selection of the most critical disparities. In the future, it can result in the development of a therapy, considering the specificity of DNA methylation aberrations in t-AML and being dedicated to radio- and chemotherapy-related AML patients. Until now, no study of the epigenome-wide DNA methylation profile of t-AML has been conducted.

This dissertation describes the research of DNA methylation in AML of various aetiology. The greatest approaches were developed, selected, and described from data preprocessing techniques, statistical analysis, mathematical modelling, and functional analysis of detected features to validation the results with different platforms.

1.2. Aim of the work

The objective of this work was to develop and select proper methods dedicated to analysis the DNA methylation data, which enables the detection of features differentiating examined patient groups. The research methodology includes an overview of biological mechanisms in acute myeloid leukaemia and epigenetic processes, existing approaches for DNA methylation data collection and analysis, and the invention and implementation of new pipelines for differential features detection. The expected results of this work comprise the development of new procedures for analysing DNA methylation data, especially coming from small samples, and their application.

Based on the motivation and aim of this thesis, the following statements were formulated:

1. The composition of methods based on mathematical modelling, comparative statistical analysis and their results integration, used for DNA methylation data analysis, enable the detection of differences in DNA methylation levels in genomic regions among examined patient groups.

2. Integration of results acquired using different experimental platforms leads to obtaining validated findings that are less susceptible to errors.

1.3. Chapter contents

The Background chapter introduces epigenetic processes and their functions in the living organism. Furthermore, the genesis, epidemiology, classification, therapy, prognosis, and DNA methylation aberrations in acute myeloid leukaemia are described. Then, therapy-related AML and literature findings regarding epigenetic alteration in this type of AML are presented. In the end, techniques for measuring DNA methylation and methods for analysing DNA methylation data are investigated.

The Materials and Methods chapter presents the description of the analysed datasets and methods used for their analysis. In each case, the methodology consists of DNA methylation level distribution analysis, detection of differentiating probes among examined patient groups, integration of related outcomes to find genomic regions characteristic for specific cases, and functional analysis. Additionally, the research investigating differences between genders in AML includes survival analysis, while different AML types research is extended into markers detection and validation and unsupervised feature selection.

The Results chapter presents the most important findings obtained in the presented analyses. First, an acute myeloid leukaemia profile examination based on the differences between AML and healthy sample distribution is presented. Next, methylation level differences between genders in specific genomic regions are investigated and compared with healthy control. Furthermore, the detection of genomic regions that can impact survival is described. Then, methylation profiles of various patient groups are characterised, as well as detection and validation of differentially methylated genomic regions in *de novo* AML, chemotherapy-related AML, and radiotherapy-related AML are presented. Finally, the integration of results from all of the analyses is reported.

The Conclusion chapter summarises the obtained outcomes and expands them with the biological interpretation.

2. Background

2.1. Epigenetics and its role

The term "epigenetics" refers to molecular processes impacting the changes in genes' functionality - e.g., their expression level, which leads to different phenotypes with persistent DNA sequence.

The term "epigenetics" was used for the first time in 1942 by Conrad Waddington [1]. He proposed the definition as "the branch of biology which studies the causal interactions between genes and their products which bring the phenotype into being [2]." In the original meaning, epigenetics refers to all molecular processes that regulate genotype expression, which become apparent as phenotype variants. Nowadays, the sense of this term has evolved into "the study of changes in gene function that are mitotically and/or meiotically heritable and that do not entail a change in the DNA sequence" [3]. However, epigenetic modifications of the genome impact not only gene expression and silencing but also non-coding regions [4].

The most important epigenetic mechanisms include DNA methylation, histone modifications, and pre- and post-transcriptional gene regulation by small non-coding RNAs.

2.1.1. DNA methylation

In mammals, DNA methylation is a modification of cytosine into 5-methylcytosine in CpG sites of the genome. CpG sites are symmetrical dinucleotides where cytosine is followed by guanine. "p" represents the phosphate group between nucleosides. "5" refers to the fifth position of methylated carbon in the cytosine ring. *De novo* DNA methylation is catalysed by two methyltransferases, Dnmt3a and Dnmt3b [5]. The process is the addition of a methyl group onto unmethylated DNA, which determines the methylation pattern. After cell division, methylation patterns are maintained by Dnmt1 methyltransferase. The scheme of the abovementioned enzymes activity is presented in Figure 2.1.

DNA methylation plays a crucial role in establishing parental imprinting during gametogenesis and also repressing retrotransposons and silencing genes on an inactivated X chromosome.

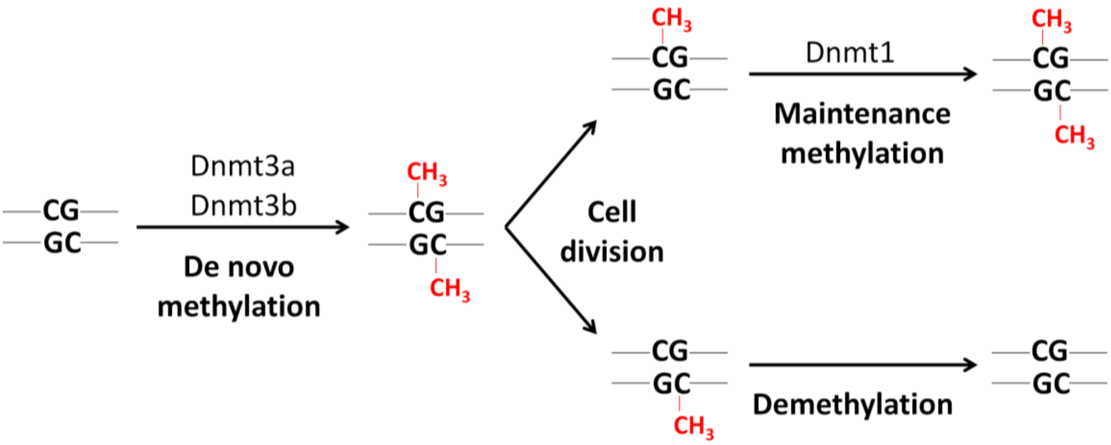

Figure 2.1 Scheme of activity of Dnmt3a and Dnmt3b in *de novo* methylation and Dnmt1 in maintenance methylation.

CpG dinucleotides occur with a frequency lower than expected across almost the whole genome. However, in some areas, their incidence is higher than expected - these areas are called CpG islands [6]. CpG islands are not equally distributed in the genome. They are mainly concentrated in gene promoter regions [7]. However, a lot of them are localised inside genes or intergenic areas. Intergenically located CpG sites can be transcription start sites for non-coding RNA [8].

2.1.2. Histone modifications

The best-examined histone modification process is histone acetylation. The histone acetylation process is binding an acetyl group onto lysine on the N-termini of histones.

Histone modifications are widespread changes in histone status. Modifications in the N-termini of H3 and H4 histones are well examined. H3 histone has lysine amino acid in several positions. Lysine can be mono-, di- or tri-methylated. The trimethylation of lysine 4 (in the fourth position) and acetylation of lysine 9 is an activation process.

It weakens histone-DNA affinity and leads to a lack of chromatin compaction. Methylation of lysine 9 enables the binding of heterochromatin protein 1 (HP1), which leads to inactivation. Genes in heterochromatin are usually inactivated. Heterochromatin structure prevents polymerase accession (Figure 2.2).

DNA methylation also has an impact on chromatin structure. A hydrogen bond between 5-methylcytosine and guanine is 1.8 times stronger than that between cytosine and guanine. Proteins that bind to methyl-CpG sequences (MECPs - methyl-CpG binding proteins) cooperate with methylated DNA. MECP2 directly blocks IIB transcription factor or forms a complex with histone deacetylases to modulate and condense chromatin structure [9].

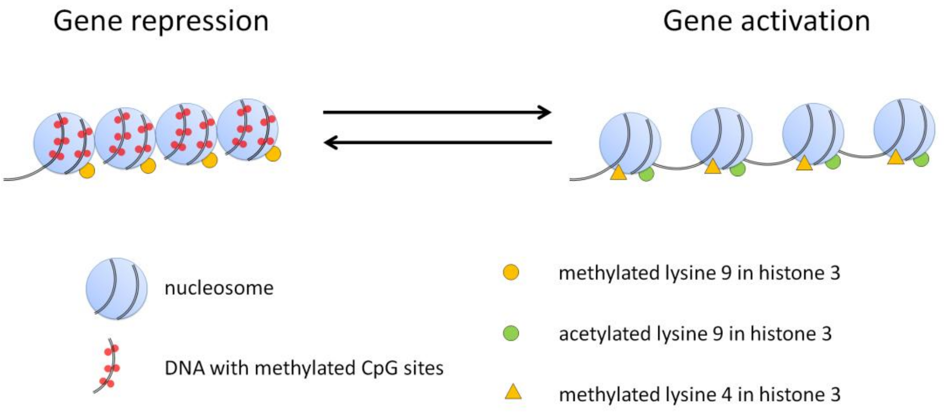
Gene repression
Gene activation
nucleosome
DNA with methylated CpG sites
methylated lysine 9 in histone 3
acetylated lysine 9 in histone 3
methylated lysine 4 in histone 3

Figure 2.2 A scheme of histone modifications during gene activation and repression.

2.1.3. Pre- and post-transcriptional gene regulation by small non-coding RNAs.

The study on *Schizosaccharomyces pombe* yeast proves that interference-RNA also participates in chromatin modification [10]. Double-stranded RNA prevents the accumulation of transcripts. Deletion of these mechanism elements can lead to impairment of centromere function and loss of histone H3 methylation. The siRNA (small interfering RNA) coming from centromere-homologous repeat initiates methylation of lysine 9 in histone 3. The next step of this process is the binding of HP1homolog (Swi6), which is essential for heterochromatin maintenance [11].

Another epigenetic process involving RNA is controlling gene expression by miRNA. One-stranded miRNA binds to specific mRNA thanks to sequence complementarity.

Complementarity in this situation is very high but not complete. Bounded mRNA cannot be translated, which decreases the expression of a particular gene [12].

Endogenous, small non-coding RNAs can also induce chromatin inactivation nearby targeted sequences of the genome [13].

2.1.4. Role of epigenetic processes

Stem cell differentiation

Two features are the most specific for stem cells: self-renewal and the ability of differentiation into cells, specific for the tissue type. Pluripotent stem cells can develop into every type of cell in the organism. Multipotent or unipotent stem cells are maintained in all tissues for their whole life. They can develop only into specific cells. During the process of cell differentiation, morphological and functional changes occur in cells. They are determined by gene expression patterns. Genes responsible for self-renewal are silenced, while genes specific for cell type are activated. The epigenetic status of stem cells and differentiated cells is relatively stable because of the mechanisms of inheritance. Initiation and maintenance of gene expression changes are connected to a unique epigenetic program, including covalent DNA and chromatin modifications. Small non-coding RNAs are also involved in pre- and post-transcriptional gene regulations [14].

Gene expression control

Genetic information is expressed through transcription, translation, and protein modifications. Every cell contains the same DNA sequence. However, cells differ according to types, functions, and, consequently, gene expression. Gene expression patterns are determined during cell differentiation and maintained during mitotic cell divisions. Hence, cells inherit genetic and epigenetic information. Inherited epigenetic information consists of cytosine methylation, post-translational histone modifications, chromatin remodelling, and RNA-based mechanisms. Epigenetic processes mainly impact transcription, but they can also regulate splicing and translation.

DNA methylation leads to two processes - inhibition of DNA recognition by some proteins as well as simplification of binding other proteins with DNA.

In vertebrates, over 80% of CpG sites are located outside CpG islands, while most CpG sites inside CpG islands are more often unmethylated [15].

The impact of cytosine methylation on gene expression involves multiple mechanisms. One of them is DNMTs and transcription factor interactions. This site-specific methylation in the gene promoter region leads to binding proteins which recognise methylated DNA [16]. Consequently, these protein clusters directly impact the transcription process or lead to chromatin structure changes.

Other proteins, i.e., methyl-binding proteins (MBPs), bind methylated cytosine with their MBD domain and repress the transcription process across hundreds of base pairs with the transcription repressive domain [17]. Alternatively, the MBP changes the chromatin condensation through binding with linker DNA and nucleosomes, which is a physical block for transcription factors. MBPs cooperate with additional enzymes, such as histone deacetylases, which also are crucial for gene expression control.

Histone protein post-translational modifications are another epigenetic processes that control transcription. These modifications impact chromatin structure, changing its conformation and cooperating with other proteins: attract effector proteins to the chromatin or repress binding of associating with chromatin proteins. In general, histone acetylation has a positive effect on transcription, while for deacetylation, it is the opposite. The effect of methylation depends on modified aminoacid. Deamination is part of the repression process.

The following gene expression control epigenetic factors are small non-coding RNAs. Several types of them can be distinguished: miRNA (microRNA), siRNA (short interfering RNA), and piRNA (piwi-interacting RNA). They develop from bigger RNA precursors. Mature miRNAs bind to target mRNA to inhibit transcription or direct mRNA degradation. Mature siRNAs behave similarly to miRNA - they can inhibit transcription or lead to mRNA degradation, depending on the complementarity level. siRNAs are also involved in the silencing of transcriptional genes, especially transposable elements. In animals, transcriptional gene silencing consists of the involvement of histone methyltransferases and heterochromatin forming. piRNA binds to piwi-family proteins in spermatozoa and oocytes. Maternal piRNA in oocytes changes descendants' phenotype. Hence they are part of the epigenetic inheritance mechanism [18].

The last type of transcription controlling particles is long non-coding RNAs (lncRNAs). lncRNA transcription impacts the transcription level of a downstream promoter by changing the recruitment of polymerase II or chromatin configuration. Alternatively, antisense and sense transcripts hybridisation can lead to alternative splicing of sense transcripts or siRNA generation. lncRNA can also interact with proteins, regulating their localisation in cells or activity. Some lncRNAs interact with modifying chromatin complexes [15].

Genome stability maintenance

Genome stability is an organism feature that maintains and transmits genetic material during cell division. It consists of DNA and RNA replication and replication mistakes repair as well as reconstruction of damaged DNA or RNA. Genome instability leads to DNA damages and mutations [19]. The parts of the genome that are most vulnerable to genome instability processes are repeated DNAs. Their recombination causes chromosome rearrangements. Chromatin condensation plays a crucial role in DNA damage recognition and repair. Demethylation of lysine 9 histone H3 leads to an increase in spontaneous DNA damages occurrence and activation of DNA repair processes [20].

2.1.5. Epigenetic modification factors

Some external processes impacting epigenetic modification can be distinguished among internal organism mechanisms. Some of them are physical environment factors. An example can be seasonal changes in DNA methylation pattern in women from Gambia [21]. The immediate factor was diet changes across the whole year. The presence of some compounds such as folic acid, vitamin B-12, choline, and betaine impact DNA methylation in women and, consequently, their children's phenotype.

The other factors are psychological experiences. In examined males, psychological anxiety, measured as the feeling of fear, depression, and hostility, was positively correlated. At the same time, happiness and life satisfaction were negatively correlated to the average methylation level of Intercellular Adhesion Molecule-1 (ICAM-1) and coagulation factor III (F3) promoter. These psychological factors were also associated with average methylation of the glucocorticoid receptor (NR3C1), interferon-γ (IFN-γ), and interleukin 6 (IL-6) promoters. Mentioned genes are involved in inflammatory processes, which are connected to coronary heart disease (CHD) [22].

Oppositely to chronic effects, DNA methylation level can change dynamically in a stress reaction. The DNA methylation level of OXTR (oxytocin receptor) was examined just before stress as well as 10 minutes and 90 minutes after stress. Directly after stress, methylation level increases and then decreases to a level lower than in the beginning [23].

2.1.6. Aberrant DNA Methylation and Cancer

Two epigenetic processes are common in cancer diseases: hypomethylation and hypermethylation. However, they occur in different DNA sequences. Most of the genome becomes hypomethylated, while hypermethylation occurs in CpG islands in gene promoters [24]. Global hypomethylation brings genome instability. Gene promoter hypermethylation leads to tumour suppressor genes inactivation. Additionally, hypermethylation regulates ncRNA (e.g., miRNA) expression, which plays a role in tumour suppression. Hypermethylated gene promoters can be considered cancer biomarkers.

DNA methylation changing factors can be considered in cancer therapies because of their reversibility [25]. Epigenetic-oriented therapy can be used in hematological malignancies treatment [26].

2.2. Acute myeloid leukaemia

2.2.1. AML genesis

Acute myeloid leukaemia (AML) is a cancer of the myeloid cell line in the bone marrow. In this malignancy, abnormal hematopoietic cells are produced and accumulated [27]. At the same time, the production of other blood cells is defective [28]. That is why most AML symptoms are associated with the dysfunctionality of blood cells [29]. A low level of erythrocytes leads to anemia with weakness, suffocation, tachycardia, and pallor of skin and mucosae. As a result of thrombocyte deficit, bleeding from the nose, gums, digestive system, vagina, central nervous system, and disseminated intravascular coagulation can occur. The lack of normal leucocytes causes immune disorders and increases the number of severe infections. Sometimes splenomegaly and adenopathy occur [30].

Anomalies can appear on different levels of cell maturation. Usually, two types of genetic mutations occur. The first concerns the genes responsible for the proliferation, and the second concerns the genes responsible for cell differentiation and self-renewal. After a cascade of genetic disorders, clones of abnormal, immature progenitor cells proliferate. The apoptosis process in these clones is disturbed, so cancer cells accumulate [29].

2.2.2. AML epidemiology

Acute myeloid leukaemia is the most common acute leukaemia in adults, however its percentage depends on population (80% in Poland [31]). It is rather rare in children - 18% of all leukaemias [32]. It occurs more often in men than in women [33].

Risk factors for AML include radiation exposure (e.g., radiotherapy) [34], exposure to certain chemical substances such as benzene [35], agricultural chemicals or chemotherapy [36], blood disorders such as myelodysplastic syndrome (MDS) [37] or myeloproliferative neoplasms (MPN) and genetic factors, e.g., Down syndrome [38].

2.2.3. AML types

In 1976 French-American-British classification of AMLs (FAB) was proposed [39]. Six types of AML were distinguished, and two types were added later:

- M0 - acute myeloblastic leukaemia, minimally differentiated,
- M1 - acute myeloblastic leukaemia, without maturation,
- M2 - acute myeloblastic leukaemia, with granulocytic maturation,
- M3 - promyelocytic or acute promyelocytic leukaemia (APL),
- M4 - acute myelomonocytic leukaemia or myelomonocytic together with bone marrow eosinophilia (M4eo),
- M5 - acute monoblastic leukaemia (M5a) or acute monocytic leukaemia (M5b),
- M6 - acute erythroid leukaemias, including erythroleukaemia (M6a) and very rare pure erythroid leukaemia (M6b),
- M7 - acute megakaryoblastic leukaemia.

The currently used classification is the one proposed by World Health Organization (WHO) in 2008 [40]. It considers morphological, immunophenotypic features as well as molecular and cytogenetic abnormalities. It distinguishes seven types of AML:

- Acute myeloid leukaemia with recurrent genetic abnormalities with nine subtypes,
- Acute myeloid leukaemia with myelodysplasia-related changes with 18 subtypes,
- Therapy-related myeloid neoplasms,
- Myeloid sarcoma,
- Myeloid proliferations related to Down syndrome,
- Blastic plasmacytoid dendritic cell neoplasm,
- AMLs not otherwise categorised with nine subtypes, which do not suit the above types and are similar to FAB categories.

Molecular AML typing has influenced diagnosis and risk assessment but cannot be used to predict specific treatment. The only type of AML with specially designed therapy is APL [41].

2.3.4. AML therapy

Treatment of AML contains three phases: induction, consolidation, and optional maintenance [42]. Induction chemotherapy is provided with anthracycline and cytarabine [28]. Consolidation therapy follows clinical and hematological remission. Allogeneic stem cell transplantation is recommended for patients with a high risk of relapse. In other cases, chemotherapy is continued. Maintenance therapy consists of observation of the malignancy followed by allogeneic stem cell transplantation if needed [43]. In APL, treatment begins with all-trans retinoic acid oral application [44]. Then, it is continued with chemotherapy.

2.2.5. Prognosis in AML

Overall survival of AML is age-dependent. Prognosis is poor for patients older than 60 years. The 5-year survival rate for them is 17%, while for younger patients, it is 32% [45]. For the youngest patients (below 20 years) 5-year survival rate is 69% [46].

2.2.6. DNA methylation in AML

Changes in DNA methylation patterns are specific for cancer, and so for acute myeloid leukaemia. Several mechanisms induce aberrant methylation in AML. The methylation pattern in young and healthy hematopoietic stem cells changes with age, similarly to cancer. The acquired methylation pattern is inhomogeneous and diversified. Genetic mutations of transcription factors disable binding them to their binding sites. They induce hypermethylation of these binding sites. This situation occurs in the mutated CEPBα transcription factor, which leads to hypermethylation of its binding sites. Another mechanism - downregulation of transcription factors by oncogenes also can result in hypermethylation of their binding sites. An exemplary transcription factor is PU.1, downregulated by PML-RARα. The last processes impacting DNA methylation are changes in chromatin conformation [47]. Some of the described processes are shown in Figure 2.3.

DNA methylation is a feature that can be used to distinguish patients with different AML types. Differentially methylated gene sets can be used as biomarkers as well as therapeutic decision and prognosis indicators [48].

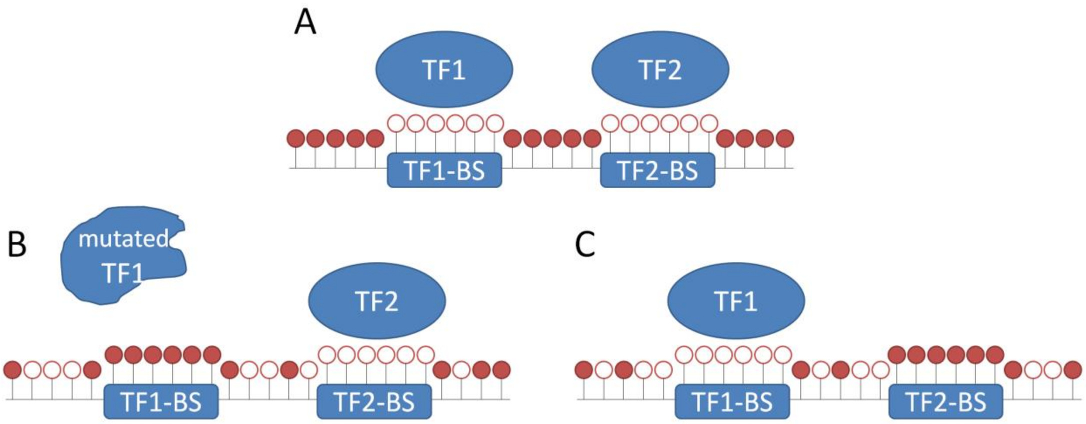

Figure 2.3 Mechanism of inducing aberrant DNA methylation in AML in comparison to healthy hematopoietic stem cell (A): mutation of transcription factor 1 (TF) induces hypermethylation of its binding site (TF1-BS) (B). Downregulation of transcription factor 2 (TF2) also induces hypermethylation of its binding site (TS2-BS) (C). In both (B and C) cases, epigenetic drift - various changes in methylation pattern, can be observed.

2.3. Therapy-related AML

Therapy-related acute myeloid leukaemia (t-AML) is AML occurring as a side effect of molecular changes (i.e., mutations, DNA methylation aberrations) developed after radiotherapy, chemotherapy, immunosuppressive therapy, or their combinations, given for pre-existing malignancy. Approximately 10% of AML cases are therapy-related AML.

The first reports about t-AML came in 1970 and were published in Lancet [50] [51]. Both described cases concerning ovarian cancer and its therapy with a drug called thiotepa. t-AML occurred about a dozen months after the therapy.

According to WHO classification, we can distinguish two types of t-AML: alkylating agent/radiation-related t-AML and t-MDS and topoisomerase II inhibitor-related AML (The WHO classification of the myeloid neoplasms). The first type usually occurs 4 - 7 years after exposure to therapy. One-third of patients have AML with myelodysplastic features, and the rest have MDS. Deletion or loss of chromosomes 5 or 7 is characteristic for this AML subtype [51]. Other hallmarks are complex karyotype and worse clinical outcomes within two t-AML types. The second type usually occurs between six months and five years (median two-three years) after initiation of therapy. Most patients with this type do not manifest preceding myelodysplastic phase but immediate acute leukaemia. This subtype of t-AML is connected to chromosome translocations i.e. 11q23 or 21q22.27,28,42-44 and other translocations: inv(16)(p13q22) or t(15;17)(q22;q12) and 11q23 and 21q22 abnormalities. The primary response to treatment and overall survival is similar to *de novo* AML with the same genetic abnormalities.

According to FAB AML classification, all AML types are represented within t-AML cases. In one study describing 37 cases of t-AML, 19 of them could be classified with FAB criteria: M2 - nine cases, M3 - one case, M4 - three cases, M5 - two cases, and M6 - four cases [53]. In this study, the time between the beginning of therapy and t-AML occurrence was 11-192 months, with a median equal to 58 (1-16 years). The median survival time was four months, regardless of getting the treatment.

t-AML characterises poor prognosis compared to *de novo* AML. A study of 200 patients with t-AML and 2,653 patients with *de novo* AML shows that four years overall survival rate for t-AML is 25.5%, while for *de novo* AML - 37.9%. Neither type of previous therapy nor latency time have an impact on survival [49].

t-AML can appear after treatment of different malignancies. In abovementioned study 37% of patients had breast cancer, 6% - thyroid cancer, 5% - gastrointestinal cancer, 4.5% - prostate cancer, 4.5% - testicular cancer. 27.5% had hematologic malignancies: 12.5% - non-Hodgkin lymphoma, 10% Hodgkin lymphoma, 3% - MDS. 1.5% of patients received cytotoxic therapy for autoimmune diseases.

The latency period between primary diagnosis and occurrence of t-AML was 0.33 - 44.14 years (median 4.04 years).

Another study of 306 patients [52] with t-AML consists of 10.5% patients with breast cancer, 5% with ovary cancer, 4% with prostate cancer, 25% with Hodgkin lymphoma, 23% with non-Hodgkin lymphoma and 7.4% with myeloma. It is also described that some individual features such as single-nucleotide polymorphisms variants can increase the risk of developing t-AML.

Some DNA methylation aberrations were observed in t-AML. In one study [54] 13 gene promoters for 11 patients were examined. Methylation level was measured with methylation-specific polymerase chain reaction (MSP). In five patients, at least one gene promoter was hypermethylated. Hypermethylated gene promoters were: CDKN2B (p15), RB1, MLH1, MGMT, PRDM2, SOCS1. CDKN2B gene is almost always hypermethylated in *de novo* AML, while RB1, MLH1, MGMT, PRDM2, and SOCS1 genes are not methylated. Latency time between the therapy of primary malignancy and developing t-AML was shorter for patients with hypermethylation than for patients without hypermethylation (49.3 months and 133.2 months, respectively). t-AML patients were classified according to FAB criteria: M0 - one patient, M1 - two patients, M2 - one patient, M4 - four patients, and M5 - two patients. One patient had chronic myeloid leukaemia in the chronic phase. There was no significant association between methylation and FAB AML type, age, gender, therapy, and survival time. It was suggested that hypermethylation of specific gene promoters could accelerate the development of t-AML.

Another study [55] confirms very frequent hypermethylation of the CDKN2B (p15) gene in t-MDS and t-AML. It was less common in the M5 subtype according to FAB criteria. Hypermethylation of the CDKN2B gene can cooperate with deletions on chromosome arm 7q in developing t-AML. Hypermethylation of this gene can be observed even two years before t-MDS/AML diagnosis [56].

Another gene that is methylated more often in t-MDS/AML than in *de novo* AML is DAPK1 [57]. Inactivation of DAPK1 by its hypermethylation leads to inhibition of the apoptosis process and metastasis probability enhancement.

None of the mentioned research was an epigenome-wide association study - in each case, only a few genes were investigated for DNA methylation level.

2.4. DNA Methylation measurement techniques

There are a lot of DNA methylation measurement techniques that facilitate the examination of methylation level genome-wide or in particular regions.

2.4.1. Genome-wide methylation level

The ratio of 5-methylcytosine and cytosine must be computed to find the whole-genome methylation level. One of the methods used for that purpose is high-performance liquid chromatography (HPLC) [58]. In this procedure, whole genomic DNA is hydrolyzed to deoxyribonucleotides, which are transformed into deoxyribonucleosides. They are separated by standard reverse-phase HPLC. Quantification of cytosine and 5-methylcytosine is provided by UV absorbance. High-performance capillary electrophoresis (HPCE) can also be used to separate deoxyribonucleosides [59]. Better sensitivity provides a combination of HPLC with mass spectrometry detection [60], which can also be used with HPCE. An alternative method can be thin-layer chromatography (TLC) [61], which uses radioactively labelled cytosine monophosphate and 5-methylcytosine monophosphate.

Another method is the SssI acceptance assay. It uses bacterial methyltransferase, which methylates unmethylated cytosines in CpG sites by the tritium-labelled donor. Then, methylated DNA is immobilised on nitrocellulose paper [62]. The amount of radioactive label can be measured with a scintillation counter. To measure the whole-genome methylation level, a chloroacetaldehyde assay can be used [63]. DNA is processed with bisulfite conversion, which changes unmethylated cytosine into uracil (Figure 2.4). After incubation of a sample with chloroacetaldehyde, fluorescent ethenocytosine from 5-methylcytosine can be quantified to find a level of methylated cytosine in the genome. Another factor that can be used to determine methylation level is monoclonal antibodies raised against 5-methylcytosine [64]. Before incubation with antibodies, DNA must be denatured and immobilised on a nitrocellulose membrane. Subsequently, the whole sample is incubated with a fluorescein-conjugated secondary antibody and scanned.

2.4.2. Gene-specific methylation analysis

To find gene-specific methylation level, DNA must be firstly amplified. To distinguish cytosine and 5-methylcytosine, DNA must be modified by methylation-sensitive restriction endonucleases (MSREs), bisulfite, hydrazine, or permanganate [65].

MSREs methods

In MSREs-based methods, two nucleases are used, where one of them is cytosine methylation insensitive (cuts methylated CpG sites) and the other is not. Products of digestion can be analysed by Southern-blotting, i.e., fractioned in electrophoresis and hybridised with a radioactive probe of the examined gene. The fraction of methylated DNA can be quantified by image processing. The second method of digested DNA processing is PCR. With primers designed for the gene of interest, it will flow only for not cut DNA. Real-time qPCR should be applied to quantify the fraction of methylated DNA in the sample.

Restriction landmark genomic scanning (RLGS) [66] is a method that enables the measurement of thousands of CpG sites methylation levels using one electrophoresis gel. After DNA digestion, it is labelled with radioactive molecules in not methylated sites and then digested a second time and put into electrophoresis in two dimensions. Resulted gel with DNA spots can be analysed with software tools.

Bisulfite conversion methods

To prepare a sample for the bisulfite reaction, DNA is digested with restriction enzymes and denatured with sodium hydroxide [67]. Then, it is treated with bisulfite at pH 5. During bisulfite conversion, each methylated cytosine remains unchanged, and each unmethylated cytosine is converted into uracil. Then, in PCR amplification, uracil is replaced by thymine. The scheme of this reaction is presented in the Figure 2.4.

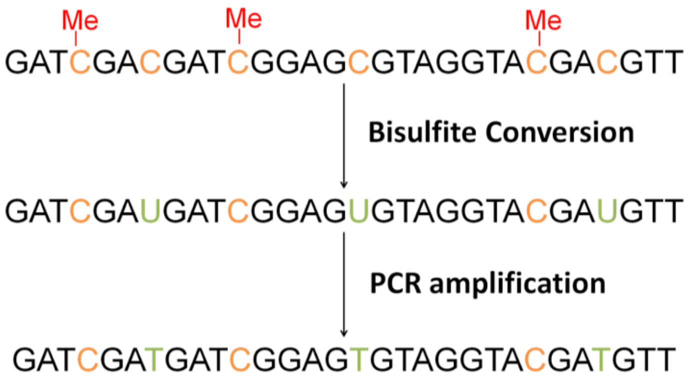
Me
Me
Me
GATCGACGATCGGAGCGTAGGTACGACGTT
Bisulfite Conversion
GATCGAUGATCGGAGUGTAGGTACGAUGTT
PCR amplification
GATCGATGATCGGAGTGTAGGTACGATGTT

Figure 2.4 DNA sequence modifications in bisulfite conversion and PCR amplification

DNA modified by bisulfite can be analysed by methylation-specific PCR. Two types of primers are included in reactions to amplify separately methylated and not methylated molecules. After that, the products of PCR are processed with electrophoresis and ethidium bromide. Methylation-specific PCR is only a qualitative method and should be verified with one of the quantitative methods. A quantitative method is methylation-sensitive single-nucleotide primer extension (Ms-SnuPE) [68]. It allows quantification of methylation level in any CpG site. This technique was originally designed for single-nucleotide mutation detection. It uses a primer, hybridising with a DNA matrix at an interesting CpG site. The amount of hybridised labelled deoxyribonucleotide is proportional to the amount of this base on the DNA matrix. Oligonucleotides used in this reaction should contain only one CpG site, which is difficult in CpG-rich regions.

The product of amplification can be quantified with electrophoresis and phosphorimager analysis or with transfer to nylon membranes.

The product of bisulfite conversion can also be sequenced. Sequence analysis provides information about methylation level in every CpG site in an interesting sequence. One of the sequencing methods is pyrosequencing.

Pyrosequencing

Pyrosequencing is a sequencing-by-synthesis system [69]. Usually, it is used to quantify single-nucleotide-polymorphisms (SNPs) [70]. In methylation level quantification, SNPs are artificially created by bisulfite conversion.

Independent primers for PCR amplification and sequencing must be designed. The amplified DNA sequence is a matrix with which primers are bounded. Particular nucleotides are added to the reaction. The synthesis of the proper nucleotide leads to a release of pyrophosphate and light emission. Light is detected by a camera in real-time. Light intensity is proportional to the number of bounded nucleotides. Each potentially methylated site is examined for the percentage of cytosine and thymine in the whole sample. The length of the analysed sequence should be no more than 200 bp. 96 samples can be tested at the same time. Pyrosequencing is a relatively fast, precise, low-cost DNA methylation quantification method.

Another sequencing method that combines both MSREs methods and bisulfite conversion is reduced representation bisulfite sequencing (RRBS). It allows for quantifying genome-wide methylation profile for each CpG site [71]. It uses restriction enzymes to perform DNA fragmentation and then bisulfite sequencing. It covers CpG sites across the whole genome, but only 10-15% of them. It does not work for CpG sites in regions without the enzyme restriction site - long DNA fragments are excluded from the analysis.

A high-sensitivity and high-throughput method of DNA methylation quantification is MethyLight [72]. It uses fluorescent probes and PCR, performed after bisulfite conversion. The number of PCR cycles with fluorescence detection is proportional to the methylation level in the CpG site. It quantifies the methylation level with high-throughput capability and high sensitivity but the low resolution (not with single CpG precision).

Another method that uses bisulfite conversion is methylation-specific single-strand conformation analysis (MS-SSCA) [73]. The bisulfite-treated DNA fragment is amplified in PCR. The products of amplification are separated with electrophoresis in a polyacrylamide gel. Differences between methylated and unmethylated DNA secondary structures lead to the formation of separated bands. They can be visualised with fluorescent gel stains. The specificity of this method is lower than 70%.

In MS-DGGE (methylation-specific denaturing gradient gel electrophoresis) method, differentially methylated DNA molecules are divided based on different melting temperatures after bisulfite conversion [74]. Different methylation level sequences will stay at different positions in the electrophoresis gel.

The last of these types of methods is MS-DHPLC (methylation-specific denaturing high-performance liquid chromatography) [75]. This mutation detection method was modified for DNA methylation detection. Differences are detected with different retention of DNA molecules at high temperatures.

2.4.3. DNA methylation arrays

The last but the most popular technique nowadays is the methylation array (bead-chip) assay. This high-throughput epigenome-wide method can be used for biomarkers identification [76]. It enables for evaluation of over 27 thousand (Illumina Infinium Human Methylation 27K) [77], over 450 thousand (Illumina Infinium Human Methylation 450K) or even 850 thousand (Illumina Infinium Human Methylation EPIC) [78] methylation sites of the genome. In sample preparation, DNA is digested with proteinase K and treated with sodium bisulfite. Then, DNA is amplified and hybridised onto a bead-chip. Two bead types, dedicated for each CpG site, are used on the array to detect methylation level. The beads are bounded to the one-strand DNA oligonucleotides (Figure 2.5).

Then, the chip is stained with antibodies in an immunohistochemical assay and scanned to measure the fluorescence intensity. The methylation value is calculated, considering both beads' values.

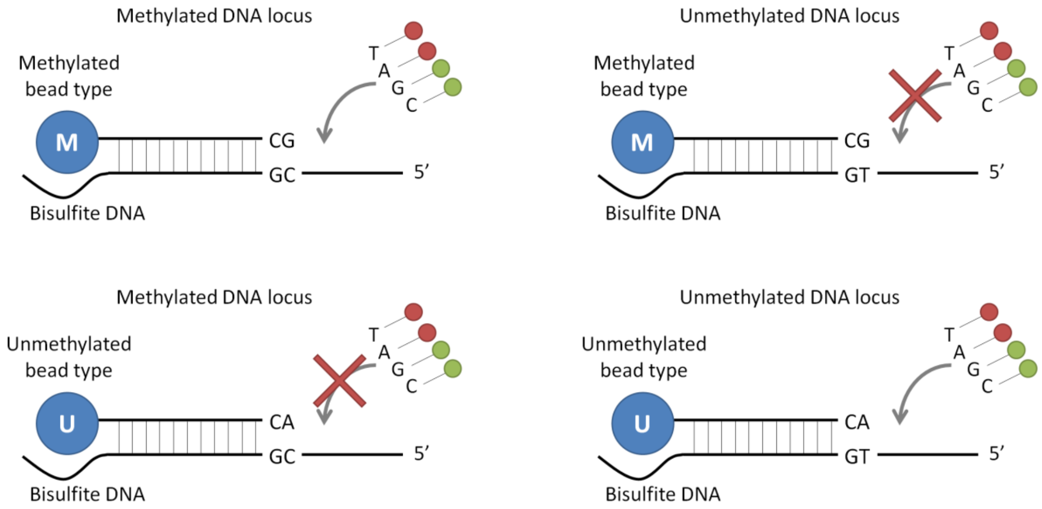

Figure 2.5 Methylated beads match the methylated CpG site, while unmethylated beads - are the unmethylated sites. In the case of a match, single-base extension occurs. The red label is biotin, and the green label is 2,4-dinitrophenol [79].

The CpG sites examined on the matrix are spread genome-wide. They cover all genomic regions, such as promoter, gene body, and intergenic regions, as well as CpG density areas: island, shore, shelf, and the open sea. Moreover, Illumina provides an annotation system where each probe is described with genome region, density area, and genome location (chromosome number and locus). The contribution of each type of probe annotation is presented in Figure 2.6.

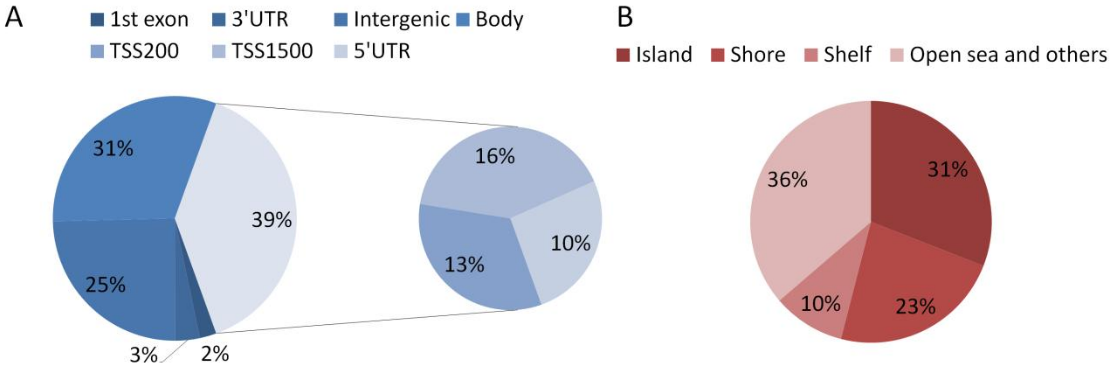

Figure 2.6 Percentage of probes belonging to genomic regions (A) and CpG density areas (B). TSS200, TSS1500, and 5'UTR form a promoter region of a gene [80].

Array methods are the compromise among costs, time, and genome coverage.

2.5. Analysis of DNA Methylation arrays data

The purpose of DNA methylation data analysis is to detect differentially methylated positions (DMP) and differentially methylated regions (DMR) [81].

2.5.1. Normalisation

The first step of data processing is normalisation. It is aimed to minimise technical variance and bias while keeping biological differences between samples and probes [80]. Regarding DNA methylation array data, two types of normalisation can be distinguished: based on β-value and based on M-value [83]. β-value represents the contribution of methylated probes, while M-value is a ratio between methylated and unmethylated probes signal.

$$\beta = \frac{m}{m+u}$$

$$M = \frac{m}{u}$$

Where:

- m is methylated probes signal;
- u is unmethylated probes signal.

The relationship between β-value and M-value is a logit transformation. Thus, the M-value can also be defined in relation to the β-value:

$$M = \frac{\beta}{1-\beta}$$

The distribution of β-value ranges from 0 to 1, while M-value distribution ranges from - to + infinity, with a mean equal to 0. β-value is easier to interpret biologically and, thanks to this, is more often chosen to represent the methylation level.

One of the normalisation methods that is based on β-value is Beta MIxture Quantile dilation (BMIQ) [84]. Type II beads have a lower dynamic range, so their signal must be adjusted to type I beads. This procedure uses a model of β distribution mixture of three states: U - unmethylated, H - hemimethylated, and M - fully methylated. For U and M states, probabilities are transformed into quantiles. For the H state, methylation-dependent dilation transformation is performed to adjust data to the gap between U and M components. BMIQ reduces technical variability and bias of type II beads signal and cuts out the bias of type I beads.

The following method is Subset-quantile Within Array Normalization (SWAN) [85]. This method contains two stages. The first one is the determination of average quantile distribution using the subset of probes described as "biologically similar" based on their CpG site content. The second step is the adjustment of the intensity of the remaining probes using linear interpolation to define new intensities. SWAN also reduces technical variability and makes type I beads' and type II beads' signal distributions more similar. SWAN also leads to better detection of differential methylation than raw data analysis.

Functional normalisation is an alternative to quantile normalisation algorithms [86]. It does not force the equal distribution of all samples. It removes only the variability explained by a set of covariates, independent of biological diversity. Covariates are calculated with the use of two main components of PCA. Information about technical variability comes from control probes. It can also be used for batch effect removal, but it is suggested to perform it anyway, after normalisation.

2.5.2. DMR detection

The next step of DNA methylation array data analysis is the detection of differentially methylated regions (DMR). The first method of DMR detection is IMA [87]. In this procedure, the representative methylation value for the whole region is calculated. It can be mean, median, or Tukey's biweight robust average. To detect DMR, a statistical examination is conducted - Wilcoxon test, t-Student test, or empirical Bayes statistics. General linear models can also be used to detect the impact of a continuous variable (e.g., age) on methylation level. The recognition of differential methylation in the region is obtained with the *p*-value of a test.

Another method of DMR detection is BumpHunter [88]. It is based on seeking spatial compartments characterised by the difference between an estimated function and 0. The estimated function can be the average difference between methylation levels in two groups. For continuous variables, it can be a slope of the regression curve for each probe. Spatial compartments are searched across the genome location.

In the Probe Lasso method [89], the probes are divided into 28 categories - based on genomic region and CpG density from Illumina annotations. For each probe, the dynamic window - a lasso is created. If enough probes are inside the lasso, a region is created - overlapping windows are joined. Each region is examined for its differential methylation, based on Stouffer's [90] *p*-value integration. *P*-values come from the statistical test for comparison, e.g., two conditions.

The last popular method, DMRcate [91], is based on M-value and does not use regions proposed in Illumina annotations. It compares two stages or groups with limma for each CpG site. Then, *p*-values are corrected with Benjamini-Hochberg procedure [92]. The minimal corrected value is representative of the examined region. According to the authors, this method characterises with higher precision than BumpHunter and Probe Lasso.

3. Materials and Methods

3.1. Acute myeloid leukaemia methylation profile

3.1.1. Data description

The first dataset used in this study was downloaded from the GEO database [93]. GEO database is a public data repository containing functional genomic data such as coming from arrays or sequencing experiments. The study regarding described dataset (GSE63409) was published before by Namyoung Jung et al. [94] and proved the existence of AML subgroups that are epigenetically various. The data consist of 19 samples: 5 samples of hematopoietic stem cells (HSC) from healthy donors (control) and 14 samples of CD34+38- cells from AML patients. The data were collected with Illumina Infinium Human Methylation 450K microarray. Data were normalised using the Illumina preprocessing method implemented in the minfi package [95] in Bioconductor. Normalisation procedure resulted in estimation of methylation level as β-value in 485 512 probes. The distribution of β-value consists of two "peaks" at low methylation and high methylation values. The number of sites with medium methylation values is the lowest.

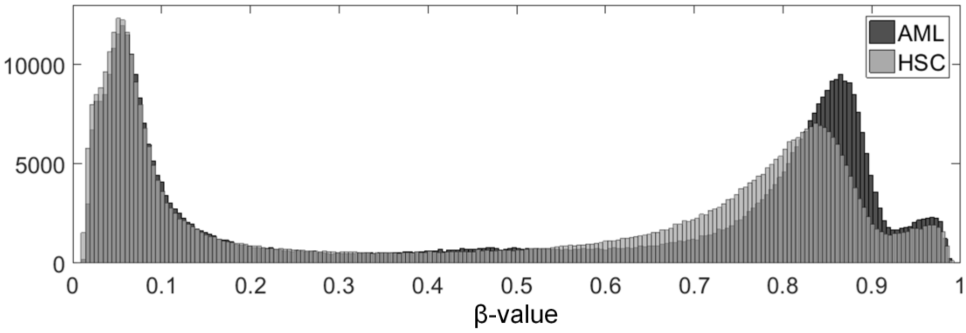

Figure 3.1 Histogram of β-value in HSC and AML samples.

According to annotations provided by Illumina, each probe is described with gene name (if it lies inside a gene), genome region (intergenic, TSS1500, TSS200, 5′UTR, 1stExon, Body, and 3′UTR), CpG density area (island, shelf, shore, and open sea) and genome location: chromosome number and locus. We decided to combine genome regions into three groups: TSS (consisting of TSS1500, TSS200, and 5′UTR sites), gene body (consisting of 1stExon, Body, and 3′UTR), and intergenic. The numerosity of each group is presented in Table 3.1.

Table 3.1 Number of probes annotated to each region group.

Region	TSS regions	Gene body regions	Intergenic regions
Number of probes	189,524	227,032	93,520

3.1.2. Methylation level distribution and profiles

Empirical cumulative distribution function (CDF) was calculated for pooled samples: control HSC and AML for the whole genome (all probes), TSS regions, gene body regions, and intergenic regions using Kaplan-Meier estimate [96].

To compare whole genome methylation profiles between AML and healthy donors with effect size, Cohen's *d* statistics [97] was used in pooled samples. Cohen's *d* statistic can be calculated according to the following formula:

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s}$$

Where:

- $\bar{x}_1$ and $\bar{x}_2$ are averages for sample 1 and sample 2, respectively
- s is pooled standard deviation, defined as:

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_1 - 2}}$$

Where:

- n_1 and n_2 are numbers of elements in sample 1 and sample 2, respectively
- s_1 and s_2 are standard deviations for sample 1 and sample 2, respectively

To verify the hypothesis about the consistency in methylation profiles, Cramer's *V* coefficient [98] and its *p*-value were calculated. Cramer's *V* coefficient can be calculated as:

$$V = \sqrt{\frac{\frac{\chi^2}{n}}{\min(k-1, r-1)}}$$

Where:

- *n* is a sample size
- *k* and *r* are numbers of subgroups for which samples are divided:
  in this case: low, medium, and high methylated subgroups for healthy control (k) and AML (r) samples
- χ^2 is the value of the statistic, calculated according to the formula:

$$\chi^2 = \sum_{i,j} \frac{\left(n_{ij} - \frac{n_i n_j}{n}\right)^2}{\frac{n_i n_j}{n}}$$

Where:

- n_i, NJ are numbers of samples element belonging to a particular subgroup
- n_{ij} is the number of samples element belonging to *i*-th as well as to *j*-th group
- *i* and *j* are subgroups indices; *i* ranges from 1 to *r*, *j* ranges from 1 to *k*

It was performed for probe subgroups, created according to their methylation level. Each sample was divided into low, medium, and high methylated groups. Healthy control and AML samples were compared. This procedure was performed for the whole genome and genome region groups.

3.1.3. Estimation of shift between β-value distributions

To estimate the shift between β-value distributions of healthy control and AML samples, Hodges-Lehmann (HL) statistics [99] was calculated. This robust and nonparametric estimator is based on the median of differences between each pair of elements from two samples.

It can be calculated according to the following formula:

$$d_{ij} = x_i - y_j$$

$$HL = median(d)$$

Where:

- *d* is a set of distances between each pair of sample 1 and sample 2
- x_i is the *i*-th element of sample 1 (*i* ranges from 1 to N_1, which is sample 1 size)
- y_j is the *j*-th element of sample 2 (*j* ranges from 1 to N_2, which is sample 2 size)

Hodges-Lehmann statistics value was calculated for each probe from the microarray. Positive values of HL statistics were obtained for hypermethylated probes, while negative for hypomethylated.

3.1.4. Gaussian Mixture Modelling

The distribution of Hodges-Lehman distance values was decomposed into Gaussian components to detect subpopulations or subgroups of probes [100]. For each subgroup, three parameters of Gaussian distribution are estimated: μ - mean, σ - standard deviation, and α - weight. Their sum composes the whole HL distribution. Let *f(x)* denote the probability density function corresponding to the analysed signal x. The Gaussian mixture model (GMM) of *f(x)* can be calculated as below:

$$f(x) = \sum_{k=1}^{K} \alpha_k f_k(x, \mu_k, \sigma_k)$$

Where:

- K is the number of Gaussian components,
- α_k are non-negative component weights; their sum equals 1,
- f_k is the probability density function of a normal distribution of the *k*-th component,
- μ_k and σ_k are *k*-th Gaussian component mean and standard deviation, respectively.

For fitting Gaussian mixture to HL values distribution, maximisation of log-likelihood function is used. The log-likelihood function can be calculated as:

$$\log L = \sum_{n=1}^{N} ln \sum_{k=1}^{K} \alpha_k f(x_n, \mu_k, \sigma_k)$$

Where:

- N is the total number of elements in a modelled vector.

Expectation-maximisation (EM) algorithm was applied to maximise the log-likelihood function. The initial values of GMM components were set according to the algorithm by Polanski et al. [101].

To find the best number of Gaussian components, Bayesian Information Criterion (BIC) [102] was used. BIC value was calculated for different mixture models with K ranging from 2 to 12. The minimal value of BIC indicates the best model:

$$BIC = -2 \log L + (3K - 1) \log N$$

Thanks to GMM of HL distance values, some subgroups of probed were distinguished. Probes can be differentiated into hypomethylated, low hypermethylated, medium hypermethylated, high or extremely high hypermethylated, and no changed. The classification of individual probes to the proper component is based on the maximum probability rule. The cut-off levels between each subgroup were determined by intersection points of probability density functions of components.

3.1.5. Defining methylation level and detection of differentially methylated probes.

To check if the individual site is low, medium, or high methylated, its methylation value was compared with β = 0.5 (null hypothesis) with one-tailed Wilcoxon tests [103]. Probes whose methylation level was significantly lower than 0.5 were considered low methylated. Probes whose methylation level was significantly higher than 0.5 were considered high methylated. Probes whose methylation level was not significantly different from 0.5 were considered medium methylated. This procedure was performed for healthy control samples and AML samples separately.

To detect differentially methylated probes in AML compared to healthy control, one-tailed Mann–Whitney tests were used [104]. Null hypotheses for these tests were:

no difference between AML and control (HL = 0) or difference is lower or greater than cut-off levels values from GMM modelling. For positive cut-off levels, right-tailed tests were performed, while for the negative - left-tailed test. Both tailed tests were performed for the null hypothesis that HL equals 0. The procedure was repeated for each probe, so multiple testing correction needed to be applied. It was done with Storey's method [105].

3.1.6. Detection of differentially methylated genomic regions

To check if the individual genomic region is differentially methylated in AML compared to healthy control, all p-values of probes belonging to this region were considered. For finding one global p-value for each region, Stouffer's method of p-value integration was applied [90]. Integrated p-value based on Z-value (from a standard normal distribution), which can be calculated as follows:

$$Z \sim \frac{\sum_{i=1}^{k} Z_i}{\sqrt{k}}$$

Where:

- $Z_i = \Phi^{-1}(1 - p_i)$
- p_i is the p-value for the i-th hypothesis test,
- Φ is the standard normal cumulative distribution function,
- k is the number of integrated p-values.

The procedure was applied for each genomic TSS and gene body region. Additionally, for each set of integrated p-values, the adjusted significance level (α) was calculated. The basic α (equals 0.025 for one-tailed tests) was transformed according to Stouffer's formula, with k equals the number of integrated p-values. Regions with integrated p-value lower or equal to adjusted α were considered differentially methylated.

3.1.7. Functional analysis

A functional analysis procedure was performed for hypo- and extremely high hypermethylated genomic regions. It was done based on Gene Ontology Terms [106]. The Gene Ontology database consists of three directed graphs of GO Terms: Biological Process, Cellular Component, and Molecular Function. The graph structure allows for the interpretation of dependencies between terms (processes or functions) in parent-child relationships. Overrepresented GO Terms for examined genomic region set were found with the *topGO* package [107] in Bioconductor.

3.2. Gender differences in DNA methylation in AML

3.2.1. Data description

The second dataset considered in this work was downloaded from the TCGA-LAML project database. The data were obtained with the same Illumina Infinium Human Methylation 450K microarray as previously. The number of all probes on the array was 485,577. However, after filtering out probes that were not significantly different from the background, repeat regions, common SNPs, and probes lying on sex chromosomes, there remained only 396,065 probes (regarding CpG sites). 2,627 (0.66%) probes had missing values, which would be difficult for imputation, so they were removed from further analysis. As the result of data preprocessing, the β-value for 393,438 probes was obtained.

The control dataset - healthy donors, was downloaded from the GEO database (GSE73103). It regards the study about changes in DNA methylation in obesity. Data were obtained with the same array and consisted of β-values for 397,615 probes.

To compose genomic region groups, Illumina's annotations were also used. However, this time probes were divided into three groups: CpG-rich regulatory sequence (RS) regions (consisting of sites annotated to TSS1500, TSS200, or 5’UTR genomic regions and island or shore CpG density areas), body regions (consisting of sites annotated to 1stExon, ExonBnd, and Body regions) and 3’UTR regions. We also have taken into account sites annotated to intergenic regions. The number of probes belonging to particular genomic regions is presented in Figure 3.2.

AML data were collected for 140 patients. In addition to DNA methylation level, the dataset also contains information about patients' gender, age at diagnosis, vital status, the number of days to death (for dead patients), or the number of days to last follow up since diagnosis (for alive patients) and information about receiving prior treatment. Information about the distribution of selected clinical features is presented in Table 3.2.

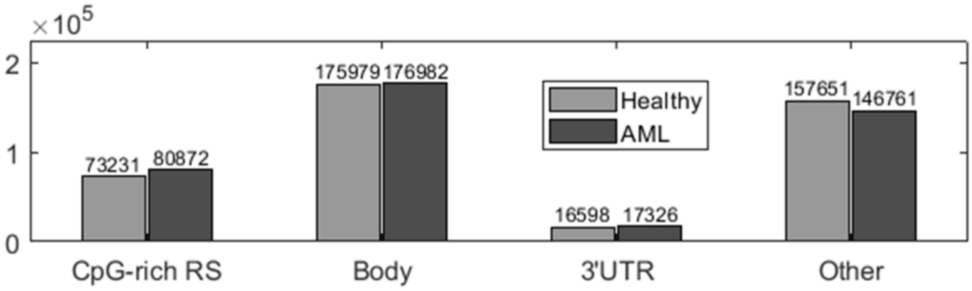

Figure 3.2 Number of probes among genomic regions groups in healthy and AML datasets

Table 3.2 Clinical factors collected for AML patients.

Feature	Values	Counts
Gender	Male	73
	Female	67
Vital status	Alive	64
	Dead	76
Race	White	127
	Black or African-American	10
	Not reported	3
Prior treatment	Yes	38
	No	102

39 females and 45 males, healthy and normal-weighted, were selected from the control dataset. All of the healthy control patients were of similar ages.

3.2.2. Comparison of clinical features values between male and female AML patients

Values of clinical features were compared between genders to ensure that none of them differentiates males and females. Vital status proportions and receiving of prior treatment proportions were examined with the χ^2 test [108], age at diagnosis with t-test and days to death with Mann-Whitney test (after checking normality of distributions). Genders were not compared according to race because of huge disproportions of this clinical feature counts.

3.2.3. Detection of differentially methylated probes and genomic regions.

Principal Component Analysis (PCA) [109] was conducted on β-values, firstly from the whole genome and then for particular genomic region groups separately. It was performed for the healthy control sample and AML sample.

The next step was outlier detection - all elements spaced more than three scaled MAD [110] from the median were detected as the outliers and replaced with median value. Then, the normality of DNA methylation level distribution was examined with the Lilliefors test [111]. Followingly, the methylation level was compared between males and females using the Mann-Whitney test. The whole procedure was performed for each probe. At the end, Benjamini-Hochberg FDR correction [92] was performed for Mann-Whitney test *p*-values. Additionally, effect size as rank biserial correlation was calculated [112].

Functional analysis of genes characterised by differentially methylated CpG-rich RS region was held in the STRING tool [113]. It provides Gene Ontology Terms, KEGG pathways, UniProt Keywords, Pfam Protein Domains, SMART Protein Domains, Reactome Pathways, InterPro Protein Domains and Features, STRING Clusters, and Reference publications. Items with a *p*-value ≤ 0.05 were considered significantly enriched.

3.2.4. Survival analysis

The methylation level for each genomic region was calculated as a median of β-values of probes, of which the regions consist. Survival curves for both genders were created and compared with a log-rank test [114].

The survival analysis was also done for each probe separately for all patients and genders. Cox proportional hazard [115] model *p*-value was found for each probe. All of the sites for which this *p*-value was lower than 0.05 were examined deeper. To check the impact of methylation level in individual probes on the survival time, patients were divided into two groups, depending on whether the methylation level in this site was lower or higher than the median. Survival curves for these groups were compared with the log-rank test. To identify impacting survival genomic regions, *p*-values from the log-rank test were integrated with Stouffer's method.

3.3. DNA methylation aberrations in *de novo* and therapy-related AML

3.3.1. Data description

Data were collected in two experiments: the primary experiment conducted with methylation array and the validation experiment conducted with pyrosequencing.

The primary experiment was first done using Illumina Infinium Human Methylation 450K Array for 12 patients: five *de novo* AML, four chemotherapy-induced AML, and three radiotherapy-induced AML.

Secondly, the experiment was repeated on Illumina Infinium Human Methylation EPIC Array with over 850K probes for the part of the abovementioned patients and, additionally, healthy persons. Data were collected in two batches: three *de novo* AML patients, three chemotherapy-induced AML patients, and two radiotherapy-induced AML patients in the first batch, as well as five healthy donors, two *de novo* AML patients, and one combined radio- and chemotherapy patient in the second batch. The scheme of the experiment design is shown in Table 3.3.

Table 3.3 The scheme of EPIC array experiment design.

Patient group	1st batch	2nd batch	Sum
Normal BM	5	-	5
De novo AML	2	3	5
Chemo-AML	-	3	3
Radio-AML	-	2	2
t-AML	1	-	1

The validation experiment - pyrosequencing was performed for the same patients and five additional *de novo* AML patients. This experiment measured methylation level for several genomic regions of interest in four or five methylation sites for each region. Primers for PCR and sequencing were designed in UK Health Security Agency or bought from the Qiagen company. The pyrosequencing experiment was conducted with PyroMark Q48 Autoprep. It results in a methylation level as a value between 0 and 1 for each examined probe.

The scheme of patients participating in the experiments labelled with their numbers is presented in the Figure 3.3.

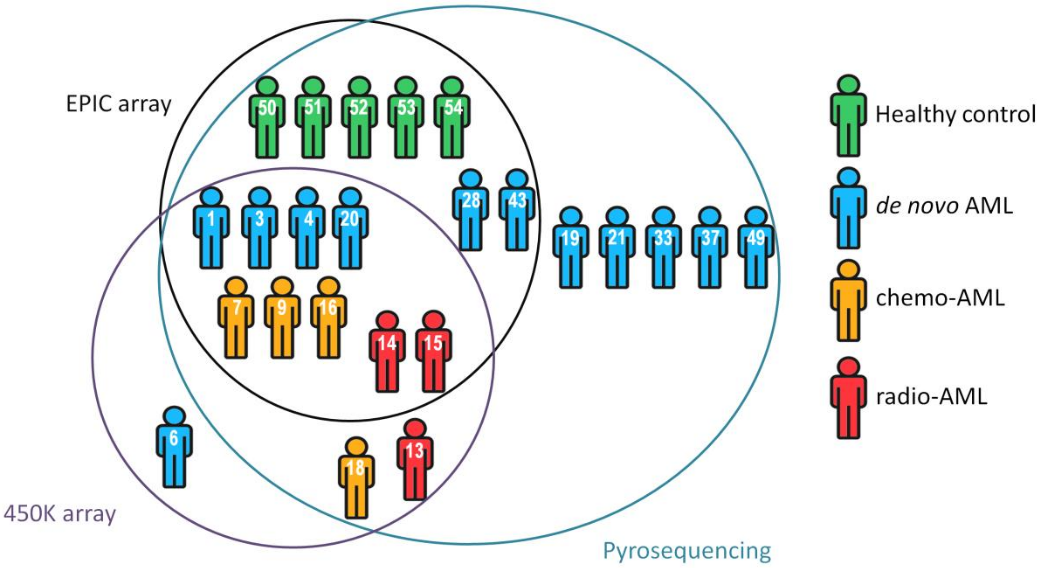

Figure 3.3 The scheme of patient groups examined in the experiments.

3.3.2. Data preprocessing

Methylation microarray data were preprocessed using methods suggested in the ChAMP library in Bioconductor [116] and considered optimal for this type of data. All preprocessing procedures were performed using the ChAMP package [117] [118] .

Loading and filtering were done using the ChAMP method. Non-CG probes, probes from X and Y chromosomes, and probes with detection *p*-value above 0.01 were filtered out. The data was loaded as β values from 0 (no methylation) to 1 (total methylation). Normalisation was performed with the Beta Mixture Quantile method (BMIQ). The ComBat procedure [119] was applied to remove the batch effect in data (collected in two batches). In this method, additive and multiplicative batch bias parameters are estimated and used for modification of DNA methylation level values.

3.3.3. Global DNA methylation profile analysis

The whole genome DNA methylation profile was analysed for data generated with an EPIC array. Firstly, the distribution of β-value was decomposed into Gaussian Mixture components. It was done to detect cut-off levels for defining probes as low, medium, and high methylated. Probes with methylation level lower than the first cut-off were considered low methylated, and those with methylation level higher than the second one were considered high methylated. The rest of the probes were considered medium methylated.

To compare global methylation profiles among patient groups and genomic regions, empirical cumulative distribution functions were plotted. It was done for pooled samples (among patient groups) and individual patients for the whole genome. The procedure was also repeated for each of the genomic region groups described below.

3.3.4. Detection of differentially methylated probes.

Many statistical methods can be used to detect differentially methylated probes. Jeanmougin et al. [120] compared them for gene expression array data. The most popular methods: limma [121] and ANOVA have very similar power, in a slight favor for limma (the difference in power of the methods equal to 0.02). Both of these methods were used, and their results were compared.

First, healthy control vs. *de novo* AML vs. chemo AML vs. radio AML comparison was performed with ANOVA. FDR was estimated from obtained *p*-value using the Benjamini-Hochberg procedure [92]. Tukey-Kramer pairwise comparisons for probes with FDR ≤ 0.05 [122] were conducted (both left and right-tailed tests). Next, probes with *p*-value ≤ 0.025 were considered differentially methylated (distinguished as hypomethylated or hypermethylated in pairwise comparisons.)

Secondly, the limma procedure was performed with the *champ.DMP* function in ChAMP package, which uses limma package. Pairwise comparisons for each pair of patient groups were performed, distinguishing probes into hyper- or hypomethylated (not only differentially methylated). Next, probes with a *p*-value ≤ 0.025 were considered significantly different.

Additionally, to examine the "inside group" diversity, each AML sample separately was compared with the control group using a t-test for one observation, according to the formula:

$$t = \frac{x - \mu}{\sigma\sqrt{\frac{N+1}{N}}}$$

Where:

- x is the observation value
- μ and σ are mean and standard deviation estimated from the control sample
- N is the control sample size.

After comparisons, features that were not differentiating in any case were removed from further analysis. No change, hyper- and hypomethylation frequency was calculated among the rest features. Also, the average frequency of no change, hypo- and hypermethylation for AML types was computed. Pielou index [123] was calculated for individual patient frequencies. It is defined as:

$$J = -\frac{\sum_{i=1}^{S} p_i \ln p_i}{\ln S}$$

Where:

- S is the number of classes
- p_i *is* the frequency for individual class

Values of the Pielou index were compared among AML types with ANOVA and then pairwise with Tukey-Kramer tests.

3.3.5. Detection of differentially methylated genomic regions

As described before, the composition of genomic regions is based on Illumina annotations. Three groups of genomic regions were constructed: gene regulatory sequence (RS) regions (related to the transcription start site (TSS1500, TSS200) and 5'UTR annotation) lying on CpG islands or shores, gene body regions (related to 1stExon, ExonBnd, and Body annotation) and 3'UTR regions. The scheme of genomic region groups is presented in Figure 3.4.

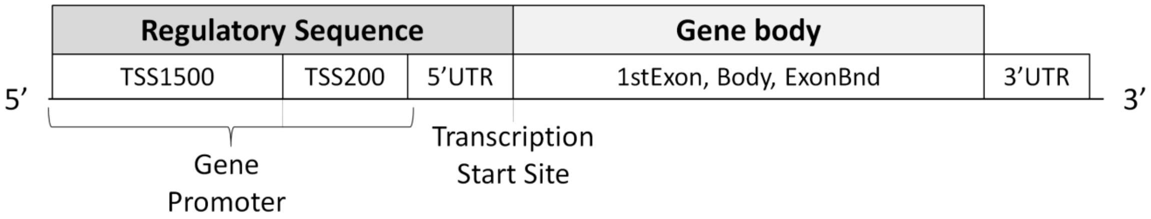

Figure 3.4 The scheme of genomic region groups based on Illumina's annotation system.

P-values of probes belonging to the same genomic region were integrated with Stouffer's method. In the first approach, *p*-values from Tukey-Kramer tests performed for each probe were integrated. In the second approach, it was *p*-values from limma pairwise comparisons.

The CpG-rich RS regions, which differentiated one of the AML types and control and did not differentiate other AML types and control, were considered epigenomic biomarkers.

3.3.6. Unsupervised feature selection

To examine the informativeness of data, unsupervised methods were applied. Methylation levels from the EPIC array for each CpG-rich RS region and patient were calculated as the mean methylation level of probes belonging to this region. Then, the variance of methylation level across healthy control and *de novo* AML samples for each CpG-rich Regulatory Sequence region was computed. The distribution of $\log_2$ of variance was decomposed to Gaussian Mixture Model [100]. The number of Gaussian components was selected according to BIC [102]. The most diverse RS regions were selected based on the maximum probability rule. The cut-off level was an intersection point between two probability density functions of components with the highest means. Mean values of selected regions [124] were used to create a hierarchical tree and a heatmap. Row standardisation was applied; the assumed distance metric was Spearman correlation. The strength of correlation of RS regions methylation level among patient groups was measured with Cramer's *V* effect size [98].

3.3.7. Functional analysis

Functional analysis was conducted with the hiPathia [125]. The hiPathia tool calculates activation scores for subpathways based on gene expression level and performs the statistical test to find deregulated pathways. Thus, the potential expression level was calculated for each CpG-rich RS region as the reverse of the mean methylation level because both processes are inversely proportional (RS region hypermethylation inhibits gene expression). The computed value was used to find pathways activation scores in hiPathia. Activation score was calculated for each patient and hiPatia subpathway. The unsupervised selection was performed based on the variance of activation scores, similarly to regulatory sequence regions. Most diverse subpathways were selected to construct a hierarchical tree and a heatmap.

3.3.8. Validation of selected biomarkers

The methylation level was measured in the pyrosequencing experiment for several identified biomarkers. Obtained methylation level values were compared among patient groups. For the one gene that was considered all AMLs biomarkers, a t-test for comparison between all AMLs and healthy control was used. ANOVA and Tukey-

Kramer post-hoc tests were applied to compare all patient groups for other potential biomarkers. In the end, *p*-values for the same gene were integrated with Stouffer's method. Additionally, Cohen's *d* [126] as an effect size measure was calculated for each performed comparison for each CpG site. Obtained *p*-values and effect size values were compared with methylation array experiment results.

4. Results

4.1. Acute myeloid leukaemia methylation profile

4.1.1. Methylation level distribution and profiles

For comparison of global methylation profiles between acute myeloid leukaemia patients (AML) and healthy donors (HSC) as well as among genomic region types, empirical cumulative distribution functions (CDF) of pooled samples were plotted. The result is presented in Figure 4.1.

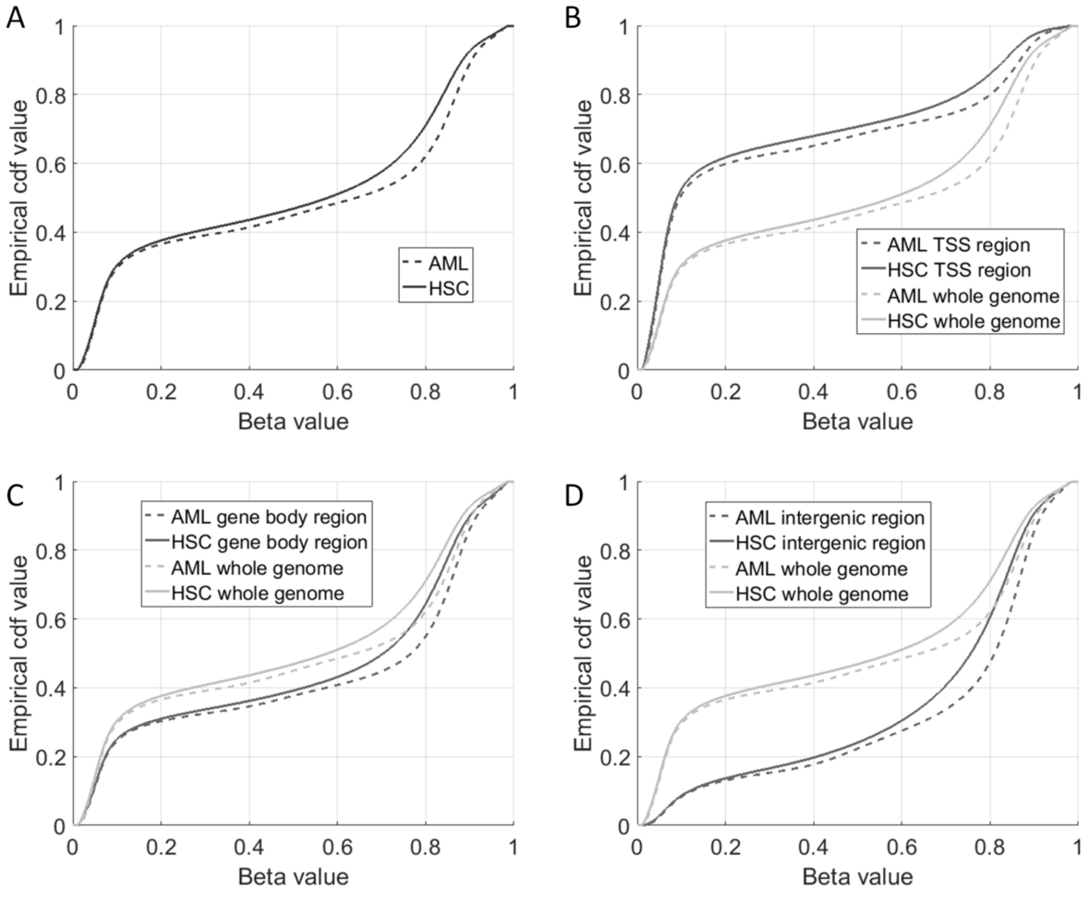

Figure 4.1 Empirical CDF of β-value in AML and HSC pooled samples in the whole genome (A), TSS regions (B), gene body regions (C), and intergenic regions (D). In all cases, CDF for the whole genome is presented for comparison.

The differences in methylation profiles between leukaemia and healthy donors can be observed in the whole genome and every region type. The effect size between healthy and AML samples is small - Cohen's *d* statistic value equals 0.2183. In each case, AML samples are higher methylated than HSC samples. Differences are slim for small β-values and increases in higher.

The differences among various regions are much higher. In TSS regions, methylation level is lower than in the whole genome and in the other regions. The methylation level in gene body regions is slightly higher than in the whole genome. In intergenic regions, methylation level is the highest among all regions.

For a deeper examination of methylation profiles, the contributions of a low, medium, and high methylated samples were calculated for the whole genome and all regions. It was done by statistically comparing the β-value with 0.5. The results for the whole genome are presented in Table 4.1

Table 4.1 The numbers of a low, medium, and high methylated probes in AML and HSC samples.

Methylation level		**AML**			
		Low	**Medium**	**High**	**Total**
HSC	**Low**	191,043	14,739	2,985	**208,767**
	Medium	5,668	11,286	33,931	**50,885**
	High	2,297	10,093	213,470	**225,860**
	Total	**199,008**	**36,118**	**250,386**	**485,512**

In HSC, almost 43% of probes are low methylated, while 46.5% are high methylated. In AML, almost 41% of probes are low methylated, and almost 52% are high methylated. More probes are medium methylated in HSC (10.5%) than in AML (7%).

The methylation profiles are generally consistent in both groups - the biggest numbers lie on the diagonal of the table. Most probes classified as low methylated in HSC were also classified as low methylated in AML (191,043 probes), the same with high methylation status (213,470 probes). Cramer's V association coefficient for obtained contingency table was equal to 0.6667 (*p*-value < 10^{-6} [108]). The detailed inspection of the table confirms global higher methylation in AML compared to HSC.

The same procedure was conducted for all genomic region types. Probes were classified as low, medium, or high methylated in each case. Results are presented in Table 4.2.

Table 4.2 The numbers of low, medium, and high methylated probes in AML and HSC for different genomic regions.

Methylation level		AML											
		TSS region				Gene body region				Intergenic region			
		Low	Medium	High	Total	Low	Medium	High	Total	Low	Medium	High	Total
HSC	Low	121,393	5,693	1,059	**128,145**	73,494	5,874	1,300	**80,668**	13,548	3,295	711	**17,554**
	Medium	2,101	3,578	8,643	**14,322**	2,417	4,639	16,438	**23,494**	1,154	3,065	9,893	**14,112**
	High	614	2,741	43,702	**47,057**	1,087	4,661	117,122	**122,870**	559	2,722	58,573	**61,854**
	Total	**124,108**	**12,012**	**53,404**	**189,524**	**76,998**	**15,174**	**134,860**	**227,032**	**15,261**	**9,082**	**69,177**	**93,520**

In TSS regions, almost 68% of probes are low methylated, and 25% are high methylated in HSC, while 65.5% are low methylated and 28% are high methylated in AML. In gene body regions, almost 35.5% of probes are low methylated, and 50% are high methylated in HSC, while 34% are low methylated and 59% are high methylated in AML. In intergenic regions, almost 19% of probes are low methylated, 66% are high methylated in HSC, 16% are low methylated, and 74% are high methylated in AML. Hence, differences among genomic region types are much higher than between HSC and AML inside the same region. Cramer's *V* statistic was calculated to examine the consistency of methylation profiles between AML and HSC. For TSS regions it was equal to 0.6692, in gene body regions - 0.6658 and in intergenic regions - 0.6119. So, in TSS and gene body regions, *V* values are similar and almost equal to the whole genome *V* value, while in intergenic regions, it is a little lower. So biggest differences in methylation profiles between AML and HSC occur in intergenic regions.

4.1.2. Estimation of shift between β-value distributions

The Hodges-Lehmann statistic was calculated for each probe to estimate the difference between distributions of β-value in AML patients and healthy donors. Its distribution is presented in Figure 4.2

Values of HL statistic greater than 0 indicate hypermethylation of particular probes, while values lower than 0 indicate hypomethylation. Based on the histogram, much more probes are hypermethylated than hypomethylated. The hypermethylation process is stronger than hypomethylation - positive values range further from 0 than negatives.

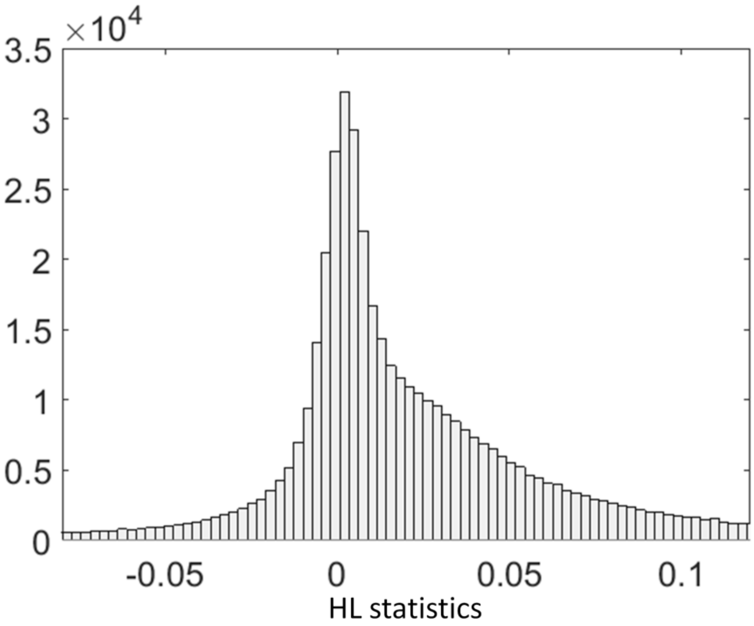

Figure 4.2 Distribution of HL statistic values across whole genome probes.

The procedure was repeated for genomic region types to inspect differences in HL statistic distribution among regions visually. Results are presented in Figure 4.3.

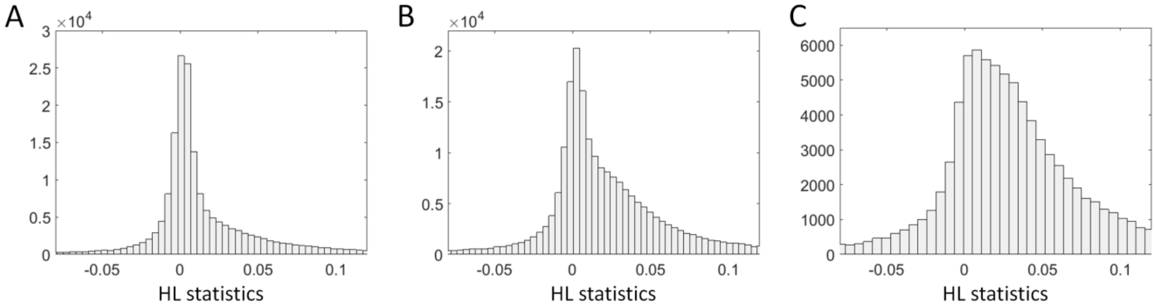

Figure 4.3 Distributions of HL statistics across TSS regions (A), gene body regions (B), and intergenic regions (C)

In each case, overrepresentation of hypermethylated probes can be observed. Based on the HL histogram shape, this process is strongest in intergenic regions and weakest in TSS regions.

4.1.3. Gaussian decomposition of Hodges-Lehmann statistic distribution

The distribution of Hodges-Lehmann statistic values was decomposed into a mixture of Gaussian components. The optimal number of components was selected with BIC. The result of Gaussian Mixture decomposition is presented in Figure 4.4.

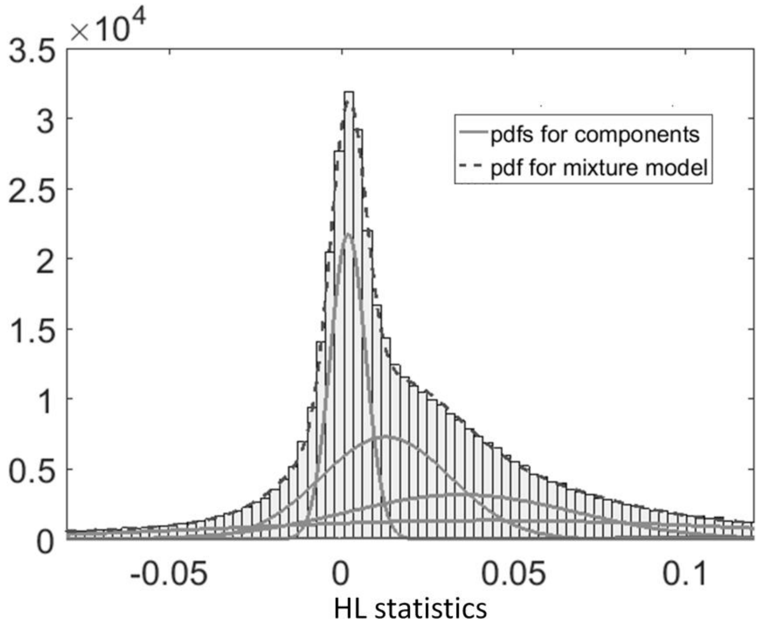

Figure 4.4 Gaussian Mixture Modelling on HL statistics distribution, solid lines represent individual components, and dash line - represents their sum.

The best number of components was nine. Their parameters - means, standard deviations, and weights are presented in Table 4.3.

Table 4.3 Parameters of Gaussian components sorted according to weight.

Component ID	Mean	Standard deviation	Weight
1	0.0128	0.0189	0.2645
2	0.0019	0.0051	0.2148
3	0.0348	0.0334	0.2045
4	0.0427	0.0748	0.1942
5	0.1792	0.1269	0.0597
6	-0.1248	0.1107	0.0358
7	-0.3006	0.1940	0.0158
8	0.3818	0.1775	0.0107

The first four components describe almost 88% of the whole signal. The last four components have lower weights but also higher standard deviations. They can be responsible for the background signal. The component closest to zero (ID = 2) describes probes with no changed methylation level in AML. The other major components have positive means and are responsible for hypermethylated probes. It confirms that hypermethylation is a stronger process than hypomethylation - several subgroups of hypermethylated probes can be distinguished. Cut-off thresholds according to the maximum probability rule were estimated to find them. They are presented in Figure 4.5.

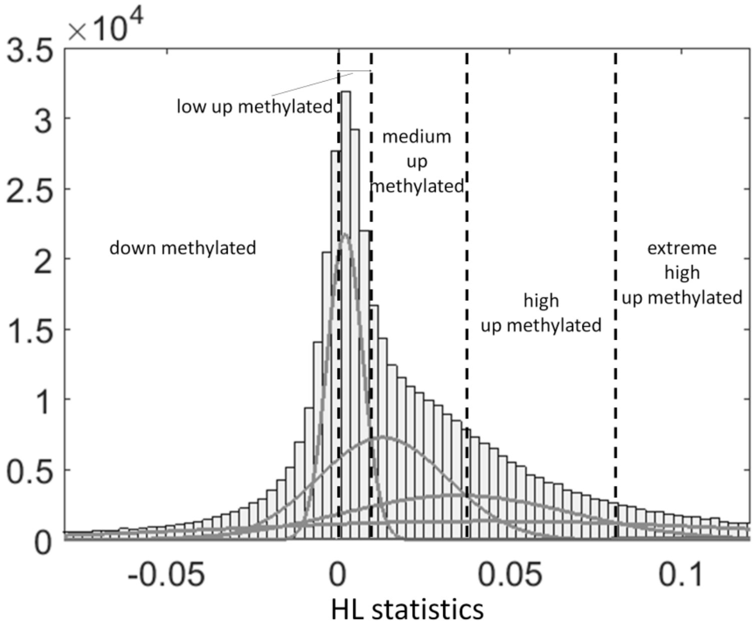

Figure 4.5 Cut-off levels for distinguishing subgroups of differentially methylated probes.

According to cut-off thresholds, hypermethylated probes can be classified as low, medium, high, or extremely high hypermethylated. Cut-off threshold values are presented in Table 4.4.

Table 4.4 Threshold values for individual hypermethylation classes.

Hypermethylation level	**low**	**medium**	**high**	**extreme high**
Threshold values	0.0000	0.0096	0.0372	0.0819

4.1.4. Detection of differentially methylated probes

Probes with HL statistic values significantly lower than 0 were considered hypomethylated, while ones with HL statistic values significantly higher than 0 were considered hypermethylated. Additionally, each probe HL statistic value was compared with threshold values with the right-tailed test to define hypermethylation degree. Probes with HL statistic significantly greater than 0.0096 were considered at least medium hypermethylated, ones with HL significantly greater than 0.0372 - at least high hypermethylated, and ones with HL significantly greater than 0.0819 - extreme high hypermethylated. The numbers of differentially methylated probes in the whole genome and individual genomic regions are presented in Table 4.5.

Table 4.5 Number and percentage of differentially methylated probes according to genomic regions and differential methylation level.

Level of AML differential methylation		Whole genome		TSS region		Gene body		Intergenic	
		N	%	N	%	N	%	N	%
Hypomethylation		15,260	3.14	5,287	2.79	7,010	3.09	3,075	3.29
Hypermethylation	At least low	84,073	17.32	28,492	15.03	39,622	17.45	19,737	21.10
	At least medium	47,659	9.82	14,196	7.49	22,177	9.77	12,738	13.62
	At least high	17,317	3.57	5,577	2.94	7,414	3.27	4,734	5.06
	Extreme high	8,149	1.86	2,716	1.43	3,477	1.53	2,142	2.29

The hypermethylation process is much stronger than hypomethylation in each genomic region type. Hypomethylation is slightly higher than the "by chance" level, especially in TSS regions (2,5%). Both processes are the weakest in TSS regions and strongest in intergenic regions. These relations remain similar independently of hypermethylation level. The percentage of medium and high hypermethylated probes is the biggest in intergenic regions and the lowest in the TSS region. The percentage of probes categorised as extreme high hypermethylated probes is lower than the "by chance" level.

Differential methylation processes can differ depending on the initial methylation level in particular probes. Relations between initial methylation level in HSC (investigated in Chapter 4.1.1) and differential methylation between AML and HSC is presented in Table 4.6.

Table 4.6 The number of differentially methylated probes according to methylation level in HSC.

AML differential methylation		HSC Low	HSC Medium	HSC High
		N	N	N
Whole genome	Hypomethylation	5,374	2,373	7,513
	No change	172,711	37,773	175,735
	Hypermethylation	30,682	10,779	42,612
TSS	Hypomethylation	2,764	774	1,749
	No change	109,047	10,694	36,014
	Hypermethylation	16,334	2,864	9,294
Body	Hypomethylation	2,189	1,046	3,775
	No change	66,433	17,439	96,528
	Hypermethylation	12,046	5,009	22,567
Intergenic	Hypomethylation	523	535	2,017
	No change	12,575	10,364	47,769
	Hypermethylation	4,456	3,213	12,068

In the whole genome, 3.3% of HSC high methylated probes are hypomethylated in AML, and 14.7% of HSC low methylated probes are hypermethylated in AML. The situation is consistent in all genomic regions. However, changes in TSS regions are weakest and intergenic regions are strongest. Two main processes are most interesting: enhancing initial methylation level (hypomethylation of low methylated probes or hypermethylation of high methylated probes) and compensation (hypermethylation of low methylated probes or hypomethylation of high methylated probes). These processes are different among genomic regions. Methylation enhancement regards 6.36% for TSS probes, rises to 10.90% for gene body located probes to almost double for the intergenic region (13.46%). Methylation compensation regards a similar percentage of probes in the body and intergenic regions (6.97% and 6.92%, respectively) and 1.5 times increases for the TSS regions (9.54% of these probes).

4.1.5. Detection of differentially methylated genomic regions

Genomic regions were detected as differentially methylated due to *p*-value integration. Methylation array probes were annotated to 21,227 different gens, from which 20,852 were measured in at least one TSS probe and 20,527 in at least one gene body probe. *P*-values from the abovementioned tests were integrated into one global *p*-value for each TSS and gene body genomic region. The procedure was firstly performed for unadjusted *p*-values and then repeated for corrected *p*-values. Results of *p*-value integration are presented in Table 4.7.

Table 4.7 The number of differentially methylated TSS and gene body regions for various differential methylation levels and integrated *p*-values.

AML-associated differential methylation at the gene level		**Unadjusted *p*-values**				**Storey's corrected *p*-values**			
		Differentially methylated TSS regions		**Differentially methylated body regions**		**Differentially methylated TSS regions**		**Differentially methylated body regions**	
		N	**%**	**N**	**%**	**N**	**%**	**N**	**%**
Hypomethylation		90	0.43	112	0.55	22	0.11	14	0.07
Hypermethylation	**At least low**	945	4.53	948	4.62	600	2.88	598	2.91
	At least medium	385	1.85	422	2.06	187	0.90	162	0.79
	At least high	105	0.50	115	0.56	53	0.25	25	0.12
	Extreme high	31	0.15	35	0.17	18	0.09	5	0.02

In a genomic regions domain, the strength of particular differential processes is consistent with changes in individual probes. All processes are stronger in gene body regions than in TSS regions, considering unadjusted *p*-values. Oppositely, after corrected *p*-values integration, almost all processes are stronger in TSS regions, but the difference is minor. Hypomethylation and extreme high hypermethylation regard a similar number of genomic regions. Hence these regions were investigated deeper in further analysis.

Hypomethylation of the genomic TSS region leads to higher gene expression. The number of genes with hypomethylated TSS region has been mentioned in the literature concerning cancer diseases. Higher expression of TRPM2 was observed in several tumour family diseases: insulinoma, hepatocellular carcinoma, prostate cancer, lymphoma, leukaemia, and lung cancer cell lines. In these cases, TRPM2 could enhance cell death [127]. ESPNL gene has been detected as hypomethylated in MDS. It is also crucial in age-related epigenetic drift in AML and MDS pathogenesis [128]. CFD is the main regulator of complement activation and may advantage leukaemia aggressiveness by suppressing the immune response to AML and regulating stem cell function [129]. NNAT gene has been described as transcriptionally silenced because of hypermethylation in paediatric AML [130].

Hypermethylation of the genomic TSS region can stop the expression of this gene. It can inhibit some tumour-suppressing processes. HTRA4 was detected as extremely high hypermethylated in TSS and gene body regions. It is confirmed to be a tumour suppressor gene and a biomarker in other cancers [131]. Extremely high hypermethylated OXT gene characterises decreased activity in chronic myeloid leukaemia [132]. MYOD1 has already been described as hypermethylated in AML [133], which is confirmed in this study. DPP6 and ID4, identified as hypermethylated in the promoter region, and downregulated in AML [134], were detected in as medium hypermethylated in the TSS region in this study.

Methylation levels of some abovementioned differentially methylated genes are presented in Figure 4.6.

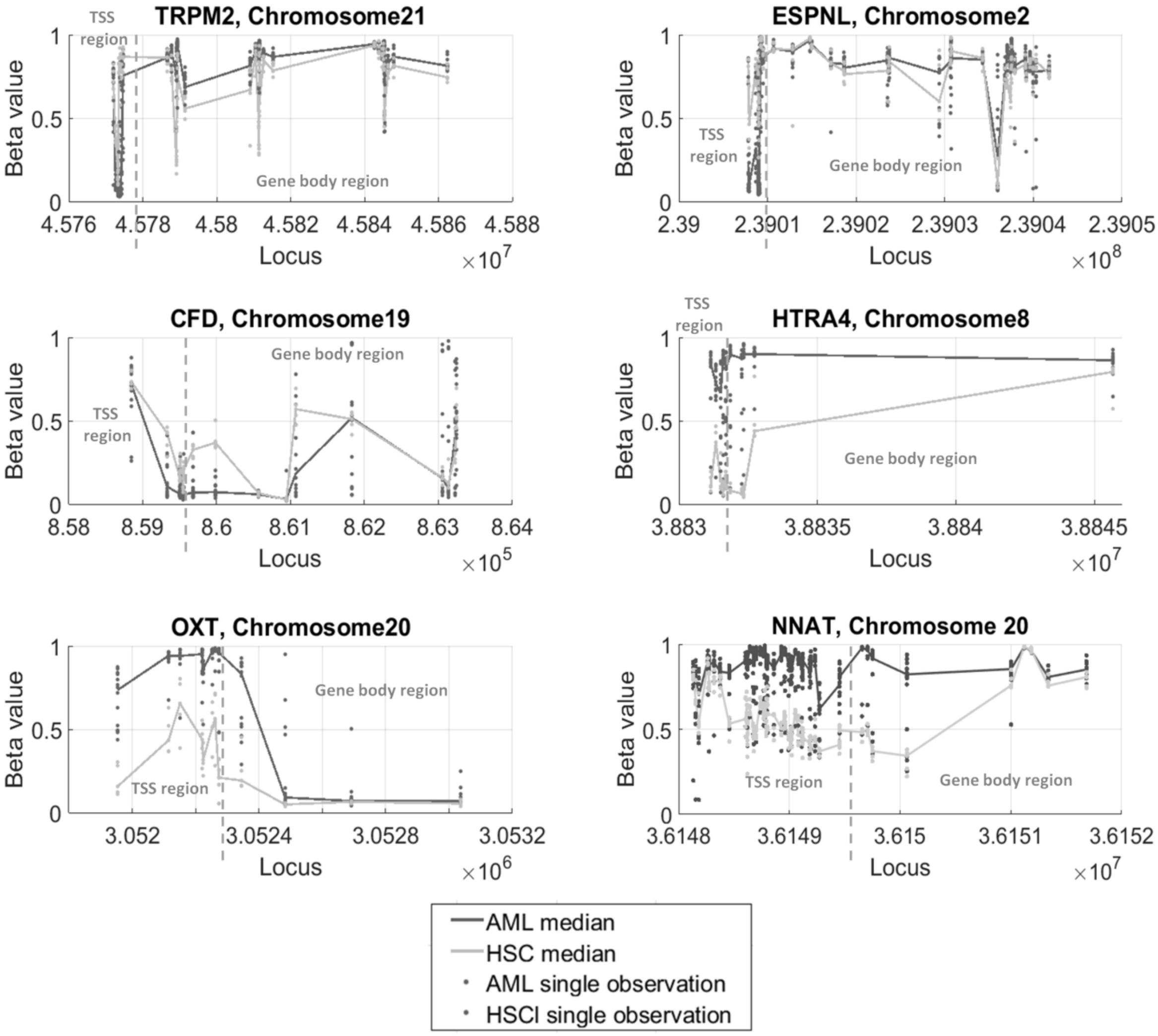

Figure 4.6 The course of methylation level in AML and HCS in selected genes.

4.1.6. Functional analysis

Functional analysis was performed for four separate gene sets: TSS hypo- and extremely high hypermethylated and gene body hypo- and extremely high hypermethylated. Overrepresented Gene Ontology Terms were looked for in three domains: Biological Process, Molecular Function, and Cellular Component. Several GO Terms found for each gene set are presented in Table 4.8.

Table 4.8 The number of overrepresented GO Terms in examined gene sets.

Gene Ontology terms	TSS hypomethylation	TSS extreme high hypermethylation	Body hypomethylation	Body extreme high hypermethylation
Biological Process	113	74	8	56
Molecular Function	13	4	7	2
Cellular Component	25	7	10	0

Some GO terms overrepresented in TSS hypomethylated genes regard calcium ion transport and sequestering (for example, GO:0051283, GO:0051282, GO:0060402, GO:0070588, GO:0060401, GO:0010857, GO:0009931) which confirms literature findings of alteration in calcium processes in AML [135]. The second group of GO terms overrepresented in this gene set regards are immune system processes, which are injured in AML [136]. Examples of these processes are leukocyte differentiation (GO:0002521), hematopoietic or lymphoid organ development (GO:0048534), regulation of interleukin-1 production (GO:0032652), negative regulation of myeloid cell differentiation (GO:0045638), regulation of cytokine secretion (GO:0050707) and many more.

Several GO Terms found for TSS extreme high hypermethylated genes are connected to hormone metabolic processes, especially estrogen (GO:0042445, GO:0032355, GO:0071391, GO:0010817, GO:0046883, GO:0009914, GO:0042562). The estrogen receptor gene was described as a cancer biomarker [137]. Some overrepresented GO Terms for the same gene set regard response for drugs and steroids, i.e., alkaloids, alcohol, cocaine (GO:0042220, GO:0008202, GO:0097305, GO:0045472, GO:0043279). The affectivity of drugs is usually bigger in tumours [138].

4.2. Gender differences in DNA methylation in AML

4.2.1. Comparison of clinical features values between male and female AML patients

Males and females were compared according to the clinical factors to ensure they are not differentiated by them. Results are presented in Table 4.9.

Table 4.9 Results of comparison between genders for clinical factors

Feature	**Vital status**		**Prior treatment**		**Age at diagnosis**	**Days to death**
Values	Alive	Dead	Yes	No	Mean (SD)	Median (MAD)
Male	36	37	19	54	54.21 (15.67)	365 (212)
Female	28	39	19	48	53.76 (16.38)	243 (212)
Statistical significance	*p*-value = 0.3720		*p*-value = 0.7567		*p*-value = 0.8709	*p*-value = 0.3844

For vital status and prior treatment, proportions for each count were compared. The mean values were investigated for age at diagnosis, while the median values - for the number of days to death. The impact of each clinical factor is not statistically significant between genders. In conclusion, none of the clinical factors would impact gender differences in DNA methylation in AML patients.

4.2.2. Detection of differentially methylated probes.

Principal Component Analysis (PCA) was done for the whole genome and genomic region groups separately. It was performed for the healthy control sample and AML sample. Results are presented in Figure 4.7.

Only whole genome PCA results and CpG-rich regulatory sequence regions PCA are presented because results for other regions are very similar to the whole genome. The only case where gender differences are observable is CpG-rich RS regions in AML patients. Such differences do not occur in healthy persons, even in CpG-rich RS regions.

In CpG-rich RS regions in AML, two separable groups are apparent. These groups represent male and female patients. Such grouping indicates that differences between genders occurring in mentioned regions in AML are bigger than in any other genomic regions and healthy donors. The methylation level of each probe was compared between genders to examine these differences. Because the Lilliefors normality test indicated that over 60% of features show non-normal distribution, the Mann-Whitney test was used for this comparison. Additionally, effect size as rank biserial correlation was calculated. The number and percentage of differentially methylated probes in healthy donors are presented in Table 4.10, while for AML patients, Table 4.11. Number represents probes that are differentially methylated in males compared to females.

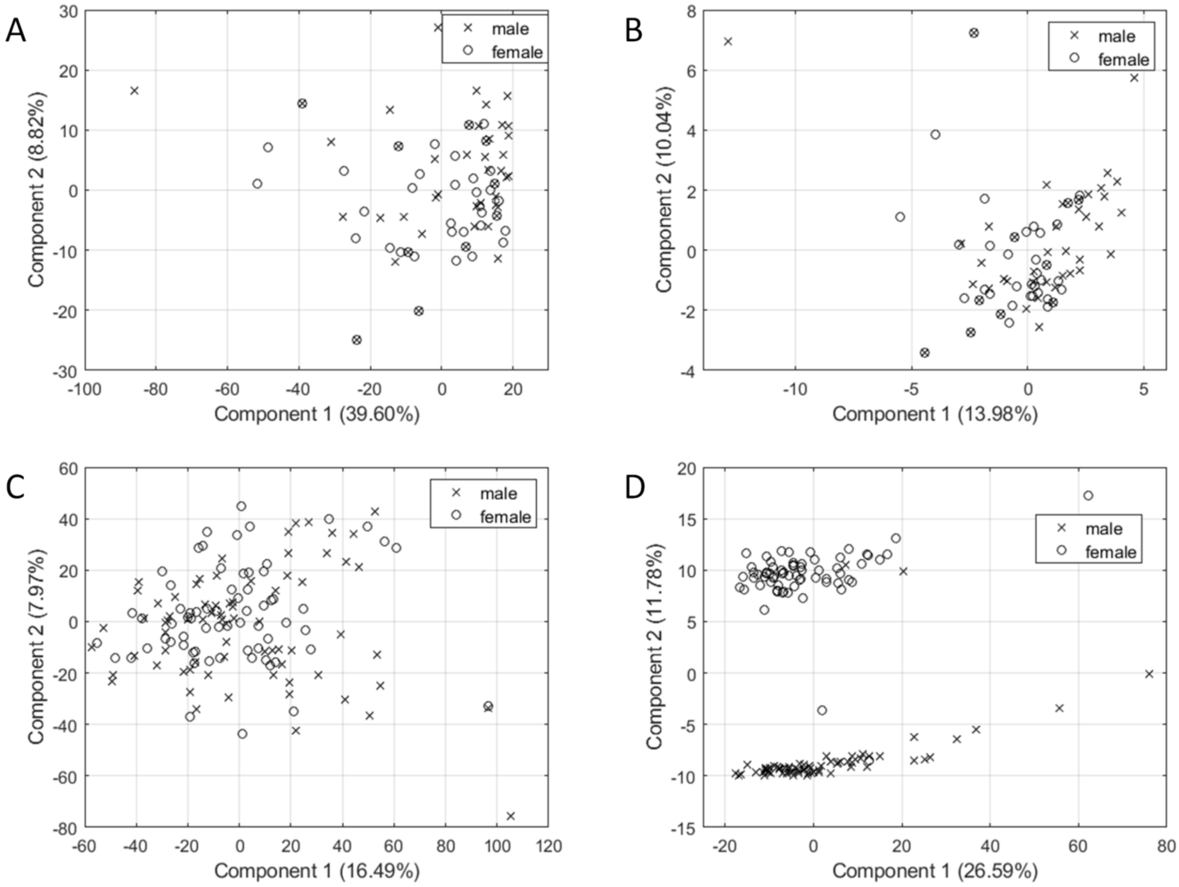

Figure 4.7 PCA on β-values of all probes (A) and CpG-rich RS regions (B) in healthy donors and of all probes (C) and CpG-rich RS regions (D) in AML patients.

Table 4.10 Number and percentage of differentially methylated probes in males in comparison to females among different genomic regions for healthy donors.

	CpG-rich RS regions		**Body regions**		**3'UTR regions**		**Other**		**Whole genome**	
	Hypome-thylated	Hyperme-thylated	Hypome-thylated	Hyperme-thylated	Hypome-thylated	Hyperme-thylated	Hypome-thylated	Hyperme-thylated	Hypome-thylated	Hyperme-thylated
p-value ≤ 0.025	7,591	4,263	14,300	23,978	1,166	3,340	14,881	21,602	35,453	51,364
	10.37%	5.82%	8.13%	13.63%	7.02%	20.12%	9.44%	13.70%	8.92%	12.92%
FDR ≤ 0.025	373	411	739	3,456	49	553	717	2,964	1731	7,166
	0.51%	0.56%	0.42%	1.96%	0.30%	3.33%	0.45%	1.88%	0.44%	1.80%
Effect size ≥0.5	233	100	442	1,045	24	183	415	844	1,016	2,108
	0.32%	0.14%	0.25%	0.59%	0.14%	1.10%	0.26%	0.54%	0.26%	0.53%

Table 4.11 Number and percentage of differentially methylated probes in males compared to females among different genomic regions for AML patients.

	CpG-rich RS regions		**Body regions**		**3'UTR regions**		**Other**		**Whole genome**	
	Hypomethylated	Hypermethylated	Hypomethylated	Hypermethylated	Hypomethylated	Hypermethylated	Hypomethylated	Hypermethylated	Hypomethylated	Hypermethylated
p-value ≤ 0.025	7,045	7,388	11,211	12,160	811	1,001	9,496	11,884	26,035	30,183
	8.71%	9.14%	6.33%	6.87%	4.68%	5.78%	6.47%	8.10%	6.62%	7.67%
FDR ≤ 0.025	2,723	2,212	2,999	3,510	162	354	2,543	3,402	7,386	8,786
	3.37%	2.74%	1.69%	1.98%	0.94%	2.04%	1.73%	2.32%	1.88%	2.23%
Effect size ≥0.5	2318	251	2093	972	91	191	1681	988	5314	2309
	2.87%	0.31%	1.18%	0.55%	0.53%	1.10%	1.15%	0.67%	1.35%	0.59%

It is hard to compare results in healthy and AML samples, looking only at unadjusted *p*-values. However, considering FDR, in a healthy sample, none of the genomic regions characterises significant hypo- or hypermethylation, except 3'UTR regions. In AML patients, a significant number (FDR greater than 2.5%) of differentially methylated probes occurs only in CpG-rich regulatory sequence regions. Both hypo- and hypermethylation processes occur. If the effect size value is considered, much more probes are hypomethylated than hypermethylated in males. Considering FDR, the number of hypo- and hypermethylated probes is similar. In total, 6.11% of probes are differentially methylated. Investigating differentially methylated probes in AML, only 5.54% of them were also differentially methylated in healthy donors.

Differentially methylated probes were marked on a diagram according to their locus in the proper chromosome to examine if differences are equally spread across the whole genome. Results are presented in Figure 4.8. The figure also presents the percentage of differentially methylated probes, which equals the sum of hypo- and hypermethylated probes in males compared to females.

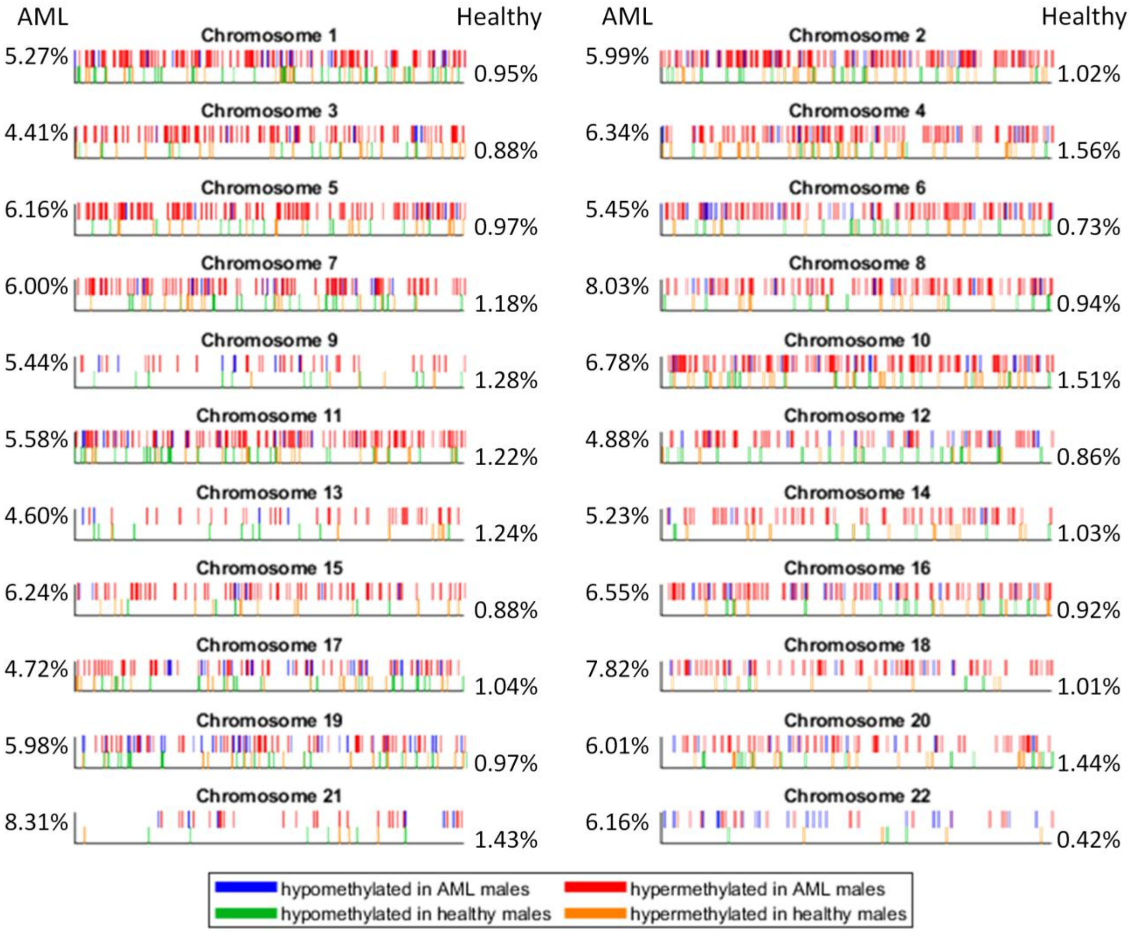

Figure 4.8 Differentially methylated probes in CpG-rich Regulatory Sequence regions and their percentage in AML and healthy groups among chromosomes.

The percentage of differentially methylated probes is rather similar among all chromosomes. However, in several chromosomes, it is two times higher than in others. An example can be chromosomes 8 and 21 with over 8% of differentially methylated probes (in AML) and chromosomes 3 and 13 with around 4.5% of differentially methylated probes. Differences among chromosomes in healthy samples are not so big - the maximal difference is around one percentage point (between 0.42% and 1.44% in chromosomes 22 and 20, respectively), but it is three times more at the same time. However, this is still less than 2.5%, so that these phenomena can occur by chance.

4.2.3. Detection of differentially methylated genomic regions.

P-values of probes belonging to the same regions were integrated to identify differentially methylated genomic regions between genders. Then, *p*-values were compared with an adjusted significance level (α), obtained by integrating basic α repeated as many times as the number of probes in the region. Regions with integrated *p*-values lower or equal to adjusted α were considered differentially methylated. Results of this procedure for both AML and healthy samples are presented in Table 4.12.

Table 4.12 Number and percentage of differentially methylated genomic regions in males compared to females, according to integrated *p*-values.

	CpG-rich RS regions		**Body regions**		**3'UTR regions**		**All regions**	
	Hypomethylated	Hypermethylated	Hypomethylated	Hypermethylated	Hypomethylated	Hypermethylated	Hypomethylated	Hypermethylated
Healthy	272	228	260	522	574	2,168	1,106	2,918
	2.37%	1.98%	1.54%	3.09%	5.50%	20.76%	2.85%	7.52%
AML	462	704	482	470	417	521	1,361	1,695
	3.82%	5.82%	2.68%	2.61%	3.80%	4.75%	3.31%	4.13%

The number of differentially methylated regions in healthy donors is almost always lower than in AML patients, except for 3'UTR regions. A big difference in DNA methylation of the 3'UTR region between genders has not been reported yet. It can happen because only one probe in the 3'UTR region is often measured. When its *p*-value is statistically significant, the whole region becomes differentiating. In CpG-rich RS and body regions, *p*-values must be consistent across the whole region to make it differentiating. Various *p*-values make the whole region irrelevant.

In AML patients, number of genomic regions which are significantly hypermethylated in males is bigger in almost all genomic regions, except body regions. The largest disparity and the highest percentage of differentially methylated regions is observed in CpG-rich RS regions. Because the methylation level of the CpG-rich regulatory sequence region has the most impact on gene transcription, about 9.64% of genes (3.82% hypomethylated and 5.82% hypermethylated in males compared to females) can be differently expressed in males and females.

Genes with differentially methylated CpG-rich RS regions were analysed for their functions in the STRING tool. The number of enriched items is presented in Table 4.13.

Some of the enriched items are specific for hypomethylated genes, while much more of the items are specific for hypermethylated ones. Several items appear only in the combined set of genes analysis: hypo- or hypermethylated in males compared to females.

Table 4.13 The number of enriched items for genes hypermethylated in CpG-rich RS regions in females, males, and a combination of these two sets.

	Females	Males	Combined
GO Biological Process	-	507	303
GO Molecular Function	-	62	34
GO Cellular Component	9	51	44
Publications	282	1025	228
String Cluster	18	81	53
Uniprot Keyword	7	45	41
Pfam	3	40	7
InterPro	3	53	11
SMART	-	14	2
KEGG	-	10	3
Reactome	-	5	3

The items enriching genes hypomethylated genes are connected to diseases linked to the X chromosome, e.g., mental disorders and homeobox. This short DNA fragment occurs in genes involved in morphogenesis and organs development.

The items enriching hypermethylated genes are mentioned in many publications about DNA methylation profiling in many different malignancies. They are expected to be tumour suppressor genes or diagnostic and prognostic markers (e.g., in pancreatic cancer [139] or bladder cancer [140]. They are also connected to homeobox and morphogenesis, different tissues and organs development, cell differentiation, and G protein pathways.

InterPro Keywords: KW-0225, KW-0818, and KW-9995 (Disease mutation, Triplet repeat expansion, and Disease) occur in both hypo- and hypermethylated genes enriched items.

4.2.4. Survival analysis

One of the clinical factors measured for AML patients was time to death, but it did not differentiate genders. Cox proportional hazards model indicates that gender has no impact on the risk (*p*-value = 0.3866). Log-rank test shows no difference between gender survival curves (*p*-value = 0.3902). The survival curves are presented in Figure 4.9. Hence, not only differences between DNA methylation levels in genders should be examined, but also the impact of DNA methylation on survival in all patients and separately in genders.

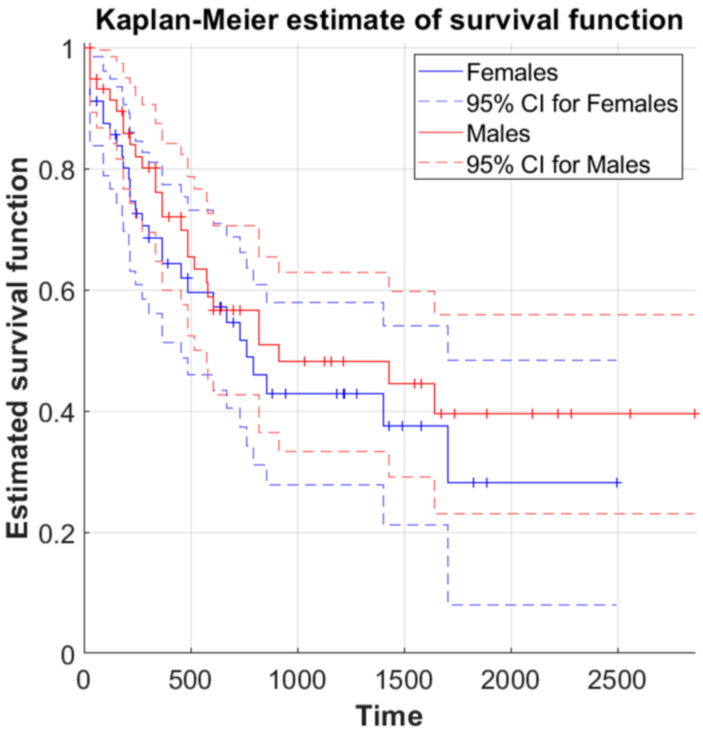

Figure 4.9 Survival curves for both genders with 95% confidence intervals (CI).

For each probe, survival time was compared between two samples: one with a methylation level lower than the median and the second with a methylation level greater or equal to the median. For this purpose, Cox proportional hazard model was estimated. A positive Cox proportional hazards coefficient indicates that survival decreases with a higher methylation level. A negative coefficient means that survival increases with a higher methylation level. Histograms of Cox proportional hazards model *p*-values are presented in Figure 4.10.

In the histograms of *p*-values of models for all patients, an accumulation of low, next to 0, values is observed. It means that factors impacting survival exist. In the case of females, the whole histogram is even. It suggests that all findings can be obtained "by chance".

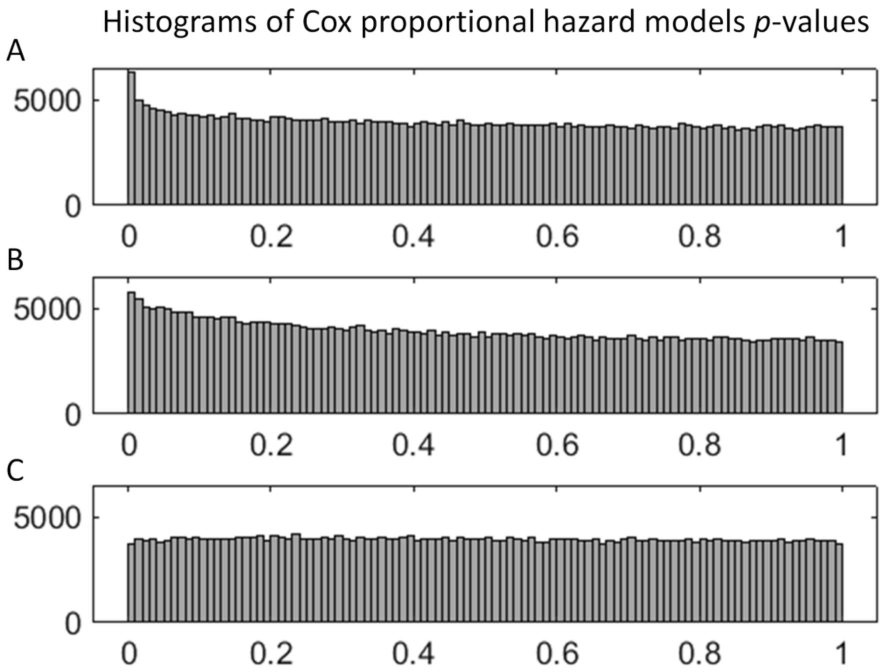

Figure 4.10 Histograms of Cox proportional hazard models for all AML patients(A), males (B), and females (C).

Probes with a significant impact on regression were examined deeper with a log-rank test. *P*-values of the log-rank test were compared with significance level α=0.025 because of distinguishing negative or positive impact on survival. Results are presented in Table 4.14.

Table 4.14 Numbers and percentages of probes impact survival among genomic regions (N is the number of significant features according to the Cox model).

	CpG-rich RS regions		Body regions		3'UTR regions		Other		Whole genome	
Impact on survival	Positive	Negative	Positive	Negative	Positive	Negative	Positive	Negative	Positive	Negative
All patients										
N	5,026		11,464		1,060		9,252		25,037	
Log-rank *p*-value ≤ 0.025	785	2,311	1,870	5,856	110	631	1,711	4,358	4,210	12,318
	15.62%	45.98%	16.31%	51.08%	10.38%	59.53%	18.49%	47.10%	16.82%	49.20%
Males										
N	8,443		11,468		955		8,317		26,340	
Log-rank *p*-value ≤ 0.025	287	5,290	1,147	5,720	125	401	779	3,921	2,226	13,543
	3.40%	62.66%	10.00%	49.88%	13.09%	41.99%	9.37%	47.14%	8.45%	51.42%
Females										
N	3,789		8,476		860		7,447		19,241	
Log-rank *p*-value ≤ 0.025	353	1,854	2,364	2,163	279	152	2,115	1,835	4,957	5,374
	9.32%	48.93%	27.89%	25.52%	32.44%	17.67%	28.40%	24.64%	25.76%	27.93%

The percentage of features impacting survival is similar among all genome regions. Almost always, the number of features with a negative impact on survival is greater than positive. The number of significant features negatively impacting males' survival is even bigger than in all patients. As mentioned, all of the findings in females were probably obtained by chance so that they can compound all patients' results. In females, the number of features impacting survival according to the proportional hazards model (N in Table 4.14) is very close to 5% (significance level). The percentage of CpG-rich RS regions with a negative impact on survival is greater than for all patients. Contrary to other examinations, in the remaining regions, more features positively impact survival than negatively. Looking at the whole genome, the percentage of features having a positive and negative impact on survival is similar.

Two features having the most impact on survival (positive or negative) for all patients were selected. Their survival curves and density are shown in Figure 4.11.

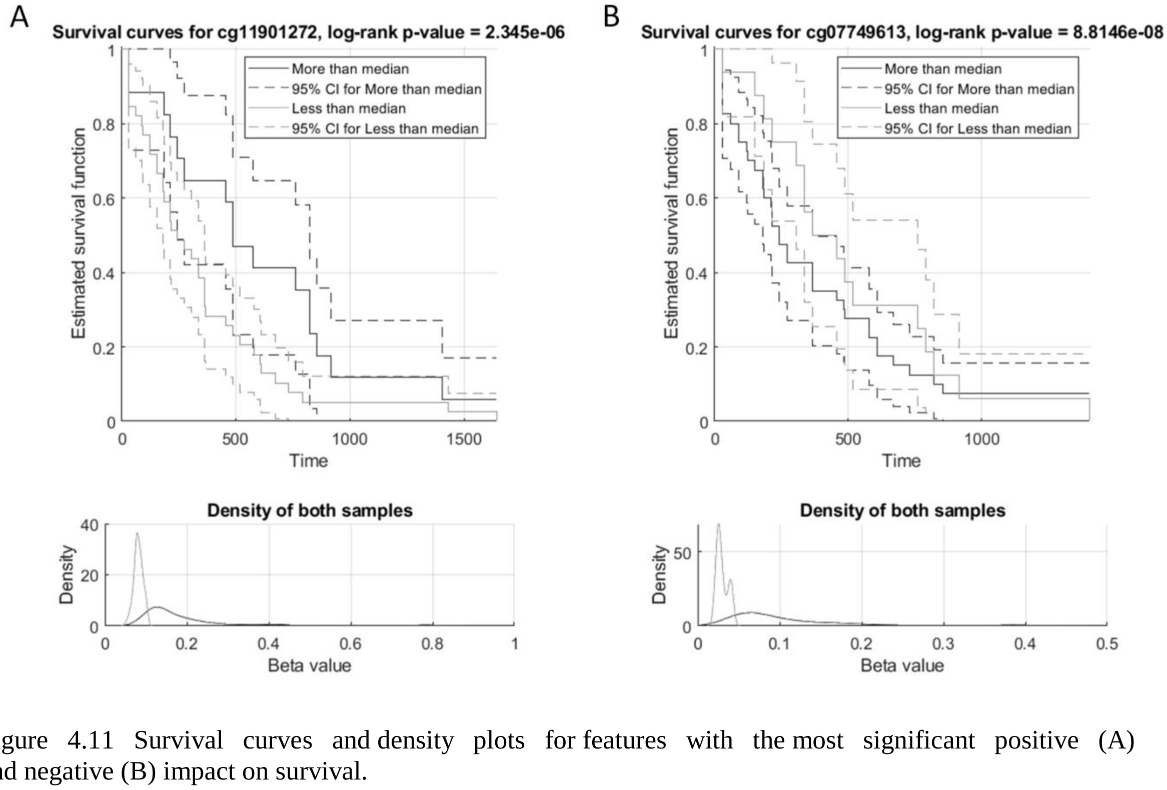

Figure 4.11 Survival curves and density plots for features with the most significant positive (A) and negative (B) impact on survival.

Feature cg11901272 lies on a CpG island in the body region of the HCG4 gene. Its methylation level positively impacts survival - higher β-values are associated with better survival. Feature cg07749613 lies in the intergenic region. Its methylation level harms survival - higher β-values are associated with worse survival. They both are low methylated probes.

Integration of log-rank test *p*-values leads to obtaining genomic regions that can impact survival. The exact number of each type of genomic regions for all patients and genders separately is presented in Table 4.15.

Table 4.15 The number and percentage of genomic regions significantly impact survival for all patients, males and females.

	CpG-rich RS regions		Body regions		3'UTR regions		All regions	
Impact on survival	Positive	Negative	Positive	Negative	Positive	Negative	Positive	Negative
All patients	67	77	45	113	104	491	216	681
	0.55%	0.64%	0.25%	0.63%	0.95%	4.48%	0.53%	1.66%
Males	8	119	43	79	145	167	196	365
	0.07%	0.98%	0.24%	0.44%	1.32%	1.52%	0.48%	0.89%
Females	11	32	30	35	167	103	208	170
	0.09%	0.26%	0.17%	0.19%	1.52%	0.94%	0.51%	0.41%

In all patients and males examinations, the number of features with a negative impact on survival was greater than ones with a positive impact; in females, oppositely. The highest percentage is observed in 3'UTR regions (probably because of a small number of probes inside a region), and the lowest is for body regions. Generally, the most features were found for all patients, than for males and finally for females. However, in CpG-rich regulatory sequence regions, the greatest number is for males. Methylation level in CpG-rich RS regions influences gene expression the most, mainly affecting survival. Venn diagram of CpG-rich RS regions for all examinations is presented in Figure 4.12.

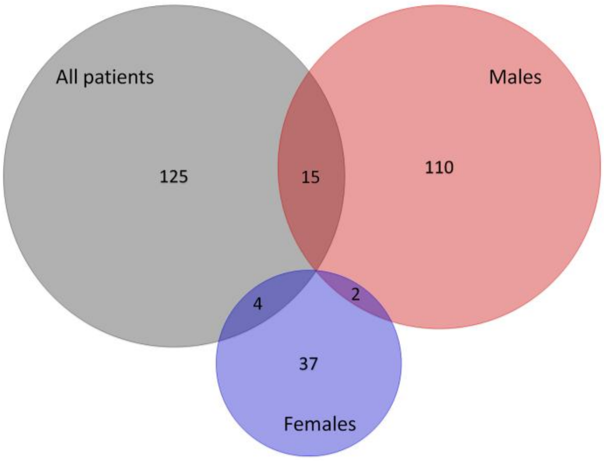

Figure 4.12 Venn diagram showing the number of common CpG-rich RS regions significantly impacting survival.

To compare the significance of features, the impact value was proposed. The impact value is an integrated *p*-value divided by adjusted α for each region. Features with an impact value lower or equal to 1 were considered significantly impacting the survival. Impact values for males and females are highly correlated (Spearman correlation coefficient equals 0.9024, *p*-value < 10^{-15}). However, this correlation occurs for not significant features. There is no relationship between male and female results for those with low impact values.

No common feature for all examinations was obtained. It means that different factors can determine survival in genders. Only two CpG-rich RS gene regions were common for males and females. One of them, the RBL2 gene, was already reported as differentiating indolent and progressive disease of B-cell chronic lymphocytic leukaemia (B-CLL) [141]. In progressive B-CLL, it was downregulated. It this study, its higher methylation level was detected as negatively impacting survival, which is consistent with the literature finding. The second one, the LTF gene, has been described as downregulated and hypermethylated in MLL (mixed lineage leukaemia) in humans and mice [142]. It this study, it was also detected as negatively impacting survival.

4.3. DNA methylation aberrations in *de novo* and therapy-related AML

4.3.1. Data preprocessing

Loading, filtering, and normalisation resulted in DNA methylation level as β-value ranging from 0 to 1, where 0 means no methylation and 1 means total methylation.

In EPIC arrays, methylation level has been measured for 821,323 probes spread genome-wide. Over half of them - 472,454 probes were assigned to genomic regions. 94,321 are located in CpG-rich regulatory sequence regions, 391,335 probes lie in the gene body regions, and 23,772 belong to the 3'UTR regions. Because of performing the experiment on two arrays (two batches), a batch effect removal procedure was necessary to perform. Visualisation of batch effect removal according to Principal Component Analysis is presented in Figure 4.13.

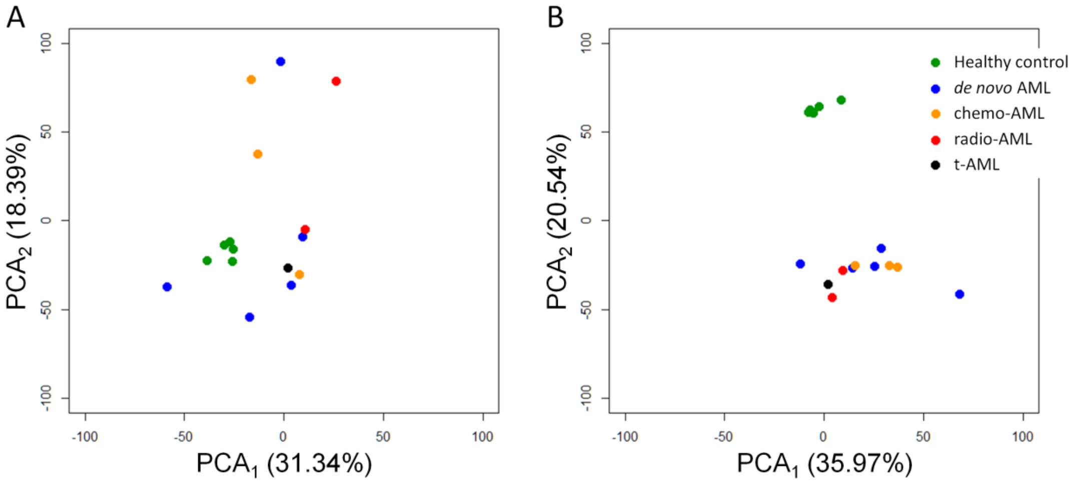

Figure 4.13 First two components from PCA before (A) and after (B) batch effect removal procedure

Batch effect removal resulted in distinguishing the control sample from AML samples. Particular AMLs became more aggregated - representing them points moved closer to each other.

In the 450K array, methylation level was measured for 485,512 probes. 140,717 were assigned to CpG-rich regulatory sequence regions, 210,470 were annotated to gene body regions, and 19,741 were on 3'UTR regions. Samples from all 12 patients were measured on the same array.

4.3.2. Global DNA methylation profile analysis

All probes' distribution on β-values was decomposed into Gaussian Mixture to detect thresholds defining low, medium, and high methylation. The result is presented in Figure 4.14.

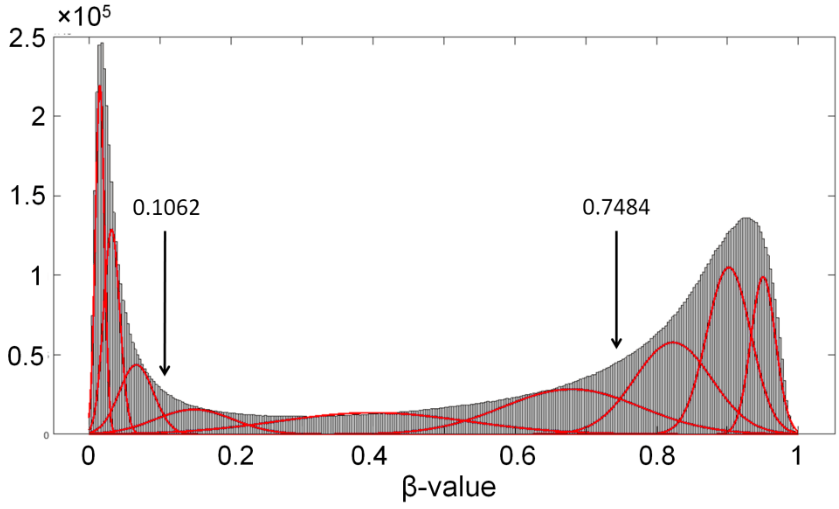

Figure 4.14 Histogram of β-values, their decomposition into Gaussian mixture, and detected thresholds.

Probes with methylation level lower than 0.1062 were considered low methylated; ones with β-value higher than 0.7484 were considered high methylated. The remaining probes were methylated at the medium level. The numbers of probes assigned to each category for every patient group are presented in Table 4.16.

Table 4.16 The number and percentage of a low, medium, and high methylated samples for each patient group.

Methylation level	low	medium	high
Healthy control	189,362	239,265	412,696
	22.51%	28.44%	49.05%
de novo AML	193,094	216,571	431,658
	22.95%	25.74%	51.31%
chemo-AML	205,784	242,888	392,651
	24.46%	28.87%	46.67%
radio-AML	179,432	260,501	401,390
	21.33%	30.96%	47.71%

Across the whole genome, most probes (around a half) are high methylated, despite the patient group. Methylation levels for the remaining probes are low and medium, equally for each quarter. The smallest number of high methylated probes and the greatest number of low methylated ones is observed in chemo-AML (46.67% and 24.46%, respectively). The greatest number of high methylated probes is observed in *de novo* AML (51.31%), but the smallest number of low methylated probes occurs in radio-AML (21.33%).

The analysis was repeated for genomic regions separately. Results are shown in Table 4.17.

The methylation profile of particular genomic regions is diversified. In CpG-rich regulatory sequence regions, most probes are low methylated, indicating good accessibility of gene promoters. The lowest number of low methylated probes occurs in radio-AML. In the body region methylation profile is similar to the whole genome profile. 3' UTR region is the region with the highest methylation.

Table 4.17 The number and percentage of a low, medium, and high methylated samples for each patient group and every genomic regions.

Region	CpG-rich RS regions			Body regions			3' UTR regions		
Methylation level	low	medium	high	low	medium	high	Low	medium	high
Healthy control	82,779	9,701	1,841	73,601	100,273	217,461	1,805	5,836	16,131
	87.76%	10.29%	1.95%	18.81%	25.62%	55.57%	7.59%	24.55%	67.86%
de novo AML	80,052	12,105	2,164	75,202	90,652	225,481	1,929	5,551	16,292
	84.87%	12.83%	2.29%	19.22%	23.16%	57.62%	8.11%	23.35%	68.53%
chemo-AML	82,291	10,013	2,017	79,836	102,780	208,719	2,101	6,436	15,235
	87.25%	10.62%	2.14%	20.40%	26.26%	53.34%	8.84%	27.07%	64.09%
radio-AML	77,991	13,910	2,420	69,886	108,174	213,275	1,617	6,313	15,842
	82.69%	14.75%	2.57%	17.86%	27.64%	54.50%	6.80%	26.56%	66.64%

Inside-group diversity is observed even if the whole genome methylation profile is not diversified among patient groups. The most diverse group is *de novo* AML. However, even the healthy control group is not consistent. It can be observed in the empirical cumulative distribution function for pooled patient groups samples and patients separately (Figure 4.15).

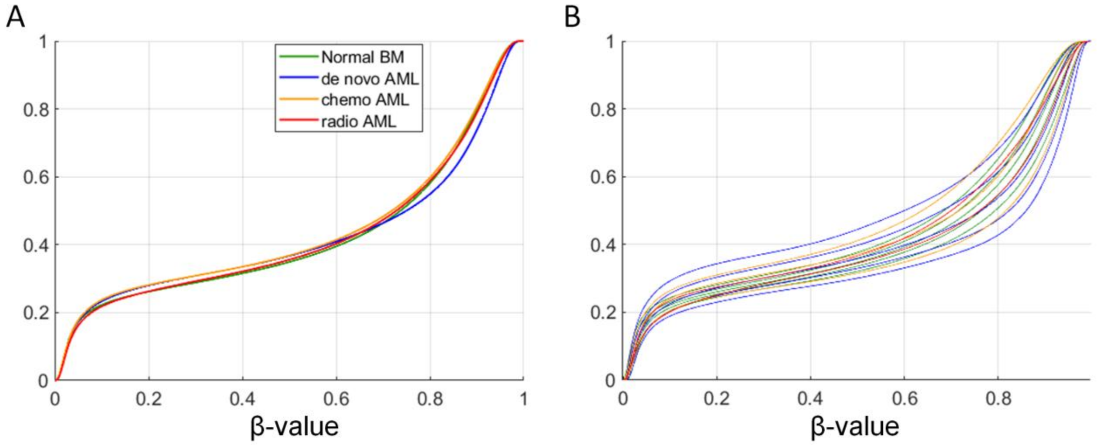

Figure 4.15 Empirical CDF plots for pooled samples (A) and patients separately (B).

Around 35% of probes in pooled samples have a methylation level lower than 0.5 in every patient group. However, among individual patients, this value ranges from around 30% to around 45% in both cases for *de novo* AML patients. Empirical CDF for patients in the same group were plotted separately to look deeper into inside-group diversity. It is presented in Figure 4.16.

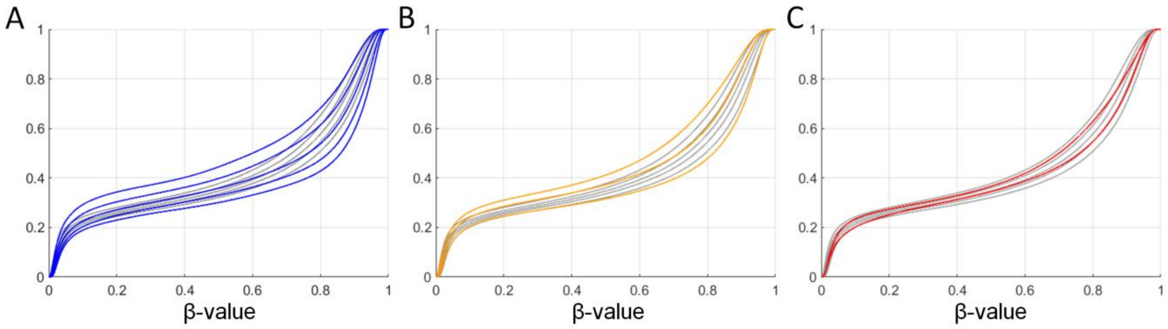

Figure 4.16 Empirical CDF plots patients separately for *de novo* AML (A), chemo-AML (B), and radio-AML (C) compared to healthy control (grey lines).

The most diversified patient group is *de novo* AML; patients from this group have methylation profiles scattered wider than other patient groups and healthy control. None of the patient groups can be easily distinguished from the control group based on these plots.

The distribution of β-values is even more diversified among genomic regions. Empirical CDF plots for CpG-rich regulatory sequence regions are presented in Figure 4.17, for body regions in Figure 4.18, and 3'UTR regions in Figure 4.19. In each case, empirical CDF is calculated for pooled samples and individual patients.

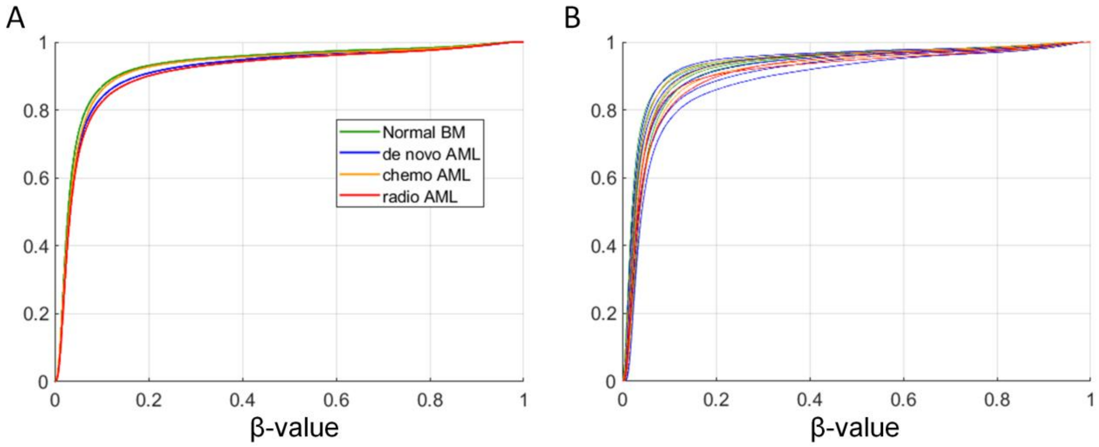

Figure 4.17 Empirical CDF plots for pooled samples (A) and patients separately (B) for CpG-rich RS regions

Probes annotated to CpG-rich RS regions are characterised by a much lower methylation level than the whole genome. Patient groups are also less diversified inside.

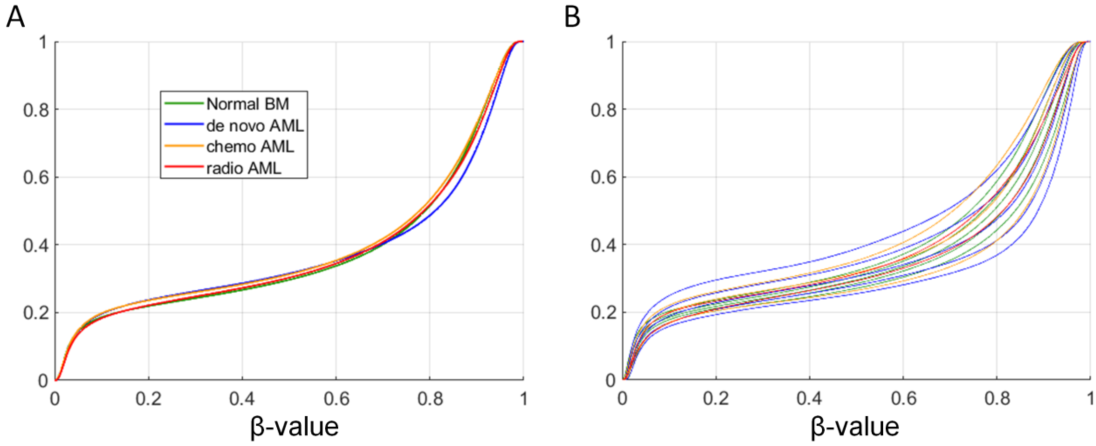

Figure 4.18 Empirical CDF plots for pooled samples (A) and patients separately (B) for body regions

Methylation level in body regions is similar to the whole genome; it is only slightly higher than in the whole genome. Patient groups are also diversified inside.

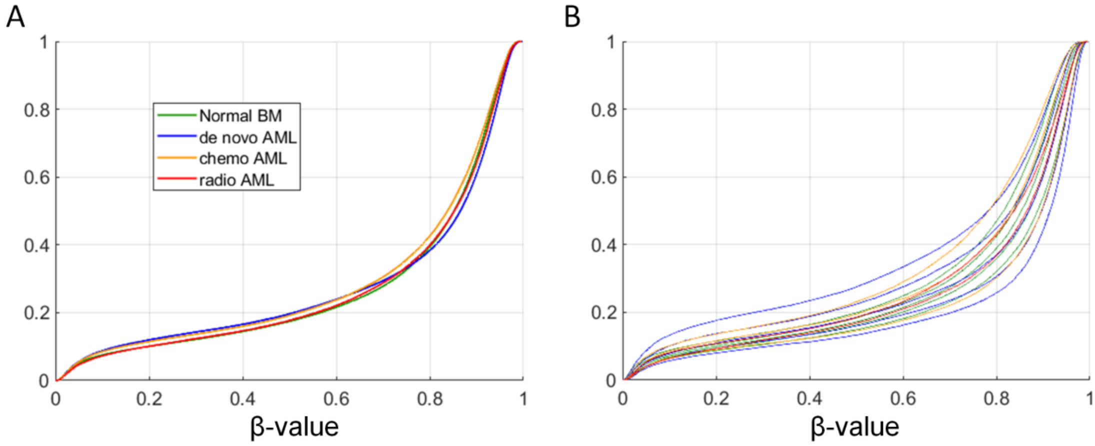

Figure 4.19 Empirical CDF plots for pooled samples (A) and patients separately (B) for 3'UTR regions

Probes annotated to 3'UTR regions have the highest methylation level of all genomic region types. The methylation profile for patients is strongly diversified inside patient groups.

Changes in the categorisation of probes between healthy control and each AML are presented in Figure 4.20.

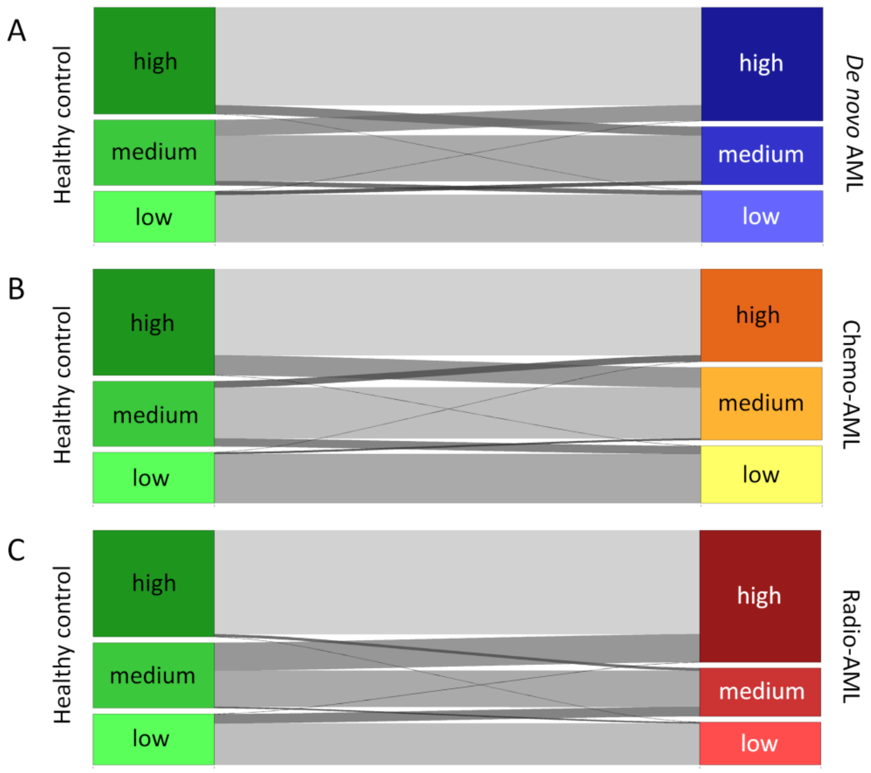

Figure 4.20 Changes in categorisation from healthy control to *de novo* AML (A), chemo-AML (B), and radio-AML (C). Probes state in control is shown on the left and AML on the right part of the diagram. Each line is proportional to the number of probes.

In the case of *de novo* AML, most probes remain in the same category; the biggest change is from medium methylation in control to high methylation. Transformations from low to high and high to low are very weakly represented. In the case of chemo-AML, the biggest transformation is from high to medium methylation. In radio-AML, generally, changes are the biggest among all AMLs, and the strongest process is a change from medium to high methylation.

4.3.3. Detection of differentially methylated probes

Detection of differentially methylated probes was performed in two ways:

- ANOVA and post-hoc pairwise comparisons with Tukey-Kramer test;
- pairwise comparisons with limma.

The number of preliminarily selected probes found with ANOVA is presented in Table 4.18.

Table 4.18 The number of significant probes obtained in ANOVA test, before and after *p*-value correction for all examined probes and each genomic region type

	Overall N = 472,454	CpG-rich RS regions N = 94,321	Body regions N = 391,335	3'UTR regions N = 23,772
No *p*-value correction (*p*-value ≤ 0.05)	118,284	27,641	95,728	5,555
	25.04%	29.31%	24.46%	23.37%
BH correction (FDR ≤ 0.05)	**38,445**	8,187	31,687	1,702
	8.14%	8.68%	8.10%	7.16%

The percentage of differentially methylated probes before *p*-value correction is highest in CpG-rich regulatory sequence regions. After the Benjamini-Hochberg correction percentage of differentiating probes is similar for every region types.

Number of probes differentiating particular pairs of patient groups is presented in Table 4.19.

Table 4.19 The number of significant probes obtained in Tukey-Kramer pairwise tests, for each pair of patient groups and genomic region type, hypo- and hypermethylation.

	Overall (ANOVA FDR≤0.05) N = 38,445		CpG-rich RS regions N = 8,187		Body regions N = 31,687		3'UTR regions N = 1,702	
	Hypomethylated	Hypermetyhlated	Hypomethylated	Hypermetyhlated	Hypomethylated	Hypermetyhlated	Hypomethylated	Hypermetyhlated
De novo AML vs. control	13,503	14,933	1,402	2,311	12,080	12,764	619	754
	35.12%	38.84%	17.12%	28.23%	38.12%	40.28%	36.37%	44.30%
Chemo-AML vs. control	14,918	15,180	1,975	4,036	13,085	12,028	688	683
	38.80%	39.48%	24.12%	49.30%	41.29%	37.96%	40.42%	40.13%
Radio-AML vs. control	11,309	18,463	1,660	4,589	9,858	14,720	472	822
	29.42%	48.02%	20.28%	56.05%	31.11%	46.45%	27.73%	48.30%
Chemo-AML vs. *de novo* AML	3,828	3,525	891	2,151	3,110	2,120	177	81
	9.96%	9.17%	10.88%	26.27%	9.81%	6.69%	10.40%	4.76%
Radio-AML vs. *de novo* AML	1,892	7,068	581	2,603	1,468	5,092	69	245
	4.92%	18.38%	7.10%	31.79%	4.63%	16.07%	4.05%	14.39%
Radio-AML vs. chemo-AML	1,255	6,235	628	1,694	824	4,864	34	255
	3.26%	16.22%	7.67%	20.69%	2.60%	15.35%	2.00%	14.98%

Particular AMLs differ more with healthy control than among each other. In AMLs, the hypermethylation process is stronger than hypomethylation; the exceptions are body regions and 3'UTR regions in chemo-AML. Hypermethylation process is also stronger in both therapy-related AMLs compared to *de novo* AML, with the same exceptions.

Limma pairwise comparisons were performed instantly in the second approach, without preliminary selection. Number of differentiating probes obtained in this examination is presented in Table 4.20.

Table 4.20 The number of significant probes obtained in limma pairwise tests, for each pair of patient groups and genomic region type, hypo- and hypermethylation.

	Overall N = 472,454		CpG-rich RS regions N = 94,321		Body regions N = 391,335		3'UTR regions N = 23,772	
	Hypomethylated	Hypermetyhlated	Hypomethylated	Hypermetyhlated	Hypomethylated	Hypermetyhlated	Hypomethylated	Hypermetyhlated
De novo AML vs. control	43,736	72,747	2,891	8,961	39,584	63,733	2,754	3,639
	9.26%	15.40%	3.07%	9.50%	10.12%	16.29%	11.59%	15.31%
Chemo-AML vs. control	54,027	41,253	5,454	9,764	47,901	33,423	3,159	1,891
	11.44%	8.73%	5.78%	10.35%	12.24%	8.54%	13.29%	7.95%
Radio-AML vs. control	34,676	62,356	3,597	13,492	30,961	50,627	1,741	3,197
	7.34%	13.20%	3.81%	14.30%	7.91%	12.94%	7.32%	13.45%
Chemo-AML vs. *de novo* AML	29,015	10,231	4,287	3,703	24,939	7,588	1,493	422
	6.14%	2.17%	4.55%	3.93%	6.37%	1.94%	6.28%	1.78%
Radio-AML vs. *de novo* AML	17,718	26,636	1,890	6,559	15,841	20,948	841	1,478
	3.75%	5.64%	2.00%	6.95%	4.05%	5.35%	3.54%	6.22%
Radio-AML vs. chemo-AML	5,408	28,725	1,276	5,733	4,439	23,464	231	1,612
	1.14%	6.08%	1.35%	6.08%	1.13%	6.00%	0.97%	6.78%

Number of differentially methylated probes in each comparison is greater than in the Tukey-Kramer test. However, percentages are lower because the reference is the number of all probes, not preliminary selected. Relations between hypo- and hypermethylation are consistent with previous analysis. A comparison of both method results is presented in Figure 4.21.

Proportions between limma and Tukey-Kramer results are consistent for every comparison. The ANOVA + Tukey-Kramer approach seems more restrictive because of the correction for multiple pairwise comparisons in the Tukey-Kramer test.

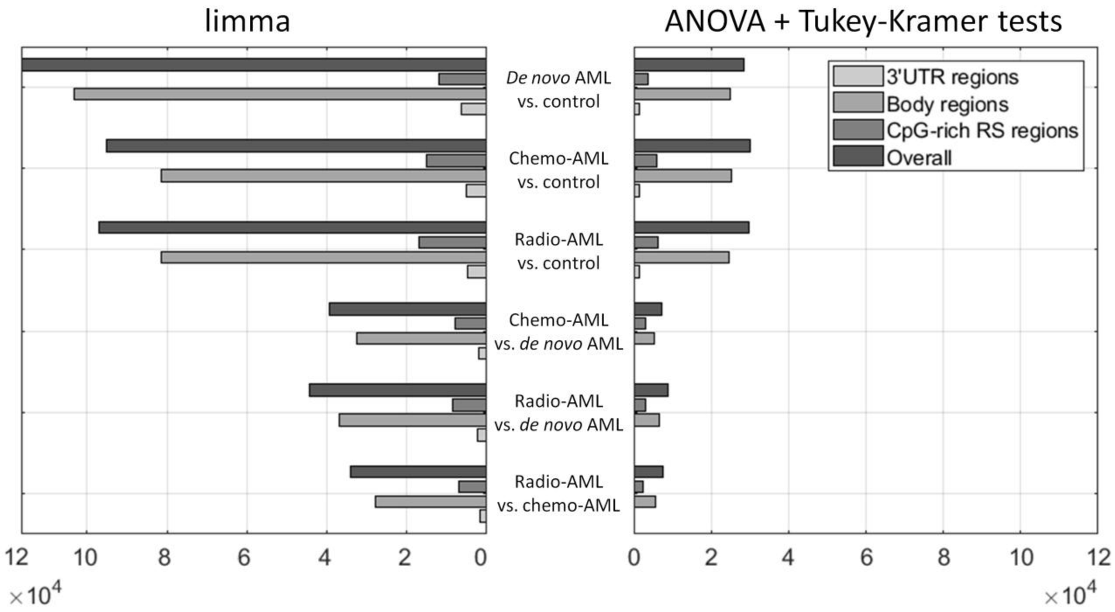

Figure 4.21 Number of differentially methylated (sum of hyper- and hypomethylated) probes in each comparison.

To check "inside-group" diversity, each AML patient was compared with healthy control. Only probes annotated to CpG-rich regulatory sequence regions were considered. 47,596 probes did not differentiate any AML patients with healthy control. For the remaining probes, frequencies for hypo- and hypermethylation and no change were computed. Results are presented in Table 4.21.

Table 4.21 Frequencies of hypo- methylation, hypermethylation, and no change for probes differentiating at least one patient and control.

Patient group	*De novo* AML					Chemo-AML			Radio-AML	
Patient's number	1	2	3	4	5	1	2	3	1	2
Hypomethylation	0.0345	0.0653	0.0373	0.0301	0.0288	0.0717	0.0403	0.0681	0.0407	0.0531
No change	0.6337	0.7013	0.7643	0.7291	0.6384	0.7192	0.5693	0.6813	0.5313	0.5488
Hypermethylation	0.3317	0.2335	0.1984	0.2408	0.3328	0.2091	0.3904	0.2506	0.4280	0.3981

Average frequencies are presented in Figure 4.22.

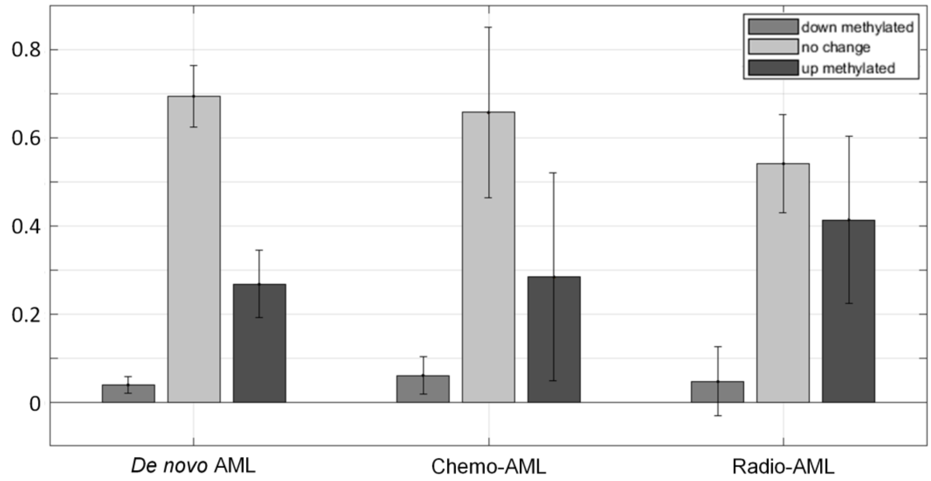

Figure 4.22 Average frequencies with confidence intervals for hypo- and hypermethylation and no change for all patient groups.

The highest frequency of hypomethylation is observed in chemo-AML, while the highest frequency for hypermethylation is for radio-AML. *De novo* AML is characterised by the highest no-change frequency. Confidence intervals are widest for chemo AML (for no change and hypermethylation) and radio-AML (for hypomethylation). The relationship between hypo- and hypermethylation frequencies for each patient is presented in Figure 4.23.

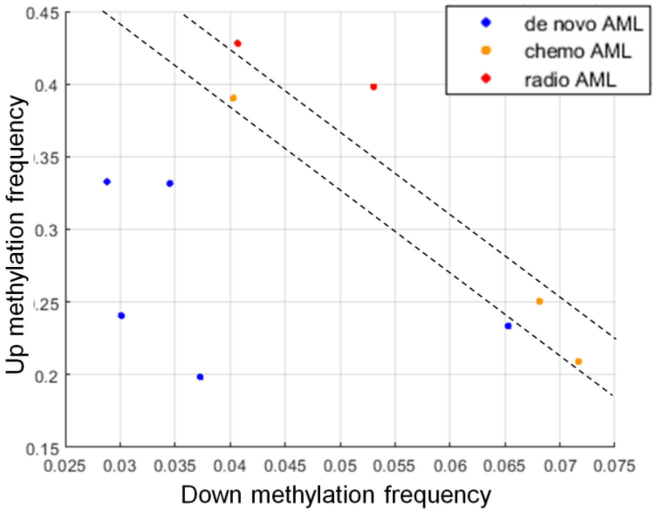

Figure 4.23 Relationship between hyper- and hypomethylation frequencies for each patient.

Frequencies are diversified even among patients in the same group. However, it is possible to separate patients from different groups according to the frequencies (dash lines).

The Pielou diversity index for each patient was calculated based on these frequencies. Results are presented in Table 4.22.

Table 4.22 Pielou diversity index values for individual patients and their averages for patient groups.

Patient group	*De novo* AML					Chemo-AML			Radio-AML	
Patient's number	1	2	3	4	5	1	2	3	1	2
Pielou index	0.7021	0.6979	0.5907	0.6177	0.6871	0.6856	0.7440	0.7202	0.7550	0.7753
Mean	0.6591					0.7166			0.7652	

The highest average diversity index is for radio-AML, and the lowest is for *de novo* AML. Pielou index values were compared among patient groups with ANOVA. ANOVA *p*-value was 0.0458. It means that the diversity of methylation changes significantly differentiates patient groups. Also, Tukey-Kramer post hoc tests for pairwise comparisons were performed. *P*-values were 0.2180, 0.0461, and 0.4587 for *de novo* AML vs. chemo-AML, *de novo* AML vs. radio-AML, and chemo-AML vs. radio-AML comparisons, respectively. Hence, only two far patient groups differ significantly according to the Pielou diversity index.

4.3.4. Detection of differentially methylated genomic regions

Differentially methylated regions were detected due to *p*-value integration. In this case, it was focused only on the 14,338 CpG-rich RS regions because they have a proven impact on gene expression. It was performed for Tukey-Kramer tests *p*-values and limma test *p*-values. Number of differentially methylated genomic regions according is presented in Table 4.23.

Number of genomic regions obtained with limma is greater than with Tukey-Kramer tests. However, the results are relatively consistent. The greatest difference is the number of genomic regions hypomethylated on chemo-AML compared to *de novo* AML: 33 in the Tukey-Kramer approach and 204 in the limma approach. Number of common genomic regions found in both approaches for each AML compared to control is presented in Figure 4.24.

Table 4.23 Number and percentage of differentially methylated CpG-rich RS regions.

	Tukey-Kramer		limma	
	Hypomethylated	Hypermetyhlated	Hypomethylated	Hypermetyhlated
De novo AML vs. control	40	236	58	745
	0.28%	1.65%	0.40%	5.20%
Chemo-AML vs. control	74	168	111	280
	0.52%	1.17%	0.77%	1.95%
Radio-AML vs. control	31	466	53	958
	0.22%	3.25%	0.37%	6.68%
Chemo-AML vs. *de novo* AML	33	29	204	42
	0.23%	0.20%	1.42%	0.29%
Radio-AML vs. *de novo* AML	9	157	27	362
	0.06%	1.09%	0.19%	2.52%
Radio-AML vs. chemo-AML	5	140	7	378
	0.03%	0.98%	0.05%	2.64%

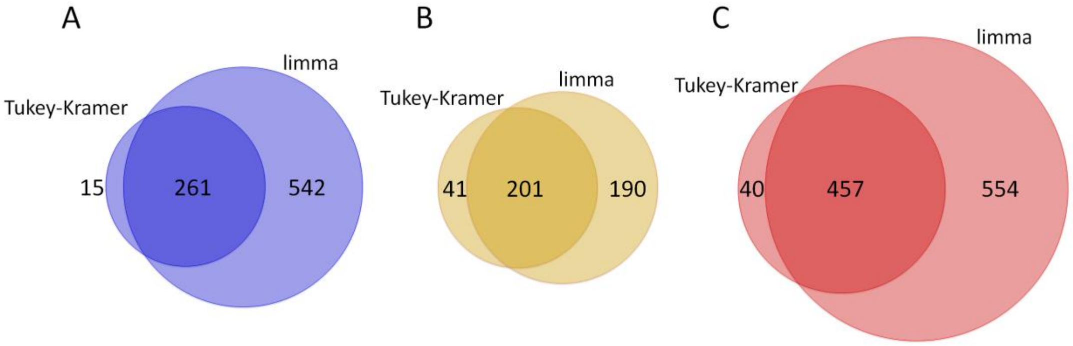

Figure 4.24 Venn diagrams presenting the number of common genomic regions obtained with *p*-value integration of Tukey-Kramer tests and limma for *de novo* AML (A), chemo-AML (B), and radio-AML (C) compared to control.

Almost all genomic regions found in integration *p*-values from Tukey-Kramer tests were also obtained with integration *p*-values from limma.

4.3.5. Unsupervised feature selection

The distribution logarithm of the variance of their methylation level was decomposed into Gaussian Mixture to select features (CpG-rich regulatory sequence genomic regions) in an unsupervised way. The visualisation of the results is shown in Figure 4.25.

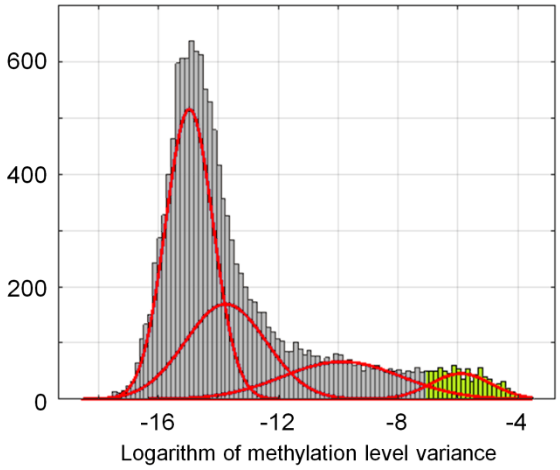

Figure 4.25 Gaussian Mixture components for distribution of the logarithm of methylation level variance. Values for genomic regions selected for further analysis are marked with green color.

Most diverse genomic regions (828) were selected to create a hierarchical tree and heatmap of patients according to the average methylation level for each CpG-rich RS genomic region. Results are presented in Figure 4.26.

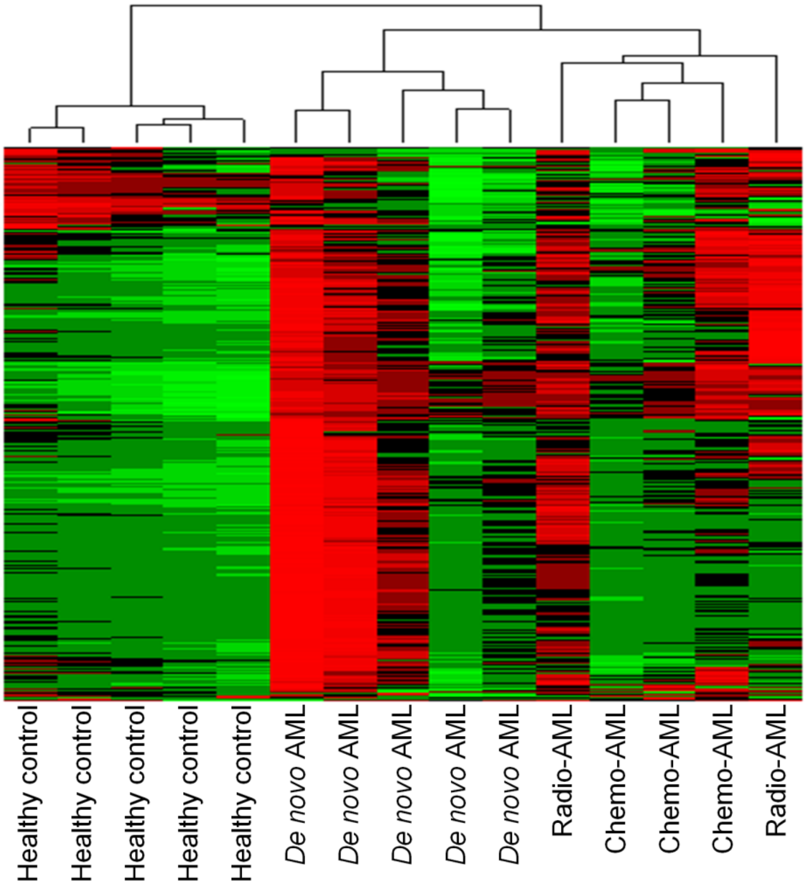

Figure 4.26 Heatmap and a hierarchical tree of patients according to mean methylation level in selected genomic regions.

The biggest similarity occurs among all healthy control samples grouped. The second main root consists of AMLs samples with separated *de novo* AML and therapy-related AMLs. Inside therapy-related AMLs, chemo-AML samples are grouped.

4.3.6. Functional analysis

The same procedure was conducted for hiPatia subpathways activation scores. Visualisation of Gaussian Mixture Model for distribution of their variance is presented in Figure 4.27.

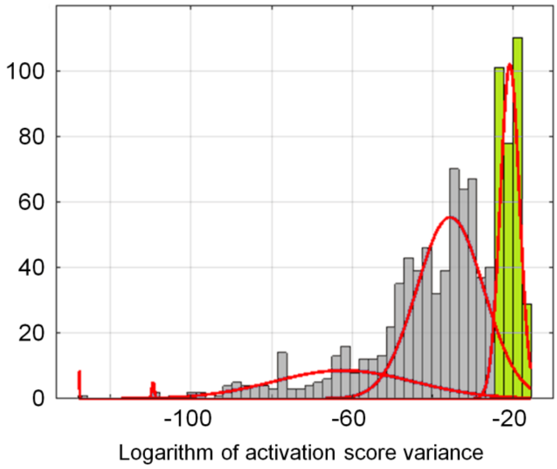

Figure 4.27 Gaussian Mixture components for distribution of the logarithm of hiPathia subpathways activity score variance. Values for subpathways selected for further analysis are marked with green color.

Most diverse subpathways (322) were selected to create a hierarchical tree and heatmap of patients according to activation score. Results are presented in Figure 4.28.

Similarly, as for RS regions, the most compact group was healthy control. AMLs are not separated so well - one of the radio-AML patients is clustered with *de novo* AML patients. However, many selected subpathways are included in the KEGG AML pathway (hsa05221): PI3K-Akt signaling pathway, Apoptosis, mTOR signaling pathway, MAPK signaling pathway, and Cell cycle. They consist of 11% of selected subpathways.

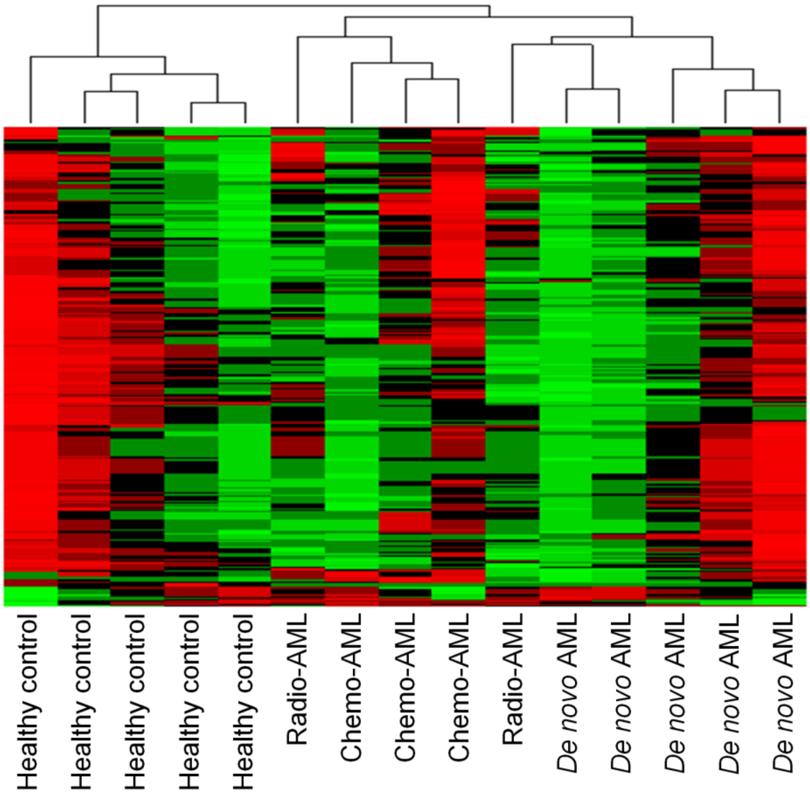

Figure 4.28 Heatmap and a hierarchical tree of patients according to subpathways' activation score.

4.3.7. Validation of selected biomarkers

A biomarker was defined as a genomic region that differentiates a particular AML from control and does not differentiate other AMLs from control. A pyrosequencing experiment examined one potential AML marker: AURKC, one potential chemotherapy-related AML marker: Mir886 (VTRNA2-1), and two potential radiotherapy-related AML markers: MEST and GATA5. All selected biomarkers were found in both approaches: ANOVA and Tukey-Kramer post hoc tests, as well as in limma. Their significance according to integrated *p*-value and effect size in both primary and validation experiments is presented in Table 4.24.

Table 4.24 Results of *p*-value integration and effect size estimation for CpG-rich genomic region of selected genes in primary and validation experiments.

Gene	Experiment	Comparison	Statistical test	Adjusted α	Integrated *p*-value	Median effect size
AURKC	Primary experiment	*De novo* AML vs. control (hypermethylation)	Tukey-Kramer	$4.00 \cdot 10^{-11}$	$1.04 \cdot 10^{-26}$	10.87 (huge)
		Chemo-AML vs. control (hypermethylation)			$1.64 \cdot 10^{-25}$	8.52 (huge)
		Radio-AML vs. control (hypermethylation)			$4.35 \cdot 10^{-24}$	9.33 (huge)
	Validation experiment	All AMLs vs. control (hypermethylation)	T-test	$4.43 \cdot 10^{-5}$	$2.33 \cdot 10^{-7}$	9.58 (huge)
		De novo AML vs. control (hypermethylation)	Tukey-Kramer		$1.56 \cdot 10^{-5}$	1.66 (very large)
Mir886 (VTRNA2-1)	Primary experiment	Chemo-AML vs. control (hypomethylation)	Tukey-Kramer	$5.86 \cdot 10^{-6}$	$4.14 \cdot 10^{-11}$	119.50 (huge)
	Validation experiment		Tukey-Kramer	$5.86 \cdot 10^{-6}$	$9.66 \cdot 10^{-5}$	1.61 (very large)
MEST	Primary experiment	Radio-AML vs. control (hypermethylation)	Tukey-Kramer	$2.67 \cdot 10^{-42}$	$7.92 \cdot 10^{-44}$	6.53 (huge)
	Validation experiment		Tukey-Kramer	$5.86 \cdot 10^{-6}$	$8.91 \cdot 10^{-5}$	1.91 (very large)
GATA5	Primary experiment	Radio-AML vs. control (hypermethylation)	Tukey-Kramer	$1.91 \cdot 10^{-20}$	$4.57 \cdot 10^{-22}$	7.06 (huge)
	Validation experiment		Tukey-Kramer	$4.43 \cdot 10^{-5}$	$7.46 \cdot 10^{-4}$	1.77 (very large)

AURKC was considered an all AMLs biomarker because it differentiated each AML with healthy control. In validation experiment analysis, its methylation level was compared between all AML patients and healthy control and between *de novo* AML and healthy control. In each case, the integrated *p*-value is lower than the adjusted significance level, and the median effect size is huge or very large (in *de novo* AML vs. control comparison in validation experiment). In the case of other genes, comparisons that resulted in a significant *p*-value in the primary experiment were repeated in validation experiment analysis. In each case, in the primary experiment median effect size is huge, and in the validation experiment, it is very large. *P*-values in the validation experiment were slightly higher than the adjusted significance level. All markers were confirmed in the validation experiment based on median effect size, while AURKC was also based on the integrated *p*-value.

Additionally, methylation level in CpG-rich regulatory sequence regions of selected biomarkers was checked in the first experiment, conducted on Human Methylation 450K array. Visualisation for the AURKC gene is shown in Figure 4.29.

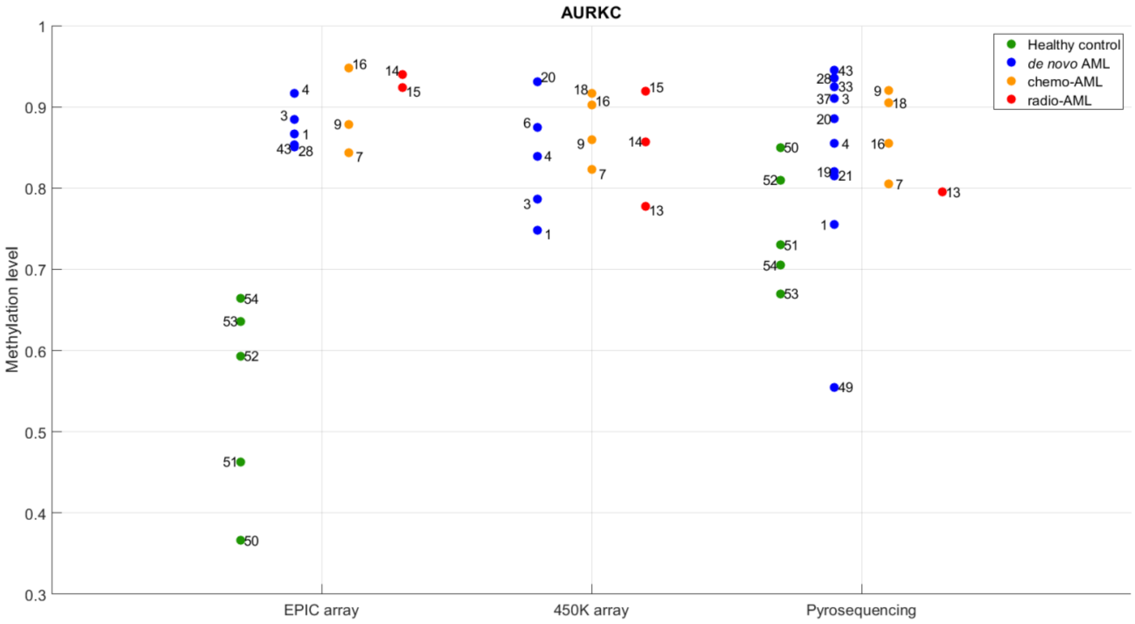

Figure 4.29 Median methylation level for AURKC CpG-rich RS region for each experiment and patient.

Measurements from all experiments are consistent. All AML patients have similar methylation level, except the 49 *de novo* AML patient, which is an outlier in this case. Differences between healthy control and AMLs are greater in the EPIC array (primary) experiment than in pyrosequencing (validation) experiment. AURKC is one of the protein kinases. It has been described as downregulated in AML, independently of gender [143].

Similar visualisation was performed for the MEST gene (Figure 4.30).

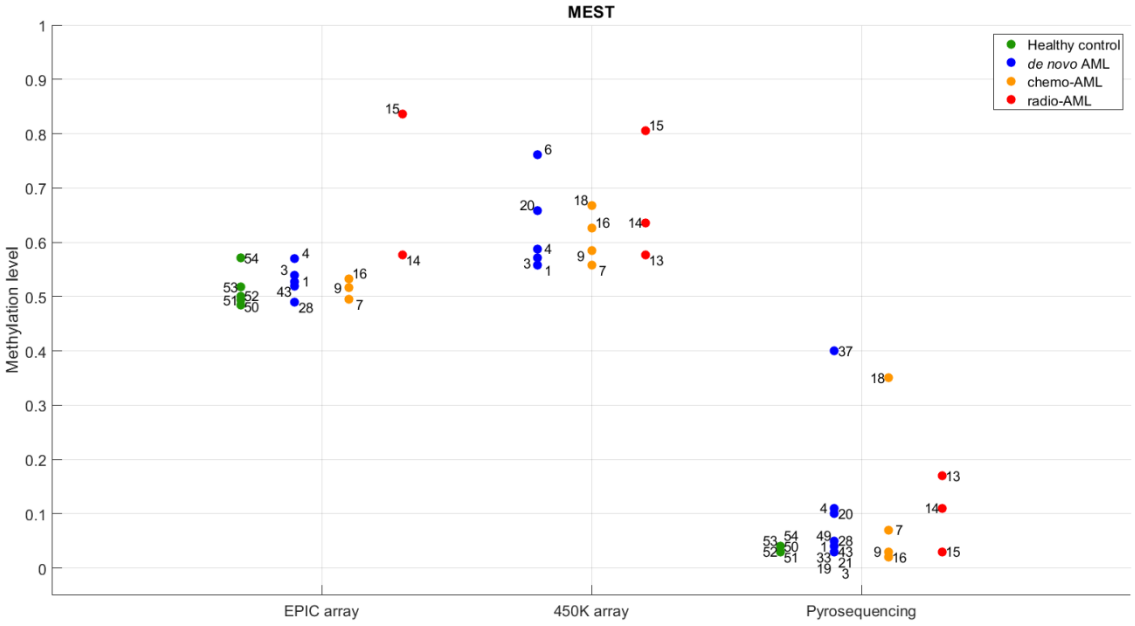

Figure 4.30 Median methylation level for MEST CpG-rich RS region for each experiment and patient.

In the case of the MEST gene, the methylation level measured in the pyrosequencing experiment is lower than in array experiments. Radio-AML patients distinguish from the rest patients, except 37 (*de novo* AML patient) and 18 (chemo-AML patient) in the pyrosequencing experiment. These patients are outliers in this case. In the 450K array experiment, radio-AML patients have similar methylation level as other AMLs patients. MEST is characterised by parental imprinting in lymphocytes. Loss of this imprinting is connected to several types of cancer. MEST is silenced by hypermethylation in AML, independently of the methylation status of the imprinting control region, and can be a tumour suppressor gene [144]. Methylation level at MEST was associated with invasive cervical cancer risk [145]. MEST lower expression is associated with NMP1 mutation in AML and correlated with worse overall survival [146].

Analogous visualisation was prepared for the GATA5 gene (Figure 4.31).

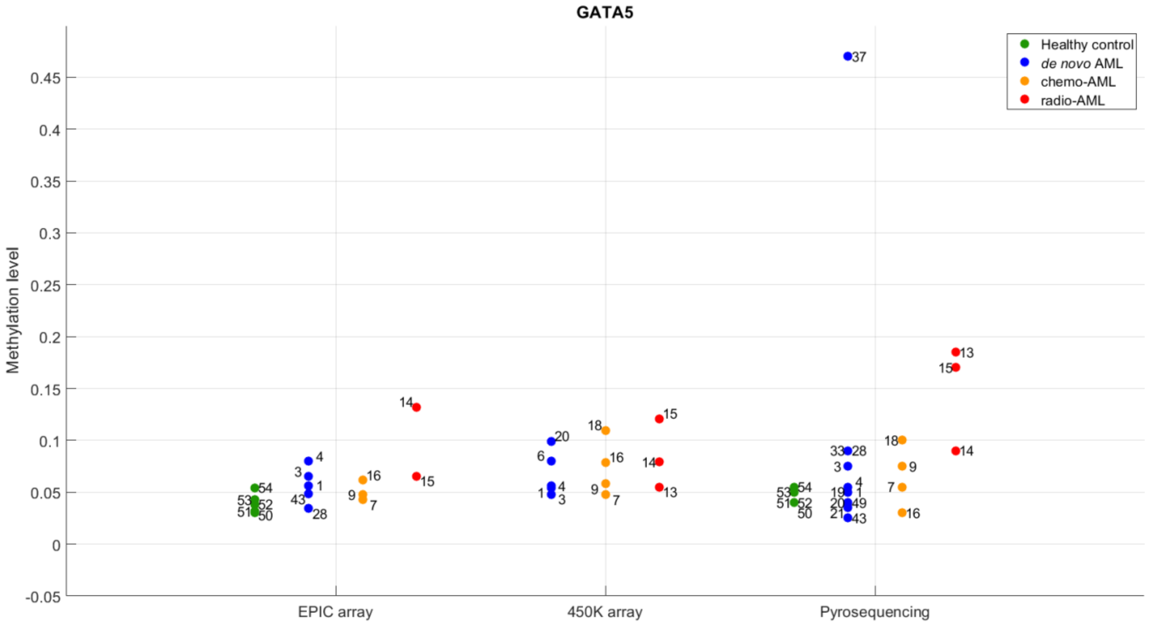

Figure 4.31 Median methylation level for GATA5 CpG-rich RS region for each experiment and patient.

The situation for the GATA5 gene is similar to the MEST gene. Radio-AML patients distinguish well from others in EPIC array and pyrosequencing experiments. In the 450K experiment, all AMLs patients have similar methylation level. GATA5 is a transcription factor, playing a crucial role in cardiovascular development. It is also a tumour suppressor gene. It impacts proliferation and colony formation ability in hepatocellular carcinoma (HCC) [147] [148]. Its expression is silenced in colorectal and lung cancers [149]. GATA5 is also involved in leukaemia inhibitory factor-responsive transcription of the β-myosin heavy chain gene in cardiac myocytes [150]. GATA5 hypermethylation has already been reported in chronic lymphocytic leukaemia (CLL) [151] and colorectal carcinoma [152]. Its hypermethylation is associated with radiation-induced lung adenocarcinoma [153].

Mir886 (VTRNA2-1) gene was not measured on 450K array, so visualisation was prepared only for EPIC array and pyrosequencing experiments. It is presented in Figure 4.32

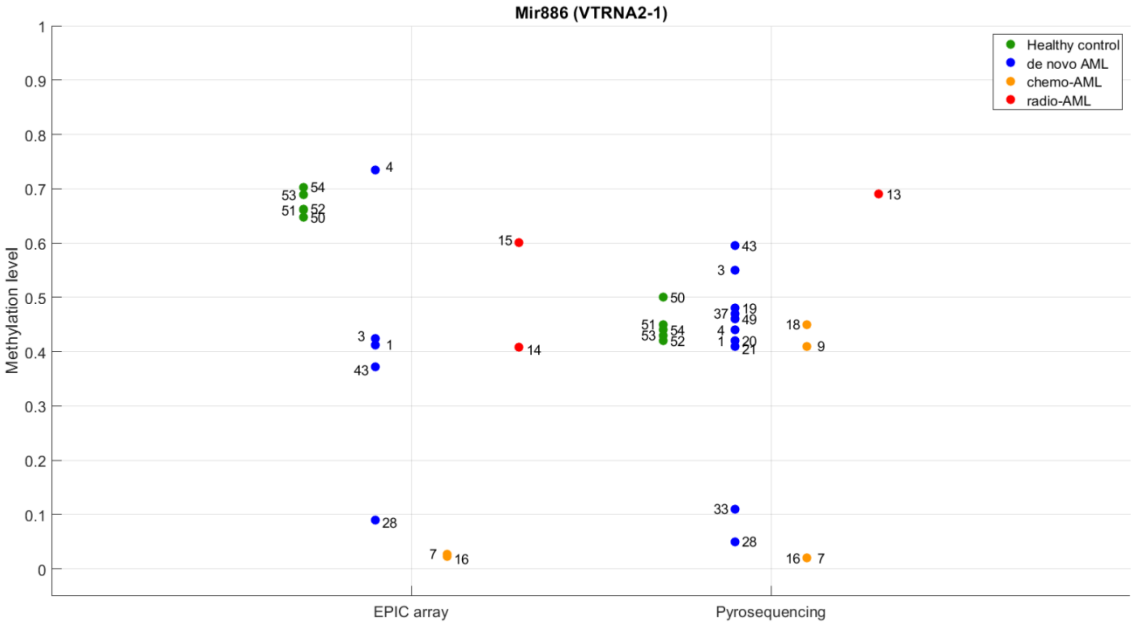

Figure 4.32 Median methylation level for Mir886 (VTRNA2-1) CpG-rich RS region for EPIC array and pyrosequencing experiments and each patient.

In the EPIC array (primary) experiment, chemo-AML patients distinguish well from other patients. In pyrosequencing (validation) experiment, the same patients (7 and 16) also differ from the rest patient groups. However, remaining chemo-AML patients (9 and 18) have similar methylation level to control and *de novo* AML patients. Hence, in this case, a smaller effect size is a result of the inside chemo-AML group variety. Mir886 (VTRNA2-1) is a tumour suppressor gene and outcome predictor in AML - lower methylation of it is related to better survival [154].

4.4. Integration of all experiments results

4.4.1 AML methylation profile

In the first described study (Chapter 4.1) AML methylation profile was analysed. A number of genomic regions that differentiate AML and healthy control was detected according to methylation level. Differentiation was categorised into five levels: hypomethylation, hypermethylation, medium hypermethylation, high hypermethylation, and extreme high hypermethylation. In the third described study (Chapter 4.3), the AML methylation profile was examined for *de novo* AML, chemotherapy-related AML, and radiotherapy-related AML. In this case, several hypo- and hypermethylated genomic regions were also found.

It has been checked if the results for both studies were consistent. TSS genomic regions from the first study are related to the CpG-rich regulatory sequence region from the third study. However, TSS regions include all probes annotated to TSS1500, TSS200, and 5'UTR regions, while CpG-rich RS regions take only those that are also annotated to CpG island or shore. The number of hypo- and hypermethylated genomic regions detected in corresponding patient groups compared to control is presented in Table 4.25.

Table 4.25 The number of differentially methylated TSS/CpG-rich RS genomic regions in AML/*de novo* AML separately and in common.

AML			*De novo* AML		Common
Differential methylation		Number of genomic regions	Differential methylation	Number of genomic regions	Number of genomic regions
Hypomethylation		90	Hypomethylation	40	1
Hypermethylation	At least low	945	Hypermethylation	236	62
	At least medium	385			37
	At least high	105			18
	Extreme high	31			3

In the first study, methylation level was measured for 20,852 TSS genomic regions, while in the third, for 14,338 CpG-rich RS regions. The number of features common for both experiments was 11,774. Dice similarity coefficient [155] for hypomethylated genomic regions equals 0.0154, while for hypermethylated genomic regions, it is 0.1050. Such low value may be the result of different definitions of TSS and CpG-rich RS genomic regions.

One common hypomethylated genomic region was LOC648691. Its integrated *p*-value in the first experiment was $8.43\cdot10^{-9}$ (adjusted significance level: $1.08\cdot10^{-7}$), and in the third experiment, it was $9.06\cdot10^{-9}$ (adjusted significance level: $5.86\cdot10^{-6}$). It was also detected as hypomethylated in chemo-AML; the integrated *p*-value was $9.63\cdot10^{-9}$.

Three genomic regions common for extreme high hypermethylation in the first experiment and hypermethylation in the third experiment were KRTCAP3, MTMR7, and SPACA1. The integrated *p*-value for KRTCAP3 in the first experiment was $2.86 \cdot 10^{-21}$ for extreme high hypermethylation (adjusted significance level: $1.12 \cdot 10^{-13}$), while in the third experiment was equal to $1.15 \cdot 10^{-12}$ (adjusted significance level: $5.86 \cdot 10^{-6}$). It was also hypermethylated in chemo-AML (integrated *p*-value: $3.50 \cdot 10^{-8}$) and in radio-AML (integrated *p*-value: $1.83 \cdot 10^{-8}$). The integrated *p*-value for MTMR7 in the first experiment was $4.18 \cdot 10^{-11}$ (adjusted significance level: $2.86 \cdot 10^{-10}$), while in the third experiment was equal to $1.37 \cdot 10^{-13}$ (adjusted significance level: $7.90 \cdot 10^{-7}$). It was also hypermethylated in radio-AML (integrated *p*-value: $2.44 \cdot 10^{-7}$). The integrated *p*-value for SPACA1 in the first experiment was $1.84 \cdot 10^{-10}$ (adjusted significance level: $7.90 \cdot 10^{-7}$), while in another experiment, it was $1.47 \cdot 10^{-11}$ (adjusted significance level: $5.86 \cdot 10^{-6}$).

None of the mentioned genes has already been reported as related to leukaemia. One of the genes detected as high hypermethylated in the first experiment and hypermethylated in the third experiment was NNAT.

4.4.2 Gender impact on methylation in AML

In the second described study (Chapter 4.2), differences in methylation level between female and male AML patients were detected. Validation for this analysis was performed using the data from the third study (Chapter 4.3): healthy control and *de novo* AML patient groups were taken into account. The three most gender differentiating probes in the AML patient group, according to FDR value, were selected. The difference between genders was also checked in healthy patients. Boxplots of their methylation levels are presented in Figure 4.33. Values from the third validation experiment are marked with red color.

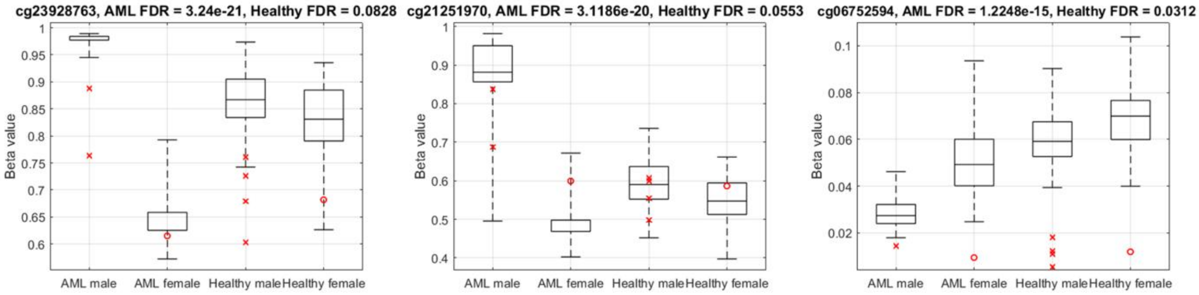

Figure 4.33 Methylation level in selected probes among all examined patient groups - boxplots for the data from the second experiment and individual values for the data from the third experiment.

In the two first probes, the relationship between methylation level in males and females is consistent between the second and the third study - methylation level in females is lower than in male AML patients. In the case of the third probe, the relationship is not preserved. In each case, no difference between healthy males and females is observed.

5. Conclusions

The objective of this dissertation, which was the detection of differentially methylated probes and genomic regions among various patient groups and the integration of the results obtained with the use of different experimental platforms, has been achieved in several aspects.

5.1. Acute myeloid leukaemia methylation profile

A novel method for methylation data analysis was proposed. It facilitates an effective detection of differentially methylated probes and differentially methylated genomic regions. AML genome-wide methylation fingerprint was identified with the use of the developed technique. The algorithm uses selected statistical methods fitted to the characteristics of the data. Additionally, it is supported by mathematical modelling. Contrary to existing approaches, it is data-driven and does not use a priori assumed cut-offs for differential methylation definition. It uses Gaussian mixture modelling of the distribution of methylation shift between groups to detect specific thresholds. They allow classifying probes as low, medium, or high hyper- or hypomethylated with the support of probability for class membership. Due to *p*-value integration, this approach enables a conclusion about differential methylation of genomic regions, such as TSS and gene body for individual genes. The study confirmed that alterations in DNA methylation occur in acute myeloid leukaemia. The AML methylation modification varies for different genomic region types: TSS, gene body, and intergenic. Much more probes and regions were detected as hypermethylated than hypomethylated. The genes in which genomic regions (especially TSS regions) were detected as hypo- or hypermethylated in AML were confirmed as directly connected to leukaemia. Functional analysis revealed the relationship between the found genes and processes alternated in AML.

5.2. Gender differences in DNA methylation in AML

The obtained results reveal differences in methylation profile between males and females in AML. Corresponding differences are not observed in healthy persons. Gender disparity in AML concerns probes in CpG-rich Regulatory Sequence genomic regions. Alterations of DNA methylation in these regions impact gene expression the most. In other genomic region types, differences between genders in AML are insignificant. The integration of *p*-values of probes annotated to the same genomic region shows that almost 10% of genes can be differentially expressed between males and females in AML. These genes are connected to many molecular processes and functions examined in functional analysis. Several enriched GO Terms, such as GO:0006935 (chemotaxis) and GO:0048870 (cell motility), are related to AML development [156]. Additionally, the expression of homeobox genes, found in the functional analysis, is correlated with epigenetic modifiers and specific to malignant hematopoiesis, suggesting their potential causal relationships [157]. Furthermore, survival analysis shows no differences in prognosis between males and females. However, it demonstrates that different prognostic markers can characterise males and females with AML. Prognostic markers - genomic regions in which methylation level significantly impacts survival - were successfully detected for all patients and males. The results obtained for females are on the "by chance" level. Identified gender-specific differences in epigenomics prognostic markers should be considered in the diagnosis and prognosis of AML.

5.3. DNA methylation aberrations in *de novo* and therapy-related AML

The results obtained in this study reveal the difficulty of analysing the data with small samples. The lack of a control sample and then putting the healthy control in one batch could lead to problems with detecting differences between healthy and AML patients. However, this imbalance was limited by preprocessing methods, such as batch effect removal. Methylation profile analysis presents differences in methylation level distribution among genomic region types. The analysis was performed using a composition of statistical methods and an Illumina annotation system, which assign probes to particular genomic regions. Detection of differentially methylated probes and genomic regions, performed with two approaches, leads to consistent results. It shows that methods with corrections for multiple pairwise comparisons are more restrictive. Compared to control, aberrations in DNA methylation level in AMLs are different for probes belonging to various genomic regions. In regulatory sequence regions, hypermethylation is prevalent, while in body regions and 3'UTR regions, hypermethylation and hypomethylation are similar. Inside-group variety analysis reveals that patient groups are various, but it is lowest for *de novo* AML and highest for radio-AML. *P*-value integration allows concluding about differential methylation of individual regulatory sequence regions. The results show which genes can be differentially expressed in *de novo* AML and therapy-related AMLs. Unsupervised feature selection confirms patients' data structure and categorisation as *de novo*, chemo- and radio-AML patients. Functional analysis of genes differentially expressed in *de novo* AML confirms the biological meaning of statistical assay. One AML, one chemo-AML, and two radio-AML markers were detected and validated. Differential methylation of the found markers is associated with AML in the literature. The found chemo- and radio-AML markers can be used as diagnostic factors in distinguishing these two types of AML.

5.4. Integration of all experiments results

The integration of the results from the first and third studies aimed to find differentially methylated features for AML. Only a few common hypo- and hypermethylated genomic regions were obtained. However, the differences in methylation level distributions among genomic region types are similar in both studies. The prevalence of hypermethylation over hypomethylation was confirmed. The low similarity of obtained results can be an effect of different definitions of compared regions (TSS in the first study and CpG-rich RS in the third study) as well as of a small number of examined patients and inside-group variety. The integration of results from the second and third studies confirmed the gender impact on methylation level in most cases. The consistency of observation was evaluated visually.

To summarise, the composition of mathematical modelling, comparative statistical analysis and their results integration enabled the detection of differentially methylated genomic regions between AML patients and healthy donors as well as among several AML patient's groups and healthy control. Integration of results acquired using methylation arrays and pyrosequencing enabled the validation of detected AML epigenomic biomarkers.

Bibliography

[1] U. Deichmann, „Epigenetics: The origins and evolution of a fashionable topic,” *Developmental biology,* tom 416, pp. 249-254, 2016.

[2] C. Waddington, “Towards a Theoretical Biology,” *Nature,* vol. 218, pp. 525-527, 1968.

[3] C.-t. Wu and J. R. Morris, „Genes, Genetics, and Epigenetics: A Correspondence,” *Science,* tom 293, pp. 1103-1105, 2001.

[4] C. Dupont, D. R. Armant and C. A. Brenner, „Epigenetics: definition, mechanisms and clinical perspective.,” *Seminars in reproductive medicine,* tom 27, nr 5, p. 351–357, September 2009.

[5] M. Okano, D. W. Bell, D. A. Haber and E. Li, „DNA Methyltransferases Dnmt3a and Dnmt3b Are Essential for De Novo Methylation and Mammalian Development,” *Cell,* tom 99, pp. 247-257, 1999.

[6] M. Gardiner-Garden and M. Frommer, „CpG Islands in vertebrate genomes,” *Journal of molecular biology,* tom 196, pp. 261-282, 1987.

[7] F. Antequera, “Structure, function and evolution of CpG island promoters,” *Cellular and Molecular Life Sciences CMLS,* vol. 60, pp. 1647-1658, 2003.

[8] R. Illingworth, A. Kerr, D. Desousa, H. Jørgensen, P. Ellis, J. Stalker, D. Jackson, C. Clee, R. Plumb, J. Rogers, S. Humphray, T. Cox, C. Langford and A. Bird, „A novel CpG island set identifies tissue-specific methylation at developmental gene loci.,” *PLoS biology,* tom 6, nr 1, p. e22, January 2008.

[9] M. Chahrour, S. Y. Jung, C. Shaw, X. Zhou, S. T. C. Wong, J. Qin and H. Y. Zoghbi, „MeCP2, a Key Contributor to Neurological Disease, Activates and Represses Transcription,” *Science,* tom 320, p. 1224, May 2008.

[10] R. J. Sims, K. Nishioka and D. Reinberg, „Histone lysine methylation: a signature for chromatin function,” *Trends in genetics : TIG,* tom 19, pp. 629-639, 2003.

[11] T. A. Volpe, C. Kidner, I. M. Hall, G. Teng, S. I. S. Grewal and R. A. Martienssen, „Regulation of Heterochromatic Silencing and Histone H3 Lysine-9 Methylation by RNAi,” *Science,* tom 297, pp. 1833-1837, September 2002.

[12] L. He and G. J. Hannon, „MicroRNAs: small RNAs with a big role in gene regulation,” *Nature Reviews Genetics,* tom 5, pp. 522-531, 2004.

[13] S. Francia, „Non-Coding RNA: Sequence-Specific Guide for Chromatin Modification and DNA Damage Signaling,” *Frontiers in genetics,* tom 6, 2015.

[14] H. Wu and Y. E. Sun, „Epigenetic Regulation of Stem Cell Differentiation,” *Pediatric research,* tom 59, pp. 21R-25R, 2006.

[15] E. R. Gibney and C. M. Nolan, „Epigenetics and gene expression,” *Heredity,* tom 105, pp. 4-13, 2010.

[16] E. Hervouet, F. M. Vallette and P.-F. Cartron, „Dnmt3/transcription factor interactions as crucial players in targeted DNA methylation.,” *Epigenetics,* tom 4, nr 7, p. 487–499, October 2009.

[17] A. Razin, „CpG methylation, chromatin structure and gene silencing—a three-way connection,” The *EMBO journal,* tom 17, pp. 4905-4908, 1998.

[18] C. D. Malone and G. J. Hannon, „Small RNAs as Guardians of the Genome,” *Cell,* tom 136, pp. 656-668, 2009.

[19] I. Kovalchuk, “Genome Stability: An Evolutionary Perspective,” in *Genome Stability,* Academic Press, 2016, pp. 1-18.

[20] J. C. Peng and G. H. Karpen, „Heterochromatic genome stability requires regulators of histone H3 K9 methylation.,” *PLoS Genetics,* tom 5, p. e1000435, March 2009.

[21] P. Dominguez-Salas, S. E. Moore, D. Cole, K.-A. da Costa, S. E. Cox, R. A. Dyer, A. J. C. Fulford, S. M. Innis, R. A. Waterland, S. H. Zeisel, A. M. Prentice and B. J. Hennig, „DNA methylation potential: dietary intake and blood concentrations of one-carbon metabolites and cofactors in rural African women.,” The *American journal of clinical nutrition,* tom 97, nr 6, p. 1217–1227, June 2013.

[22] D. Kim, L. D. Kubzansky, A. Baccarelli, D. Sparrow, A. Spiro, L. Tarantini, L. Cantone, P. Vokonas and J. Schwartz, „Psychological factors and DNA methylation of genes related to immune/inflammatory system markers: the VA Normative Aging Study.,” *BMJ open,* tom 6, nr 1, p. e009790, January 2016.

[23] E. Unternaehrer, P. Luers, J. Mill, E. Dempster, A. H. Meyer, S. Staehli, R. Lieb, D. H. Hellhammer and G. Meinlschmidt, „Dynamic changes in DNA methylation of stress-associated genes (OXTR, BDNF) after acute psychosocial stress,” *Translational psychiatry,* tom 2, pp. e150-e150, 2012.

[24] M. Ehrlich, „DNA methylation in cancer: too much, but also too little,” *Oncogene,* tom 21, pp. 5400-5413, 2002.

[25] M. Kulis and M. Esteller, *DNA Methylation and Cancer,* 2010, pp. 27-56.

[26] Y. Cheng, C. He, M. Wang, X. Ma, F. Mo, S. Yang, J. Han and X. Wei, „Targeting epigenetic regulators for cancer therapy: mechanisms and advances in clinical trials.,” *Signal transduction and targeted therapy,* tom 4, p. 62, 2019.

[27] H. Döhner, D. J. Weisdorf and C. D. Bloomfield, „Acute Myeloid Leukemia,” *England journal of medicine,* tom 373, pp. 1136-1152, 2015.

[28] W. Blum and C. D. Bloomfield, „Acute Myeloid Leukemia,” w *Harrison's Principles of Internal Medicine, 20e.,* McGraw Hill, 2018, pp. 743-745.

[29] V. Hoffbrand and P. Moss, „Acute myeloid leukemia,” w *Hoffbrand's essential hematology,* Chichester, John Wiley & Sons, 2016, pp. 146-149.

[30] J. Liesveld and M. Lichtman, „Acute Myelogenous Leukemia,” w *Williams Hematology, 9e,* McGraw-Hill Education, 2016.

[31] I. Seferyńska, „Epidemiologia zachorowań na ostre białaczki u ludzi dorosłych w Polsce w latach 2004–2006,” *Postępy Nauk Medycznych,* 2007.

[32] S. Puumala, J. Ross, R. Aplenc and L. Spector, “Epidemiology of Childhood Acute Myeloid Leukemia,” *Pediatr Blood Cancer,* vol. 60(5), pp. 728-733, 2014.

[33] Greenlee, R. T., Hill-Harmon, M. B., Murray, T., and Thun, M., “Cancer statistics, 2001.,” *CA: a cancer journal for clinicians,* vol. 51(1), p. 15–36, 2001.

[34] M. J. Thirman and R. A. Larson, „Therapy-related myeloid leukemia,” *Hematology/Oncology Clinics,* tom 10, pp. 293-320, 1996.

[35] A. Khalade, M. S. Jaakkola, E. Pukkala and J. J. K. Jaakkola, „Exposure to benzene at work and the risk of leukemia: a systematic review and meta-analysis.,” *Environmental health : a global access science source,* tom 9, p. 31, June 2010.

[36] B. Deschler and M. Lübbert, *Acute Myeloid Leukemia: Epidemiology and Etiology,* 2008, pp. 47-56.

[37] V. Visconte, R. V. Tiu and H. J. Rogers, „Pathogenesis of myelodysplastic syndromes: an overview of molecular and non-molecular aspects of the disease.,” *Blood research,* tom 49, nr 4, p. 216–227, December 2014.

[38] A. C. Xavier, Y. Ge and J. W. Taub, „Down syndrome and malignancies: a unique clinical relationship: a paper from the 2008 william beaumont hospital symposium on molecular pathology.," The *Journal of molecular diagnostics : JMD,* tom 11, nr 5, p. 371–380, September 2009.

[39] J. M. Bennett, D. Catovsky, M. T. Daniel, G. Flandrin, D. A. Galton, H. R. Gralnick and C. Sultan, „Proposals for the classification of the acute leukaemias. French-American-British (FAB) co-operative group.," *British journal of haematology,* tom 33, nr 4, p. 451–458, August 1976.

[40] J. W. Vardiman, J. Thiele, D. A. Arber, R. D. Brunning, M. J. Borowitz, A. Porwit, N. L. Harris, M. M. Le Beau, E. Hellström-Lindberg, A. Tefferi and C. D. Bloomfield, „The 2008 revision of the World Health Organization (WHO) classification of myeloid neoplasms and acute leukemia: rationale and important changes.," *Blood,* tom 114, nr 5, p. 937–951, July 2009.

[41] P. Fenaux, "Acute promyelocytic leukemia: biology and treatment," *Seminars in Oncology,* vol. 24, pp. 92-102, 1997.

[42] M. F. Fey and C. Buske, „Acute myeloblastic leukaemias in adult patients: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up," *Ann Oncol,* tom 24, pp. vi138-vi143, 2013.

[43] H. Döhner, E. Estey, S. Amador and F. Appelbaum, "Diagnosis and management of acute myeloid leukemia in adults: recommendations from an international expert panel, on behalf of the European LeukemiaNet," *Blood,* vol. 115 (3), p. 453–474, 2010.

[44] A. K. Burnett, D. Grimwade, E. Solomon, K. Wheatley and A. H. Goldstone, „Presenting white blood cell count and kinetics of molecular remission predict prognosis in acute promyelocytic leukemia treated with all-trans retinoic acid: result of the Randomized MRC Trial.," *Blood,* tom 93, nr 12, p. 4131–4143, June 1999.

[45] J. M. Rowe, „Will new agents impact survival in AML?," *Best practice & research. Clinical haematology,* tom 32, nr 4, p. 101094, December 2019.

[46] „Leukemia - Acute Myeloid - AML: Statistics," [Online]. Available: https://www.cancer.net/cancer-types/leukemia-acute-myeloid-aml/statistics. [Available at: 01.07. 2022].

[47] T. Schoofs, W. E. Berdel and C. Müller-Tidow, „Origins of aberrant DNA methylation in acute myeloid leukemia," *Leukemia,* tom 28, pp. 1-14, 2014.

[48] X. Yang, M. P. M. Wong and R. K. Ng, „Aberrant DNA Methylation in Acute Myeloid Leukemia and Its Clinical Implications," *International Journal of Molecular Sciences,* tom 20, p. 4576, September 2019.

[49] S. Kayser, K. Döhner, J. Krauter, C.-H. Köhne, H. A. Horst, G. Held, M. von Lilienfeld-Toal, S. Wilhelm, A. Kündgen, K. Götze, M. Rummel, D. Nachbaur, B. Schlegelberger, G. Göhring, D. Späth, C. Morlok, M. Zucknick, A. Ganser, H. Döhner, R. F. Schlenk and G.-A. AMLSG, „The impact of therapy-related acute myeloid leukemia (AML) on outcome in 2853 adult patients with newly diagnosed AML.," *Blood,* tom 117, nr 7, p. 2137–2145, February 2011.

[50] W. S. A. Allan, „ACUTE MYELOID LEUKAEMIA AFTER TREATMENT WITH CYTOSTATIC AGENTS," *Lancet,* tom 296, p. 775, 1970.

[51] C. G. S. Smit and L. Meyler, „Acute myeloid leukaemia after treatment with cytostatic agents," *Lancet,* tom 296, pp. 671-672, 1970.

[52] L. A. Godley and R. A. Larson, „Therapy-Related Myeloid Leukemia," *Semin Oncol.,* tom 35, pp. 418-429, 2008.

[53] S. D. Michels, R. W. McKenna, D. C. Arthur and R. D. Brunning, „Therapy-related acute myeloid leukemia and myelodysplastic syndrome: a clinical and morphologic study of 65 cases.," *Blood,* tom 65, nr 6, p. 1364–1372, June 1985.

[54] E. Uehara, S. Takeuchi, T. Tasaka, Y. Matsuhashi, Y. Yang, M. Fujita, T. Tamura, M. Nagai and H. P. Koeffler, „Aberrant methylation in promoter-associated CpG islands of multiple genes in therapy-related leukemia.," *International journal of oncology,* tom 23, nr 3, p. 693–696, September 2003.

[55] D. H. Christiansen, M. K. Andersen and J. Pedersen-Bjergaard, „Methylation of p15INK4B is common, is associated with deletion of genes on chromosome arm 7q and predicts a poor prognosis in therapy-related myelodysplasia and acute myeloid leukemia.," *Leukemia,* tom 17, nr 9, p. 1813–1819, September 2003.

[56] W. Y. Au, A. Fung, C. Man, S. K. Ma, T. S. Wan, R. Liang and Y. L. Kwong, „Aberrant p15 gene promoter methylation in therapy-related myelodysplastic syndrome and acute myeloid leukaemia: clinicopathological and karyotypic associations.," *British journal of haematology,* tom 120, nr 6, p. 1062–1065, March 2003.

[57] M. T. Voso, F. D'Alò, M. Greco, E. Fabiani, M. Criscuolo, G. Migliara, L. Pagano, L. Fianchi, F. Guidi, S. Hohaus and G. Leone, „Epigenetic changes in therapy-related MDS/AML.," *Chemico-biological interactions,* tom 184, nr 1-2, p. 46–49, March 2010.

[58] K. C. Kuo, R. A. McCune, C. W. Gehrke, R. Midgett and M. Ehrlich, „Quantitative reversed-phase high performance liquid chromatographic determination of major and modified deoxyribonucleosides in DNA.," *Nucleic acids research,* tom 8, nr 20, p. 4763–4776, October 1980.

[59] M. F. Fraga, E. Uriol, L. Borja Diego, M. Berdasco, M. Esteller, M. J. Cañal and R. Rodríguez, „High-performance capillary electrophoretic method for the quantification of 5-methyl 2'-deoxycytidine in genomic DNA: application to plant, animal and human cancer tissues.," *Electrophoresis,* tom 23, nr 11, p. 1677–1681, June 2002.

[60] R. d. Gaudio, R. D. Giaimo and G. Geraci, „Genome methylation of the marine annelid worm *Chaetopterus variopedatus*: methylation of a CpG in an expressed H1 histone gene," *FEBS letters,* tom 417, pp. 48-52, 1997.

[61] F. Schmitt, E. J. Oakeley and J. P. Jost, „Antibiotics Induce Genome-wide Hypermethylation in Cultured Nicotiana tabacum Plants," The *Journal of biological chemistry,* tom 272, pp. 1534-1540, 1997.

[62] J. Wu, J.-P. Issa, J. Herman, J. Bassett, B. D. Nelkin and S. B. Baylin, „Expression of an exogenous eukaryotic DNA methyltransferase gene induces transformation of NIH 3T3 cells.," *Proceedings of the National Academy of Science,* tom 90, pp. 8891-8895, October 1993.

[63] E. J. Oakeley, F. Schmitt and J. P. Jost, „Quantification of 5-methylcytosine in DNA by the chloroacetaldehyde reaction.," *BioTechniques,* tom 27, nr 4, pp. 744–6, 748-50, 752, October 1999.

[64] E. J. Oakeley, A. Podestà and J.-P. Jost, „Developmental Changes in DNA Methylation of the Two Tobacco Pollen Nuclei during Maturation," *Proceedings of the National Academy of Science,* tom 94, pp. 11721-11725, October 1997.

[65] C. Dahl and P. Guldberg, "DNA methylation analysis techniques," *Biogerontology,* vol. 4, pp. 233-250, 2003.

[66] L. J. Rush and C. Plass, „Restriction landmark genomic scanning for DNA methylation in cancer: past, present, and future applications.," *Analytical biochemistry,* tom 307, nr 2, p. 191–201, August 2002.

[67] M. Frommer, L. E. McDonald, D. S. Millar, C. M. Collis, F. Watt, G. W. Grigg, P. L. Molloy and C. L. Paul, „A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands.," *Proceedings of the National Academy of Science,* tom 89, pp. 1827-1831, March 1992.

[68] M. L. Gonzalgo and P. A. Jones, „Rapid quantitation of methylation differences at specific sites using methylation-sensitive single nucleotide primer extension (Ms-SNuPE)," *Nucleic Acids Res.,* tom 25, pp. 2529-2531, 1997.

[69] J. Tost and I. G. Gut, „DNA methylation analysis by pyrosequencing," *Nature Protocols,* tom 2, pp. 2265-2275, 2007.

[70] C. Delaney, S. K. Garg and R. Yung, *Analysis of DNA Methylation by Pyrosequencing,* 2015, pp. 249-264.

[71] A. Meissner, A. Gnirke, G. W. Bell, B. Ramsahoye, E. S. Lander and R. Jaenisch, „Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis.," *Nucleic acids research,* tom 33, nr 18, p. 5868–5877, 2005.

[72] C. A. Eads, K. Kawakami, L. B. Saltz, C. Blake, D. Shibata, P. V. Danenberg and P. W. Laird, "MethyLight: a high-throughput assay to measure DNA methylation," *Nucleic Acids Res.*, vol. 28, 2000.

[73] T. Bianco, D. Hussey and A. Dobrovic, „Methylation-sensitive, single-strand conformation analysis (MS-SSCA): A rapid method to screen for and analyze methylation.," *Human mutation,* tom 14, nr 4, p. 289–293, 1999.

[74] Aggerholm, A., Guldberg, P., Hokland, M., and Hokland, P, "Extensive intra- and interindividual heterogeneity of p15INK4B methylation in acute myeloid leukemia," *Cancer Res.,* vol. 59, p. 436–441, 1999.

[75] W. Xiao and P. Oefner, "Denaturing high-performance liquid chromatography: A review," *Human mutation,* vol. 17, pp. 439-474, 2001.

[76] Dedeurwaerder, S., Defrance, M., Calonne, E., Denis, H., Sotiriou, C., and Fuks, F., "Evaluation of the Infinium Methylation 450K technology.," *Epigenomics,* vol. 3(6), p. 771–784, 2011.

[77] M. Bibikova, J. Le, B. Barnes, S. Saedinia-Melnyk, L. Zhou, R. Shen and K. L. Gunderson, „Genome-wide DNA methylation profiling using Infinium® assay," *Epigenomics,* tom 1, pp. 177-200, 2009.

[78] S. Moran, C. Arribas and M. Esteller, „Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences.," *Epigenomics,* tom 8, nr 3, p. 389–399, March 2016.

[79] F. J. Steemers, W. Chang, G. Lee, D. L. Barker, R. Shen and K. L. Gunderson, „Whole-genome genotyping with the single-base extension assay," *Nature methods,* tom 3, pp. 31-33, 2006.

[80] J. Sandoval, H. Heyn, S. Moran, J. Serra-Musach, M. A. Pujana, M. Bibikova and M. Esteller, „Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome.," *Epigenetics,* tom 6, nr 6, p. 692–702, June 2011.

[81] L. M. Spindola, M. L. Santoro, P. M. Pan, V. K. Ota, G. Xavier, C. M. Carvalho, F. Talarico, P. Sleiman, M. March, R. Pellegrino, E. Brietzke, R. Grassi-Oliveira, J. J. Mari, A. Gadelha, E. C. Miguel, L. A. Rohde, R. A. Bressan, D. R. Mazzotti, J. R. Sato, G. A. Salum, H. Hakonarson and S. I. Belangero, „Detecting multiple differentially methylated CpG sites and regions related to dimensional psychopathology in youths.," *Clinical epigenetics,* tom 11, nr 1, p. 146, October 2019.

[82] P. Yousefi, K. Huen, R. Aguilar Schall, A. Decker, E. Elboudwarej, H. Quach, L. Barcellos and N. Holland, „Considerations for normalization of DNA methylation data by Illumina 450K BeadChip assay in population studies.," *Epigenetics,* tom 8, nr 11, p. 1141–1152, November 2013.

[83] A. Kibbe Warren, J. Nadereh, H. Chiang-Ching, Z. Xiao, D. Pan, H. Lifang and M. Lin Simon, „Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis," *BMC Bioinformatics,* tom 11, p. 587, November 2010.

[84] A. E. Teschendorff, F. Marabita, M. Lechner, T. Bartlett, J. Tegner, D. Gomez-Cabrero and S. Beck, „A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data.," *Bioinformatics (Oxford, England),* tom 29, nr 2, p. 189–196, January 2013.

[85] J. Maksimovic, L. Gordon and A. Oshlack, „SWAN: Subset-quantile Within Array Normalization for Illumina Infinium HumanMethylation450 BeadChips," *Genome Biology,* tom 13, p. R44, 2012.

[86] J.-P. Fortin, A. Labbe, M. Lemire, B. W. Zanke, T. J. Hudson, E. J. Fertig, C. M. Greenwood and K. D. Hansen, „Functional normalization of 450k methylation array data improves replication in large cancer studies.," *Genome biology,* tom 15, nr 12, p. 503, December 2014.

[87] D. Wang, L. Yan, Q. Hu, L. E. Sucheston, M. J. Higgins, C. B. Ambrosone, C. S. Johnson, D. J. Smiraglia and S. Liu, „IMA: an R package for high-throughput analysis of Illumina's 450K Infinium methylation data," *Bioinformatics,* tom 28, pp. 729-730, 2012.

[88] A. E. Jaffe, P. Murakami, H. Lee, J. T. Leek, M. D. Fallin, A. P. Feinberg and R. A. Irizarry, „Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies.,” *International journal of epidemiology,* tom 41, nr 1, p. 200–209, February 2012.

[89] L. M. Butcher and S. Beck, „Probe Lasso: a novel method to rope in differentially methylated regions with 450K DNA methylation data.,” *Methods (San Diego, Calif.),* tom 72, p. 21–28, January 2015.

[90] Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A., and Williams Jr, R. M., „The american soldier: Adjustment during army life.,” *Studies in social psychology in World War II,* tom 1, 1949.

[91] T. J. Peters, M. J. Buckley, A. L. Statham, R. Pidsley, K. Samaras, R. V Lord, S. J. Clark and P. L. Molloy, „De novo identification of differentially methylated regions in the human genome.,” *Epigenetics & chromatin,* tom 8, p. 6, 2015.

[92] Y. Benjamini and Y. Hochberg, „Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing,” *Journal of the Royal Statistical Society,* tom 57, pp. 289-300, 1995.

[93] R. Edgar, M. Domrachev and A. E. Lash, „Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.,” *Nucleic acids research,* tom 30, nr 1, p. 207–210, January 2002.

[94] N. Jung, B. Dai, A. J. Gentles, R. Majeti and A. P. Feinberg, „An LSC epigenetic signature is largely mutation independent and implicates the HOXA cluster in AML pathogenesis,” *Nature Communications,* tom 6, p. 8489, October 2015.

[95] M. J. Aryee, A. E. Jaffe, H. Corrada-Bravo, C. Ladd-Acosta, A. P. Feinberg, K. D. Hansen and R. A. Irizarry, „Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays.,” *Bioinformatics (Oxford, England),* tom 30, nr 10, p. 1363–1369, May 2014.

[96] R. R. Barton and L. W. Schruben, *Uniform and bootstrap resampling of empirical distributions,* 1993.

[97] O. I. Hedges LV, Statistical methods for meta-analysis, Orlando: Academic Press, 1985.

[98] H. Cramér, Mathematical methods of statistics., Princeton: Princeton University Press, 1946.

[99] J. L. j. Hodges and E. L. Lehmann, "Estimates of location based on rank tests," *Annals of Mathematical Statistics,* vol. 34, p. 598–611, 1963.

[100] P. D. McLachlan G, Finite mixture models., New York: Wiley, 2004.

[101] A. Polanski, M. Marczyk, M. Pietrowska, P. Widlak and J. Polanska, „Initializing the EM Algorithm for Univariate Gaussian, Multi-Component, Heteroscedastic Mixture Models by Dynamic Programming Partitions,” *International Journal of Computational Methods,* tom 15, p. 1850012, 2018.

[102] H. N. Claeskens G, Model selection and model averaging, Cambridge University Press: Cambridge, 2008.

[103] F. Wilcoxon, „Individual Comparisons of Grouped Data by Ranking Methods,” *Biometrics Bulletin,* tom 39, p. 80–83, 1945.

[104] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *Annals of Mathematical Statistics,* vol. 18, p. 50–60, 1947.

[105] J. D. Storey, "A direct approach to false discovery rates," *Journal of the Royal Statistical Society. Series B. Statistical Methodology,* vol. 64, p. 479–498, 2002.

[106] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A .P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. and Sherlock, G., “Gene ontology: tool for the unification of biology,” *Nature Genetics,* vol. 25 (1), pp. 25-29, 2000.

[107] R. J. Alexa A, “topGO: enrichment analysis for gene ontology.,” *R package version 2.28.0,* 2016.

[108] K. Pearson, "On the criterion, that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling.," The *London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, V. Series,* vol. 50, p. 157–175, 1900.

[109] K. Pearson, "On lines and planes of closest fit to systems of points in space.," The *London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, VI. Series,* vol. 2, p. 559–572, 1901.

[110] R. D. Martin and V. J. Yohai, *Bias robust estimation of autoregression parameters,* 1991.

[111] H. W. Lilliefors, „On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown,” *Journal of the American Statistical Association,* tom 62, pp. 399-402, 1967.

[112] E. E. Cureton, "Rank-biserial correlation," *Psychometrika,* vol. 21, p. 287–290, 1956.

[113] C. v. Mering, „STRING: a database of predicted functional associations between proteins," *Nucleic Acids Res.,* tom 31, pp. 258-261, 2003.

[114] R. Peto and J. Peto, *Asymptotically efficient rank invariant test procedures,* tom 2, Oxford: Blackwell, 1972, pp. 185-207.

[115] D. R. Cox and D. Oakes, *Analysis of Survival Data,* 2018.

[116] W. Huber, V. J. Carey, R. Gentleman, S. Anders, M. Carlson, B. S. Carvalho, H. C. Bravo, S. Davis, L. Gatto, T. Girke, R. Gottardo, F. Hahne, K. D. Hansen, R. A. Irizarry, M. Lawrence, M. I. Love, J. MacDonald, V. Obenchain, A. K. Oleś, H. Pagès, A. Reyes, P. Shannon, G. K. Smyth, D. Tenenbaum, L. Waldron and M. Morgan, „Orchestrating high-throughput genomic analysis with Bioconductor," *Nature methods,* tom 12, pp. 115-121, 2015.

[117] Y. Tian, T. J. Morris, A. P. Webster, Z. Yang, S. Beck, A. Feber and A. E. Teschendorff, „ChAMP: updated methylation analysis pipeline for Illumina BeadChips," *Bioinformatics,* tom 33, pp. 3982-3984, 2017.

[118] T. J. Morris, L. M. Butcher, A. Feber, A. E. Teschendorff, A. R. Chakravarthy, T. K. Wojdacz and S. Beck, „ChAMP: 450k Chip Analysis Methylation Pipeline," *Bioinformatics,* tom 30, pp. 428-430, 2014.

[119] W. E. Johnson, C. Li and A. Rabinovic, „Adjusting batch effects in microarray expression data using empirical Bayes methods.," *Biostatistics (Oxford, England),* tom 8, nr 1, p. 118–127, January 2007.

[120] M. Jeanmougin, A. de Reynies, L. Marisa, C. Paccard, G. Nuel and M. Guedj, „Should we abandon the t-test in the analysis of gene expression microarray data: a comparison of variance modeling strategies.," *PloS one,* tom 5, nr 9, p. e12336, September 2010.

[121] G. K. Smyth, *limma: Linear Models for Microarray Data,* pp. 397-420.

[122] J. W. Tukey, „Comparing Individual Means in the Analysis of Variance," *Biometrics,* tom 5, p. 99, 1949.

[123] E. C. Pielou, „The measurement of diversity in different types of biological collections," *Journal of Theoretical Biology,* tom 13, pp. 131-144, January 1966.

[124] M. Marczyk, R. Jaksik, A. Polanski and J. Polanska, „Adaptive filtering of microarray gene expression data based on Gaussian mixture decomposition.,” *BMC bioinformatics,* tom 14, p. 101, March 2013.

[125] M. R. Hidalgo, C. Cubuk, A. Amadoz, F. Salavert, J. Carbonell-Caballero and J. Dopazo, „High throughput estimation of functional cell activities reveals disease mechanisms and predicts relevant clinical outcomes,” *Oncotarget,* tom 8, pp. 5160-5178, 2017.

[126] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences.*

[127] D. Klumpp, M. Misovic, K. Szteyn, E. Shumilina, J. Rudner and S. M. Huber, „Targeting TRPM2 Channels Impairs Radiation-Induced Cell Cycle Arrest and Fosters Cell Death of T Cell Leukemia Cells in a Bcl-2-Dependent Manner.,” *Oxidative medicine and cellular longevity,* tom 2016, p. 8026702, 2016.

[128] S. Maegawa, S. M. Gough, N. Watanabe-Okochi, Y. Lu, N. Zhang, R. J. Castoro, M. R. H. Estecio, J. Jelinek, S. Liang, T. Kitamura, P. D. Aplan and J.-P. J. Issa, „Age-related epigenetic drift in the pathogenesis of MDS and AML,” *Genome Res.,* tom 24, pp. 580-591, 2014.

[129] I. Laverdière, M. Boileau, T. Herold, J. Rak, W. E. Berdel, B. Wörmann, W. Hiddemann, K. Spiekermann, S. K. Bohlander and K. Eppert, „Complement cascade gene expression defines novel prognostic subgroups of acute myeloid leukemia.,” *Experimental hematology,* tom 44, nr 11, p. 1039–1043.e10, November 2016.

[130] S. J. Kuerbitz, J. Pahys, A. Wilson, N. Compitello and T. A. Gray, „Hypermethylation of the imprinted NNAT locus occurs frequently in pediatric acute leukemia.,” *Carcinogenesis,* tom 23, nr 4, p. 559–564, April 2002.

[131] S. C. Lueck, A. C. Russ, U. Botzenhardt, R. F. Schlenk, K. Zobel, K. Deshayes, D. Vucic, H. Döhner, K. Döhner, S. Fulda and L. Bullinger, „Smac mimetic induces cell death in a large proportion of primary acute myeloid leukemia samples, which correlates with defined molecular markers,” *Oncotarget,* tom 7, pp. 49539-49551, 2016.

[132] J. R. Engler, A. Frede, V. A. Saunders, A. C. W. Zannettino, T. P. Hughes and D. L. White, „Chronic myeloid leukemia CD34+ cells have reduced uptake of imatinib due to low OCT-1 activity.,” *Leukemia,* tom 24, nr 4, p. 765–770, April 2010.

[133] M. Toyota, K. J. Kopecky, M.-O. Toyota, K.-W. Jair, C. L. Willman and J.-P. J. Issa, „Methylation profiling in acute myeloid leukemia,” *Blood,* tom 97, pp. 2823-2829, 2001.

[134] M. H. Saied, J. Marzec, S. Khalid, P. Smith, T. A. Down, V. K. Rakyan, G. Molloy, M. Raghavan, S. Debernardi and B. D. Young, „Genome wide analysis of acute myeloid leukemia reveal leukemia specific methylome and subtype specific hypomethylation of repeats.," *PloS one,* tom 7, nr 3, p. e33213, 2012.

[135] Y. K. Chae, A. Dimou, S. Pierce, H. Kantarjian and M. Andreeff, „The effect of calcium channel blockers on the outcome of acute myeloid leukemia.," *Leukemia & lymphoma,* tom 55, nr 12, p. 2822–2829, December 2014.

[136] A. Rambaldi, M. Torcia, S. Bettoni, E. Vannier, T. Barbui, A. R. Shaw, C. A. Dinarello and F. Cozzolino, „Modulation of cell proliferation and cytokine production in acute myeloblastic leukemia by interleukin-1 receptor antagonist and lack of its expression by leukemic cells," *Blood,* tom 78, pp. 3248-3253, 1991.

[137] S.-G. Rota, A. Roma, I. Dude, C. Ma, R. Stevens, J. MacEachern, J. Graczyk, S. M. G. Espiritu, P. N. Rao, M. D. Minden, E. Kreinin, D. A. Hess, A. C. Doxey and P. A. Spagnuolo, „Estrogen Receptor β Is a Novel Target in Acute Myeloid Leukemia.," *Molecular cancer therapeutics,* tom 16, nr 11, p. 2618–2626, November 2017.

[138] S. M. Garrido, F. R. Appelbaum, C. L. Willman and D. E. Banker, „Acute myeloid leukemia cells are protected from spontaneous and drug-induced apoptosis by direct contact with a human bone marrow stromal cell line (HS-5).," *Experimental hematology,* tom 29, nr 4, p. 448–457, April 2001.

[139] L. G. Kondratyeva, A. A. Sveshnikova, E. V. Grankina, I. P. Chernov, M. R. Kopantseva, E. P. Kopantzev and E. D. Sverdlov, „Downregulation of expression of mater genes SOX9, FOXA2, and GATA4 in pancreatic cancer cells stimulated with TGFβ1 epithelial-mesenchymal transition.," *Doklady. Biochemistry and biophysics,* tom 469, nr 1, p. 257–259, July 2016.

[140] J. I. López, J. C. Angulo, A. Martín, M. Sánchez-Chapado, A. González-Corpas, B. Colás and S. Ropero, „A DNA hypermethylation profile reveals new potential biomarkers for the evaluation of prognosis in urothelial bladder cancer.," *APMIS : acta pathologica, microbiologica, et immunologica Scandinavica,* tom 125, nr 9, p. 787–796, September 2017.

[141] S. Fält, M. Merup, G. Gahrton, B. Lambert and A. Wennborg, „Identification of progression markers in B-CLL by gene expression profiling.," *Experimental hematology,* tom 33, nr 8, p. 883–893, August 2005.

[142] Z. Li, R. T. Luo, S. Mi, M. Sun, P. Chen, J. Bao, M. B. Neilly, N. Jayathilaka, D. S. Johnson, L. Wang, C. Lavau, Y. Zhang, C. Tseng, X. Zhang, J. Wang, J. Yu, H. Yang, S. M. Wang, J. D. Rowley, J. Chen and M. J. Thirman, „Consistent deregulation of gene expression between human and murine MLL rearrangement leukemias.," *Cancer research,* tom 69, nr 3, p. 1109–1116, February 2009.

[143] F. Yazarloo, R. Shirkoohi, M. B. Mobasheri, A. Emami and M. H. Modarressi, „Expression analysis of four testis-specific genes AURKC, OIP5, PIWIL2 and TAF7L in acute myeloid leukemia: a gender-dependent expression pattern.," *Medical oncology (Northwood, London, England),* tom 30, nr 1, p. 368, March 2013.

[144] A. R. Poetsch, R. Claus, L. Bullinger, T. Witte, M. Lübbert, M. Rehli, K. Döhner and C. Plass, „Genetic and Epigenetic Silencing of Mesoderm Specific Transcript (MEST) In Acute Myelogenous Leukemia.," *Blood,* tom 116, pp. 3639-3639, 2010.

[145] A. C. Vidal, N. M. Henry, S. K. Murphy, O. Oneko, M. Nye, J. A. Bartlett, F. Overcash, Z. Huang, F. Wang, P. Mlay, J. Obure, J. Smith, B. Vasquez, B. Swai, B. Hernandez and C. Hoyo, „PEG1/MEST and IGF2 DNA methylation in CIN and in cervical cancer.," *Clinical & translational oncology : official publication of the Federation of Spanish Oncology Societies and of the National Cancer Institute of Mexico,* tom 16, nr 3, p. 266–272, March 2014.

[146] Á. Nagy, Á. Ősz, J. Budczies, S. Krizsán, G. Szombath, J. Demeter, C. Bödör and B. Győrffy, „Elevated HOX gene expression in acute myeloid leukemia is associated with NPM1 mutations and poor survival.," *Journal of advanced research,* tom 20, p. 105–116, November 2019.

[147] L. Xia, Y. Gong, A. Zhang, S. Cai and Q. Zeng, „Loss of GATA5 expression due to gene promoter methylation induces growth and colony formation of hepatocellular carcinoma cells.," *Oncology letters,* tom 11, nr 1, p. 861–869, January 2016.

[148] M. Mžik, M. Chmelařová, S. John, J. Laco, O. Slabý, I. Kiss, L. Bohovicová, V. Palička and J. Nekvindová, „Aberrant methylation of tumour suppressor genes WT1, GATA5 and PAX5 in hepatocellular carcinoma.," *Clinical chemistry and laboratory medicine,* tom 54, nr 12, p. 1971–1980, December 2016.

[149] R. Zheng and G. A. Blobel, „GATA Transcription Factors and Cancer.," *Genes & cancer,* tom 1, nr 12, p. 1178–1188, December 2010.

[150] T. Morimoto, K. Hasegawa, S. Kaburagi, T. Kakita, H. Masutani, R. N. Kitsis, A. Matsumori and S. Sasayama, „GATA-5 is involved in leukemia inhibitory factor-responsive transcription of the beta-myosin heavy chain gene in cardiac myocytes.," The *Journal of biological chemistry,* tom 274, nr 18, p. 12811–12818, April 1999.

[151] B. Ribau, J. Jorge, R. Alves, P. I. Ribeiro, A. C. Gonçalves, I. M. Carreira and A. B. Sarmento-Ribeiro, „Epigenetic modifications as targets to new therapies for Chronic Lymphocytic leukaemia – A preliminary study," *Porto biomedical journal,* tom 2, p. 223, 2017.

[152] D. M. E. I. Hellebrekers, M. H. F. M. Lentjes, S. M. van den Bosch, V. Melotte, K. A. D. Wouters, K. L. J. Daenen, K. M. Smits, Y. Akiyama, Y. Yuasa, S. Sanduleanu, C. A. J. Khalid-de Bakker, D. Jonkers, M. P. Weijenberg, J. Louwagie, W. van Criekinge, B. Carvalho, G. A. Meijer, S. B. Baylin, J. G. Herman, A. P. de Bruïne and M. van Engeland, „GATA4 and GATA5 are potential tumor suppressors and biomarkers in colorectal cancer.," *Clinical cancer research : an official journal of the American Association for Cancer Research,* tom 15, nr 12, p. 3990–3997, June 2009.

[153] C. M. Lyon, D. M. Klinge, K. C. Liechty, F. D. Gentry, T. H. March, T. Kang, F. D. Gilliland, G. Adamova, G. Rusinova, V. Telnov and S. A. Belinsky, „Radiation-induced lung adenocarcinoma is associated with increased frequency of genes inactivated by promoter hypermethylation.," *Radiation research,* tom 168, nr 4, p. 409–414, October 2007.

[154] M. B. Treppendahl, X. Qiu, A. Søgaard, X. Yang, C. Nandrup-Bus, C. Hother, M. K. Andersen, L. Kjeldsen, L. Möllgård, L. Möllgaard, E. Hellström-Lindberg, J. Jendholm, B. T. Porse, P. A. Jones, G. Liang and K. Grønbæk, „Allelic methylation levels of the noncoding VTRNA2-1 located on chromosome 5q31.1 predict outcome in AML.," *Blood,* tom 119, nr 1, p. 206–216, January 2012.

[155] L. R. Dice, „Measures of the Amount of Ecologic Association Between Species," *Ecology,* tom 26, pp. 297-302, 1945.

[156] J. Chen, C. Li, Y. Zhu, L. Sun, H. Sun, Y. Liu, Z. Zhang and C. Wang, „Integrating GO and KEGG terms to characterize and predict acute myeloid leukemia-related genes," *Hematology,* tom 20, pp. 336-342, 2015.

[157] K. S. Kramarzova, K. Fiser, E. Mejstrikova, K. Rejlova, M. Zaliova, M. Fornerod, H. A. Drabkin, M. M. van den Heuvel-Eibrink, J. Stary, J. Trka and J. Starkova, „Homeobox gene expression in acute myeloid leukemia is linked to typical underlying molecular aberrations," *Journal of Hematology & Oncology,* tom 7, 2014.

List of Figures

List of Tables

Acknowledgements

This work was financially supported by:

- European Social Fund grant POWR.03.02.00-00-I029,
- NCN Harmonia grant 2013/08/M/ ST6/924,
- POIG.02.03.01-24-099/13 grant funding and infrastructure: "GeCONiI - Upper Silesian Center for Computational Science and Engineering",
- SUT’s grants for Support and Development of Research Potential.

Analysed dataset were obtained thanks to:

- Collaboration with Cancer Mechanisms and Biomarkers group, Radiation Effects Department, UK Health Security Agency,
- GEO database,
- TCGA project database.

List of Author's publications

JCR Indexed Articles

1. O'Brien G., **Cecotka A**., Manola K. N., Pagoni M., Polańska J., Badie C.: Epigenetic Signature of Ionising Radiation In Therapy-Related AML Patients. Available at SSRN: https://ssrn.com/abstract=4213757 or http://dx.doi.org/10.2139/ssrn.4213757 (iScience, Under Review)

2. **Cecotka A**., Polanska J.: Region-specific methylation profiling in acute myeloid leukemia. Interdisciplinary Sciences - Computational Life Sciences, 2018, 10(1):33-42, doi: 10.1007/s12539-018-0285-4

3. Badie C., **Blachowicz (Cecotka) A**., Barjaktarovic Z., Finnon R., Michaux A., Sarioglu H., Brown N., Benotmane M.A., Tapio S., Polanska J., Bouffler S.D.: Transcriptomic and proteomic analysis of mouse radiation–induced acute myeloid leukaemia (AML). Oncotarget, 2016, 7(26):40461-80, doi:10.18632/oncotarget.9626

Monograph Chapters

4. **Cecotka A**., Krol L., O'Brien G., Badie C., Polanska J. May gender have an impact on methylation profile and survival prognosis in Acute Myeloid Leukemia? In: Rocha M., Fdez-Riverola F., Mohamad M. S., Casado-Vara R. (eds) Practical Applications of Computational Biology & Bioinformatics, 15th International Conference (PACBB 2021). PACBB 2021. Lecture Notes in Networks and Systems, vol 325, pp 126-135. Springer, Cham doi.org/10.1007/978-3-030-86258-9

5. **Cecotka A**., Polanska J.: Novel Method of Identifying DNA Methylation Fingerprint of Acute Myeloid Leukaemia. In: Fdez-Riverola F., Mohamad M., Rocha M., De Paz J., Pinto T. (eds) 11th International Conference on Practical Applications of Computational Biology & Bioinformatics. PACBB 2017. Advances in Intelligent Systems and Computing, vol 616. pp 189-196, (2017) Springer, Cham ISBN:978-3-319-60815-0

Conference Proceedings Abstracts

6. O'Brien G., **Cecotka A**., Manola K. N., Pagoni M., Polańska J., Badie C.: Epigenetic signature of ionizing radiation in therapy-related AML patients, European Radiation Protection Week 2022 (ERPW-2022), 9-14.10.2022, Estoril, Portugal

7. **Cecotka A**., O'Brien G., Manola K, Pagoni M, Badie C, Polanska J. AURKC - epigenomic biomarker of AML, 20th Asia Pacific Bioinformatics Conference (APBC), 28-28.04.2022

8. **Cecotka A**., O'Brien G., Badie C., Polańska J. DNA methylation alterations in de novo and therapy-related AML in several genomic regions. ISMB/ECCB2021. Online:25-30.07.2021

9. **Cecotka A**., O'Brien G., Badie C., Polanska J., Zróżnicowanie zmian w profilu metylacji DNA w obszarze promotorów genowych u pacjentów z ostrą białaczką szpikową. Krajowa Konferencja: Onkologia Obliczeniowa i spersonalizowana medycyna – Tu i teraz! COPM 2021, 21.04.2021 (p.29)

10. **Cecotka A**., Polanska J.: Different methylation profiles and their alterations among gene associated and intergenic genome regions in AML. XXI Gliwice Scientific Meetings, Book of abstracts, p. 71. 17-18.11.2017, Gliwice, Poland.

11. **Cecotka A**., Polanska J.: Zmiany poziomu metylacji DNA w różnych regionach genomu w ostrej białaczce szpikowej. Śląskie Spotkania Naukowe IV SSN, 24-25.03.2017, Ustroń, Polska, p.12

12. **Cecotka A**, Manning G, Badie C, Bouffler S, Polanska J: Demethylation level diversification among different genome regions in human AML, Book of Abstracts, p.104, XX Gliwice Scientific Meetings, Gliwice, November 18-19, 2016

13. **Błachowicz (Cecotka) A**., Manning G., Badie C., Bouffler S., Polanska J.: Detekcja regionów promotorowych genów o profilu metylacji istotnie różniącym osoby zdrowe i cierpiące na AML. Śląskie Spotkania Naukowe, 3-4.06.2016, Dzierżno

14. **Błachowicz (Cecotka) A**., Badie C., Bouffler S., Polanska J, Detection of Differentially Methylated Regions of Genome in Human Leukaemias, XIX Gliwice Scientific Meetings, Book of abstracts, p. 60. 20-21.11.2015, Gliwice, Poland

15. **Błachowicz (Cecotka) A**, Tapio S, Barjaktarovic Z, Benotmane R, Finnon R, Badie C, Bouffler S, Polanska J: Does the way of induction of AML among mouse has influence on protein signature in cells? 9th Central and Eastern European Proteomics Conference CEEPC 2015, p.40, 15-18.06.2015 Poznań

16. **Błachowicz (Cecotka) A**, Tapio S, Barjaktarovic Z, Benotmane R, Finnon R, Badie C, Bouffler S and Polańska J: Data driven estimation of cutoffs in searching for differentially expressed proteins, 2015. 19th Annual International Conference on Research in Computational Molecular Biology (RECOMB 2015), 12/04/2015-15/04/2015, Warsaw, Poland.

17. Badie C, Tapio S, Barjaktarovic Z, Finnon R, Brown N, **Błachowicz (Cecotka) A**, Kotas J, Polanska J, Benotmane R, Candeias S, Bouffler S: Integrated systems level analysis of myeloid leukaemogenesis and evaluation of the molecular alterations of antigenic T-cell receptor repertoire following X-irradiation. 15th International Congress of Radiation Research ICRR 2015, 25-29.05.2015, Kyoto, Japan

18. **Błachowicz (Cecotka) A**, S Tapio, Z Barjaktarovic, R Benotmane, C Badie, S Bouffler, J Polanska: Comparison study of transcriptomic and proteomic analyses of three mouse radiation induced AMLs. XVIII Gliwice Scientific Meetings, 21-22 Nov 2014, Gliwice, p.56

19. **Błachowicz (Cecotka) A**: Analiza wieloplatformowa biomarkerów białaczki indukowanej promieniotwórczością. Śląskie Spotkania Naukowe, Szczyrk 24-26.10.2014