**Cândida Patrícia Valente Cardoso**

# Uncovering differential DNA methylation alterations in subtypes of Acute Myeloid Leukemia

UNIVERSIDADE DO ALGARVE

Departamento de Ciências Biomédicas e Medicina

2020

**Cândida Patrícia Valente Cardoso**

# Uncovering differential DNA methylation alterations in subtypes of Acute Myeloid Leukemia

**Master in Oncobiology - Molecular Mechanisms of Cancer**

**This work was done under the supervision of:**

**Pedro Castelo-Branco, Ph.D**

**Mónica Teotónio Fernandes, Ph.D**

UNIVERSIDADE DO ALGARVE

Departamento de Ciências Biomédicas e Medicina

2020

# Uncovering differential DNA methylation alterations in subtypes of Acute Myeloid Leukemia

## Declaração de autoria do trabalho

Declaro ser a autora deste trabalho, que é original e inédito. Autores e trabalhos consultados estão devidamente citados no texto e constam da listagem de referências incluída.

"*I declare that I am the author of this work, that is original and unpublished. Authors and works consulted are properly cited in the text and included in the list of references.*"

**(Cândida Cardoso)**

## Agradecimentos

Em primeiro lugar gostaria de agradecer ao meu orientador Professor Pedro Castelo-Branco por me ter aceite como membro do seu grupo de investigação, pelos desafios propostos ao longo desta jornada que permitiram tornar-la mais enriquecidora, e pela ajuda e orientação prestadas ao longo da realização da dissertação.

Gostaria também de agradecer à minha orientadora Professora Mónica Fernandes por ter estado sempre disponível para me ajudar, por toda a orientação e paciência dadas ao longo deste percurso e por partilhar comigo o seu conhecimento permitindo uma melhor integração.

Agradeço também à Professora Ana Marreiros pela ajuda prestada ao longo da tese e pela sua boa disposição e simpatia.

Deixo também um obrigada aos restante colegas de laboratório.

Aos meus pais deixo um grande obrigada por todo o apoio incondicional sempre prestado, por investirem sempre em mim e no meu progresso. Agradeço por terem tornado tudo isto possível.

Agradeço também ao Daniel Pestana por todo o seu apoio incondicional, dedicação, por ter sido sempre um bom ouvinte e um excelente conselheiro. Agradeço-lhe por me ter ajudado sempre a ultrapassar as minhas dificuldades e a criar novos objetivos mais ambiciosos e desafiantes.

Às minhas amigas, Mariana Santos, Marta Afonso e Sofia Rodrigues, por toda a sua amizade, carinho e boa disposição.

## Abstract

Acute myeloid leukemia (AML) is a hematological malignancy where the hematopoietic stem cells or progenitor cells accumulate epigenetic and genetic alterations, losing their differentiation ability and gain proliferative advantage. AML is classified based on the cytogenetic abnormalities detected in the patient's leukemic cells. The World Health Organization (WHO) classification system is the most used and more current, distinguishing six subgroups. Moreover, the French-American-British (FAB) classification system also classifies AML, distinguishing seven subtypes (M0, M1, M2, M4, M5, M6, M7). The cytogenetic abnormalities and mutations in specific genes also allow the AML stratification into three prognostic risk groups: favorable, intermediate, and adverse. However, not all AML patients' leukemic cells exhibit chromosomal arrangements or gene mutations with prognostic impact, being categorized in the intermediate prognostic risk group. These patients show high clinical heterogeneity, being the treatment decision a current problem. Our goal was to identify potential prognostic biomarkers of gene expression and DNA methylation that could predict survival in AML patients that were categorized in the intermediate prognostic risk group. Thus, we developed an R-based algorithm that evaluates the prognostic potential of each gene and CpG site, available on the TCGA LAML cohort, in AML patients classified as FAB M1, M2, M4, and M5 subtypes. The algorithm was also performed in a group of patients with AML classified as FAB M0, M1, M2, M4 and M5 together. Our results suggest that there are some genes whose expression and/or DNA methylation are able to subdivide the AML patients categorized in the intermediate prognostic risk group into two subgroups with distinct overall survival. In conclusion, although the patients categorized in the intermediate prognostic risk group show a heterogeneous prognosis, they can be segregated by some candidate prognostic biomarkers of gene expression and DNA methylation, which can help to decide the best therapy for them.

**Keywords:** Acute myeloid leukemia, intermediate prognostic risk group, prognostic biomarkers, gene expression, DNA methylation

## Resumo

A leucemia mielóide aguda (LMA) é um grupo de cancros hematológicos heterogéneos que resultam da transformação de células estaminais hematopoiéticas ou progenitoras através da acumulação de alterações epigenéticas e genéticas, que conferem uma maior capacidade proliferativa e bloqueiam a sua diferenciação em células sanguíneas mais especializadas e funcionais. Por sua vez, as células transformadas (células leucémicas) acumulam-se na medula óssea e sangue periférico conduzindo a falhas ao nível da medula óssea e da hematopoiese.

Quanto à epidemiologia da doença, a LMA é o tipo de leucemia mais frequente em adultos e pode ser desenvolvida em indivíduos de todos os grupos etários. Contudo, a doença é mais frequente em indivíduos mais velhos com uma idade média de diagnóstico aos 68 anos, sendo a idade aumentada o principal factor de risco da doença.

A LMA é diagnosticada pela presença de pelo menos 20% de blastos miéloides na medula óssea ou através da deteção de alterações citogenéticas ou moleculares que já foram associadas com o desenvolvimento da doença. Estas alterações incluem: a t(8;21) (RUNX1-RUNX1T1), inv(16) ou t(16;16) (CBFB-MYH11) e a t(15;17) (PML-RARA), que caracteriza a leucemia promielócitica aguda. Além disso, as alterações citogenéticas e moleculares detetadas nas células leucémicas dos pacientes e usadas para diagnosticar a LMA permitem também a classificação da doença. O sistema de classificação proposto pela Organização Mundial da Saúde é atualmente usado para classificar a LMA, permitindo distinguir seis subtipos de leucemia: (1) LMA com alterações citogenéticas recurrentes, (2) LMA com alterações relacionadas com mielodisplasia, (3) LMA relacionada com terapia, (4) LMA não especificada, (5) sarcoma mieloide, e (6) proliferação mielóide de síndrome de Down. Relativamente aos casos de LMA do subtipo não especificado, a LMA é ainda classficada através do *French-American-British (FAB) classification system* que distingue sete subtipos (M0, M1, M2, M4, M5, M6, e M7) de LMA, com base na morfologia e características citoquímicas das células leucémicas.

Após a identificação do subtipo de LMA, os pacientes são ainda estratificados em três grupos de prognóstico: favorável, intermédio e adverso. Esta estratificação é feita com base nas alterações citogenéticas e mutações génicas detetadas nas células leucémicas dos pacientes e apresenta um papel importante na escolha do melhor tipo de tratamento para o paciente. Por exemplo, a t(8;21) [RUNX1/RUNX1T1] permite a estratificação da doença no grupo de risco

favorável. Por sua vez, a deteção de mutações no *RUNX1* estratifica a doença no grupo de risco adverso. Contudo, a maioria dos casos de LMA são categorizados no grupo de risco intermédio, sendo este definido por LMA com cariótipo normal, t(9;11)(p22;q23) [MLLT3-MLL9] ou pela presença de anormalidades citogenéticas que não são incluídas nos grupos de risco favorável ou adverso. LMA com cariótipo normal, ou seja, quando as células leucémicas não apresentam alterações citogenéticas ou mutações com valor de prognóstico conhecido, representam a maioria dos casos de LMA categorizada como risco intermédio. Estes pacientes são ainda caracterizados por uma grande heterogeneidade clínica, o que dificulta a escolha do melhor tipo de tratamento.

Assim, o objetivo do nosso estudo é identificar potenciais biomarcadores de prognóstico de expressão genética e metilação de DNA que sejam capazes de prever sobrevida em pacientes com prognóstico intermédio e sem alterações citogenéticas ou mutações com valor de prognóstico conhecido. Para tal, desenvolvemos um algoritmo que compreende quatro fases: preparação dos dados para análise, identificação dos primeiros candidatos para biomarcadores de prognóstico, calibração da idade como fator de confusão, e seleção restritiva dos candidatos finais. A metodologia desenvolvida faz uso de técnicas de inferência estatística para avaliar o potencial de prognóstico dos níveis de expressão de cada gene e metilação de DNA de cada CpG, disponíveis no *The Cancer Genome Atlas* (TCGA) *Acute Myeloid Leukemia* (LAML) cohort, em pacientes cuja LMA foi classificada nos subtipos FAB M1, M2, M4, e M5 e categorizada no grupo de risco intermédio com cariótipo normal. O algoritmo foi também aplicado a um grupo de pacientes com os subtipos FAB M0, M1, M2, M4 e M5 agrupados.

Os nossos resultados sugerem que a expressão e/ou metilação de certos genes podem subdividir os pacientes com subtipos de LMA categorizados no grupo de risco intermédio estudados em dois subgrupos com sobrevidas distintas. Por exemplo, de acordo com os nossos resultados, os níveis de expressão do gene *MCM4* são capazes de diferenciar sobrevida em pacientes com o subtipo FAB M1 categorizado no grupo de risco intermédio, sendo que os pacientes do subgrupo com pior prognóstico exibem baixos níveis de expressão do potencial biomarcador. No mesmo grupo de pacientes, a metilação do promotor do gene *SCIN* é um exemplo de um potencial biomarcador de prognóstico de metilação de DNA, sendo a metilação do mesmo no  promotor relacionada com pior prognóstico nos grupo de pacientes estudados.

Comparando os biomarcadores identificados nos grupos de pacientes com FAB M1, M2, M4 e M4 categorizados no grupo de risco intermédio, o algoritmo identificou maior número de biomarcadores candidatos de expressão genética e de metilação de DNA no grupo de pacientes com o subtipo FAB M2.

Existem potenciais biomarcadores identificados cujo o seu valor de prognóstico já foi documentado em pacientes diagnosticados com AML, como por exemplo o gene *ABCB1* como biomarcador de prognóstico de expressão, e o gene DLX4 como biomarcador candidato de metlação de DNA em pacientes do subtipo FAB M2 estudados.

Além disso, com base nas subdivisões geradas por cada potencial biomarcador de expressão genética identificado, averiguámos se os subgrupos de pacientes com pior prognóstico identificados compartilham alterações na regulação de conjuntos de genes relacionados com processos biológicos em comparação com os subgrupos de pacientes com melhor prognóstico. Verificámos que, apesar de a maioria dos potenciais biomarcadores de prognóstico identificados distribuírem os pacientes de forma distinta, a maioria dos subgrupos com pior prognóstico parecem compartilhar sobreregulações e subregulações de conjuntos de genes relacionados com processos biológicos distintos. Por exemplo, as células leucémicas da maioria dos pacientes dos subgrupos de pior prognóstico com o subtipo FAB M1 categorizado no grupo de risco intermédio, parecem subregular conjuntos de genes que estão relacionados com o processo de catabolismo do peróxido de hidrogénio quando comparados com os subgrupos de melhor prognóstico do mesmo subtipo de LMA. Estes processos biológicos poderão estar a influenciar o prognóstico dos pacientes com LMA.

Em conclusão, apesar de pacientes com células leucémicas sem rearranjos cromossomais ou mutações génicas com valor de prognóstico conhecido serem categorizados no mesmo subgrupo de risco, estes pacientes podem apresentar diferentes prognósticos que podem ser previstos através de potenciais biomarcadores de expressão e/ou metilação de DNA. Esta subdivisão poderá ajudar na decisão de um melhor tipo de tratamento para pacientes nas condições estudadas.

**Palavras-chave:** leucemia mielóide aguda, grupo de risco intermédio, biomarcadores de prognóstico, expressão génica, metilação de DNA.

## Index of Contents

**Index of figures**

## Index of tables

## Index of Annexes

## Abbreviations

**5caC** - 5-carboxylcytosine

**5fC** - 5-formylcytosine

**5hmC** - 5-hydroxymethylcytosine

**5mc** - 5-methycytosine

**APL** - Acute promyelocytic leukemia

**B2M** - Beta-2-Microglobulin

**CLPs** - Common Lymphoid Progenitors

**CMPs** - Common Myeloid Progenitors

**CR** - complete remission

**CXCL12 -** stromal cell-derived factor 1

**DNA** - deoxyribonucleic acid

**DNMTs** - DNA Methyltransferases

**EVI1** - Ecotropic Viral Integration Site 1

**GMPs** - granulocyte/macrophage progenitors

**HATs -** histone acetyl transferases

**HDACs** - histone deacetylases

**HMAs** - Hypomethylating agents

**HMTs** – Histone methyltransferases

**HSCs** - hematopoietic stem cells

**lncRNAs** - long noncoding RNAs

**MCM4 -** Minichromosome Maintenance Complex Component 4

**MEPs** - megakaryocyte/erythroid progenitors

**miRNA** – microRNA

**MKL1-** Megakaryoblastic Leukemia-1

**MPPs** - multipotent progenitor cells

**ncRNAs** - noncoding RNAs

**NK** - Natural Killer

**NPM1** - nucleophosmin 1

**piRNAs** - piwi-interacting RNAs

**PLA2G4A** - phospholipase A2 Group IVA

**PTMs** - posttranslational modifications

**RBM15 -** RNA-binding motif protein-15

**RISC -** RNA-induced silencing complex

**RPN1**- Ribophorin I

**RUNX1** - RUNT-related transcription factor 1

**siRNAs** - short interfering RNA

**SNAP23 -** Synaptosome Associated Protein 23

**TET** - ten-eleven translocation

**CN-AML** – AML with normal cytogenetics

**AML** – Acute Myeloid Leukemia

**TAM** - Transient abnormal myelopoiesis

**NOS** - AML not otherwise specified

**WHO** - World Health Organization

**FAB** - French-American-British

**NUP214 -** Nucleoporin 214

***PML* -** promyelocytic leukemia

***RARA*** - retinoic acid receptor alpha

**CBF -** core binding factor

***FLT3* -** FMS-like tyrosine kinase 3

**NK** - natural killer

**TCGA** - The Cancer Genome Atlas

**UCSC -** University of California, Santa Cruz cancer

**cDNAs -** complementary DNAs

**GSA -** gene set analysis

**GAGE** - Generally Applicable Gene-set Enrichment

# CHAPTER 1

# INTRODUCTION

## 1.1 Epigenetics and regulation of gene expression

All somatic cells of an organism have the same genetic information stored in the deoxyribonucleic acid (DNA) and it is transmitted from one cell to its daughter cells during cellular division.[1,2] There are some proteins called histones that allow the organization and compaction of the DNA in the nucleus, forming the chromatin.[3] The basic units of chromatin are the nucleosomes composed by 147 bp of DNA and a histone octamer with two copies of each of the four histones H2A, H2B, H3, and H4.[4] There is also a linker histone (H1 histone family) integrated where the DNA enters and exits the nucleosome. Subsequently, the chromatin fibers form the chromosome (Figure 1.1).[3] Through epigenetic control mechanisms that establish, regulate and maintain specialized gene expression patterns, different types of cells with different phenotypes can originate from the translation of the same DNA sequence.[4]

Epigenetics corresponds to alterations in gene expression without occurring any change in the underlying DNA sequence.[1] The epigenetic pattern of a cell is stable and transmitted to the daughter cells during cell division, maintaining the cell-type specific phenotype.[5] Furthermore, as the epigenetic alterations are reversible, they constitute a potential therapeutic target for treatment of diseases associated with epigenetic defects.[5] There are different mechanisms of epigenetic regulation that modulate the gene expression by mediating the access of the translational machinery (e.g., transcription factors and cofactors to specific genomic regions).[3] Through epigenetic regulation, the chromatin can adopt different conformations, euchromatin or heterochromatin, that influence the gene expression.[4] In the euchromatin conformation, the DNA is more relaxed and the gene expression is active.[4] In contrast, the heterochromatin state is associated with gene repression, since the DNA is supercoiled.[4] The referred genomic regions include enhancers (to improve transcription), promotors (to initiate transcription), gene body/ open reading frames (to be transcribed and translated into proteins), and silencers (to inactivate transcription).[3] The epigenetic mechanisms include posttranslational modifications of histones, chromatin remodeling, noncoding RNAs, and DNA methylation (Figure 1.1). [3]

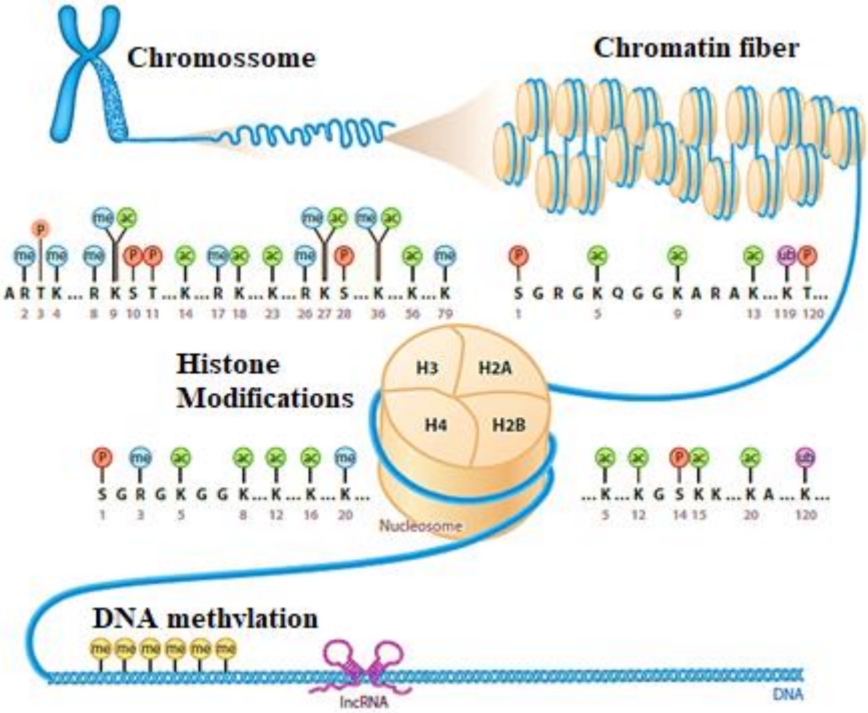


**Figure 1.1 DNA compaction levels and epigenetic mechanisms that regulate gene expression.** In the nucleus, the DNA is associated with an octamer with two copies of histones H2A, H2B, H3, and H4, forming the nucleosome. Together, various nucleosomes form the chromatin fiber and the chromosome. The epigenetic mechanisms of DNA methylation and posttranslational modifications of histones such as acetylation (ac), methylation (me), phosphorylation (P), and ubiquitination (ub), occur at the chromatin level. Long noncoding RNAs (lncRNAs) have been associated with regulation of gene expression, cell differentiation and chromatin remodeling. Illustration adapted from Chen et al., 2017.

### 1.1.1 Posttranslational modifications of histones and chromatin remodeling

The histones are highly conserved proteins constituted by a globular domain and a flexible unstructured amino terminal tail (the histone tail).[6] These proteins can undergo posttranslational modifications (PTMs) in the amino acid residues mainly present in the histone tail and some present in the globular domain, in particular of the histones H3 e H4.[6] The PTMs include acetylation, phosphorylation, methylation, and ubiquitination and they modulate gene transcription by controlling DNA accessibility (Figure 1.2).[3]

The histone acetylation is mediated by  histone acetyl transferases (HATs) and occurs in specific lysine residues.[3] This process is often associated with active chromatin regions that allow the transcriptional machinery to access the DNA and initiate transcription.  For example, the acetylation of histone H3 at lysine 27 (H3K27ac) is linked to active transcription regions.  The acetyl group can be removed by histone deacetylases (HDACs).[3] In contrast, the phosphorylation of histones in specific serine, threonine, or tyrosine residues by enzymes of the histone kinase family seems to be usually associated with transcriptional silencing, linked to condensed chromatin regions. [3]

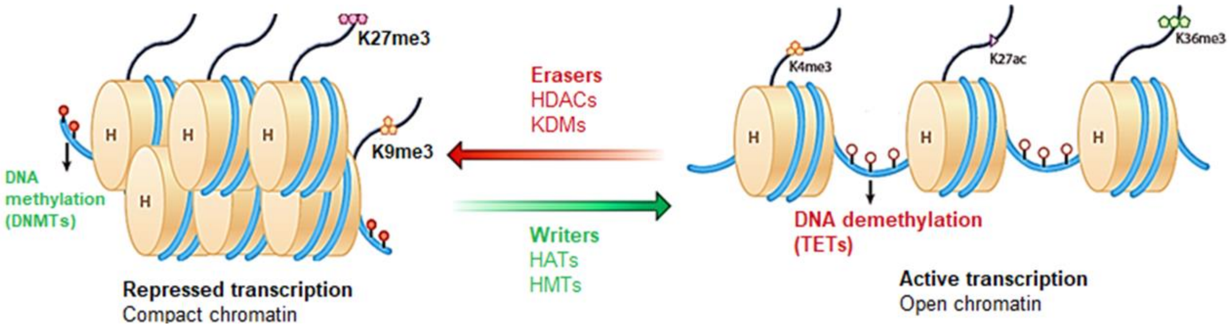


**Figure 1.2 Effect of histone posttranslational modifications in gene expression.** At the histone tail of H3 and H4, there are amino acids that can undergo posttranslational modifications, such as acetylation, methylation, phosphorylation, and ubiquitination, which influence the chromatin structure and subsequent gene expression. The acetylation of histone H3 at lysine 27 mediated by HATs as well as the trimethylation of histone 3 at lysine 4 and the trimethylation at lysine 36 mediated by HMTs, are marks of active transcription. These events lead to the opening of chromatin, making the DNA accessible to the transcriptional machinery. Illustration adapted from Chen *et al.*, 2017.

The histone methylation occurs in the specific lysine and arginine residues by histone lysine methyltransferases and arginine methyltransferases (HMTs).[7,8] The residues can be monomethylated (me1), dimethylated (me2), or trimethylated (me3). For instance, the di- or trimethylation of histone H3 at lysine 4 (H3K4me2 and H3K4me3, respectively) and monomethylation of H3K9 are associated with active gene expression. In contrast, di- and trimethylation of H3K9 and H3K27 are marks of inactive transcription.[7,8]

### 1.1.2 Noncoding RNAs

Noncoding RNAs (ncRNAs) are RNAs transcribed from the mammalian genome that are not translated into proteins, being functional RNAs.[3] Depending on their size, the ncRNAs are categorized into long ncRNAs (lncRNAs) if their sequences are longer than 200 nucleotides, and small ncRNAs, characterized by less than 200 nucleotides.[9] The lncRNAs work as regulators of gene expression through the modulation of nuclear architecture and transcription in the nucleus, and through the modulation of mRNA stability, translation and post-translational modifications in the cytoplasm. Furthermore, they are also implicated in cell differentiation, invasion and metastasis in cancer, and chromatin remodeling.[9]

The small ncRNAs also include short interfering RNA (siRNAs), microRNA (miRNA), and piwi-interacting RNAs (piRNAs) categorized depending on their length, biogenesis, and effector proteins.[1] These RNAs can modulate gene expression in a sequence-specific manner, since they are guides to the recognition of target RNAs. The siRNA is a double-stranded RNA that can target a complementary mRNA for degradation, leading to gene silencing.[1]

MicroRNAs are small, highly conserved, single-stranded non-coding RNA molecules that regulate gene expression by the RNA-induced silencing complex (RISC).[1] After their biogenesis, the miRNAs form RISC that can bind to a mRNA specified by base-pairing with the miRNA. The mRNA targeted can undergo cleavage and later degradation, or its transduction can be inhibited.[1]

Lastly, the piRNAs are formed by 21-35 nucleotides. This class of RNA molecules binds to PIWI proteins, guiding them to a target RNA to be cleaved.[10] Moreover, the piRNAs can also participate in  heterochromatin assembly and DNA methylation .[10]

### 1.1.2 DNA Methylation

The DNA methylation is one of the most studied epigenetic mechanisms in the mammalian genome and it is mediated by the covalent addition of a methyl group to the carbon in the 5-position of the pyrimidine ring of cytosine’s in the DNA.[3,11]  Usually, the cytosines to which a methyl group is added are adjacent to guanines by means a phosphate group, known as CpG dinucleotides.[1] In the mammalian genomes, the majority of CpG dinucleotides (70%) are in the methylated state. The remaining CpG dinucleotides are localized in clusters, known as

CpG islands, and are in the unmethylated state. The CpG islands are located near gene transcription start sites as well as intragenically.[1]

DNA Methyltransferases (DNMTs) are the class of enzymes responsible for the cytosine's methylation and there are three isoforms playing different roles in the cell.[3,11] DNMT1 is associated with the maintenance of DNA methylation.[3] During cell division, DNMT1 methylates the cytosines in the newly synthesized strand that are methylated in the complementary strand, ensuring that the daughter cells maintain the DNA methylation pattern of the original cell.[1] Thus, methylation patterns are stable. On the other hand, DNMT3A and DNMT3B are responsible for *de novo* methylation, since they methylate cytosines that were not previously methylated on either DNA strand. Thus, a new pattern of methylation can be created.[3,11]

The effect of DNA methylation on gene expression is dependent on the location of methylated CpG dinucleotides.[3] In general, when DNA methylation occurs in promotor regions, the downstream genes are silenced, since the transcription factors cannot interact with the DNA and promote transcription. The CpG dinucleotides can also be located in the gene body, and this is usually associated with activation of  gene transcription.[3] Nevertheless, this is not always the case.[12] For example, it has been described that *TERT* promotor hypermethylation is associated with increased *TERT* expression in cancer.[12] Each cell type has stable and unique DNA methylation patterns.[11]

The same way that the methylation can be added *de novo* to cytosines and maintained, the methyl group can also be removed, making DNA methylation a reversible process.[3] The ten-eleven translocation (TET) family proteins are responsible for the removal of the methyl group and this process is known as DNA demethylation. In this process, the 5-methycytosine (5mc) is converted by successive oxidation steps into 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC), and 5-carboxylcytosine (5caC).[3]

Since the DNA methylation has a temporal (e.g., developmental or differentiation stages) and spatial (e.g., specific DNA region) precision, it plays a central role in cellular processes like hematopoiesis, where it is necessary to activate and stabilize gene expression patterns during cell fate decision that allow for the differentiation of cells.[3]

## 1.2 Adult Hematopoiesis

The human body is composed by different organ systems that interact together in order to maintain the homeostasis throughout the life of the individual. The circulatory system is one of the organ systems and it is responsible for the transport of the blood for the whole body. Blood is a fluid formed by plasma and several different types of cells such as white blood cells, red blood cells and platelets, with different functionalities.[13] White blood cells or leukocytes participate in the inflammatory reaction and immune response and include the granulocytes (neutrophils, basophils, eosinophils, and mast cells), lymphocytes (T cells, B cells, and natural killer cells), monocytes/macrophages, and dendritic cells.[2] Red blood cells or erythrocytes are responsible for the transport of oxygen from the lungs to the tissues and carbon dioxide removal.[2] Platelets have a crucial role in wound healing and blood clotting.[13]

However, since the blood cells have a short half-life ranging from hours to days, it is necessary to form new blood cells daily to replace the dying ones.[13] The process of new blood cells formation is called hematopoiesis.[14] According to the classical model, hematopoiesis is a highly organized, dynamic, and regulated process that follows a hierarchy of cellular differentiation states from stem to progenitor to precursor to mature cells. The hematopoietic hierarchy has at the starting point a rare population of hematopoietic stem cells (HSCs) localized mostly in the bone marrow of axial bones. These cells have two relevant functional characteristics: multipotent differentiation and self-renewal.[14] The multipotent differentiation refers to the capacity that HSCs have to generate all mature blood cells, including erythrocytes, platelets, lymphocytes, monocytes, and granulocytes.[14] Transcription factors such as RUNT-related transcription factor 1 (RUNX1) and GATA3, expressed in HSCs, are responsible for maintaining the self-renewal and multipotency abilities.[8] On the other hand, the self-renewal is the process by which each HSC generates at least one more HSC by cellular division, ensuring the presence of a rare HSCs pool throughout life.[15] The self-renewal capacity is maintained by the Ikaros and E2A transcription factors.[8] The HSCs are quiescent and upregulate drug-detoxifying enzymes as the ATP-binding cassette (ABC), making them resistant to most chemotherapy agents. Furthermore, they reside in a highly protective microenvironment able to inactivate cytotoxins due to high expression of cytochrome P450 enzymes.[16]

The HSCs can give rise to specialized mature blood cells through differentiation into a series of progenitor intermediates.[17] Therefore, there is a balance between self-renewal and differentiation.[17]

During differentiation, the HSCs loose self-renewal capacity and give rise to multipotent progenitor cells (MPPs) through changes in gene expression that are influenced by transcription factors and epigenetic modifications in gene regulatory regions. The MPPs continue to have a multipotent differentiation ability, but a restricted self-renewal capacity.[2] On the other hand, the MPPs give rise to two different committed progenitors downstream that will originate two different cell lineages: the Common Lymphoid Progenitors (CLPs) that originate the entire lymphoid lineage, and the Common Myeloid Progenitors (CMPs), responsible for the development of the myeloid lineage.[2] Regarding the lymphoid lineage, the CLPs differentiate into B-cell precursors and the earliest thymic progenitors that will differentiate into T and Natural Killer (NK) cells that participate in adaptative and innate immune response.[2] In the myeloid lineage, CMPs differentiate into the granulocyte/macrophage progenitors (GMPs) that in turn give rise to granulocytes, monocytes and macrophages, and into the megakaryocyte/erythroid progenitors (MEPs), which can originate erythroid and megakaryocyte cells.[2] The committed progenitors are oligopotent and have a limited self-renewal capacity.

### 1.2.1 Regulation of myeloid differentiation

Myeloid cells are not only of great importance to innate immunity but are also key players in the regulation of the adaptative immune response.[8] The regulation of gene expression is the key point for the differentiation of progenitors and intermediate cells into mature myeloid cells.[8], During the process, genes and their products that contribute to the undifferentiated state are downregulated, whereas the genes that allow the cellular differentiation are upregulated.[8] The differentiation process of myeloid cells initiates in the bone marrow through cytokine signals released by stromal cells that activate progressively the transcriptional program that confers the myeloid identity.[8] Afterwards, the process terminates in the blood or peripheral tissues, where the precursor cells are exposed to cytokines, antigens and other factors to form fully differentiated myeloid cells.[8]

The timely regulation of gene expression required for myeloid differentiation is controlled by epigenetic mechanisms such as histone posttranslational modifications and DNA methylation in collaboration with lineage-specific transcription factors, upstream pathways signaling, and external/microenvironmental factors.[8] The DNA methylation regulates in part the self-renewal capacity of HSCs and facilitates the commitment to a lymphoid and myeloid lineage.[8] The methylation patterns of the myeloid lineage are different from the lymphoid lineage.[8] Whereas overall DNA methylation levels increase with lymphoid differentiation, regarding the myeloid differentiation the levels of overall DNA methylation decrease throughout the process (Figure 1.3).[8] DNMT1 prevents the premature activation of transcriptional programs associated with cellular differentiation in the HSCs.[18] Also, DNMT3A and DNMT3B are associated with the repression of the transcription factors RUNX1 and GATA3.[19] Without the expression of these transcription factors, the HSCs can differentiate into MPPs.[19] In case of infection and inflammation, the macrophage colony-stimulating factor (M-CSF, also known as CSF1) is released and activates the transcription factor PU.1 that in turn promotes the activation of genes that confer the myeloid phenotype.[5]

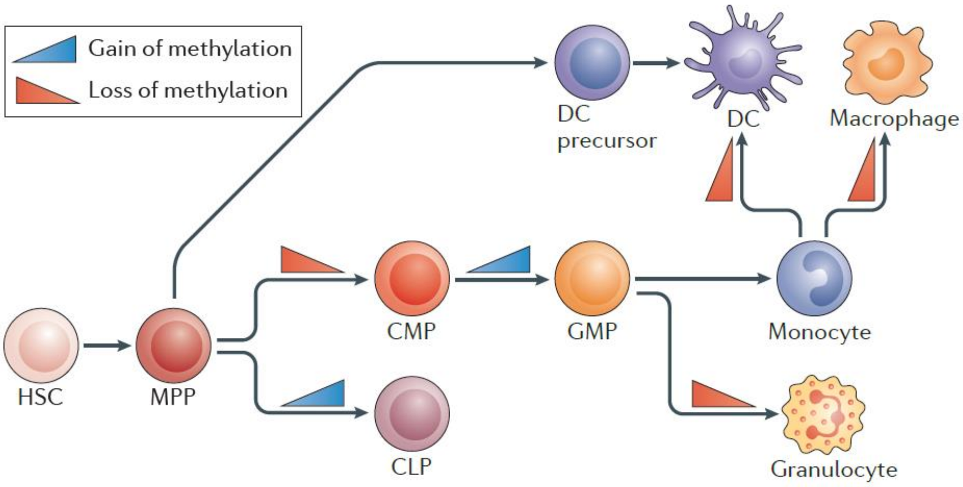


**Figure 1.3 Dynamic DNA methylation during myeloid differentiation.** Whereas hypermethylation is characteristic of lymphoid differentiation, hypomethylation is present during myeloid differentiation. However, the levels of DNA methylation in myeloid cells are dynamic. An increase in DNA methylation in MPP, promotes these cells to differentiate into CMP. In the transition of CMP to GMP a decrease in DNA methylation occurs, followed by a decrease in DNA methylation to form the granulocytes. The overall increase of DNA methylation is represented by the blue triangles and the overall decrease in methylation is represented by the red triangles. Adapted from Wouters and Delwel, 2016.

During differentiation of MPPs into CMPs occurs the simultaneous expression of PU.1 and GATA-1, and a decrease of overall methylation levels.[8] In the CMPs, both transcription factors display low expression levels, making these cells able to originate the GMP and MEP lineages.[20] When the CMPs express high levels of PU.1, they differentiate into GMPs.[20] In addition to PU.1, the C/EBP transcription factors downregulate HDAC expression and the commitment of GMPs to originate granulocytes. Furthermore, the transition of CMP to GMP is accompanied by an overall gain in methylation levels (Figure 1.3).[8]

In contrast, the upregulation of GATA-1 sustains HDAC expression and the CMPs are committed to give rise to the myeloid cells of the MEP lineage. Moreover, the differentiation of MEPs into erythrocytes is determined by the combination of GATA-1 expression and FOG-1. However, when the GATA-1 is associated to AML-1, the MEPs differentiate into megakaryocytes.

For the maturation of myeloid progenitors into monocytes, macrophages and dendritic cells it is required the expression of the transcription factor PU.1 in a concentration-dependent manner. In HSCs, this transcription factor presents low levels of expression, while in CMPs it is highly expressed. Another important transcription factor for myeloid differentiation is CCAAT/enhancer-binding protein α (C/EBPα).

The normal process of myeloid differentiation can be blocked, and the hematopoietic cells arrest in their immature forms, leading to the development of leukemia, a type of hematological cancer.[17]

## 1.3 Cancer

Cancer is a group of complex diseases that arise from the transformation of normal cells into malignant cells by the accumulation of genetic and epigenetic modifications.[21] This transformation is a multistep process whereby the cells are getting biological characteristics that give a selective advantage over normal cells, such as, sustaining proliferative signaling, evading growth suppressors, resisting cell death, enabling replicative immortality, inducing angiogenesis, activating invasion and metastasis, reprogramming of energy metabolism and evading immune destruction.[22] These biological capabilities are known as the hallmarks of cancer. Moreover, the oncogenic process is also characterized by genome instability responsible for genetic diversity, and inflammation.[22]

The first hallmark of cancer is the sustaining proliferative signaling.[22] Whereas the normal cells regulate the signals that promote cell growth in order to maintain a normal cell number and the normal tissue architecture, the cancer cells become independent of external signals to proliferate. Evading growth suppressors is the second defined cancer hallmark, which means that the cancer cells tend to inactivate growth suppressor genes, responsible for the negative regulation of cell growth and proliferation. Moreover, resisting cell death is the third defined hallmark of cancer. For uncontrolled growth, the cancer cells develop mechanisms to evade apoptosis. Normal cells have a limited number of division cycles they can undergo (Hayflick limit), after which they enter a state of senescence. Cancer cells overcome this limit by enabling replicative immortality, the fourth hallmark of cancer, which allows the tumor cells to divide indefinitely. The fifth hallmark of cancer, inducing angiogenesis, refers to the fact that cancer cells can stimulate growth of new blood vessels that will allow the irrigation to the tumor and support tumor growth. The activating invasion and metastasis was defined as the sixth hallmark of cancer and describes that cancer cells can gain the ability to invade neighboring tissues and, further, can spread to other distant locals through blood and lymphatic vessels, forming metastases. The seventh cancer hallmark is reprogramming of energy metabolism, which consists in the ability of cancer cells to dysregulate the energy metabolism, restricting their metabolism to mainly glycolysis, where glucose is metabolized in order to produce energy in the form of ATP that is important to support cell growth and division. Cancer cells can metabolize much more glucose than the normal cells, and thus produce more energy.  Finally, evading immune destruction is the eighth hallmark of cancer and describes the fact that cancer cells develop mechanisms that allow them to escape from detection and destruction mediated by the immune system.[22]

Nowadays, more than 277 different types of cancer are already identified.[23] According to the tissue and cell type of origin, the cancers are classified in three major classes.[24] Carcinomas originate from the epithelial cells and are the most common cancers, accounting for 80% of cases. Sarcomas derive from the connective tissue or muscle cells. Leukemias and lymphomas arise from blood cells and their precursors (hematopoietic cells).[24]

In 2018, were estimated 17.0 million of new cancer cases and 9.5 million of death-related cancer in worldwide.[25] The type of cancer that was most diagnosed and with more cancer deaths was the lung cancer. Female breast cancer, prostate cancer, and colorectal cancer were the following with more incidence, and the colorectal cancer, stomach cancer, and liver cancer with more mortality cases.[25]

### 1.3.1 Hematological cancers

HSCs are defined by their self-renewal ability and multipotent differentiation capacity, certifying the residence of a population of HSCs in the bone marrow and, at same time, the generation of all mature blood cells functionally different through a cascade of differentiation.[26] In normal conditions, there is a dynamic balance between the self-renewal and the multipotent differentiation of HSCs.[26] However, this equilibrium can be disrupted and dysregulated, leading to the development of hematological disorders.[26] Hematological malignancies or liquid cancers can arise from blood cells at any stage of differentiation with consequences in the production and the functionality of blood cells, leading to infections and/or uncontrolled bleeding.[26] When the hematopoietic differentiation is disrupted, it can lead to the development of one of the three types of hematological malignancies, such as leukemia, lymphoma, and myeloma.[26] Leukemia is characterized by an increased production of abnormal white blood cells in the bone marrow, leading to the circulation of leukemic cells in the blood.[26] They are classified based on the origin of leukemic cells and the time of clinical course into chronic lymphoid leukemia (CLL), chronic myeloid leukemia (CML), acute lymphocytic leukemia (ALL), and acute myeloid leukemia (AML).[26] In lymphomas, occur the transformation of  B or  T lymphocytes or natural killer (NK) cells that, posteriorly, infiltrate the secondary lymphoid organs, such as the lymph nodes or spleen.[26] The myeloma results from the malignant transformation of plasma cells that accumulate primarily in the bone marrow.

Myeloid malignancies include myeloproliferative neoplasms, myelodysplastic syndromes, and acute leukemia.[27] Myeloproliferative neoplasms are characterized by an excessive production of one or more blood cell types in the bone marrow and circulating blood.[28] Myelodysplastic syndromes are characterized by defects in the process of maturation associated with an ineffective hematopoiesis. Both types of myeloid malignancies can evolve into AML.[28]

## 1.4 Acute Myeloid Leukemia

Acute Myeloid Leukemia (AML) is a group of complex, dynamic, and clonal hematological malignancies that arise from the transformation of hematopoietic stem cells or myeloid progenitors.[29,30,31] These cells gain a proliferative advantage and lose their differentiation ability, resulting in clonal expansion of poorly differentiated myeloid progenitors (blasts) that accumulate in the bone marrow and peripheral blood.[32] Infections, anemia and hemorrhages are usually present in AML patients as the result of bone marrow failure and impairment of hematopoiesis.[30] AML is also characterized by genetic and phenotypic heterogeneity, that is reflected in heterogeneous clinical outcomes.[28]

### 1.4.1 Epidemiology

Acute Myeloid Leukemia is the most common type of leukemia in adults, accounting for 15 to 20% of acute leukemia cases in children and 80% in adults.[28] In the United States, the AML incidence has exceeded the incidence of the other subtypes of leukemia (ALL, CML, and CLL) until 2017.[33] It is estimated that approximately 19,940 adults will be diagnosed with AML in 2020. Most leukemic deaths (60%) are caused by AML in comparison with other leukemia subtypes.[33] In adults, AML is related to the shortest survival (5-year survival of 24%).[33]

AML can develop in individuals of any age group, but it is more frequent in older adults with a median age at diagnosis of 68 years.[30] The principal risk factor is indeed the increasing age. In the majority of the cases, AML arises in individuals previously healthy as a *de novo* malignancy, but it can be a consequence of the exposure to cytotoxic chemotherapy with alkylating agents or topoisomerase inhibitors, or to ionizing radiation as therapy of another primary malignancy.[32] In addition, this disease can also develop in patients with an underlying hematological disorder, such as myeloproliferative neoplasms and myelodysplastic syndromes, as a result of genomic instability and an additional gain of mutations.[30,28]

### 1.4.2 AML Leukemogenesis

In the last 15 years, the application of high-throughput sequencing techniques by genomic discovery studies has enabled significant advances in the understanding of mechanisms involved in the pathogenesis of AML.[29] The leukemogenesis is not yet completely understood.[30] However, it is known that in AML, the hematopoietic stem cells or myeloid progenitors undergo oncogenic transformation, acquiring chromosomal abnormalities and mutations in genes involved in proliferation, survival and cellular differentiation.[30,28] As a result, these cells gain a proliferative and survival advantage over the normal cells and lose their differentiation ability. Thus, the leukemic cells undergo clonal expansion and a population of poorly differentiated blasts accumulates in the bone marrow and peripheral blood.[30]

The chromosomal abnormalities and gene mutations that contribute to the malignant transformation in AML can be classified by the two hits model of leukemogenesis proposed by Gilliland.[34] The model suggests that AML is the result of the cooperation between at least two classes of mutations. The class I mutations activate the proliferation and survival of leukemic cells. Examples of class I mutations are FMS-like tyrosine kinase 3 (*FLT3),* *K/NRAS*, *TP53*, and *c-KIT*. On the other hand, the class II mutations include *NMP1* and *CEBPA*, which inactivate the normal hematopoietic differentiation and apoptosis. According to this model, it is necessary that the two classes of mutations occur in conjunction to develop AML. Recently, a third class of mutations has been considered and constitutes the mutations in genes that encode epigenetic modifiers.[34] Mutations in genes associated with DNA methylation like *DNMT3A*, *TET2*, and *IDH-1* and *IDH-2* have been identified in more than 40% of AML cases.[32]

### 1.4.2.1 Chromosomal Abnormalities in AML Leukemogenesis

Chromosomal abnormalities such as translocations, deletions, insertions, inversions, monosomies, trisomies, among others have been identified in 55% of patients diagnosed with AML.[28] The translocations create gene fusions that generate abnormal and dysfunctional proteins.[28] One of the most well-characterized chromosomal translocation in AML is the t(8;21) (q22;q22) in which the Run-related transcription factor 1 (*RUNX1*) is fused to *RUNX1T1* gene, resulting in the RUNX1-RUNX1T1 chimeric transcript (Figure 1.4).[35] The RUNX1 is a transcription factor that belongs to the family of core binding factor (CBF) involved in hematopoietic ontogeny. As a consequence of the t(8;21), the normal function of the CBF is disrupted, that in turn, disrupt the normal differentiation and maturation of the hematopoietic cells. These translocations recruit transcription repressors that block the expression of genes related to hematopoiesis and impair apoptosis. Nevertheless, this translocation alone does not cause AML. To this end, it must occur with cooperative mutations, such as those affecting *KRAS*, *NRAS*, *ASXL1*, and *KIT*. The previous translocation is present in approximately 5-10% of AML cases. Furthermore, the majority of AML cases with t(8;21) develop as *de novo*, and only 5% occur as a consequence of prior therapies.[35]

Another well-documented cytogenetic alteration in AML is the inv(16)(p13q22).[36] In this case, fusion between the *CBFB* and *MYH11* genes occurs, creating a chimeric protein product. The created fusion protein impairs the differentiation process of myeloid leukemic cells. However, this fusion is not sufficient for the development of AML. It is necessary that additional mutations occur to the disease development. Both t(8;21) and inv(16) affect the CBF complex.[36]

Acute promyelocytic leukemia (APL), a subtype of AML, is associated with the t(15;17)(q24;q21), being present in approximately 98% of APL cases. In this translocation, a fusion between the promyelocytic leukemia (*PML*) and the retinoic acid receptor alpha (*RARA*) genes occurs, resulting in the expression of the chimeric protein PML-RARA.[28] In normal conditions, the transcription factor and tumor suppressor PML participates in the regulation of cell cycle progression and can induce cell death. On the other hand, RARA forms a heterodimer with the retinoid X receptor and recruits the nuclear corepressor complex histone deacetylase, which in turn promotes the formation of nucleosomes, silencing several genes. The differentiation of promyelocytes entails the activation of several genes which happens through the binding of retinoic acid to RARA. The chimeric protein PML-RARA promotes the same effect as normal RARA when unbound to its ligand, however the previous needs a higher concentration of retinoic acid to silence gene expression. For this reason, the majority of APL patients respond to trans-retinoic acid treatment that leads to transcription and thus cell maturation.

The t(9;11)(p22;q23) is another chromosomal abnormality associated with AML development with monocytic features, and comprises the fusion of MLL3 and MLL genes.[28] The MLL gene codifies a histone methyltransferase that regulates gene transcription. MLL is a positive regulator of the HOX genes ‘expression and transcription factors involved in the development of the hematopoietic system. When the t(9;11) occurs, the MLL domain responsible for H3K4 methylation is lost and, in association with other transcription factors, promotes the HOX genes transcription and then, cell proliferation and self-renewal capacity.

The recurrent chromosomal translocation involving the fusion of the C-terminal region of Nucleoporin 214 (*NUP214*) and DEK was also identified as a driver event in leukemogenesis. The NUP214 is a part of the nuclear pore complex, responsible for protein and mRNA nuclear transport between the nucleus and the cytoplasm. As result of the translocation, a DEK-NUP214 fusion protein leads to the dysregulation of the nuclear transport.

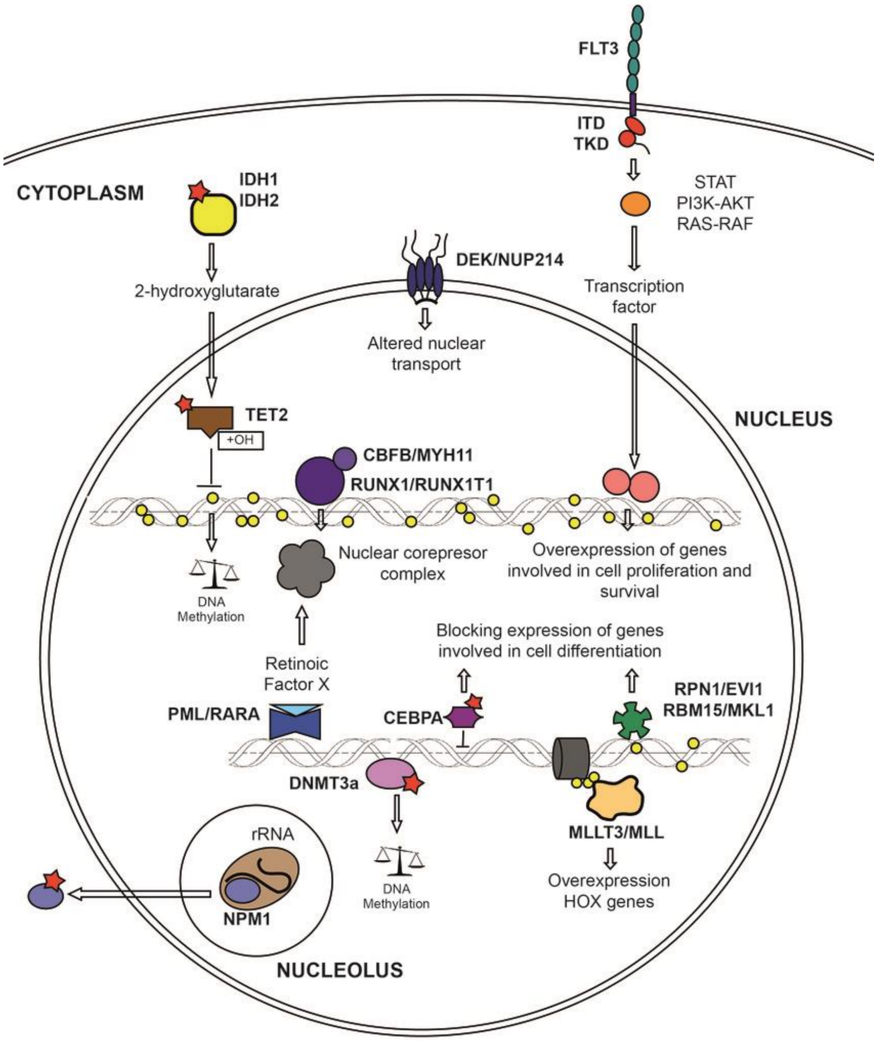


**Figure 1.4 Illustration showing the chromosomal abnormalities and gene mutations that contribute to AML development.** Cytogenetic alterations such as t(8;21) (q22;q22) [RUNX1/RUNX1T1], inv(16) (p13q22) [CBFB/MYH11], t(15;17) (q24;q21) [PML/RARA], t(9;11) (p22;23) [MLLT3/MLL], t(6;9) (p23;q24) [DEK/NUP214], inv(3) (q21;q26) [RPN1/EVI1] and t(1;22) (p13;q13) [RBM15/MKL1] are associated with the development of AML. On the other hand, gene mutations in FLT3, NMP1 and CEBPA can also contribute to AML. The disease is also characterized by aberrant methylation patterns associated with mutations in epigenetic modifiers, including mutations in DNMT3a, TET2, IDH1 and IDH2. Yellow circles represent methyl groups and the red stars represent mutations. Adapted from Lagunas-Rangel *et al.*, 2017.

The inv(3) (q21q26) is another cytogenetic alteration associated with AML, which affects the Ribophorin I (*RPN1*) and Ecotropic Viral Integration Site 1 (*EVI1*) genes. The EVI1 is a transcription factor expressed in HSCs, being essential for regulating cell self-renewal. EVI1 can also interact with histone deacetylases and chromatin-modifying enzymes, leading to epigenetic modifications that cause the silencing of certain genes. In this context, the expression of EVI1 is enhanced by RPN1, thus the fusion gene promotes cell proliferation and blocks cell differentiation, contributing to leukemogenesis.

Lastly, the t(1;22) (p13;q13) creates the fusion gene RNA-binding motif protein-15 (*RBM15*)/ Megakaryoblastic Leukemia-1 (*MKL1*) (BM15/MKL1) which affects chromatin remodeling and promotes HOX overexpression, thus affecting differentiation.[28]

### 1.4.2.2 Gene Mutations in AML leukemogenesis

Besides chromosomal translocations, gene mutations can also contribute and have an impact on the AML biology and phenotype, response to therapy and risk of relapse, thus having an impact on prognosis.[30] The molecular changes are present in more than 97% of AML cases.[32] In contrast with the other types of cancer, AML presents fewer number of mutations per cell.[29]

The pattern of mutations associated with the development of AML seems to occur in a temporal order.[29] The early phase of leukemogenesis is characterized by mutations in genes that encode epigenetic modifiers, such as *DNMT3A*, *ASXL1*, *TET2*, *IDH1*, and *IDH2*.[29] This is supported by studies that show the presence of these mutations in preleukemic cells, decades before the development of AML.[30] Mutations in the epigenetic modifiers mentioned provide a selective advantage for clonal expansion and subsequent progression to AML. The subsequent events in leukemogenesis are characterized by mutations involving nucleophosmin 1 (*NPM1)* or signaling molecules as *FLT3* and *RAS*.[29]

The most frequent mutation in AML is the *NPM1* mutation, being present in 25-30% of AML cases, and more prevalent in females.[37] In normal conditions, the NPM1 protein is mainly present in the nucleolus and is involved in ribosome biogenesis, genomic stability, DNA repair and molecular chaperoning. As a result of the mutations, the aberrant protein is more localized in the cytoplasm than the nucleus, promoting proliferation and leukemia development.[37]

The HSCs and the progenitor cells express the transmembrane FLT3 receptor tyrosine kinase that when activated by the FLT3 extracellular ligand, promotes cell survival, proliferation, and differentiation, through the activation of a signaling cascade involving PI3K, RAS, and STAT5.[38] Mutations in FLT3 receptor were identified in approximately 30% of AML cases.[38]

The mutations can occur in the juxtamembrane domain, the FLT3-ITD mutations (in approximately 25% of cases), or in the tyrosine kinase domain, the FLT3-TKD mutations (in 7 to 10% of cases).[38] Both lead to the constitutive activation of the FLT3 kinase, contributing to the proliferation and survival of leukemic cells.[38]

IDH1 and IDH2 are enzymes that participate in the epigenetic regulation and Krebs' cycle at the mitochondria level.[39] Mutations in these genes are found in approximately 20% of adult AML cases.[39] Genetic alterations lead to amino acid changes in conserved residues, resulting in neomorphic enzymatic function and production of an oncometabolite, 2-hydroxyglutarate that promotes DNA hypermethylation, aberrant gene expression, cell proliferation and blocked differentiation of hematopoietic progenitor cells.[40]

Mutations in the gene encoding the transcription factor CEBPA have been also associated with AML. CEBPA has a crucial role in the early stages of myeloid differentiation.[28] The *CEBPA* gene is expressed in myelomonocytic cells, being upregulated in granulocytic differentiation. As it presents two different AUG start sequences in the same reading frame, it can encode two proteins, an isoform of 42 KDa (p42) and another isoform of 30 KDa (p30). The ratio p42/p30 is regulated, so when the growth conditions are favorable, the transcription of p30 is promoted by the transcription factor elF2α and elF4E, leading to cell proliferation. In contrast, when the levels of elF2α and elF4E are low, p42 transcription occurs and also cell differentiation. The point mutations that can affect the *CEBPA* lead to alterations in the p42 transcription and overexpression of p30 isoform.[28] Mutations in *CEBPA* are found in 6-10% of AML cases.[37]

Currently, the genetic alterations associated with AML in combination with the cytogenetic abnormalities are incorporated in disease classification, risk stratification, and clinical care of patients.

#### 1.4.2.3 Aberrant DNA methylation in AML

Abnormalities in DNA methylation have been recognized as an important event in tumorigenesis of multiple cancer types. The epigenome of cancer is characterized by global hypomethylation that promotes genetic instability and the active transcription of oncogenes.[21] Hypermethylation is also found in cancer cells and occurs in promotors of specific genes, leading to their silencing.[21] In cancer research, DNA methylation studies have been focused on the CpG island promotor methylation.

In AML, it is also observed an aberrant distribution of the cytosine methylation pattern with clinical and prognostic relevance.[41] Furthermore, the epigenetic alterations emerge more frequently and recurrently than the genetic changes. The hypo- as well as the hypermethylation of CpG islands has also been reported in leukemogenesis.[41] Recurrent mutations in DNMT3A are identified in 6% to 36% of AML patients and can be truncating or missense.[5] Studies suggest that loss-of-function mutations in DNMT3A give a self-renewal advantage to HSCs. These are identified in older individuals before the clinical development of AML, suggesting that they are early events in leukemogenesis. The impact of DNMT3A mutations in leukemogenesis is not yet completely understood.[28] However, it is known that around 60% of all DNMT3A mutations in AML patients occur at the enzyme's methyltransferase catalytic domain.[42] This mutation not only leads to a loss of the methyltransferase activity, but also acts as a dominant negative mutation that decreases the methyltransferase activity of the wild-type DNMT3A by over 80%. Moreover, this reduction in DNMT3A activity in AML patients seems to lead to DNA hypomethylation, which promotes de binding of histone modifiers at the enhancer elements, ultimately leading to activation of leukemic stemness genes, like *Hoxa* genes. In addition, DNMT3A mutations seem to be associated with FLT3-ITD and NMP1 mutations in AML.[42] Clinically, this mutation has proved to be important in patient's stratification, conferring a poor prognosis to AML patients, an increased risk of relapse, and a decreased overall survival.[28]

Gain-of-function mutations in IDH1 and IDH2 genes are also found in AML, essentially at highly conserved arginine residues. As consequence, IDH1 and IDH2 gain the capability to transform the α-ketoglutarate to 2-hydroxyglutarate, leading to its accumulation.[28] The 2-hydroxyglutarate is a competitive inhibitor of TET2, which demethylates DNA. In addition, loss-of-function mutations in TET2 are found in 8% to 27% of AML patients and 18% to 23% patients with normal cytogenetics, giving a worse prognosis.[5] The reduction of TET2 activity results in hypermethylated DNA regions located mainly in gene regulatory elements, which leads to a deregulation of genes related to self-renewal and differentiation, such as *Gata1*.[42] Moreover, it is also known that the DNA hypermethylation caused by impaired TET2 activity affects approximately a quarter of all enhancer elements, most of which are associated with tumor suppressor genes.[42]

### 1.4.3 Diagnosis and Classification

The diagnosis of AML is made based on the morphological assessment of peripheral blood or bone marrow with the presence of 20% or more myeloid blasts in the bone marrow.[30] The myeloid origin of the malignant blast is also identified by immunophenotyping by means flow cytometry that will help to classify the AML subtype.[30] Moreover, the disease can be diagnosed by the presence of recurrent karyotypic or molecular alterations that are associated with leukemogenesis.[30] These alterations include the t(8;21) (RUNX1-RUNX1T1), inv(16) or t(16;16) (CBFB-MYH11) and the t(15;17) (PML-RARA), which characterizes the acute promyelocytic leukemia.[30]

The morphological variability of leukemic cells and the degree of differentiation led to the establishment of classification systems that allow to identify different subtypes of AML. The French-American-British (FAB) classification system established in 1976 was the first system to classify different subtypes of AML (Table 1.1).[43] The classification is made based on the morphological appearance and cytochemical characteristics of the leukemic cells (blasts) and defines eight AML subtypes (M0 to M7).[43]

**Table 1.1 The FAB AML classification system.**

| | |
|---|---|
| M0 | Acute myeloid leukemia without differentiation |
| M1 | Acute myeloid leukemia with minimal differentiation |
| M2 | Acute myeloid leukemia with differentiation |
| M3 | Acute promyelocytic leukemia hipergranular or typical |
| M3v | Acute promyelocytic leukemia hipogranular |
| M4 | Acute myelomonocytic leukemia |
| M4v | Acute myelomonocytic leukemia with bone marrow eosinophilia |
| M5 | Acute monocytic leukemia |
| M6 | Acute erythroid leukemia (Erythroleukemia) |
| M7 | Acute Megacariocytic leukemia |

However, this classification is limited since it does not have in consideration the genetic and clinical diversity of the disease.[28] Currently, AML is classified based on the World Health Organization (WHO) classification system.[44] This system, was updated in 2016 and incorporates genetic information with morphology, immunophenotype and clinical presentation, defining mainly six subgroups of AML: (1) AML with recurrent genetic abnormalities, (2) AML with myelodysplasia-related changes, (3) therapy related AML, (4) AML not otherwise specified (NOS), (5) myeloid sarcoma, and (6) myeloid proliferation of Down syndrome (Table 1.2). In the AML NOS group, the classification is generally based on the FAB classification.[44]

**Table 1.2 AML subtypes and related neoplasms based on WHO classification.**

| Subtype | Genetic abnormalities |
|---|---|
| AML with recurrent genetic abnormalities | AML with t(8;21)(q22;q22.1); *RUNX1 – RUNX1T1* |
| | AML with inv(16)(p13.1q22) or t(16;16)(p13.1;q22); *CBFB-MYH11* |
| | APL with *PML-RARA* |
| | AML with t(9;11)(p21.3;q23.3); *MLLT3-KMT2A* |
| | ML with t(6;9)(p23;q34.1); *DEK-NUP214* |
| | AML with inv(3)(q21.3q26.2) or t(3;3)(q21.3;q26.2); *GATA2*, *MECOM* |
| | AML (megakaryoblastic) with t(1;22) (p13.3; q13.3); *RBM15-MKL1* |
| | AML with BCR-ABL1 (provisional entity) |
| | AML with mutated NPM1 |
| | AML with biallelic mutations of CEBPA |
| | AML with mutated RUNX1 (provisional entity) |
| AML with myelodysplasia-related changes | |
| Therapy-related myeloid neoplasms | |
| AML, NOS | AML with minimal differentiation |
| | AML without maturation |
| | AML with maturation |
| | Acute myelomonocytic leukemia |
| | Acute monoblastic/monocytic leukemia |
| | Pure erythroid leukemia |
| | Acute megakaryoblastic leukemia |
| | Acute basophilic leukemia |
| | Acute panmyelosis with myelofibrosis |
| Myeloid sarcoma | |
| Myeloid proliferations related to Down syndrome | Transient abnormal myelopoiesis (TAM) |
| | Myeloid leukemia associated with Down syndrome |

### 1.4.5 Risk Stratification

In addition to a highly genetic heterogeneity, patients with AML present very distinct clinical outcomes.[30] Therefore, an accurate prognosis is crucial to the right management of AML patients. The patients are stratified based on the risk of treatment resistance or treatment-related mortality that will guide the decision between standard or increased treatment intensity, consolidation chemotherapy or allogenic hematopoietic stem cell transplant, or even between established or investigational therapies.[32]

The individual prognosis of AML patient is influenced by patient- and disease-related factors.[30] The patient-related factors that include age, coexisting clinical conditions, and poor performance status will predict treatment-related early death.[30] Increased age and poor performance status are associated with lower rates of complete remission and decreased overall survival. In oncology, the performance status is one important factor to have into account in cancer care and is represented by a score that reflects the ability of the patient to perform daily activities without the help of others, such as dressing, eating, bathing, among others.[45] On the other hand, the disease-related factors such as white cell counts, prior myelodysplastic syndrome or cytotoxic therapy for another malignancy, and leukemic cell genetic changes can predict resistance to current standard therapy.[31] Therapy-related AML and AML associated with a prior hematologic disorder confers a worse prognosis. Advances in clinical care of patients have contributed to a decreased of risk of treatment-related death that is lower than the risk of resistance to the treatment.

Furthermore, the cytogenetic profile of each case has an important role as a prognostic marker for complete remission and overall survival in AML.[32] Cytogenetic analysis consists in the identification of abnormalities at the chromosomal level like translocations, deletions, insertions, inversions, among others, in samples of blood or bone marrow.[28] Thus, based on the chromosomal abnormalities identified, the patients are stratified into three prognostic risk groups: favorable, intermediate, or adverse (Table 1.3).[28,46] The presence of chromosomal rearrangements like t(8;21) [RUNX1/RUNX1T1], t(15;17) [PML/RARA] or inv(16) [CBFB/MYH11] confer a favorable prognostic risk characterized by a good response to treatment and complete remission.[28] The adverse prognostic risk group is characterized by the presence of a complex karyotype (defined as the presence of three or more chromosomal abnormalities in the absence of any of the recurrent genetic abnormalities identified in the WHO 2008 classification), monosomy 5 or 7, t(6;9), or inv(3).[32] In this prognostic group, the disease is more aggressive, and the patients have a poor response to the treatment.[28] In some cases, the patients have normal cytogenetics (CN-AML), that is, they do not have any chromosomal abnormality.[47] The majority of these patients are categorized in the intermediate prognostic risk group.[47] The intermediate prognostic group represents about 45% of AML cases and is characterized by a highly clinical heterogeneity, making difficult their stratification and decision of the best treatment option.[47]

**Table 1.3 Risk stratification based on genetics.**

| Prognostic-risk group | Cytogenetic profile |
|---|---|
| Favorable | t(8;21) (q22;q22); *RUNX1-RUNX1T1* |
| | inv(16)(p13.1q22) or t(16;16)(p13.1;q22); *CBFB-MYH11* |
| | t(15;17) (q22;q12) |
| Intermediate | CN-AML |
| | t(9;11)(p22;q23) |
| | Cytogenetic abnormalities not included in the favorable or adverse prognostic risk groups |
| | |
| Adverse | t(6;9)(p23;q34); *DEK-NUP214* |
| | t(v;11q23.3); KMT2A rearranged |
| | t(9;22)(q34.1;q11.2); *BCR-ABL1* |
| | inv(3)(q21.3q26.2) or t(3;3)(q21.3;q26.2); *GATA2, MECOM(EVI1)* |
| | 25 or del(5q); –7; –17/abn(17p) |
| | Complex karyotype, monosomal karyotype |
| | Wild-type *NPM1* and *FLT3-ITD* |
| | Mutated *RUNX1* |
| | Mutated *ASXL1* |
| | Mutated *TP53* |

### 1.4.6 Treatment

In recent years, advances in the therapeutic approaches for AML patients have been made with approval of new therapeutic drugs.[30] Even with these advances, the clinical outcomes of AML patients are still disappointing, with more than half of patients dying from this disease.[30] Since AML is a group of heterogeneous diseases, it is necessary to select different treatment options. Therefore, the patient is evaluated in order to determine what treatment option is the best for him. Patient-related factors such as age are relevant for the therapeutic decision. In general, the standard treatments for AML are induction chemotherapy and allogeneic stem cell transplant for eligible candidates. As most older patients are unable to receive these treatments, since they have a lower tolerance to an intensive chemotherapy, they have a worse prognosis comparing to younger patients.[30] Moreover, their poor prognosis is also justified by the fact that AML in older patients is more often characterized by the presence of cytogenetic and molecular abnormalities characteristic of the adverse prognostic risk group.[30]

The induction chemotherapy refers to the first line treatment that the patient receives, with the objective to achieve and maintain complete remission.[32] If all signs of AML disappear in response to treatment, the patient has achieved complete remission. However, minimal residual disease often persists in complete remission, leading to relapse. For this reason, it is crucial that the patient receives post-remission therapy in order to eliminate any residual disease.[32] In younger patients, induction chemotherapy consists mainly in cytarabine and anthracyclines.[28] For intermediate prognosis patients, the treatment is more intensive with higher doses of cytarabine. The older patients are subject to comorbidities and have less tolerance to intensive chemotherapy, and thus they receive lower doses of the drugs. The patients categorized in the favorable prognostic risk group have relatively good outcomes with overall survival rates of approximately 60%. In contrast, the outcomes of patients with intermediate and adverse prognostic risk are still unsatisfactory.[30]

The allogeneic stem cell transplant is used in AML patients after the induction chemotherapy as post-remission consolidation treatment.[48] In patients with AML categorized in the favorable risk group, the transplant is not usually necessary in their first complete remission, since the risk of relapse is lower than the risk of transplant-related mortality.[48] In these cases, the transplant can be indicated only in the second complete remission, following a relapse.[48]

As the patients categorized in the adverse prognostic risk group have a high risk of relapse, the allogenic stem cell transplant should be performed in order to enhance their survival probability.[48] It must be indicated in the first complete remission, since their clinical outcome after the transplant in second remission is poorer than in first remission.[48]

In last, for patients categorized in the intermediate prognostic risk group, the criteria to perform a transplant are less clear.[48] However, nowadays the majority of these patients are evaluated for transplant in their first complete remission.[48]

Despite the use of intensive chemotherapy as well as the stem cell transplantation in the treatment of AML, the disease is still fatal and it is necessary to develop therapies more specific and less toxic for the patient.[37] In opposition to genetic alterations, the epigenetic changes are in the majority reversible, providing an opportunity for the development of targeted therapies with specific inhibitors.

### 1.4.6.1 Epigenetic therapy in AML treatment

The development of drugs for AML treatment that target epigenetic alterations have been studied in preclinical trials with some drugs already available to be used in the clinical practice.[49] Hypomethylating agents (HMAs), such as azacytidine and decitabine, and isocitrate dehydrogenase inhibitors, including ivosidenib and enasidenib, are already approved and used in AML treatment.[49]

Hypomethylating agents are DNA methyltransferase inhibitors that lead to a transient and variable DNA hypomethylation.[50] They are used in the treatment of AML patients that cannot receive intensive induction chemotherapy and for patients diagnosed with myelodysplastic syndromes.[49] The clinical responses to this type of treatment are still heterogeneous and rarely sustained.[50] The decitabine is converted into 5-aza-dCTP, an active tri-phosphorylated metabolite, by successive phosphorylation's through intracellular kinases.[50] Then the 5-aza-dCTP is incorporated into the DNA during the cell cycle, and binds to DNMT1, promoting its degradation.[50] Therefore, a DNA hypomethylation is promoted after each cell cycle, and the expression of tumor suppressor genes associated with senescence and apoptosis is activated.[50] Moreover, the differentiation of leukemic cells is also promoted.[50] In contrast the azacytidine is incorporated into the RNA, and the mRNA and protein metabolism are disrupted and the malignant proliferation is inhited.[51]

On the other hand, the isocitrate dehydrogenase inhibitors have as purpose to enhance the acetylation of histones, promoting the transcription of several genes involved in cell differentiation, cell cycle regulation and apoptosis.[49]

As single agents, both hypomethylating agents as well as isocitrate dehydrogenase inhibitors demonstrate to have a limited efficacy in the treatment of AML. Their combination with other therapies is been studied in clinical trials. [49]

# CHAPTER 2

# OBJECTIVES

Nowadays, cytogenetic analysis is still an important tool for prognostic assessment and therapeutic decision in patients with AML, allowing the patients stratification into three prognostic risk groups: favorable, intermediate and adverse risk group. However, most AML patients categorized in the intermediate prognostic group, have leukemic cells that do not exhibit any type of cytogenetic abnormality or gene mutations that allow their stratification into the favorable or adverse risk group, which causes difficulties in predicting these patient’s prognosis and in understanding the molecular mechanisms of this disease. These patients are characterized by having a high clinical heterogeneity, which is a problem when deciding the optimal treatment. Therefore, our main goal is to identify potential prognostic biomarkers based on gene expression and DNA methylation that could allow to predict survival of intermediate-risk AML patients.

To achieve this aim, we will perform the following steps:

1. The development of an algorithm that will iteratively assess the prognostic potential of every gene and CpG probe with available data in the TCGA-LAML, in intermediate-risk AML patients. Subsequently, we aim to identify candidate prognostic biomarkers that could predict survival in intermediate-risk AML patients of each FAB-subtypes.
2. Secondly, we aim to get further insights about what kind of cellular mechanisms could be impacting prognosis in AML patients studied. Therefore, we will systematically perform gene expression comparisons between intermediate-risk AML patients with worse prognosis with the intermediate-risk AML patients with better prognosis. Through these comparisons we will then examine which biological processes could be altered between the subgroups.

# CHAPTER 3

# METHODS

## 3.1. The Cancer Genome Atlas (TCGA)

The Cancer Genome Atlas (TCGA) database (http://cancergenome.nih.gov/) is a project created by the National Cancer Institute and the National Human Genome Research Institute that provides publicly available clinical and molecular data about more than 10000 patients and over than 30 tumor types. The purpose was to catalogue and discover genomic, epigenomic, transcriptomic, and proteomic alterations that contribute to cancer development, which may improve the diagnosis, treatment and prevention of cancer. TCGA contains datasets regarding methylation, whole genome sequencing, whole exome sequencing, RNA expression, proteomics and clinical observations. [52]

## 3.2. Dataset collection

Datasets regarding gene expression and whole-genome DNA methylation of AML patients (LAML cohort), publicly available at TCGA database (http://cancergenome.nih.gov/), were collected through the use of University of California, Santa Cruz cancer (UCSC) Xena Public Data Hubs (https://tcga.xenahubs.net) (Accessed 1/19/2019; 4:57 PM). Both datasets were imported to the R environment through the *read* function from *readr* package.

### 3.2.1. Gene expression dataset

The gene expression dataset has expression values of 20530 different genes determined experimentally using the *Illumina HiSeq 2000 RNA sequencing* platform from samples of peripheral blood of AML patients (n=173). Each patient is represented by only one sample. The data is level 3 and the values of gene-level transcription estimates are presented in $\log_2(x+1)$ transformed RSEM normalized count. Illumina RNA sequencing is a next-generation sequencing technology that enables to characterize and quantify the RNA transcripts present in a sample.[53] In general, the RNA transcripts are reverse transcribed into complementary DNAs (cDNAs). These are then randomly fragmented, and the end of each cDNA fragment is ligated to adapters. Then, the cDNA fragments are sequenced and aligned to a reference genome database.[53]

### 3.2.2. DNA methylation dataset

The DNA methylation dataset contains methylation values of 485577 CpGs, determined experimentally in peripheral blood samples of AML patients (n=194) using the *Illumina Infinium HumanMethylation450* platform. This technology allows quantifying the methylation status of more than 450 000 CpG sites located in the genome, using two types of probes: the Infinium I probes and the Infinium II probes.[54] Two Infinium probes target each CpG site, being one of them to detect the "methylated" (M) intensity and the other for "unmethylated" (U) intensity.[55] The Infinium I probes consists of two probes: one for the methylated allele and the other for the unmethylated allele.[54] In contrast, only one Infinium II probe is used to detect the "M" and "U" intensity by distinct dye colors (green and red).[55] Therefore, the methylation level is estimated and represented as a beta value (β).[55] The β is a continuous variable whose values vary between 0 and 1. So, higher levels of methylation (hypermethylation) are represented by higher βs (closer to 1) and lower levels of methylation (hypomethylation) are characterized by lower β's (closer to 0).

### 3.2.3. Patient clinical data

UCSC Xena Public Data Hubs data regarding clinical information (e.g., age at diagnosis, prognosis classification, overall survival time) from 200 AML patients were also collected. The samples are exclusively cancerous, since normal samples are not available. The clinical variables used in our analysis are described in Table 3.1. As we only considered patients with expression and methylation data, our sample was reduced to 171 AML patients. Patients classified with M6 (n=2) and M7 (n=3) FAB subtypes were also removed from our analysis due to insufficient number of patients. Furthermore, patients without information about any clinical variable of interest were also removed from our analysis and the sample was reduced to 148 AML patients. Afterwards, the patients were subdivided according the FAB AML subtype. To note that in our analysis, we only considered the AML patients whose disease was categorized in the intermediate risk group of prognoses, and for this reason the final sample was reduced to 89 AML patients. Moreover, the patients whose AML was categorized as FAB M3 subtype categorized in the intermediate risk group were also removed from our analysis.

**Table 3.1 Clinical Characteristics of AML patients categorized in the intermediate risk prognosis group.**

| **Group** | **M0**<br>**(n=6)** | **M1**<br>**(n=24)** | **M2**<br>**(n=19)** | **M4**<br>**(n=24)** | **M5**<br>**(n=16)** |
|---|---|---|---|---|---|
| Age ± sd | 55 ± 17 | 54 ± 17 | 56 ±18 | 58 ± 15 | 54 ± 16 |
| < 60 years | 3 (50%) | 14 (58%) | 8 (42%) | 12 (50%) | 9 (56%) |
| ≥ 60 years | 3 (50%) | 10 (42%) | 11 (58%) | 12 (50%) | 7 (44%) |
| cytogenetics | Normal karyotype | Normal karyotype | Normal karyotype | Normal karyotype | Normal karyotype |
| Vital Status | | | | | |
| Alive | 2 | 10 | 6 | 4 | 5 |
| Dead | 4 | 14 | 13 | 20 | 11 |
| Survival Time[1] ± sd | 745 ± 763 | 592 ± 576 | 684 ± 744 | 497 ± 342 | 475 ± 614 |

sd, standard deviation; 1, Survival mean time in days

### 3.3. Algorithm for the identification of potential prognosis biomarkers

In order to achieve our main goal, we developed an algorithm that allows the identification of potential prognostic biomarkers. The algorithm consists in four phases. In the first phase, the dataset is prepared for the analysis. In the second phase, the first candidate biomarkers capable to predict survival in the population of study are identified. The third phase consists in the elimination of patient's age as a confounding factor. Finally, in the fourth phase, the best potential biomarkers are non-randomly selected based on established criteria. The algorithm was applied using R language through the R software. Each phase of the algorithm, as well as the R packages used, are described below.

### 3.3.1. R language

R is a type of programming language and environment that allows the application of a variety of statistical (e.g., classical statistical tests) and graphical methods and can be easily extended through the installation of packages. This computational language is a free and open source and is commonly used for statistical inference and data analysis in the research community. The R code was developed using the R Studio software, which is also open source.

### 3.3.2. Phase 1 – Data preparation for analysis

#### 3.3.2.1. Outliers treatment

Before initiating the analysis, it is recommended to detect and remove outliers present in the dataset, in order to avoid bias of the results. The outliers are defined as the observations that are much smaller or much larger than the majority of observations.[56] These extreme values can significantly affect the statistical analysis, leading to overestimated or underestimated values.[56] The outlier's detection method incorporated in our algorithm is the boxplot method (Figure 3.1). A boxplot is a graphical representation of the distribution of the values, allowing the visualization of how the observations are spread.[56] According to this method, the outliers are the data points that lie outside the upper or lower fence lines.[56] In our dataset, the outliers were detected and removed from our analysis using the *boxplot.stats* function available in R studio.

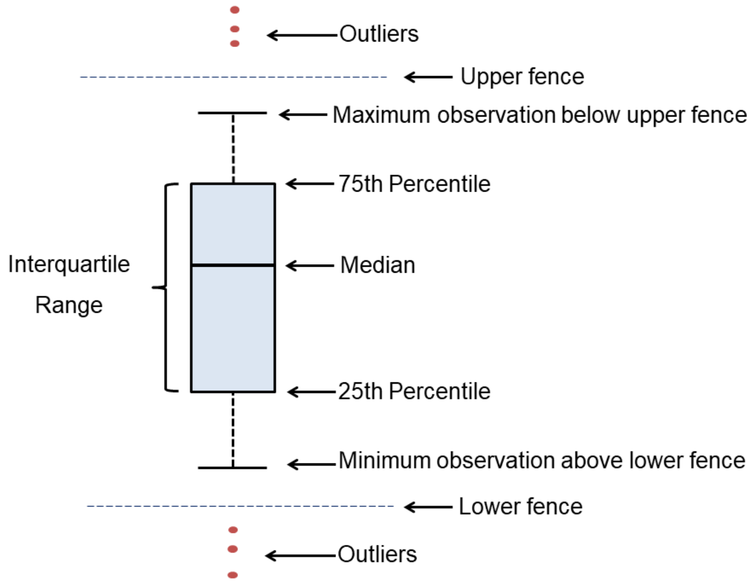


**Figure 3.1 Boxplot Method for detection and elimination of outlier values.** The observations that are more and less than the 75th and 25th percentiles are represented between the upper and the lower fences. Any value (observation) represented above the upper fence or below the lower fence are considered outliers and must be removed from the analyzing dataset. Adapted from Kwak and Kim, 2017.

#### 3.3.2.2. Missing data treatment

Missing data or missing values are a very common problem in real datasets and can be defined as a value that is not recorded for a variable in the observation of interest. Non-treated missing data are associated with reduced statistical power and can lead to biased estimates and thus, to invalid conclusions. So, to avoid an incorrect inference analysis, we handled missing data by applying the most frequently used method known as listwise or case deletion. According to this method, those variables or cases with missing data should be removed from the analysis. However, only variables with more than half of missing values were excluded in order to avoid losing extensive information, leading also to a significant reduction of the sample.

### 3.3.3. Phase 2 – Identification of first candidate prognostic biomarkers

#### 3.3.3.1. Determination of the optimal cutpoint

Since we wanted to perform survival curves to evaluate the prognostic value of a continuous or ordinal variable of interest (e.g., gene, cg), it is necessary to determine a cut point (cut off) to classify the observations into two distinct groups and then, compare their overall survival.

Our algorithm determines the optimal cut point for a variable by the maximally select rank statistics (maxstat) method. This method was already used in a variety of published studies and can be easily used in R by using the *survminer* package. The maxstat method determines the exact optimal cut point through several methods and approximations that maximize the separation of the observations.[57] Furthermore, the discrimination power of the value is also estimated and evaluated by calculating a *p- value*.[57] The optimal cut point serves to classify the overall survival observations in one group with the values lower than or equal to the cut point, and in another group with the values greater than the cut point.[57]

#### 3.3.3.2. Survival curves analysis

The survival analysis is a time-to-event analysis that involves a set of statistical approaches used to evaluate the length of time until an event of interest occurs. The time variable represents the years, months, or days from the beginning of the follow-up of a subject until the event occurs. The event variable can mean death, relapse from remission, among others. The majority of survival data is processed using non-parametric methods. The Kaplan-Meier method is the most used non-parametric method in this type of analysis. The Kaplan-Meier survival curve is a visual representation of survival data that describes the probability of surviving in a given period of time while considering time in many small intervals. In the graph, the Y axis represents the cumulative survival probability and the X axis represents the follow-up time in units of time.

#### 3.3.3.3. Comparison of survival curves

Once the Kaplan-Meier survival curves for each variable are performed, the survival of the individuals of the two or more groups can be compared by statistical tests. The most common statistical test used to compare the survival curves is the *log-rank test*. The log-rank test is performed to test the null hypothesis that the probability to occur an event (in this case, death) at any time point is not different between the groups. In our algorithm, we rejected the null hypothesis when the log-rank *p-value* is lower than 0.05, meaning that the groups of individuals had significant differences in overall survival.

However, sometimes the Kaplan-Meier curves intersect each other and in these cases the log-rank test is not recommended since the test may not detect the survival differences between the groups.[58] When intersection of survival curves occurred, we performed the two-stage method because this test demonstrates higher power and greater stability in comparison with other methods, independently if the survival curves intersect at an early, middle, or late time.[58] The two-stage method was developed by Qiu and Sheng and involves two steps.[59] In the first step, a conventional log-rank test is performed.[58] If a difference in survival between groups is detected by the conventional log-rank test (*p-value* < 0.05), then the two-stage process stops, and we can assume that the overall survival is different between the groups. However, if the log-rank test *p-value* is higher than 0.05, then either there is not a survival difference in both groups, or the survival curves cross and the conventional log-rank test is not able to detect the differences. In this case, the second step is initiated, and a weighted log-rank test is performed. The weights are chosen so that the signal changes before and after a potential crossing of hazards. Then, a new *p-value* is generated, which we considered to be statistically significant if it was lower than 0.05.[58] The two-stage test was performed in R, using the *twostage()* function from the TSHRC package.

### 3.3.3. Phase 3 - Confounding factor treatment

Our main goal is to identify potential biomarkers of prognosis to allow the subdivision of the AML patients with an intermediate prognostic risk. As the age of the AML patient is an independent risk factor, the median age of patients can constitute a confounding factor in our analysis. That is, if one group of patients is formed mostly by older patients and assuming that the older patients have an higher risk of not surviving than the younger patients, then this group will be classified with a worse prognosis in comparison with the other group. Thus, to ensure that the median age of patients is not a confounding factor in our analysis, the population analyzed was subdivided into two groups. The young group, which is formed by AML patients with less than 60 years old and the older group formed by AML patients with 60 years or more (Figure 3.2).

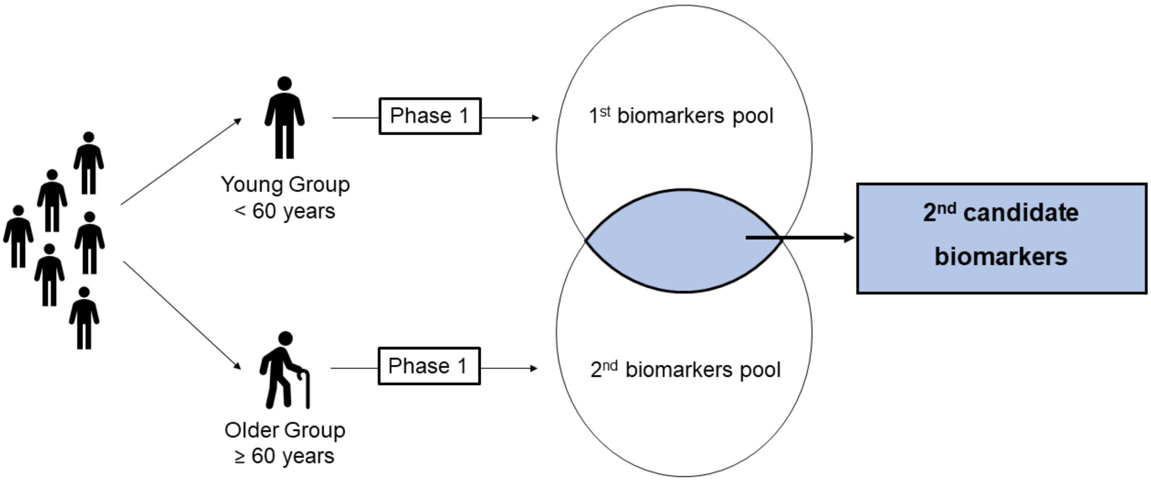


**Figure 3.2 Cofounding factor treatment.** To certify that an advanced median age of patients is not responsible for the worse prognosis of one group, the patients were divided into young (< 60 years) and older groups (≥ 60 years). For both populations we performed the phase 2 of our algorithm, resulting two pools of candidate biomarkers. One pool resultant from the analysis in the young population, and another pool resultant from the older population analysis. We only considered the candidate biomarkers present in the intersection of the two biomarker pools. These were called the second candidate biomarkers.

In both populations, young and older, the steps of phase 2 described above were performed independently. Thus, we obtained two pools of candidate biomarkers. One that was able to predict survival in the population of patients with less than 60 years, and another with predictive prognosis value in patients with 60 years or more. As we were interest in identifying the potential prognostic biomarkers that could predict survival independently of the age of the patient, we only considered the potential biomarkers resultant from the intersection of the two pools, referred as the second candidate biomarkers. In order to confirm that the median age of the two groups of patients was not significantly different, we performed a non-parametric Wilcoxon-Mann-Whitney test. We chose the Wilcoxon-Mann-Whitney test, since it does not make the assumption that the data is normally distributed. The null hypothesis was rejected when the *p-value* was lower than 0.05. This test was termed as age-test, and was performed in R through the *wilcox.test*() function from the stats package.

### 3.3.4. Phase 4 – Selection criteria

All the previously identified second candidate biomarkers were able to discriminate the patients into two groups with distinct overall survival (log-rank or two-stage *p-value* < 0.05), and statistically non-significant differences regarding patients' age (Mann-Whitney test *p-*

*value* < 0.05). However, we noted that some of these candidate biomarkers subdivided the initial population into two subgroups with a substantial difference in the number of patients in each group (e.g., 2 patients in the first group vs 12 patients in the second group), leading to an unbalanced analysis. Therefore, to select the best potential biomarkers, we established a selection criterion that had in consideration the number of patients in both of the generated groups. So, we proceeded the analysis with only the candidate biomarkers that were able to divide the initial population into two subgroups with an established a cut-off based on the ratio of patients calculated in each survival curve. In an ideal case, the number of patients in the first group would be equal to the number of patients in the second group. So, the ideal ratio would be equal to 1. However, in our analysis this situation does not occur frequently, thus we considered a ratio between 0.60 and 1.67.

In addition, we wanted to select the biomarkers that were differently expressed or methylated between the generated risk subgroups (intermediate-favorable vs intermediate-poor), since the biomarkers that meet this criterion are more reliable in predicting prognosis. For that, in the following part of the algorithm, a set of statistical tests were included.

Firstly, a *Shapiro-Wilk test* is performed. This is a highly recommended test to evaluate the null hypothesis that a variable comes from a normal distributed population. For those variables that the Shapiro-Wilk *p-value* was lower than 0.05, we rejected the null hypothesis and the variable was categorized as having a non-normal distribution. Therefore, we obtained two groups of data: normally distributed data and non-normally distributed data. Distinct statistical tests were performed for each group.

For normally distributed data, the difference of expression or methylation, between the generated patient subgroups, was tested by a parametric *t*-test. When the *t*-test *p*-value was lower than 0.05, we rejected the null hypothesis. The independent samples t-test assumes the analyzed groups have equal variances. When the samples present variances non-significantly different, the Welch-Satterthwaite method is performed. To know when to use this adjustment, we performed the levene test. The null hypothesis associated with this test is that the variances are equal in the analyzed groups. When the levene *p-value* was lower than 0.05, the null hypothesis was rejected, and the Welch-Satterthwaite adjustment was performed in the independent samples t-test.

The independent samples t-test was performed in R by the *t.test*() function from the stats package, and the levene test was performed using *leveneTest*() function from the car package.

For non-normally distributed data, the difference of expression or methylation, between the generated patient subgroups, was tested by a non-parametric Wilcoxon-Mann-Whitney test, since it does not make the assumption that the data is normally distributed. The null hypothesis was rejected when the *p-value* was lower than 0.05.

### 3.4. Gene Set Enrichment Analysis

After identifying the potential biomarkers of gene expression that were able to distinguish the patients with AML intermediate-risk AML patients into two subgroups with distinct survival distributions, we aimed to understand which cellular processes were systematically different between the subgroups with worse prognosis (intermediate-poor) and the subgroups with better prognosis (intermediate-favorable). These identified candidate biomarkers divide the intermediate-risk AML patients in different ways, and for this reason the distribution of patients between the intermediate-poor and intermediate-favorable clusters varies depending on which candidate biomarker is used. Therefore, to understand which biological processes were constantly different between the two prognostic-clusters, we performed a gene set analysis (GSA) between the intermediate-poor and intermediate-favorable subgroups for every distribution of patients our candidate biomarkers generated. Subsequently, we were able to analyze which cellular processes were systematically enriched, regardless of which candidate biomarker was used.

A GSA methodology examines the enrichment of gene ontology (GO) terms between two analyzed subgroups, which are sets of functionally related genes that describe biological functions, biological programs, and cellular locations where these take place. As such, three categories of GO terms exist: 1) biological processes, that describe the biological objective in which the group of genes participate (e.g., DNA repair); 2) molecular functions, that represent the activities that the genes products execute (e.g. kinase activity); and 3) cellular component, that describes the location in the cell where a given molecular activity or cellular process takes place. In our study, our main interest was in biological processes, since the other two categories may lead to ambiguous and unclear conclusions.

In comparison with individual gene analysis, GSA has several advantages such as an increased sensitivity, robustness and is more relevant from a biological point of view. In this study, we performed the Generally Applicable Gene-set Enrichment (GAGE), because it allows to study samples with different sizes. The GAGE methodology assesses whether certain predefined gene sets (GO terms), which describe specific cellular mechanisms, are differentially expressed between two groups. This is achieved by calculating a log-based fold change for each available gene, and then applying a two-sample t-test between the mean fold change of a given gene set and the mean fold change of the entire background of genes (which in this study is all the genes for which we had available gene expression data). For every gene set, a *p-value* from the two-sample t-test in then generated, which used to reject the null-hypothesis that the gene set is not differentially expressed between the two analyzed groups.

In this study, we iteratively performed the GAGE analysis for every patient-distribution, using all the GO terms that were available through the "gage" R package, which was the chosen tool to complete this analysis. Furthermore, each gene set was considered to be differentially expressed, in a statistically significant way, between the intermediate-poor and intermediate-favorable subgroups when the two-sample t-test *p-value* was lower than 0.05. Finally, we examined which of these GO terms were systematically enriched in all of the candidate biomarker-generated patient divisions.

### 3.5. Bibliographic analysis

Once determined the candidate genes that have a potential prognosis value in patients with AML categorized in the intermediate prognosis risk group, we performed a bibliographic analysis in order to know what genes had already described in the PubMed literature. For that, the OncoScore package was used to evaluate the genes had already described in literature and cancer-related articles. The OncoScore is an internet-based tool that measure the association of a term to cancer, depending on the citation frequency on PubMed articles.[60]

Moreover, we also were interested to know if the genes had already documented particularly in leukemia and AML-related articles. To achieve this, we developed an R-based function that sequentially queried the PubMed database, using as search terms the identified gene and the disease. The function would then tally the number of search results and store the number. Furthermore, since usually there are several terms that refer to the same disease, this function was built to query the target database using a set of predefined terms, rather just one.

### 3.6. Expression data analysis

Initially we imported the dataset with the expression values of 20530 genes and selected the 148 patients of interest (as described in section 3.2.3). In addition, these patients were subdivided based on their FAB AML subtype into 6 subgroups: M0, M1, M2, M4, and M5. As our main goal was to identify potential prognostic biomarkers for the patients categorized with an intermediate prognostic risk, for each FAB subtype we only selected these patients. As the FAB subtypes M0 and M3 included only 6 and 1 patient, respectively, these were excluded from our analysis. For each subgroup of patients classified in the intermediate prognosis risk group (M1, M2, M4, and M5), we applied the previously described algorithm (section 3.3).

### 3.7. Methylation data analysis

Regarding the methylation analysis, we imported the methylation dataset and selected the same 148 patients. The patients were also subdivided by FAB subtype and posteriorly, only the patients classified in the intermediate prognosis risk group were selected, excluding the patients classified with M0 and M3 FAB subtypes. The developed algorithm was applied in each subgroup of FAB subtypes selected (M1, M2, M4, and M5), and with the M0, M1, M2, M4, and M5 FAB subtypes together. After the algorithm application, the delta beta ($\Delta\beta$) value was also calculated for each CpG site identified as a potential prognostic biomarker, and we only selected the cg's with a $\Delta\beta > 0.2$. Finally, we identified, for each FAB subtype selected, the set of cg's that can predict survival for patients categorized within the intermediate prognosis risk group, independently of their age.

# CHAPTER 4

# RESULTS

## 4.1 FAB M1 AML SUBTYPE

From the TCGA database, we collected the datasets regarding gene expression, DNA methylation and clinical information from AML patients (TCGA LAML cohort). As we were interested in analyzing each AML FAB subtype independently, our first population of study was the patients with AML classified as the FAB M1 subtype (AML-M1) categorized within the intermediate prognostic risk group. The TCGA LAML cohort included 24 patients under these conditions. Next, we applied the developed algorithm independently to both gene expression and DNA methylation datasets. The gene expression results are described in the 4.1.1 section, and the DNA methylation results are described in the 4.1.2 section.

### 4.1.1 Gene expression as a potential prognostic biomarker in patients with AML-M1 categorized in the intermediate prognostic risk group

The expression values of 20530 genes were analyzed for patients with AML-M1 categorized in the intermediate prognostic risk group. After the first phase of the algorithm, known as data preparation for analysis, the 20530 genes were reduced to 17527 genes. These genes were posteriorly submitted to the second phase of the algorithm, consisting of the identification of the first potential prognostic biomarkers, remaining 1095 genes. From the third phase of the algorithm described as the confounding factor treatment, the 1095 genes were reduced to 90 genes. Lastly, after the fourth phase known as selection criteria, our algorithm identified 11 candidate genes whose expression appeared to be a potential biomarker to predict survival in patients with AML-M1 classified in the intermediate prognostic risk group (Table 4.1 and Annex I).

The optimal cutpoint described in the Table 4.1. is the cutpoint value mentioned throughout the results chapter explanation, and it is the value used to stablish the two subgroups (low vs high) in the Kaplan-Meier overall survival curves.

**Table 4.1 List of the 11 genes whose expression was able to predict survival in patients with AML-M1 categorized in the intermediate prognostic risk group.**

| **Gene** | ***p-value*** | **Optimal cutpoint** | **Group1 (n)** | **Group2 (n)** | **Age test (*p-value*)** | **HR** |
|---|---|---|---|---|---|---|
| *TATDN3* | 0.00858 | 7.8936 | 13 | 9 | 0.08806 | 9.53508 |
| *PBDC1* | 0.00305 | 8.0404 | 13 | 11 | 0.1727 | 7.15919 |
| *CLIC1* | 0.00875 | 12.2505 | 13 | 10 | 0.43742 | 5.97406 |
| *SNAP23* | 0.01412 | 10.4537 | 10 | 13 | 0.21405 | 4.07169 |
| *NNMT* | 0.04064 | 2.6515 | 14 | 10 | 0.76936 | 0.33026 |
| *CLEC4G* | 0.04198 | 2.2924 | 11 | 13 | 0.05932 | 0.30684 |
| *ARHGAP23* | 0.02927 | 9.553 | 11 | 13 | 0.33836 | 0.26506 |
| *ABCC9* | 0.02149 | 2.2692 | 11 | 13 | 0.05552 | 0.25617 |
| *ZBED3* | 0.00765 | 7.0613 | 13 | 11 | 0.07677 | 0.22755 |
| *MCM4* | 0.01577 | 11.7508 | 10 | 13 | 0.11315 | 0.16847 |
| *SLC10A4* | 0.01217 | 2.2692 | 10 | 13 | 0.57608 | 0.16753 |

HR, hazard ratio

In Figure 4.1, we show representative top 4 Kaplan-Meier overall survival curves with the highest hazard ratio value, obtained using the expression cutpoints of *TATDN3*, *PBDC1*, *CLIC1*, and *SNAP23* genes, respectively. Interestingly, low expression levels of the four genes was related to the subgroups with a better prognosis. Relatively to the 7 remaining identified genes, the better prognosis was related with high expression (Annex I).

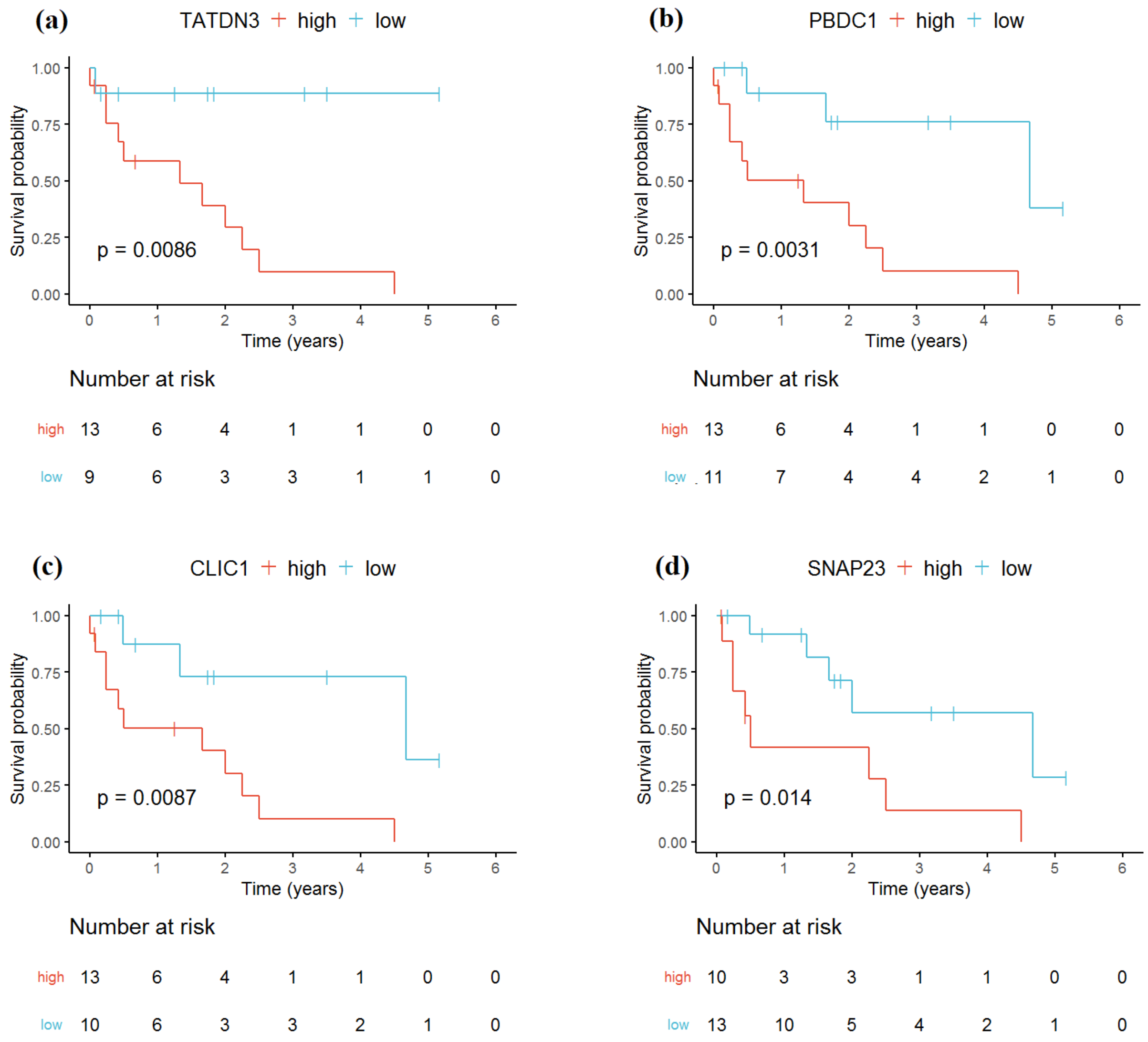


**Figure 4.1 Kaplan-Meier overall survival curves for four out of the eleven potential prognostic biomarkers identified for patients with AML-M1 categorized in the intermediate prognostic risk group.** Kaplan-Meier curve obtained using the expression cutpoint of the (a) *TATDN3* gene (p = 0.0086, log-rank test), (b) *PBDC1* gene (p = 0.0031, log-rank test), (c) *CLIC1* gene (p = 0.0087, log-rank test), (d) *SNAP23* gene (p = 0.014, log-rank test). The number at risk corresponds to the number of patients, at the indicated time point, that are still alive and whose follow-up continues.

Moreover, the survival curve obtained based on the expression cutpoint of *TATDN3* gene has the highest hazard ration value of 9.54 (Figure 4.1.a). This result means that the patients categorized in the worse prognosis subgroup have 9.54 times more risk of dying than the subgroup with the better prognosis. The hazard ratio values of the survival curves obtained using the *PBDC1*, *CLIC1*, *SNAP23* expression cutpoints were 7.16, 5.97, and 4.07, respectively (Figure 4.1. b, c, d).

We have also observed that the subdivision of patients with AML-M1 categorized in the intermediate prognostic risk group was specific for each gene expression cutpoint identified. For example, the expression cutpoint determined for the *TATDN3* gene did not generated the same patient subgroups as the expression cutpoint determined for the *PBDC1* gene.

However, despite these different subdivisions according to the expression cutpoints, we were interested to know if there were common gene sets, related to biological processes, enriched in the identified subgroups with worse prognosis (intermediate-poor subgroups) in comparison with the identified subgroups with a better prognosis (intermediate-favorable subgroups). Our results suggest that, in the majority of the intermediate-poor subgroups generated by each expression cutpoint, there were some biological processes gene sets downregulated and upregulated when compared with the intermediate-favorable subgroups (Annex II). In Figure 4.2, are shown the top 5 biological processes gene sets down and upregulated in the majority of the intermediate-poor subgroups.

Vascular endothelial growth factor signaling pathway, hydrogen peroxide catabolic process, cellular response to vascular endothelial growth factor stimulus, skeletal system development, and gas transport are examples of these downregulated gene sets in the majority of intermediate-poor versus intermediate-favorable subgroups. In contrast, regulation of natural killer cell mediated immunity, podosome assembly, positive regulation of response to biotic stimulus, regulation of innate immune response, positive regulation of innate immune response are some biological processes gene sets that seemed to be upregulated in the majority of intermediate-poor subgroups (Figure 4.2).

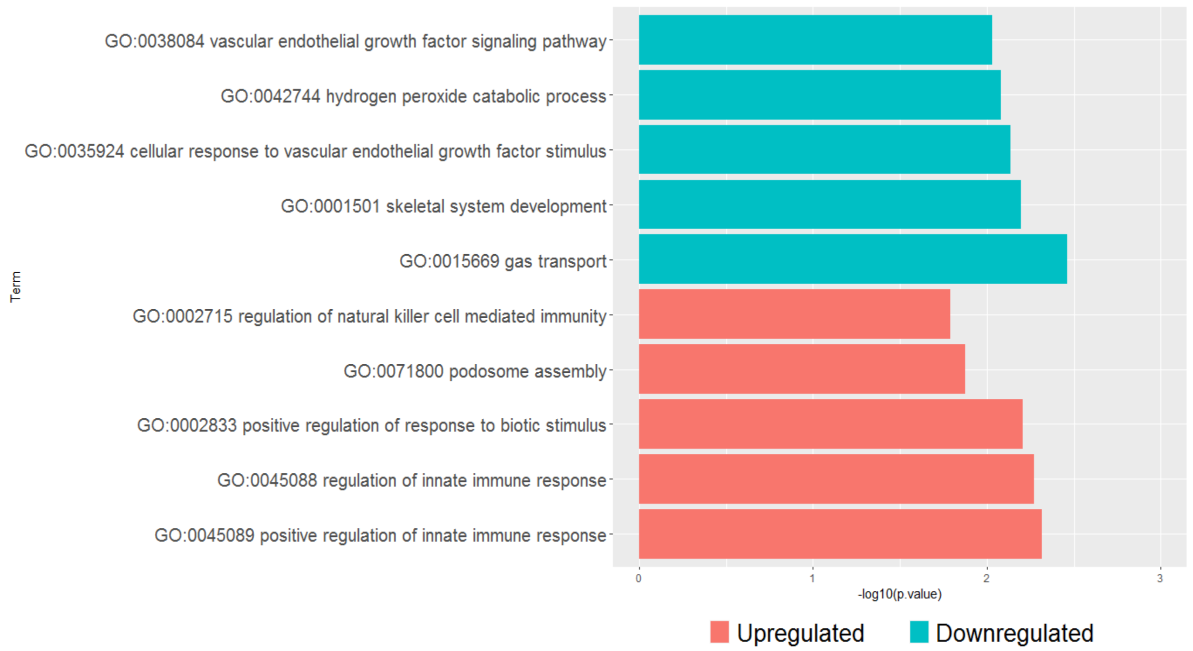


**Figure 4.2 Top 5 gene sets related to biological processes down and upregulated in most of the intermediate-poor in comparison with the intermediate-favorable subgroups with AML-M1.** The gene set analysis was performed in the biological processes category to identify the GO terms that were differently enriched between the subgroups analyzed. In blue are represented the GO terms that are downregulated and in red are represented the GO terms that were upregulated between the intermediate-poor and the intermediate-favorable subgroups.

Having identified the candidate genes, we performed a bibliographical analysis to find out if the 11 genes were already described in the literature (Annex I). We found that one gene, *TATDN3* was not previously referred in the PubMed literature. Five of the 11 identified genes (~ 46%) (*ZBED3*, *CLEC4G*, *SLC10A4*, *PBDC1*, and *ABCC9*) were already linked to other cancer types than leukemia. Three of the 11 genes (~ 27%) (*ARHGAP23*, *CLIC1*, and *NNMT*) were already associated to leukemia, but not in AML related publications. Finally, only 2 of the 11 genes (18.2%) (*MCM4* and *SNAP23*) were previously described in AML-related articles.

### 4.1.2 DNA methylation as a potential prognostic biomarker in patients with AML-M1 categorized in the intermediate prognostic risk group.

We were also interested into analyzing the potential of DNA methylation as a prognostic tool in patients with AML-M1 categorized in the intermediate prognostic risk group. For this, we selected the DNA methylation values of 485577 CpG sites, collected from the 24 AML patients (the same population used for the gene expression analysis). Our algorithm was applied to identify which CpG site methylation could predict survival. The 485577 CpG sites were reduced to 395845 CpG sites, after the application of the first phase of the algorithm (data preparation for analysis). Afterwards, the second phase of the algorithm (identification of the first potential prognostic biomarkers) reduced the 395845 CpG sites to 26606. Then, only 2051 CpG sites were selected by the third phase of the algorithm (confounding factor treatment). With the final phase completed (selection criteria), we finally identified 137 candidate CpG sites, localized in 130 genes, whose DNA methylation appeared to predict survival of AML-M1 patients categorized in the intermediate prognostic risk group (Annex III and IV). In the Table 4.2 are represented the top 15 of the identified CpG sites with a higher hazard ratio value.

**Table 4.2 List of 15 out of the 137 CpG sites identified whose DNA methylation appeared to predict survival in patients with AML-M1 categorized in the intermediate prognostic risk group.**

| **CpG sites** | ***p-value*** | **Optimal cutpoint** | **Group1 (n)** | **Group2 (n)** | **Age test (*p*-value)** | **Gene** | **HR** |
|---|---|---|---|---|---|---|---|
| cg06577205 | 0.00061 | 0.4191 | 10 | 14 | 0.1593 | *FBXL7* | 9.35265 |
| cg12523924 | 0.00099 | 0.2775 | 11 | 13 | 0.83907 | *HTR1A* | 8.85568 |
| cg27344859 | 0.00146 | 0.078 | 14 | 10 | 0.06057 | *MIR124-3* | 8.05598 |
| cg21385821 | 0.00323 | 0.2155 | 13 | 11 | 0.09243 | *CA10* | 7.08546 |
| cg00662963 | 0.00621 | 0.9328 | 12 | 12 | 0.24749 | *PRR23B* | 6.37657 |
| cg00664792 | 0.00189 | 0.5001 | 13 | 11 | 0.16374 | *SMOC2* | 6.34213 |
| cg08692733 | 0.00398 | 0.3127 | 10 | 14 | 0.197 | *RBM20* | 6.12141 |
| cg20708909 | 0.01303 | 0.7944 | 14 | 10 | 0.05296 | *OPCML* | 6.08624 |
| cg23385847 | 0.00242 | 0.9131 | 12 | 12 | 0.41821 | *CAMK4* | 6.05951 |
| cg07674139 | 0.00459 | 0.659 | 10 | 14 | 0.21813 | *NRK* | 5.71912 |
| cg25406755 | 0.00436 | 0.2133 | 10 | 14 | 0.22929 | *TFIP11* | 5.33466 |
| cg23073879 | 0.01856 | 0.454 | 14 | 10 | 0.66006 | *GALNT17* | 5.21274 |
| cg25024717 | 0.00910 | 0.4746 | 13 | 11 | 0.79401 | *HOXC13* | 5.00575 |
| cg23876072 | 0.01083 | 0.7818 | 12 | 12 | 0.1483 | *ANO1* | 4.86751 |
| cg10965508 | 0.01866 | 0.3989 | 14 | 10 | 0.09466 | *TTBK1* | 4.80965 |

HR, hazard ratio

As examples, Figure 4.3 shows the top 4 Kaplan-Meier overall survival curves with the highest hazard ratio value, obtained using the methylation cutpoints of 4 identified CpG sites localized in *FBXL7*, *HTR1A*, *MIR124-3*, and *CA10* genes, respectively. The survival curves hazard ratio values of the remaining CpG sites demonstrated localized in the *HTR1A*, *MIR124-3*, and *CA10* genes were 8.86, 8.06, and 7.09, respectively (Figure 4.3. b, c, d).

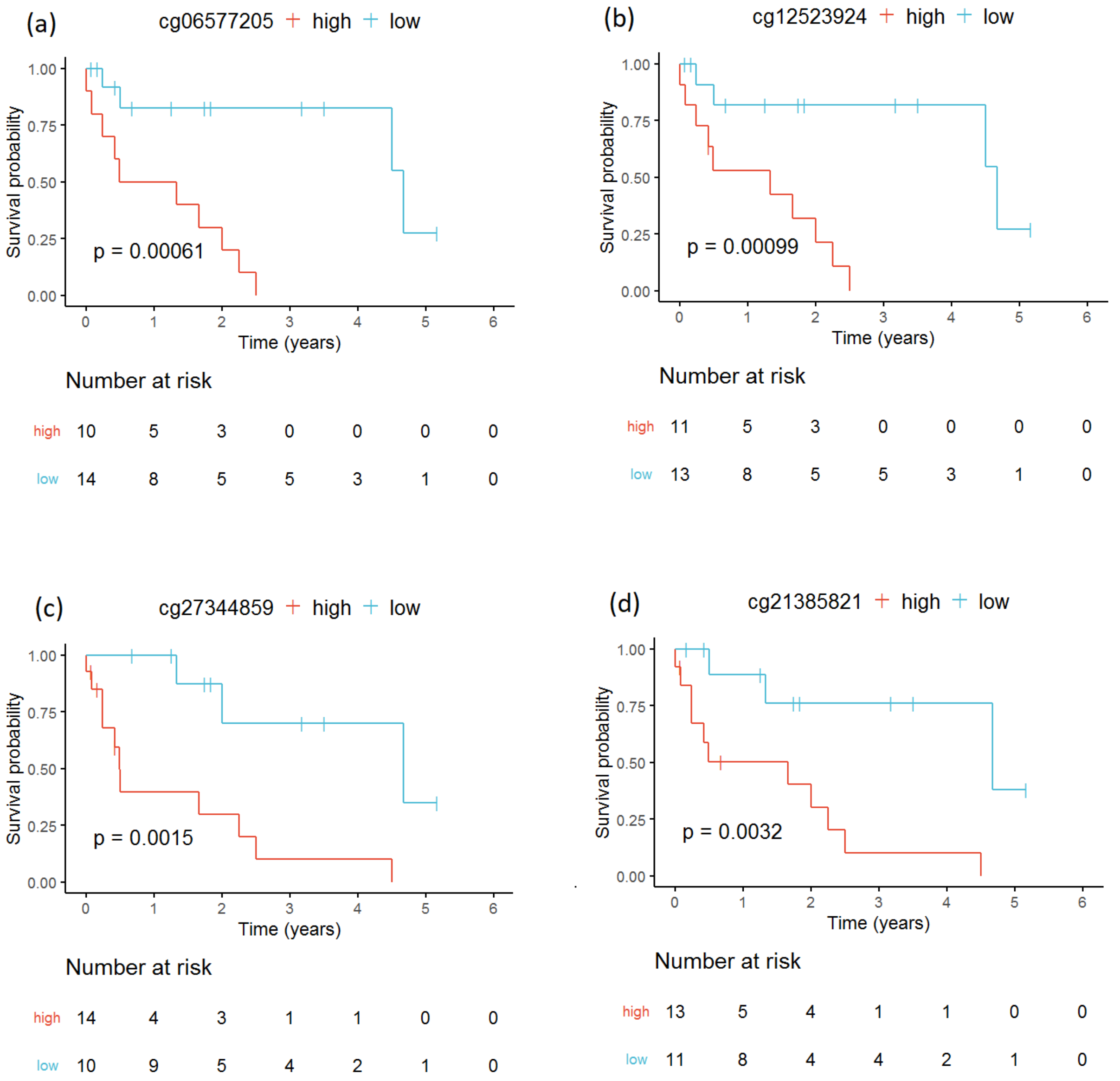


**Figure 4.3 Kaplan-Meier overall survival curves for four out of the 137 potential DNA methylation prognostic biomarkers identified for patients with AML-M1 categorized in the intermediate prognostic risk group.** Kaplan-Meier curves obtained using the methylation cutpoint of the (a) cg06577205 localized in the *FBXL7* gene (p = 0.00061, log-rank test) , (b) cg12523924 localized in the *HTR1A* gene (p = 0.00099, log-rank test), (c) cg27344859 localized in the *MIR124-3* gene (p = 0.0015, log-rank test), (d) cg21385821 localized in the *CA10* gene (p = 0.0032, log-rank test). The number at risk corresponds to the number of patients, at the indicated time point, that are still alive and whose follow-up continues.

Interestingly, the hypomethylation of each identified CpG site demonstrated in the Figure 4.3 was related with the subgroup of patients with better prognosis. In the remaining identified CpG sites, hypomethylation was related to the subgroup with better prognosis in 26 CpG sites. On the other hand, in 106 of the analyzed CpG sites, hypermethylation was related to the subgroup with better prognosis (Annex IV).

As the identified 137 CpG sites were localized in 130 genes, it means that there were genes with several CpG sites that were able to differentiate survival in the AML-M1 patients studied. In Table 4.3, are shown the 6 genes with more than one CpG site with prognostic value, the genomic location of the CpG as well as the methylation status of the CpG that was related to the subgroup with a better prognosis.

**Table 4.3 The 6 genes with more than one CpG site with prognostic value in AML-M1 patients categorized in the intermediate prognostic risk group.**

| Gene | CpG site | CpG location | Methylation status related to better prognosis |
|---|---|---|---|
| ***FOXK1*** | cg15176413 | Intron | hypermethylation |
| | cg26800803 | Intron | hypermethylation |
| ***HOXC13*** | cg17410650 | Distal Intergenic | hypomethylation |
| | cg25024717 | Distal Intergenic | hypomethylation |
| ***OPCML*** | cg10966440 | 1st Intron | hypermethylation |
| | cg20708909 | Promoter (<=1kb) | hypomethylation |
| ***PALLD*** | cg14069287 | Intron | hypermethylation |
| | cg05259872 | Promoter (<=1kb) | hypermethylation |
| ***PRDM1*** | cg04309234 | Distal Intergenic | hypermethylation |
| | cg03942932 | Distal Intergenic | hypermethylation |
| ***SIX3*** | cg11218954 | Distal Intergenic | hypermethylation |
| | cg10963518 | Distal Intergenic | hypermethylation |

Our results suggest that the *FOXK1* gene has two CpG sites localized in intronic regions, and hypermethylation of each of these CpG sites was related with the subgroup with better prognosis. Moreover, the *HOXC13* gene was another one with two CpG sites with a prognostic value in AML-M1 group of patients studied, both localized in the distal intergenic region, and the hypomethylation of each CpG site was related with the intermediate-favorable subgroup. The *OPCML* gene also seemed to have two CpG sites with a prognostic value. The cg10966440 localized in the first intron, for which hypermethylation was related with the subgroup with a better prognosis, and the cg20708909 localized in the promotor, for which hypomethylation was related with the subgroup with a better prognosis. In addition, the *PALLD* gene had the cg14069287 in an intronic region, and the cg05259872 in the promotor. The hypermethylation of each of these CpG sites was related to the better prognosis subgroup. Moreover, the *PRDM1* gene appeared to have two CpG sites localized in the distal intergenic region, for which hypermethylation was related with the subgroup with better prognosis in both cases. The same appeared to occur with the *SIX3* gene.

Analyzing the generated subgroups by each methylation cutpoint, we observed that most analyses generated a unique subdivision of AML patients. Nonetheless, the methylation cutpoint of two CpG sites (cg02436098 in the *TNFAIP8L1* gene, and cg11218954 in the *SIX3* gene) generated the same AML patients' subgroups.

We were also interested in knowing the genomic localization of the 137 identified CpG sites. We observed that the CpG's of interest were localized in promotors, 3'UTR, introns, exons, and in distal intergenic regions (Figure 4.4 and Annex IV). The promotor was the genomic localization where the majority of the identified CpG's, 62%, appeared to be located.

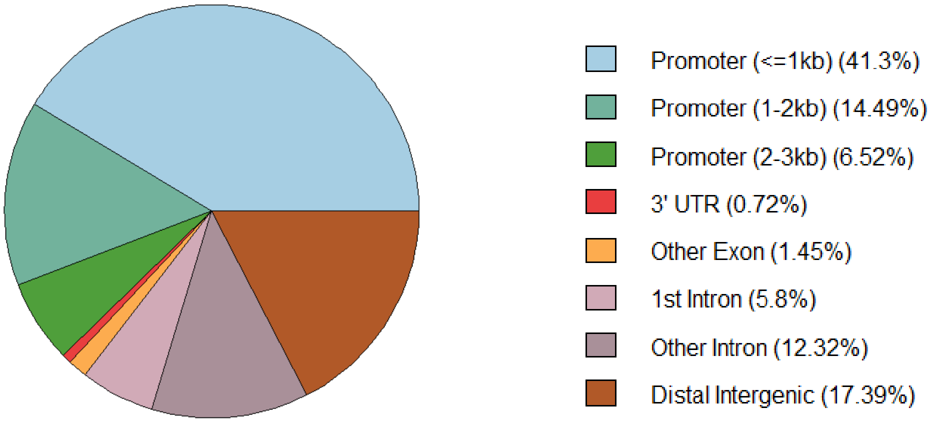


**Figure 4.4 Location of the identified CpG sites within the genes, whose methylation appeared to be a potential biomarker in AML-M1 patients categorized in the intermediate prognostic risk group.** The majority of the 137 CpG sites identified by our algorithm were localized in the promotor region (~ 62%). The second genomic region with more identified CpG's was the distal intergenic region. About 12% of the CpG's were in other intronic regions. The 3'UTR, other exons and first intron were the genomic regions with less identified CpG's (0.7%, 1.45%, and 5.8%, respectively).

To know if the 130 genes, where the 137 CpG sites identified were localized, were already described in the literature, a bibliographic analysis was performed (Annex IV). We found that 6 genes of interest (*MIR147A*, *REX1BD*, *FAM169B*, *PRR23B*, *EFCAB10*, and *C1orf53*) were never described in the PubMed literature. Moreover, 51 of the 130 identified genes (~39%) had already been studied in other cancer types than leukemia. Twenty-two of the 130 genes (~17%) had already been described in leukemia, but not in AML related articles. Finally, 46 of the 130 genes (~35%) had been described in AML related articles. The genes that we found to be cited in AML literature are: *ALK*, *TCF3, SKI*, *BMF*, *SOCS3*, *MICA*, *EBF1*, *PGF*, *HLX*, *TBL1XR1*, *HOXA1*, *CD1C*, *NRK*, *NIN*, *EVL*, *G0S2*, *USP18*, *BLNK*, *PRDM1*, *CA10*, *IRX2*, *NFIA*, *GBX2*, *PDLIM4*, *PTCH1*, *SCIN*, *SOX18*, *ADAMTS9*, *SETD1B*, *UGP2*, *DSCAML1*, *ARF6*, *PCDH9*, *DOK7*, *NRG1*, *PALLD*, *SNUPN*, *FGF6*, *HOXC9*, *CAPG*, *PRUNE2*, *CHIT1*, *PTPRG*, *PRKG1*, *WDR43*, and *EN1* genes.

Finally, we verified that the genes identified in the expression analysis (section 4.1.1) were different of the genes identified in the methylation analysis (section 4.1.2)

## 4.2. FAB M2 AML SUBTYPE

Our second population of interest was the group of patients with AML classified as FAB M2 subtype (AML-M2) categorized in the intermediate prognostic risk group. The TCGA LAML cohort contained gene expression, DNA methylation as well as clinical information data from 19 patients with this condition. The gene expression results are described in 4.2.1 section, and the DNA methylation results are described in 4.2.2 section.

### 4.2.1 Gene expression as a potential prognostic biomarker in patients with AML-M2 categorized in the intermediate prognostic risk group

The gene expression values for 20530 genes from the 19 patients with AML-M2 categorized in the intermediate prognostic risk group were analyzed by the developed algorithm in order to determine the genes whose expression could have a potential prognostic value in the population of study. Once the first algorithm phase (data preparation for analysis) was performed, the total number of genes was reduced from 20530 to 17280 genes. From these genes, just 1420 resulted after the application of the second algorithm phase (identification of the first potential prognostic biomarkers). Afterwards, the third phase (confounding factor treatment) led to the selection of 178 genes. In the final of the fourth phase, the algorithm identified 58 candidate genes whose expression appeared to predict survival in the population of patients with AML-M2 categorized in the intermediate prognostic risk group (Annex V and VI). The Table 4.4 shows the top 15 of the 58 identified genes with the highest hazard ratio value.

**Table 4.4 List of 15 of the 58 genes whose expression was able to predict survival in patients with AML-M2 categorized in the intermediate prognostic risk group.**

| Gene | *p-value* | Optimal cutpoint | Group1 (n) | Group2 (n) | Age-test (*p-value*) | HR |
|---|---|---|---|---|---|---|
| *OSM* | 0.00011 | 7.49130 | 9 | 10 | 0.08613 | 14.64881 |
| *MAP1LC3B2* | 0.00224 | 7.59950 | 10 | 7 | 0.24099 | 13.10614 |
| *RNPEP* | 0.00020 | 10.40060 | 8 | 11 | 0.74096 | 12.52281 |
| *EIF1* | 0.00036 | 12.97910 | 10 | 8 | 0.50472 | 11.07338 |
| *SPATA2L* | 0.00253 | 6.70250 | 10 | 7 | 0.18714 | 8.21557 |
| *SF1* | 0.00080 | 12.81370 | 10 | 9 | 0.39085 | 7.44932 |
| *DNAJC1* | 0.00140 | 9.26950 | 11 | 8 | 0.12628 | 7.10914 |
| *FOSL1* | 0.00545 | 7.20840 | 11 | 8 | 0.23078 | 5.52843 |
| *EXOSC6* | 0.00288 | 8.43780 | 10 | 9 | 0.08613 | 5.48530 |
| *MAFF* | 0.00295 | 7.15470 | 9 | 10 | 0.25258 | 5.33656 |
| *IL3RA* | 0.00748 | 8.36750 | 11 | 8 | 0.59114 | 5.26753 |
| *HSD11B1L* | 0.00385 | 4.30000 | 10 | 9 | 0.13057 | 5.22570 |
| *CYP27B1* | 0.00430 | 4.54000 | 9 | 10 | 0.71306 | 5.07248 |
| *DDX27* | 0.00894 | 10.13880 | 10 | 9 | 0.53994 | 4.61328 |
| *NDUFS4* | 0.01547 | 8.48790 | 11 | 8 | 0.21510 | 4.42246 |

HR, hazard ratio

The top 4 Kaplan-Meier overall survival curves with the highest hazard ratio values are depicted in the Figure 4.5. The Kaplan-Meier curve obtained using the expression cutpoint of the *OSM* gene had the highest hazard ration value of 14.65 (Figure 4.5.a). The hazard ratio values of the survival curves obtained using the expression cutpoint of the *MAP1LC3B2*, *RNPEP*, and *EIF1* genes were 13.11, 12.52, and 11.07, respectively (Figure 4.5.b, c, d).

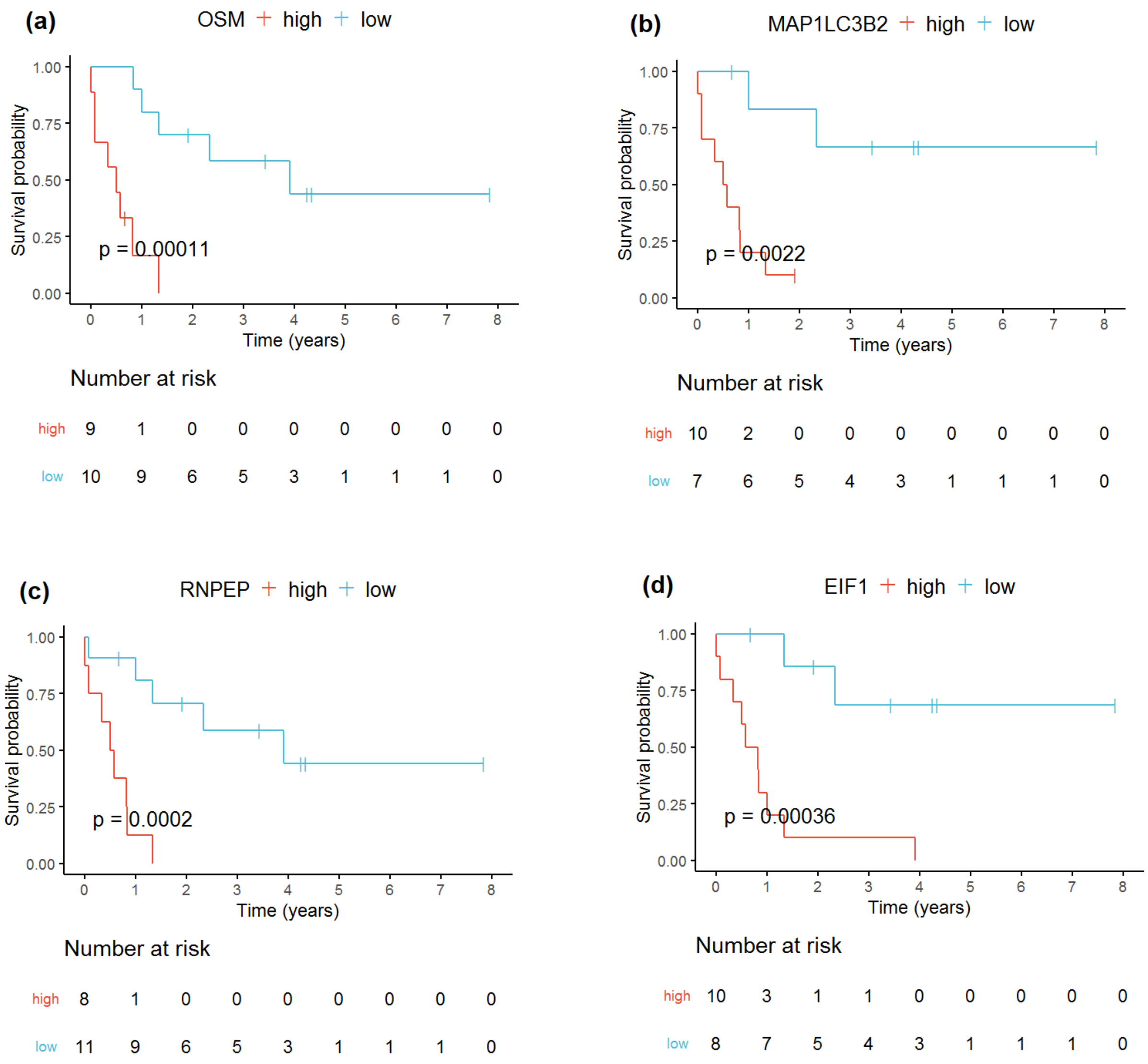


**Figure 4.5 Kaplan-Meier overall survival curves for four out of the 58 potential gene expression prognostic biomarkers identified for patients with AML-M2 categorized in the intermediate prognostic risk group.** Kaplan-Meier curve obtained using the expression cutpoint of the (a) the *OSM* gene (p = 0.00011, log-rank test) , (b) the *MAP1LC3B2* gene (p = 0.0022, log-rank test), (c) the *RNPEP* gene (p = 0.0002, log-rank test), (d) the *EIF1* gene (p = 0.00036, log-rank test). The number at risk corresponds to the number of patients, at the indicated time point, that are still alive and whose follow-up continues.

Relatively to the expression status, for each case, the low expression levels was related with the subgroup with better prognosis. For the remaining genes, the low expression levels were displayed by the better prognosis subgroup in 12 identified genes. In contrast, high expression levels were related with the better prognosis subgroup in 42 identified genes (Annex VI).

Most of the determined expression cutpoints subdivided the AML patients in a unique way. However, the expression cutpoints of two pairs of genes: *KIAA1217* and *EPB41L1* genes, and *EYA1* and *KLHL13* genes, generated the same subgroups of AML-M2 patients.

As made in the FAB M1 subtype analysis, the subgroups generated by the determined expression cutpoints were compared by the gene set analysis. The goal was to identify biological processes that were differentially enriched in the majority of the subgroups with worse prognosis (intermediate -poor) in comparison with the subgroups with better prognosis (intermediate – favorable) (Annex VII). In the Figure 4.6 are shown the top 5 gene sets that appeared to be down and upregulated in the majority of the intermediate-poor subgroups versus the intermediate-favorable subgroups.

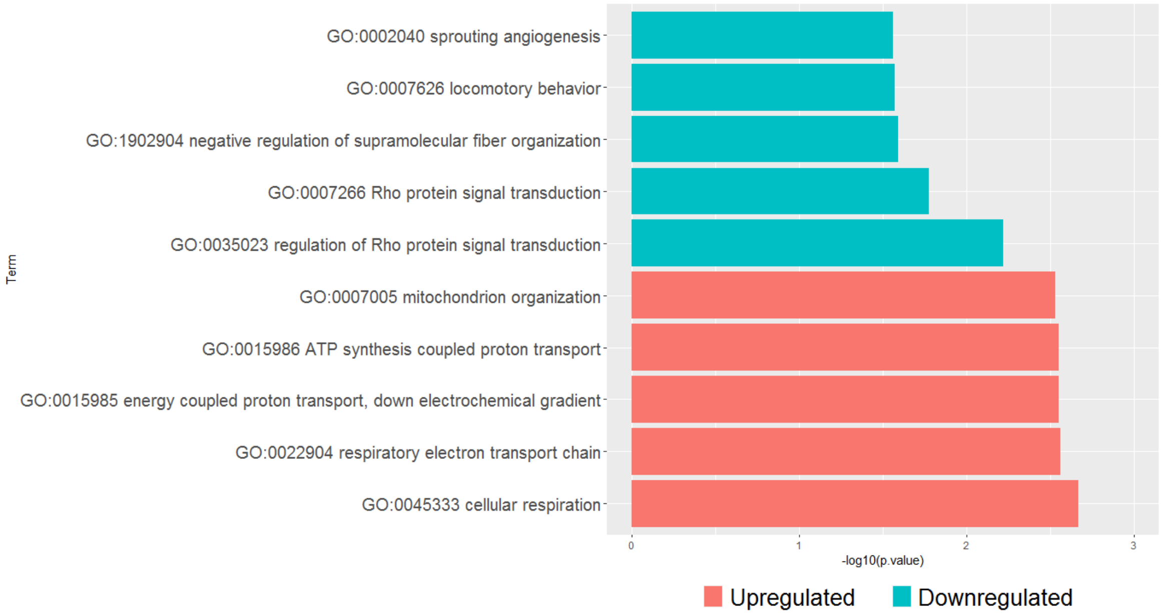


**Figure 4.6 Top 5 gene sets down and upregulated related to biological processes in most of the intermediate-poor in comparison with the intermediate-favorable subgroups with AML-M2.** Gene sets about biological processes were analyzed in order to identify which GO terms were differently enriched in the identified subgroup with worse prognosis in comparison with the identified subgroups with better prognosis. In blue are represented the GO terms that are downregulated and in red are represented the GO terms that are upregulated between the subgroups analyzed.

Our results suggest that biological processes related to sprouting angiogenesis, locomotor behavior, negative regulation of supramolecular fiber organization, Rho protein signal transduction, and regulation of Rho protein signal transduction are some examples of gene sets that appeared to be downregulated in the intermediate-poor when comparing with the intermediate-favorable subgroups.

On the other hand, mitochondrion organization, ATP synthesis coupled proton transport, energy coupled proton transport down electrochemical gradient, respiratory electron transport chain and cellular respiration were biological processes gene sets that seemed to be upregulated in the intermediate-poor in comparison with the intermediate-favorable subgroups.

Bibliographic analysis was performed to know if the 58 identified candidate genes were already described in the literature (Annex VI). We noticed that 4 identified genes (*ZNF732, ANKRD20A4*, *ZNF684*, and *SPATA2L*) were never described in the PubMed literature. Twenty-five of the 58 genes (~ 43%) already been described in other cancer types than leukemia. Seven of the 58 genes (~12%) had already described in leukemia, but not in AML related articles. Finally, 19 of the 58 genes (~33%) had been described in AML related articles. Sorting by the number of citations, the identified genes already described in the AML related articles are: *ABCB1*, *IL3RA*, *FLT1*, *OSM*, *SAMD9L*, *SF1*, *EPB41L1*, *AS3MT*, *CYP27B1*, *DNAJC1*, *SLA*, *FOSL1*, *RGS5*, *DPYD*, *DDO*, *EYA1*, *MAFF*, *PARP9*, and *CYBRD1* gene.

### 4.2.2 DNA methylation as a potential prognostic biomarker in patients with AML-M2 categorized in the intermediate prognostic risk group

Our algorithm was also applied in the DNA methylation dataset with DNA methylation values of 485577 CpG sites from the 19 patients with AML-M2 categorized in the intermediate prognostic risk group (same patients used in the previous analysis). Our goal was to determine which CpG sites that DNA methylation can be a potential prognostic biomarker to predict survival in AML-M2 patients with intermediate prognostic risk group. The first phase of the algorithm (data preparation for analysis) reduced the initial number of CpG sites to 395830. In its turn, the second algorithm phase (identification of the first potential prognostic biomarkers) decreased the CpG site number to 27067. Moreover, the third algorithm phase (confounding factor treatment) selected only 2519 CpG sites. In last, after the fourth algorithm phase (selection criteria), we identified 691 CpG sites whose DNA methylation seemed to have a predictive survival value in patients with AML-M2 categorized in the intermediate prognostic risk group (Annex VIII and Annex IX). We also verified that these CpG sites were localized in 592 genes. In the Table 4.5 are represented the top 15 identified CpG sites with the highest hazard ratio value.

**Table 4.5 List of 15 out of the 691 CpG sites whose DNA methylation was able to predict survival in patients with AML-M2 categorized in the intermediate prognostic risk group.**

| CpG sites | *p-value* | Optimal Cutpoint | Group1 (n) | Group2 (n) | Age test (*p-value*) | HR | Gene |
|---|---|---|---|---|---|---|---|
| cg00645383 | 0.00003 | 0.70540 | 10 | 9 | 0.59529 | 26.71490 | *SPACA7* |
| cg12899423 | 0.00006 | 0.55290 | 10 | 9 | 0.19103 | 26.67841 | *ALX4* |
| cg02939781 | 0.00009 | 0.66840 | 9 | 10 | 0.30701 | 23.69147 | *KLHL6* |
| cg23357198 | 0.00008 | 0.37990 | 11 | 8 | 0.77238 | 23.24957 | *PTPRT* |
| cg17264618 | 0.00008 | 0.55310 | 11 | 8 | 0.77238 | 23.24957 | *ENTPD3* |
| cg25824217 | 0.00011 | 0.56860 | 10 | 9 | 0.26992 | 22.94440 | *HLA-DPA1* |
| cg14178336 | 0.00011 | 0.36630 | 10 | 9 | 0.26992 | 22.94440 | *CALN1* |
| cg00622702 | 0.00011 | 0.54610 | 10 | 9 | 0.26992 | 22.94440 | *IFNAR1* |
| cg06321596 | 0.00011 | 0.27450 | 10 | 9 | 0.26992 | 22.94440 | *XYLT1* |
| cg08708961 | 0.00011 | 0.06960 | 10 | 9 | 0.26992 | 22.94440 | *PSEN2* |
| cg12224030 | 0.00011 | 0.13330 | 10 | 9 | 0.26992 | 22.94440 | *DLX4* |
| cg14396892 | 0.00009 | 0.26320 | 8 | 11 | 0.96704 | 22.58527 | *MIR4291* |
| cg08198176 | 0.00009 | 0.34150 | 9 | 10 | 0.90244 | 22.10625 | *ITGA9* |
| cg03572859 | 0.00009 | 0.31330 | 9 | 10 | 0.90244 | 22.10625 | *SORBS3* |
| cg17142470 | 0.00009 | 0.56250 | 9 | 10 | 0.90244 | 22.10625 | *SORBS3* |

HR, hazard ratio

In the Figure 4.7 are shown as examples, the top 4 Kaplan-Meier overall survival curves with the highest hazard ratio value, obtained using the methylation cutpoint of 4 identified CpG sites localized in the *SPACA7, ALX4, KLHL6,* and *PTPRT* genes, respectively. The survival curve obtained using the methylation cutpoint of the cg00645383 had the highest hazard ratio value of 26.71 (Figure 4.7.a). The hazard ratio values associated with the survival curves of the remaining demonstrated CpG sites localized in the *ALX4, KLHL6,* and *PTPRT* genes were 26.68, 23.69, and 23.25, respectively (Figure 4.7. b, c, d).

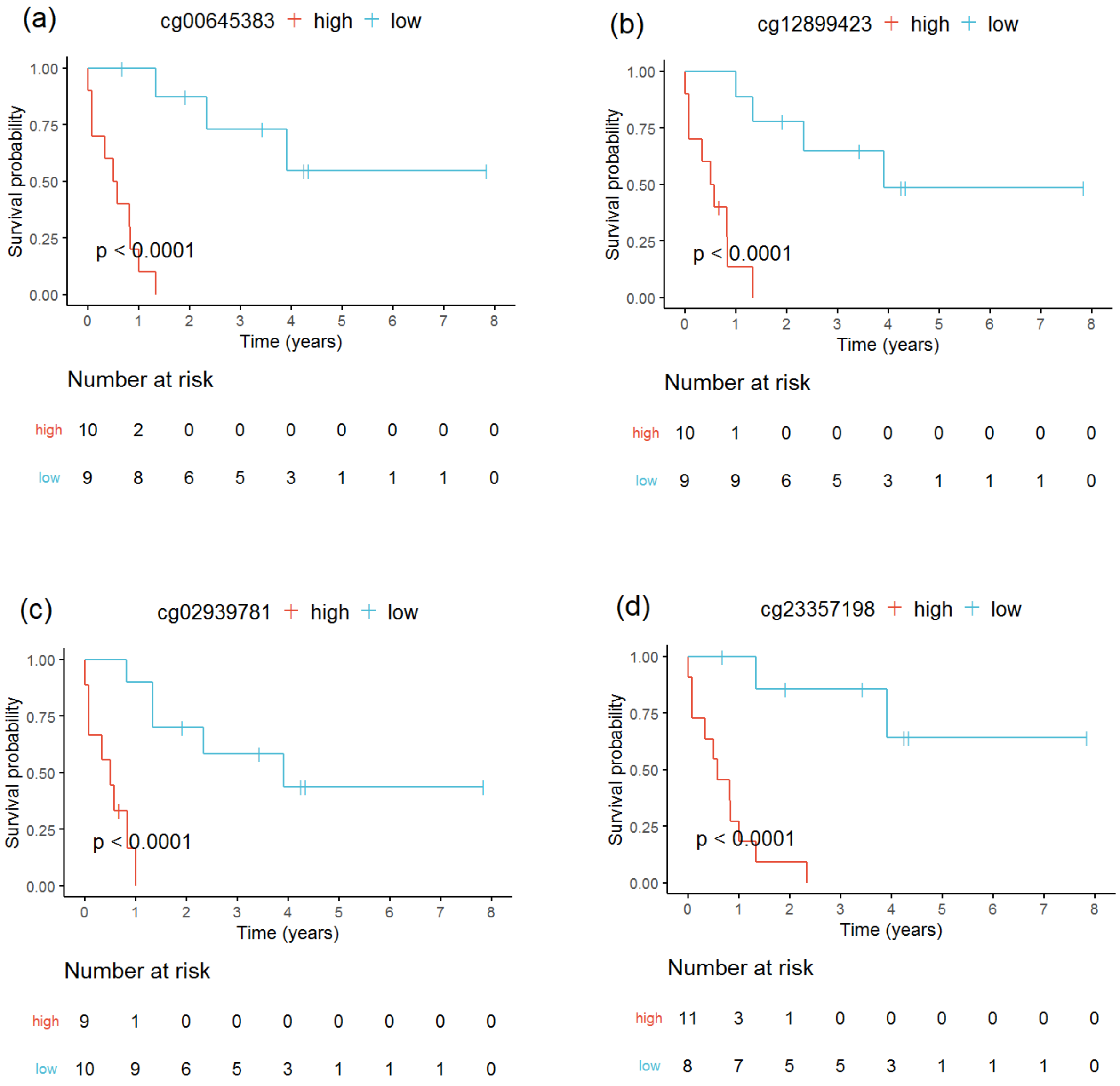


**Figure 4.7 Kaplan-Meier overall survival curves for four out of the 691 potential DNA methylation prognostic biomarkers identified for patients with AML-M2 categorized in the intermediate prognostic risk group.** Kaplan-Meier curve obtained using the methylation cutpoint of the (a) cg00645383 localized in the *SPACA7*gene (p = 0.00003, log-rank test), (b) cg12899423 localized in the *ALX4* gene (p = 0.00006, log-rank test), (c) cg02939781 localized in the *KLHL6* gene (p = 0.00009, log-rank test), (d) cg23357198 localized in the *PTPRT* gene (p = 0.00008, log-rank test). The number at risk corresponds to the number of patients, at the indicated time point, that are still alive and whose follow-up continues.

In each demonstrated survival curve, the hypomethylation levels were related with the subgroup with a better prognosis (Figure 4.7). In the remaining cases, hypomethylation was displayed by the subgroup with a better prognosis, in 191 identified CpG sites. On the other hand, the hypermethylation was related to the subgroup with a better prognosis, in 496 identified CpG sites (Annex IX).

As mentioned, the 691 identified CpG sites with a prognostic value were localized in 592 genes, which means that there are genes with more than one CpG site that were able to predict survival in the AML-M2 patients studied. We observed that there were 90 genes with more than one relevant CpG site (Annex X). In the majority of these cases the several CpG sites localized in the same gene, their methylation status that confer a better prognosis was the same. For example, the cg25880242 and cg21943117 were both localized in the distal intergenic region of *ACTRT2* gene, and the hypermethylation of each CpG site was related with the better prognosis subgroup.

**Table 4.6 Examples of 5 genes with more than one CpG site with prognostic value in AML-M2 patients categorized in the intermediate prognostic risk group.**

| Gene | CpG site | Location of the CpG site | Methylation status related to better prognosis |
|---|---|---|---|
| ***ACTRT2*** | cg25880242 | Distal Intergenic | hypermethylation |
| | cg21943117 | Distal Intergenic | hypermethylation |
| ***AHRR*** | cg12806681 | Intron | hypermethylation |
| | cg03991871 | Intron | hypermethylation |
| ***AXL*** | cg03247049 | Promoter (<=1kb) | hypomethylation |
| | cg14892768 | Promoter (<=1kb) | hypomethylation |
| ***AZU1*** | cg14663914 | Promoter (<=1kb) | hypermethylation |
| | cg17823175 | Promoter (<=1kb) | hypermethylation |
| | cg16643542 | Promoter (<=1kb) | hypermethylation |
| | cg02147126 | Promoter (<=1kb) | hypermethylation |
| | cg15610437 | Promoter (<=1kb) | hypermethylation |
| ***B3GALT4*** | cg21618521 | Promoter (<=1kb) | hypomethylation |
| | cg17103217 | Promoter (<=1kb) | hypomethylation |

Moreover, we were also interested to know the genomic localization of the 691 identified CpG sites. We observed that the identified CpG sites were localized in genomic regions such as promotor, 5´UTR, 3’UTR, other exon, first intron, other intron, downstream, and distal intergenic (Figure 4.8 and Annex IX). The promotor was the genomic region with more CpG sites (~ 47%), and the distal intergenic follows with more CpG sites (~ 22%).

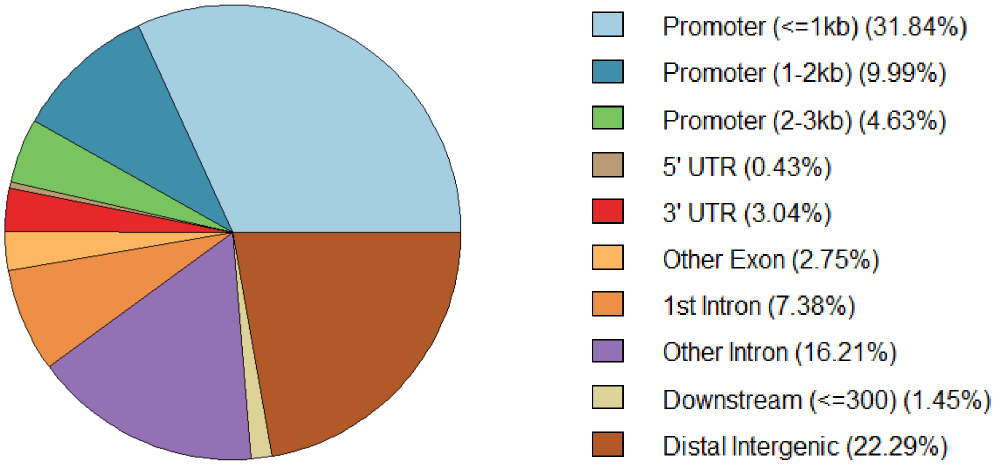


**Figure 4.8 Location of the identified 691 CpG sites within the genes, whose methylation appeared to be a potential biomarker in AML-M2 patients categorized in the intermediate prognostic risk group.** The majority of the 691 CpG sites identified by our algorithm were localized in the promotor region (~ 47%). The second genomic region with more identified CpG's was the distal intergenic region. About 16% of the CpG's were the other intron regions. The first intron, 3'UTR, other exon, downstream, and 5'UTR were the genomic regions with less identified CpG's (7.38%, 3.04%, 2.75%, 1.45%, and 0.43%, respectively).

In addition, we performed a bibliographical analysis to evaluate if the identified 592 genes, where the identified CpGs sites were located, were already described in the literature (Annex IX). We found that 22 genes (*MIR4710*, *LRRC14B*, *ZNF678*, *CPNE9*, *PNMA8B*, *SNORD115-37*, *C17orf102*, *SNORA63*, *MTRNR2L7*, *OR11H12*, *SNORD115-40*, *LCE2A*, *FAM216B*, *MIR4278*, *TMEM211*, *C9orf62*, *MIR4291*, *SPEM2*, *MIR4472-1*, *C2orf27B*, *C11orf88*, and *CEP170B*) were never described in the PubMed literature. Also, 235 of the 592 genes (~ 40%) had already been described in other cancer types than leukemia. Moreover, 118 of the 592 genes (~20%) had already described in leukemia, but not in AML related articles. Finally, 180 of the 592 genes (~30%) had been described in AML related articles.

Comparing the expression and DNA methylation results, we observed that were two genes whose expression and DNA methylation were able to predict survival in the AML-M2 patients. These genes were FAM234A and KIAA1217 genes.

### 4.3 FAB M4 AML SUBTYPE

Our third population of interest was the group of patients with AML classified as FAB M4 subtype (AML-M4) categorized in the intermediate prognostic risk group. Thereby, we extracted the gene expression, DNA methylation and the clinical information datasets referring to the patients with the desired characteristics, representing our population of study with 24 patients. Both gene expression and DNA methylation datasets were analyzed by the developed method individually. The gene expression results are described in 4.3.1 section, and the DNA methylation results are described in 4.3.2 section.

#### 4.3.1 Gene expression as a potential prognostic biomarker in patients with AML-M4 categorized in the intermediate prognostic risk group

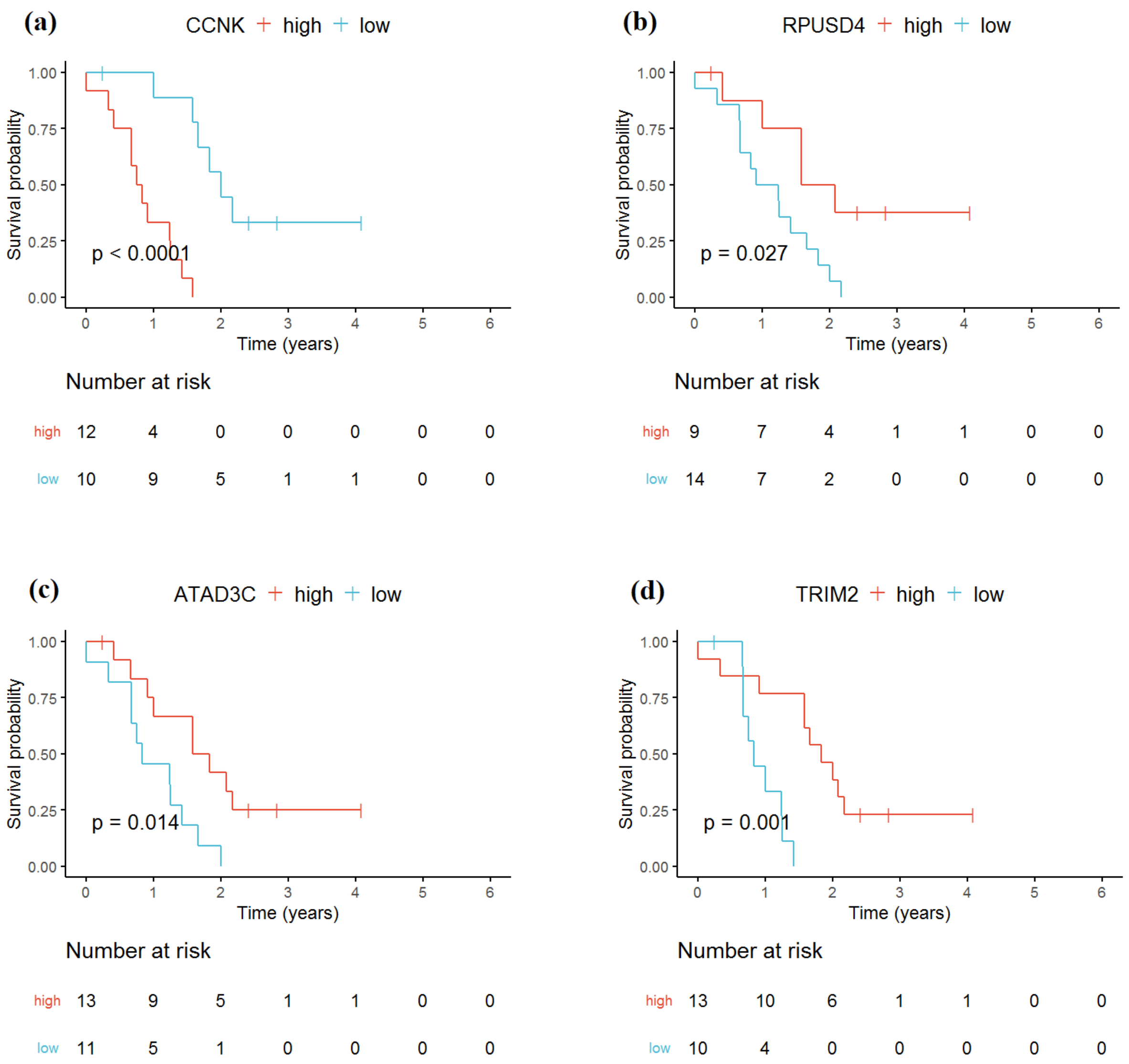
The 20530 initial genes were reduced to 17347 genes after performing the first algorithm phase (data preparation for analysis). After that, the second algorithm phase (identification of first potential prognostic biomarkers) was applied and 1173 genes were selected. Furthermore, the third algorithm phase (confounding factor treatment) reduced the 1173 genes to 115. The fourth algorithm phase (selection criteria) allowed the identification of 4 candidate genes whose expression appeared to predict survival in patients with AML-M4 categorized in the intermediate prognostic risk group (Table 4.7 and Annex XI).

**Table 4.7 List of the 4 genes whose expression was able to predict survival in patients with AML-M4 categorized in the intermediate prognostic risk group.**

| Gene | *p-value* | Optimal cutpoint | Group1 (n) | Group2 (n) | Age-test (*p-value*) | HR |
|---|---|---|---|---|---|---|
| *CCNK* | 0.00003 | 10.2570 | 12 | 10 | 0.2762 | 24.6115 |
| *RPUSD4* | 0.02652 | 8.8173 | 9 | 14 | 0.1225 | 0.3192 |
| *ATAD3C* | 0.01391 | 2.6992 | 13 | 11 | 0.9538 | 0.3091 |
| *TRIM2* | 0.00100 | 5.2940 | 13 | 10 | 0.5554 | |

HR, hazard ratio

The Figure 4.9 shows the top 4 Kaplan-Meier overall survival curves with the highest hazard ratio values, obtained using the determined expression cutpoints of *CCNK*, *RPUSD4,* *ATAD3C*, and *TRIM2* genes, respectively. The survival curve of *CCNK* gene had the highest hazard ratio value of 24.6. The hazard ratio values of the survival curves obtained using the expression cutpoint of the *RPUSD4* and *ATAD3C* genes were 0.32 and 0.31, respectively (Figure 4.9.b, c). The patient subgroups generated based on the expression cutpoint of the *TRIM2* gene did not have proportional hazards. For this reason, the hazard ratio was not calculated.



(a) Number at risk

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| high | 12 | 4 | 0 | 0 | 0 | 0 | 0 |
| low | 10 | 9 | 5 | 1 | 1 | 0 | 0 |

(b) Number at risk

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| high | 9 | 7 | 4 | 1 | 1 | 0 | 0 |
| low | 14 | 7 | 2 | 0 | 0 | 0 | 0 |

(c) Number at risk

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| high | 13 | 9 | 5 | 1 | 1 | 0 | 0 |
| low | 11 | 5 | 1 | 0 | 0 | 0 | 0 |

(d) Number at risk

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| high | 13 | 10 | 6 | 1 | 1 | 0 | 0 |
| low | 10 | 4 | 0 | 0 | 0 | 0 | 0 |

**Figure 4.9 Kaplan-Meier overall survival curves for the four potential prognostic biomarkers identified for patients with AML-M4 categorized in the intermediate prognostic risk group.** Kaplan-Meier curve obtained using the expression cutpoint of the (a) *CCNK* gene (p = 0.00003, log-rank test), (b) *RPUSD4* gene (p = 0.027, log-rank test), (c) *ATAD3C* gene (p = 0.014, log-rank test), (d) *TRIM2* gene (p = 0.001, two-stage test). The number at risk corresponds to the number of patients, at the indicated time point, that are still alive and whose follow-up continues.

Then, the low expression of *CCNK* gene seemed to be related with a better prognosis (Figure 4.9. a). The other identified genes, *RPUSD4*, *ATAD3C*, and *TRIM2*, showed high expression in the subgroup of AML patients with a better prognosis (Figure 4.9. b, c, d).

In addition, we also evaluated how the patients were subdivided based on each determined expression cutpoint. We saw that the patients are categorized in the intermediate-favorable or intermediate-poor subgroups depending on the expression cutpoint used. That is, different cutpoints do not generate the same patient´s subgroups with significant differences in overall survival.

As performed in the FAB M1 and FAB M2 AML subtypes analysis, we also evaluated if there were genes sets related to biological processes that were commonly enriched in the intermediate-poor subgroups in comparison with the intermediate-favorable subgroups, generated based on each determined expression cutpoint. We identified some biological processes that seemed to be down and upregulated in the intermediate-poor in comparison with the intermediate-favorable subgroups (Annex XII). In Figure 4.10 the top 5 down and upregulated gene sets are shown.

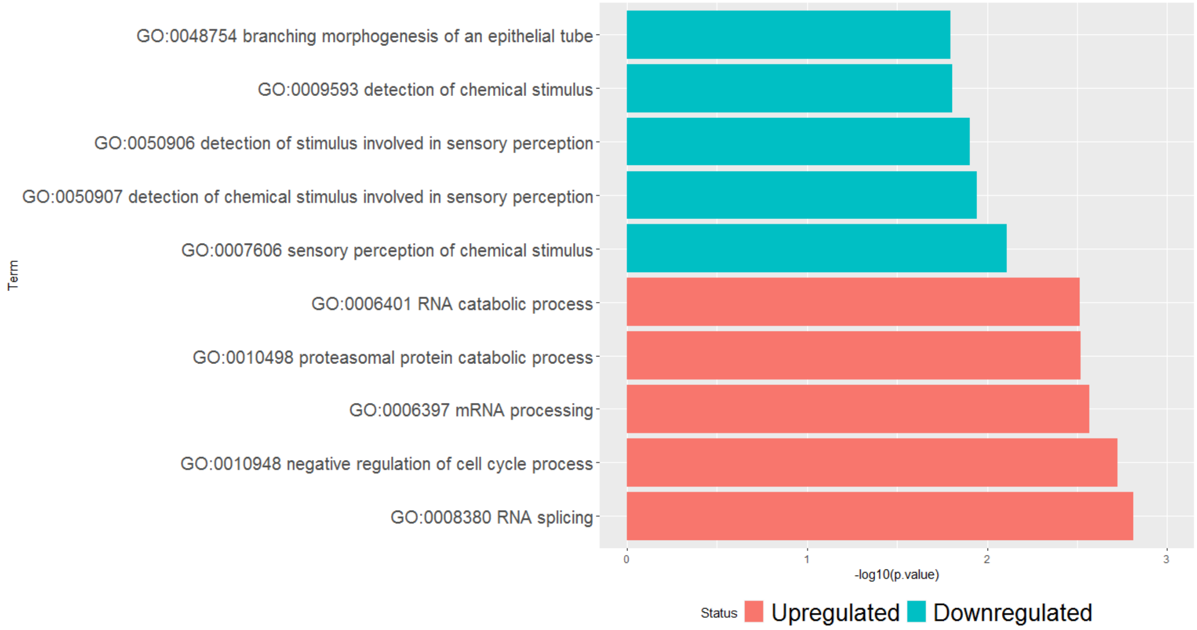


**Figure 4.10 Top 5 GO terms, in biological processes category, down and upregulated between the intermediate-poor and the intermediate-favorable groups with AML-M4.** Biological processes category was studied based on the gene set analysis to identify the GO terms that were differently enriched between the subgroups analyzed. In blue are represented the GO terms that are downregulated and in red are represented the GO terms that are upregulated between the intermediate-poor and the intermediate-favorable identified subgroups.

Branching morphogenesis of an epithelial tube, detection of chemical stimulus, detection of stimulus involved in sensory perception, detection of chemical stimulus involved in sensory perception, and sensory perception of chemical stimulus are the top 5 biological processes that seemed to be downregulated in the intermediate-poor subgroups. In contrast, RNA catabolic process, proteasomal regulation of cell cycle process, negative regulation of cell cycle process, and RNA splicing seemed to be upregulated in the intermediate-poor subgroups versus the intermediate-favorable subgroups.

Furthermore, we performed a bibliographical analysis to evaluate if the 4 identified candidate genes were already described in the literature (Annex XI). We found that all candidate genes had already been referred in the PubMed literature, specifically in cancer related articles but only 1 of the 4 genes (*CCNK*) had already been described in leukemia and AML related articles.

### 4.3.2 DNA Methylation as a potential prognostic biomarker in patients with AML-M4 categorized in the intermediate prognostic risk group

The DNA methylation dataset with records of 485577 CpG sites from the 24 patients with AML-M4 categorized in the intermediate prognostic risk group was analyzed by the developed algorithm. In the first algorithm phase (data preparation for analysis), the 485577 CpG sites were reduced to 395987. Next, by performing the second algorithm phase (identification of first potential prognostic biomarkers), 24998 CpG sites were selected. Moreover, the third algorithm phase (confounding factor treatment) reduced the CpG sites to 1869. As the last step, the fourth algorithm phase (selection criteria) was performed and 375 CpG sites were identified as potential prognostic biomarkers whose DNA methylation appeared to predict survival in patients with AML-M4 categorized in the intermediate prognostic risk group (Annex XIII and Annex XIV). The 375 identified CpG sites were localized in 330 genes. The top 15 identified CpG sites with highest hazard ratio value are shown in the Table 4.8.

**Table 4.8 List of 15 out of the 375 CpG sites whose DNA methylation was able to predict survival in patients with AML-M4 categorized in the intermediate prognostic risk group.**

| CpG site | *p-value* | Optimal cutpoint | Group1 (n) | Group2 (n) | Age-test (*p-value*) | Gene | HR |
|---|---|---|---|---|---|---|---|
| cg25914522 | 0.00221 | 0.8594 | 10 | 7 | 0.49399 | *RBMY1F* | 13.90550 |
| cg04929022 | 0.00011 | 0.7961 | 11 | 12 | 0.30939 | *VGLL1* | 12.25562 |
| cg06294373 | 0.00013 | 0.5250 | 10 | 11 | 0.62173 | *UMOD* | 11.97244 |
| cg06394109 | 0.00028 | 0.4588 | 14 | 9 | 0.94972 | *C1QTNF8* | 11.52809 |
| cg14312439 | 0.00015 | 0.2317 | 14 | 9 | 0.23089 | *CCR3* | 11.48455 |
| cg17560693 | 0.00023 | 0.7736 | 11 | 11 | 0.39277 | *TBX22* | 11.21738 |
| cg16803185 | 0.00037 | 0.7361 | 13 | 10 | 0.35176 | *PLS3* | 10.66215 |
| cg15467834 | 0.00005 | 0.8216 | 13 | 10 | 0.82799 | *CRYAA* | 10.60031 |
| cg19720260 | 0.00006 | 0.2361 | 13 | 10 | 0.66389 | *NXPE2* | 10.11071 |
| cg21230162 | 0.00003 | 0.6727 | 9 | 14 | 0.05445 | *CRIPAK* | 9.98054 |
| cg21329507 | 0.00003 | 0.7483 | 9 | 14 | 0.15597 | *TMEM255A* | 9.68735 |
| cg25946304 | 0.00008 | 0.5053 | 10 | 13 | 0.05830 | *S1PR5* | 9.55146 |
| cg13185308 | 0.00008 | 0.5694 | 10 | 13 | 0.05830 | *ABCC8* | 9.55146 |
| cg13149245 | 0.00075 | 0.3050 | 14 | 9 | 1.00000 | *PIF1* | 9.36223 |
| cg22405653 | 0.00003 | 0.5706 | 9 | 14 | 0.52833 | *KRTAP21-1* | 9.21881 |

HR, hazard ratio

In Figure 4.12 are shown the top 4 Kaplan-Meier overall survival curves with the highest hazard ratio value, obtained using the methylation cutpoint of 4 identified CpG sites localized in the *RBMY1F*, *VGLL1*, *UMOD*, and *C1QTNF8* genes, respectively. The survival curve obtained using the methylation cutpoint of the cg25914522 localized in the *RBMY1F* gene had the hazard ratio value of 13.91 (Figure 4.11. a). The hazard ratio values associated with the survival curves of the remaining demonstrated CpG sites localized in the *VGLL1*, *UMOD*, and *C1QTNF8* genes were 12.26, 11.97, and 11.53, respectively (Figure 4.11. b, c, d).

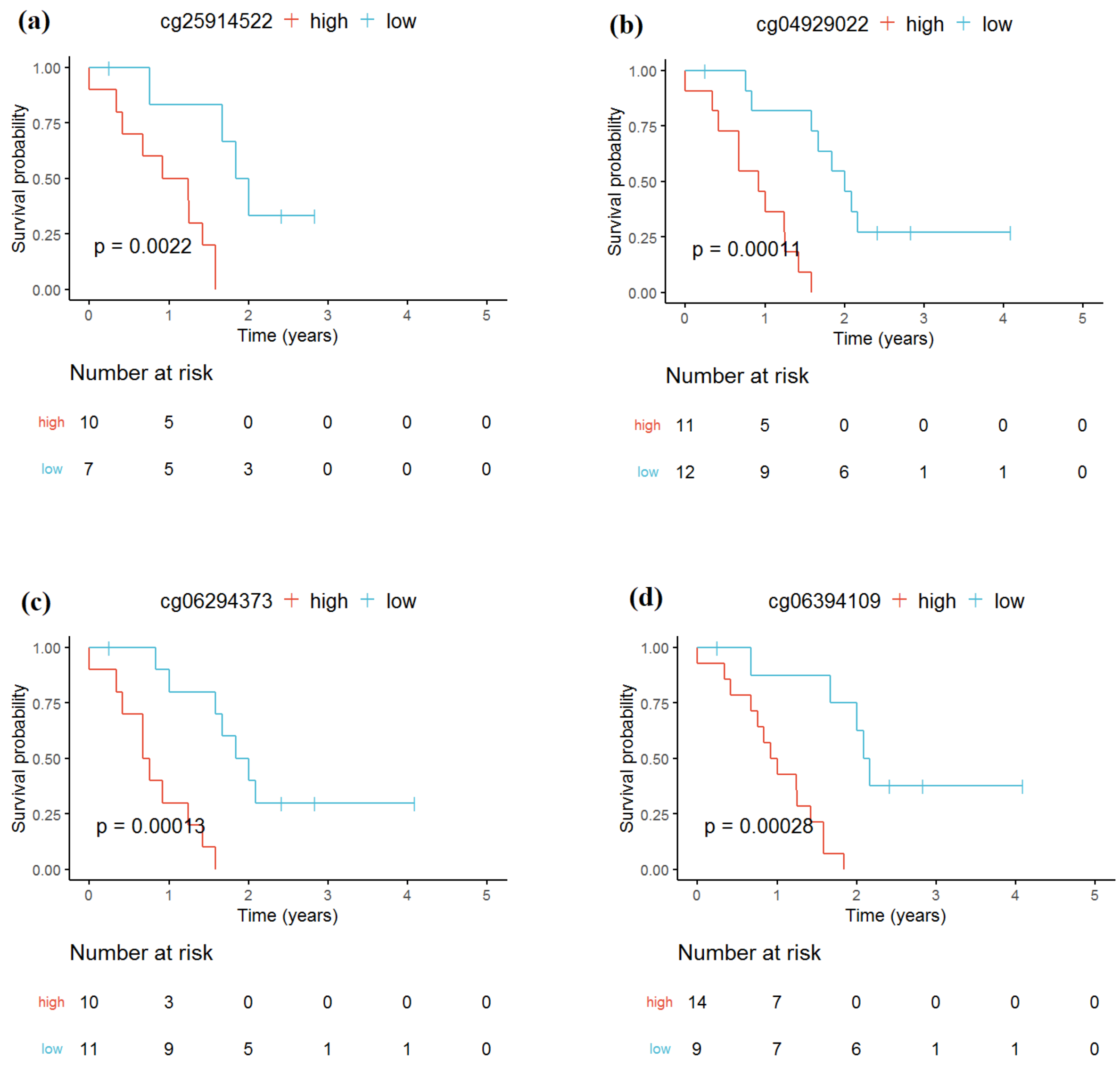


**Figure 4.11 Kaplan-Meier overall survival curves for four out of the 375 potential DNA methylation prognostic biomarkers identified for patients with AML-M4 categorized in the intermediate prognostic risk group.** Kaplan-Meier curve obtained using the methylation cutpoint of the (a) cg25914522 localized in the *RBMY1F* gene (p = 0.0022, log-rank test), (b) cg04929022 localized in the *VGLL1* gene (p = 0.00011, log-rank test), (c) cg06294373 localized in the *UMOD* gene (p = 0.00013, log-rank test) and (d) cg06394109 localized in the *C1QTNF8* gene (p = 0.00028, log-rank test). The number at risk corresponds to the number of patients, at the indicated time point, that are still alive and whose follow-up continues.

The hypomethylation of each CpG site demonstrated in the Figure 4.11 was related with the better prognostic group. In the remaining identified CpG sites, hypomethylation was related with the better prognostic risk group in 191 CpG. Instead, the hypermethylation was related with the better prognostic group in 179 CpG sites (Annex XIV).

We also observed that there were some identified CpG sites that their methylation cutpoint generated the same subgroups of patients (Annex XIV). For example, the cg25946304 localized in the *S1PR5* gene and the cg13185308 localized in the *ABCC8* gene, based on their methylation cupoints, the AML patients were subcategorized in the same two subgroups. Another example was the cg14988083 localized in the *FGFR3* gene, and the cg02593884 localized in the *FLYWCH1* gene whose methylation cutpoint generated the same patients subgroups.

As previously mentioned, the 375 identified CpG sites were localized in 330 genes. We observed that there were 45 genes with more than one CpG site with a prognostic value in AML-M4 patients (Annex XIV). In the majority of these genes, the methylation status of the various CpG localized in the same gene related to the subgroup with better prognosis was the same. In the Table 4.9 are shown 5 genes as examples. For instance, the hypermethylation of the cg16639692 in the promotor region and the cg10994149 in the downstream region of the *ANXA2R* gene are both related with the better prognostic group.

**Table 4.9 Examples of 5 genes with more than one CpG site as a potential prognostic biomarker in AML-M4 patients.**

| Gene | CpG site | CpG location | Methylation status related to better prognostic |
|---|---|---|---|
| *ANXA2R* | cg16639692 | Promoter (2-3kb) | hypermethylation |
| | cg10994149 | Downstream (<=300) | hypermethylation |
| *AR* | cg05019001 | Promoter (2-3kb) | hypermethylation |
| | cg05786601 | Promoter (<=1kb) | hypermethylation |
| *ARHGAP6* | cg03536032 | Promoter (<=1kb) | hypermethylation |
| | cg27166673 | Promoter (<=1kb) | hypermethylation |
| *ARX* | cg06943593 | Intron | hypermethylation |
| | cg02938958 | Exon | hypermethylation |
| | cg16414561 | Promoter (<=1kb) | hypermethylation |
| *BCOR* | cg24508310 | 1st Intron | hypomethylation |
| | cg07764473 | Promoter (<=1kb) | hypermethylation |
| | cg03161453 | 1st Intron | hypomethylation |
| | cg23496314 | Promoter (<=1kb) | hypermethylation |

The genomic region of the 375 identified CpG sites were also evaluated. We observed that the identified CpG sites were localized in promotor, 5´UTR, 3’UTR, first exon, other exon, first intron, other intron, downstream, and distal intergenic (Figure 4.12 and Annex XIV). The promotor was the first genomic region with more CpG sites located (~ 69%), and the distal intergenic the second genomic region with more CpG located (~ 14%).

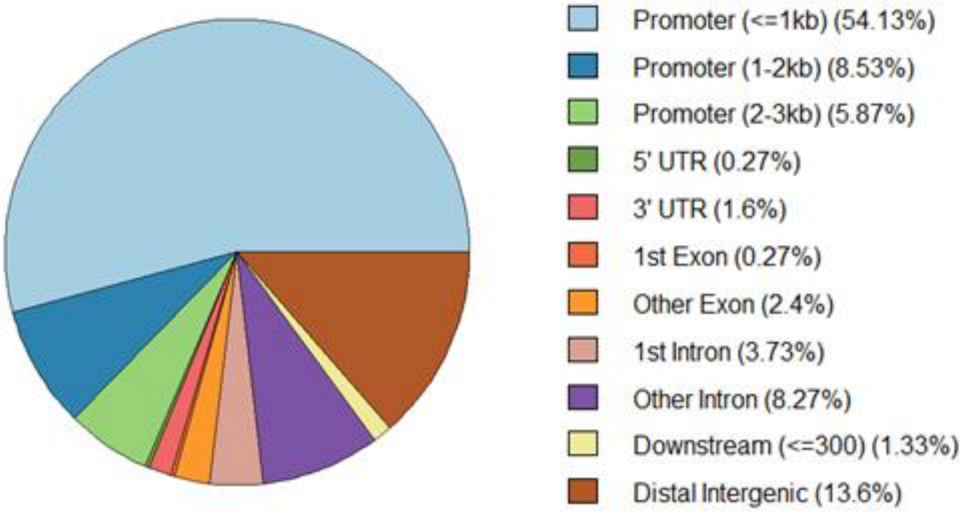


**Figure 4.12 Location of the identified CpG sites within the genes, whose methylation appeared to be potential biomarker in AML-M4 patients categorized in the intermediate prognostic risk group.** The majority of the 375 CpG sites identified by our algorithm were localized in the promotor region (~ 69%). The second genomic region with more identified CpG’s was the distal intergenic region. About 8% of the CpG’s were other intronic regions. The first intron, other exons, 3’UTR, downstream, 5’UTR, first exon were the genomic regions with less identified CpG’s (3.7%, 2.4%, 1.6%, 1.33%, 0.27%, 0.27%, and respectively). The other intron means that the CpG’s were localized in other introns than the first intron. The other exon region refers exons than the first exon.

Furthermore, we also performed bibliographic analysis to evaluate if the 330 genes, where the identified CpG sites were located, were already described in the literature (Annex XIV). We found that 4 of 330 genes (~ 1%) (*GLOD5*, *C10orf95*, *LONRF3*, and *TCEAL9*) were not already referred in PubMed articles. In addition, 138 of the 330 genes (~ 42%) had already been described in other cancer types than leukemia. Moreover, 72 of the 330 genes (~22%) had already been described in leukemia, but not in AML related articles. At last, 99 of the 330 genes (~30%) had already been referred in AML related articles. The top 10 identified genes most cited in AML-related articles were *TNF*, *RUNX1T1*, *AR*, *F3*, *TRNA*, *AKT1*, *HPRT1*, *BCOR*, *HCK*, and *CCNA1* genes.

To conclude, we also compared the genes identified in the expression analysis described in 4.3.1 section with the candidate genes identified in the methylation analysis. We found that the *ATAD3C* was present in both analyses. That is, the gene expression as well as the DNA methylation of the *ATAD3C* seemed to be potential prognostic biomarkers in patients with AML-M4 categorized in the intermediate prognostic risk group.

## 4.4 FAB M5 AML SUBTYPE

As for our fourth population of study, we extracted the gene expression, DNA methylation as well as the clinical data information from the 16 patients with AML classified as M5 FAB AML subtype (AML-M5) categorized in the intermediate prognostic risk group. The developed algorithm was applied in gene expression and DNA methylation datasets independently. The gene expression results are described in the 4.5.1 section, and the DNA methylation results are described in the 4.5.2 section.

### 4.4.1 Gene expression as a potential prognostic biomarker in patients with AML-M5 categorized in the intermediate prognostic risk group

The gene expression dataset with records of 20530 genes from the 16 patients with AML-M5 categorized in the intermediate prognostic risk group was submitted to our algorithm. In the first algorithm phase (data preparation for analysis) 17390 genes were selected. Next, the second algorithm phase (identification of the first potential biomarkers) reduced the 17390 to 953 genes. After the third algorithm phase (confounding factor treatment), the genes were reduced to 32. At last, the fourth algorithm phase (selection criteria) identified 32 genes whose gene expression appeared to be able to predict survival in patients with AML-M5 categorized in the intermediate prognostic risk group (Annex XV). In Table 4.10 the top 15 identified genes with higher the hazard ratio value are shown.

**Table 4.10 List of 15 out of the 32 genes whose expression was able to predict survival in patients with AML-M5 categorized in the intermediate prognostic risk group.**

| **Gene** | ***p-value*** | **Optimal cutpoint** | **Group1 (n)** | **Group2 (n)** | **Age-test (*p-value*)** | **HR** |
|---|---|---|---|---|---|---|
| *TTC30B* | 0.00043 | 6.4346 | 8 | 8 | 0.91617 | 18.19290 |
| *FNTB* | 0.00043 | 9.1821 | 8 | 8 | 0.91617 | 18.19290 |
| *CDKN2AIP* | 0.00047 | 8.8936 | 7 | 9 | 0.52444 | 11.16033 |
| *VANGL1* | 0.00200 | 8.6638 | 7 | 8 | 0.60155 | 8.63472 |
| *ATN1* | 0.00224 | 10.5690 | 9 | 7 | 1.00000 | 8.51887 |
| *AMACR* | 0.00224 | 7.5548 | 9 | 7 | 1.00000 | 8.51887 |
| *PLA2G4A* | 0.00281 | 9.4859 | 7 | 7 | 0.40520 | 8.43606 |
| *TBX4* | 0.00145 | 0.0000 | 6 | 9 | 1.00000 | 7.67311 |
| *ABRAXAS2* | 0.00254 | 9.0384 | 6 | 9 | 0.95289 | 6.90494 |
| *C9orf3* | 0.00229 | 7.2003 | 8 | 8 | 1.00000 | 6.80033 |
| *RTTN* | 0.00860 | 7.9537 | 9 | 7 | 0.79084 | 6.77667 |
| *ZNF625* | 0.01115 | 3.9707 | 8 | 7 | 0.77174 | 6.28059 |
| *FANK1* | 0.01227 | 4.3070 | 8 | 7 | 0.95373 | 6.06089 |
| *CDH24* | 0.02950 | 5.1656 | 9 | 7 | 0.91551 | 4.08281 |
| *ZSCAN20* | 0.02950 | 6.1707 | 9 | 7 | 1.00000 | 4.08281 |

HR, hazard ratio

In the Figure 4.13 are shows the top 4 Kaplan-Meier overall survival curves with the highest hazard ratio value, obtained using the determined expression cutpoints of the *TTC30B, FNTB, CDKN2AIP*, and *VANGL1* genes, respectively. The Kaplan-Meier survival curve obtained using the expression cutpoint of the *TTC30B* gene had the highest hazard ratio value of 18.19 (Figure 4.13. a). The hazard ratio values of the survival curves obtained using the expression cutpoint of *FNTB, CDKN2AIP*, and *VANGL1* genes were 18.19, 11.16, and 8.63, respectively (Figure 4.13.b, c, d).

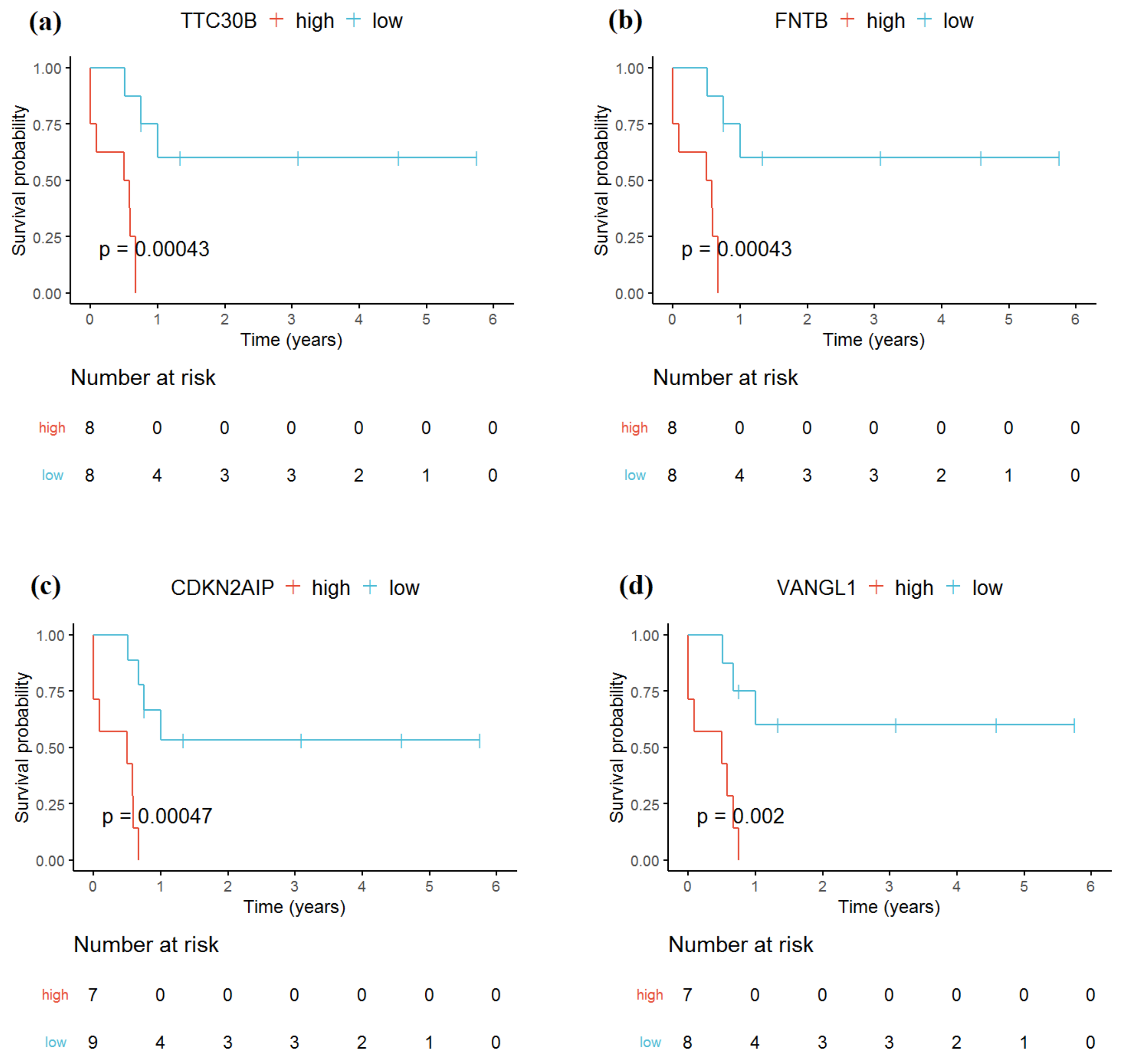


**Figure 4.13 Kaplan-Meier overall survival curves for four out of the 32 potential prognostic biomarkers identified for patients with AML-M5 categorized in the intermediate prognostic risk group.** Kaplan-Meier curve obtained using the expression cutpoint of the (a) *TTC30B* gene (p = 0.00043, log-rank test), (b) *FNTB* gene (p = 0.00043, log-rank test), (c) *CDKN2AIP* gene (p = 0.00047, log-rank test), (d) *VANGL1* gene (p = 0.002, log-rank test). The number at risk corresponds to the number of patients, at the indicated time point, that are still alive and whose follow-up continues.

In the case of each gene represented in the Figure 4.13, low expression values were related with the subgroup with a better prognosis. For the remaining identified genes, low expression values were also related to the subgroup with a better prognosis in 15 identified genes. In contrast, high expression levels were related to the subgroup with a better prognosis in 13 identified genes (Annex XVI).

Most of the determined expression cutpoints subcategorized the AML patients in a unique way. However, the expression cutpoints of two pairs of genes: 1) *FNTB* and *TTC30B* genes, 2) *AMACR* and *ATN1* genes, generated the same subgroups of patients.

Posteriorly, the subgroups generated by the expression cutpoint of the 32 identified genes were compared by gene set enrichment analysis to evaluate if there were gene sets related to biological processes differentially enriched between the subgroups with worse and better prognosis. We identified that the distinct intermediate-poor generated subgroups had in common some downregulation and upregulation of gene sets related to biological processes when comparing with the intermediate-favorable subgroup (Annex XVII). In the Figure 4.14 are shown the top 5 gene sets that we identified to be down and upregulated in the intermediate-poor subgroups in comparison with the intermediate-favorable are shown.

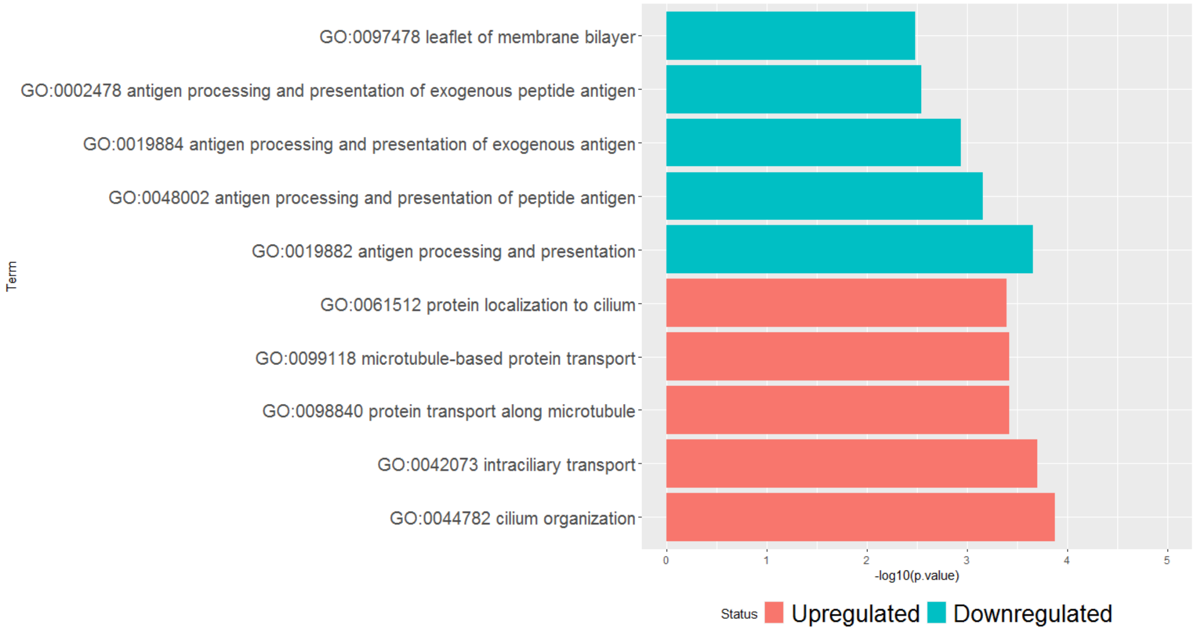


**Figure 4.14 Top 5 GO terms, in the biological processes category, down and upregulated between intermediate-poor and intermediate-favorable identified subgroups with AML-M5.** The gene set analysis was performed in biological processes category to identify the GO terms that were differently enriched between the subgroups analyzed. In blue are represented the GO terms that are downregulated and in red are represented the GO terms that are upregulated between the intermediate-poor and the intermediate-favorable identified subgroups.

Biological processes such as leaflet of membrane bilayer, antigen processing and presentation of exogenous peptide antigen, antigen processing and presentation of exogenous antigen, antigen processing and presentation of peptide antigen, antigen processing and presentation seemed to be downregulated in the majority of the intermediate-poor subgroups.

In contrast, protein localization to cilium, microtubule-based protein transport, protein transport along microtubule, intraciliary transport, and cilium organization are some examples of biological gene sets that seemed to be upregulated in the intermediate-poor in comparison with the intermediate-favorable subgroups.

Bibliographic analysis was performed to know if the 32 identified candidate genes were already described in the literature (Annex XVI). We noticed that 1 identified gene (*CTAGE6*) was never described in the PubMed literature. We also found that 17 of the 32 genes (~ 53%) had already been described in other cancer types than leukemia. Also, 7 of the 32 genes (~22%) had already been described in leukemia, but not in AML related articles. Finally, 5 of the 32 genes (~16%) (*B2M*, *PLA2G4A*, *SARAF*, *ABAT*, and *CSMD1*) had been described in AML related articles.

### 4.4.2 DNA methylation as a potential prognostic biomarker in patients with AML-M5 categorized in the intermediate prognostic risk group

The DNA methylation dataset with data regarding 485577 CpG sites from the 16 patients with AML-M5 categorized in the intermediate prognostic risk group was also analyzed by our algorithm. In the first algorithm phase (data preparation for analysis), the 485577 CpG sites were reduced to 395801. In the second algorithm phase (identification of the first potential prognostic biomarkers), 129187 were selected for the next analysis step. From the third algorithm phase (confounding factor treatment) resulted 1821 CpG sites. At last, the fourth phase of the algorithm (selection criteria) identified the final 26 candidate CpG sites whose DNA methylation appeared to predict survival of AML-M5 patients categorized in the intermediate prognosis risk group (Annex XVIII and XIX). In Table 4.11 are shown the top 15 identified CpG sites with highest hazard ratio value.

**Table 4.11 List of 15 out of the 26 CpG sites identified whose DNA methylation appeared to predict survival in patients with AML-M5 categorized in the intermediate prognostic risk group.**

| CpG sites | *p-value* | Optimal cutpoint | Group1 (n) | Group2 (n) | Age-test (*p-value*) | Gene | HR |
|---|---|---|---|---|---|---|---|
| cg22212237 | 0.00029 | 0.1786 | 7 | 9 | 0.52444 | *C1QL4* | 11.89617 |
| cg20183094 | 0.00047 | 0.2167 | 7 | 9 | 0.52444 | *FEZF1* | 11.16033 |
| cg08004278 | 0.00047 | 0.1872 | 7 | 9 | 0.52444 | *FEZF1* | 11.16033 |
| cg05422647 | 0.00117 | 0.1620 | 7 | 9 | 0.79084 | *GLRA1* | 10.03001 |
| cg17652792 | 0.00381 | 0.0850 | 9 | 7 | 0.91551 | *PCDHB16* | 7.92216 |
| cg27462975 | 0.00338 | 0.6960 | 6 | 8 | 0.84595 | *MIR573* | 6.70370 |
| cg16759787 | 0.01115 | 0.6922 | 8 | 7 | 0.77174 | *RN7SK* | 6.28059 |
| cg20362634 | 0.00357 | 0.7837 | 8 | 8 | 0.63576 | *GABRG3* | 6.15490 |
| cg03154226 | 0.01097 | 0.5625 | 9 | 7 | 0.67132 | *OR5D14* | 6.15171 |
| cg11155432 | 0.01603 | 0.8812 | 9 | 7 | 0.87357 | *CDH9* | 4.83441 |
| cg02171545 | 0.02461 | 0.4851 | 8 | 8 | 0.29255 | *SNRPN* | 3.93214 |
| cg12436427 | 0.02461 | 0.2290 | 8 | 8 | 0.37097 | *WASHC2A* | 3.93214 |
| cg18232125 | 0.02461 | 0.7407 | 8 | 8 | 0.22612 | *TENM2* | 3.93214 |
| cg16368763 | 0.02461 | 0.2669 | 8 | 8 | 0.49388 | *TRIM67* | 3.93214 |
| cg26577836 | 0.02911 | 0.5729 | 8 | 8 | 0.39977 | *TNXB* | 0.26247 |

HR, hazard ratio

In Figure 4.15, we show as examples the top 4 Kaplan-Meier overall survival curves with the highest hazard ratio value, obtained using the methylation cutpoint of 4 identified CpG sites localized in the *C1QL4*, *FEZF1*, and *GLRA1* genes, respectively. The survival curve obtained using the methylation cutpoint of the cg22212237 located in the *C1QL4* gene had the highest hazard ratio value of 11.90. Additionally, the hazard ratio values of the survival curves obtained using the methylation cutpoint of the cg20183094, cg08004278, and cg05422647 localized in the *FEZF1*, and *GLRA1* were 11.16, 11.16, and 10.03, respectively (Figure 4.15. b, c, d).

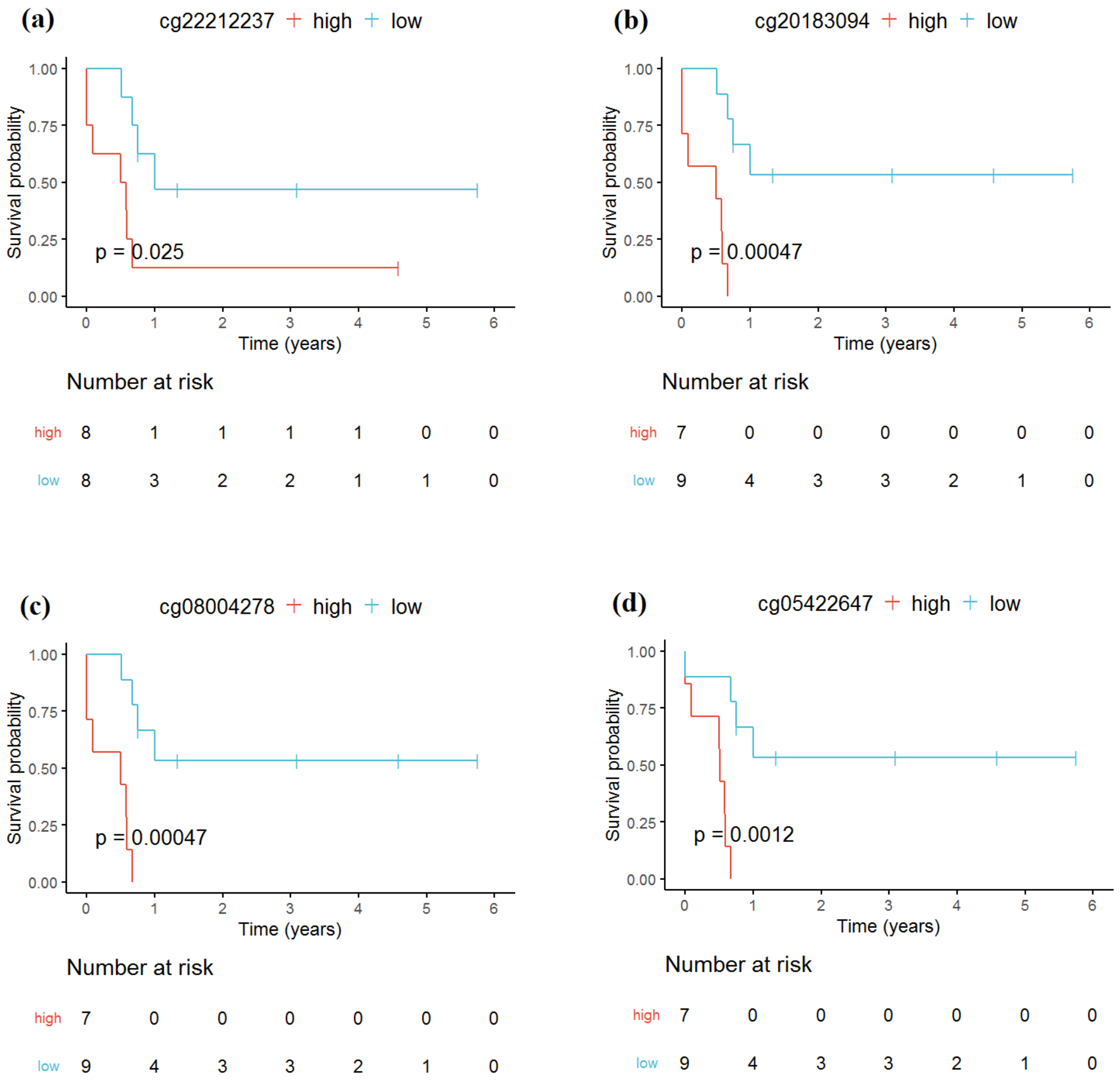


**Figure 4.15 Kaplan-Meier overall survival curves of four out of the 26 potential DNA methylation prognostic biomarkers identified for patients with AML-M5 categorized in the intermediate prognostic risk group.** Kaplan-Meier curve obtained using the methylation cutpoint of the (a) cg22212237 localized in the *C1QL4* gene (p = 0.025, log-rank test), (b) cg20183094 localized in the *FEZF1* gene (p = 0.00047, log-rank test), (c) cg08004278 localized in the *FEZF1* gene too (p = 0.00047, log-rank test), and (d) cg05422647 localized in the *GLRA1* gene (p = 0.0012, log-rank test). The number at risk corresponds to the number of patients, at the indicated time point, that are still alive and whose follow-up continues.

As we can observe, in each case presented in the Figure 4.15, the hypermethylation was related with the subgroup of AML-M5 patients with worse overall survival. In the remaining 22 identified CpG sites, the hypomethylation characterized the subgroup with a better prognosis (IXIIX).

The 26 identified CpG sites were localized in 25 genes. This result means that there were 2 CpG sites localized in the same gene with predictive prognostic value in the patient's sample studied. The cg20183094 and the cg08004278 were both localized in the distal intergenic region of the *FEZF1* gene. Interestingly, both Kaplan-Meier survival curves obtained using the methylation cutpoint of each these CpG site, had the same HR value of 11.16 (Figure 4.15 b and c).

The genomic region where the 26 identified CpG sites were located were also evaluated. We observed that the identified CpG sites were localized in promotor, 3'UTR, first intron, other introns, exons, and distal intergenic (Figure 4.16 and Annex XIX). The promotor was the first genomic region with more identified CpG sites located (~ 46%), and the distal intergenic the second genomic region with more identified CpG located (~ 27%).

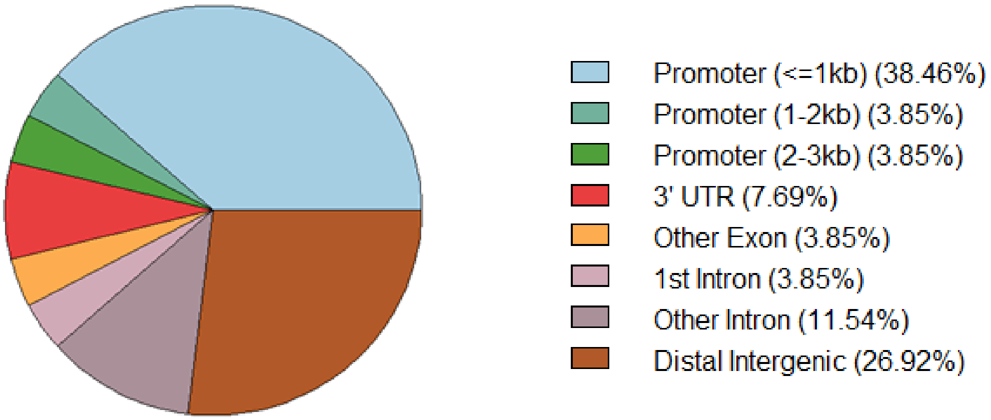


**Figure 4.16 Location of the identified CpG sites within the genes, whose methylation appeared to be potential biomarker in AML-M5 patients categorized in the intermediate prognostic risk group.** The majority of the 26 CpG sites identified by our algorithm were localized in the promotor region (~ 46%). The second genomic region with more identified CpG's was the distal intergenic region. About 12% of the CpG's were other intron regions. The 3'UTR, other exons, and first intron were the genomic regions with less identified CpG's (7.69%, 3.85%, and 3.85%, respectively). The other intron means that the CpG's were localized in other introns than the first intron. The other exon region refers exons than the first exon.

Bibliographic analysis was performed to know if the 25 identified candidate genes were already described in the literature (Annex XIX). We noticed that 2 identified genes (*OR5D14*, and *KIRREL3-AS3*) were never described in the PubMed literature. Moreover, 14 of the 25 genes (~ 56%) had already been described in other cancer types than leukemia. Four of the 25 genes (~16%) had already described in leukemia, but not in AML related articles. Finally, 4 of the 25 genes (~16%) (*SNRPN*, *FGFR2*, *SHH*, *GRB10*) had been described in AML related articles.

After all, we intersected the genes resultant of the expression (section 4.4.1) and the DNA methylation (section 4.4.2). We observed there were not genes present in both analyses.

## 4.5 M0, M1, M2, M4, AND M5 FAB AML SUBTYPES

Once analyzed the patients within each FAB subtype (M1, M2, M4 and M5) studied individually, we were also interested in analyzing the patients with M0, M1, M2, M4 and M5 FAB AML subtype together. Thus, our fifth population of studied was the patients with AML classified with M0, M1, M2, M4 and M5 FAB subtype (AML-M0-M1-M2-M4-M5) categorized in the intermediate prognostic risk group (n=89). We extracted the gene expression and DNA methylation values, as well as the clinical information data from the patients mentioned above. Next, we applied the developed methodology to both gene expression and DNA methylation datasets independently. The gene expression results are described in 4.5.1 section, and the DNA methylation results are described in 4.5.2 section.

### 4.5.1 Gene expression as a potential prognostic biomarker in patients with AML-M0-M1-M2-M4-M5 categorized in the intermediate prognostic risk group

The gene expression dataset with expression values of 20530 genes from the 89 patients with AML-M0-M1-M2-M4-M5 categorized in the intermediate prognostic risk group was analyzed by the developed algorithm. In the first algorithm phase (data preparation for analysis), 17368 genes were selected. Furthermore, performing the second algorithm phase (identification of the first potential biomarkers), the 17368 genes were reduced to 1171. Next, the third algorithm phase (confounding factor treatment) allowed the selection of 280 genes. Performing the fourth algorithm phase (selection criteria), we identified 176 candidate genes whose expression appeared to be potential prognostic biomarker to predict survival in patients with AML-M0-M1-M2-M4-M5 categorized in the intermediate prognostic risk group (Annex XX and Annex XXI). The Table 4.12 shows the top 15 identified genes with the highest hazard ratio value.

**Table 4.12 List of 15 out of the 176 genes whose expression was able to predict survival in patients with AML-M0-M1-M2-M4-M5 categorized in the intermediate prognostic risk group.**

| Gene | *p-value* | Optimal cutpoint | Group1 (n) | Group2 (n) | Age-test (*p-value*) | HR |
|---|---|---|---|---|---|---|
| *CCT6B* | 0.000004 | 5.2369 | 34 | 55 | 0.905847 | 3.211710 |
| *ITGB1BP1* | 0.000024 | 9.2228 | 40 | 43 | 0.794962 | 3.069339 |
| *PDE7B* | 0.000085 | 5.1779 | 54 | 35 | 0.432021 | 3.045263 |
| *ITGA11* | 0.000028 | 2.2561 | 37 | 50 | 0.308758 | 2.998898 |
| *PBDC1* | 0.000213 | 8.0404 | 54 | 33 | 0.551736 | 2.976073 |
| *NQO1* | 0.000108 | 6.2389 | 49 | 40 | 0.098854 | 2.785669 |
| *HSD17B7* | 0.000252 | 7.4822 | 33 | 53 | 0.587825 | 2.722778 |
| *TOMM40L* | 0.000332 | 8.5908 | 51 | 34 | 0.332359 | 2.657828 |
| *DRC7* | 0.000304 | 3.1340 | 48 | 41 | 0.178146 | 2.553137 |
| *TBC1D29P* | 0.000279 | 3.0231 | 34 | 55 | 0.115111 | 2.546112 |
| *ASCC1* | 0.000330 | 9.0165 | 43 | 46 | 0.983615 | 2.511365 |
| *CCND3* | 0.000395 | 11.5099 | 48 | 41 | 0.184873 | 2.504012 |
| *IQCG* | 0.000965 | 7.1977 | 53 | 34 | 0.982640 | 2.464832 |
| *SH3TC2* | 0.000896 | 7.5208 | 44 | 43 | 0.310077 | 2.406249 |
| *S100A1* | 0.000685 | 4.5428 | 43 | 45 | 0.757306 | 2.402793 |

HR, hazard ratio

As examples, in the Figure 4.17 are shown the top 4 Kaplan-Meier overall survival curves with the highest hazard ratio value, obtained using the expression cutpoints of *CCT6B*, *ITGB1BP1*, *PDE7B*, and *ITGA11* genes, respectively. The survival curve obtained using the expression cutpoint of *CCT6B* genes had the highest hazard ratio value of 3.21 (Figure 4.17. a). The hazard ratio values of the remaining demonstrated survival curves obtained using the expression cutpoint of the *ITGB1BP1*, *PDE7B*, and *ITGA11* genes were 3.07, 3.05, and 3, respectively.

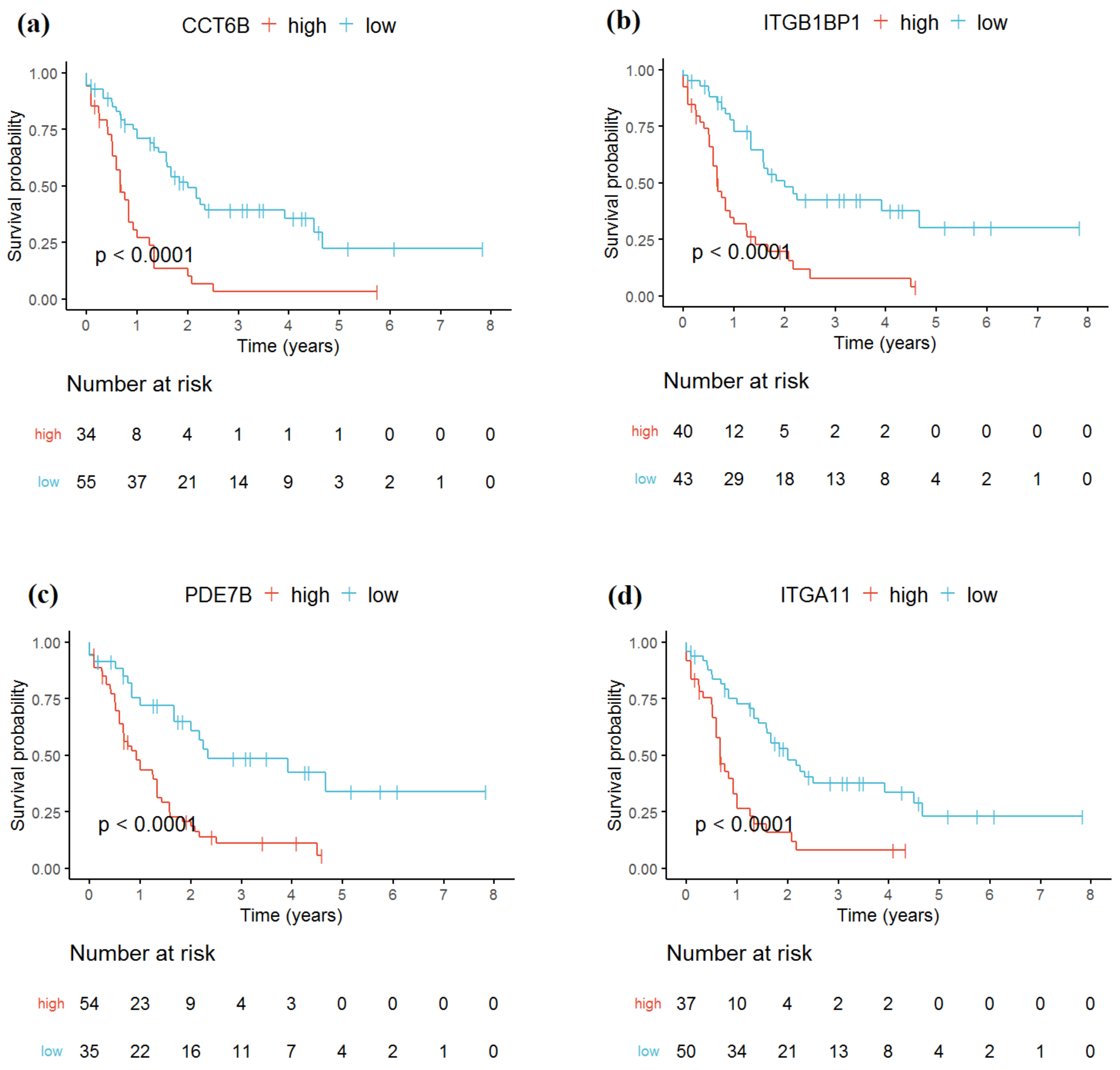


**Figure 4.17 Kaplan-Meier overall survival curves for four out of the 176 potential prognostic biomarkers identified for patients with AML-M0-M1-M2-M4-M5 categorized in the intermediate prognosis risk group.** Kaplan-Meier curve obtained using the expression cutpoint of the (a) *CCT6B* gene (p = 0.000004, log-rank test), (b) *ITGB1BP1* gene (p = 0.000024, log-rank test), (c) *PDE7B* gene (p = 0.000085, log-rank test) and (d) *ITGA11* gene (p = 0.000028, log-rank test). The number at risk corresponds to the number of patients, at the indicated time point, that are still alive and whose follow-up continues.

For each of the four demonstrated cases, the low expression seemed to be related with the subgroup with a better prognosis. In addition, the low expression levels were also displayed by the subgroup with better prognosis in 40 identified genes. In contrast, for the remaining 131 identified genes, high expression levels were related with the subgroup with better prognosis (Annex XXI).

We also analyzed the subgroups of AML patients generated by each determined expression cutpoint, and we observed that the subcategorization of AML-M0-M1-M2-M4-M5 patients was dependent on the expression cutpoint applied. That is, different expression cutpoints subdivided the patients in a unique way.

The gene set analysis about biological processes was also performed to know if there were genes sets commonly enriched in all intermediate-poor in comparison with all intermediate-favorable identified subgroups. Our results suggest that the majority of the intermediate-poor subgroups shared some biological processes genes sets that appeared to be downregulated and upregulated in comparison with the intermediate-favorable subgroups (Annex XXII). In Figure 4.18 are shown the top 5 gene sets that we identified to be down and upregulated in the intermediate-poor subgroups in comparison with the intermediate-favorable are shown.

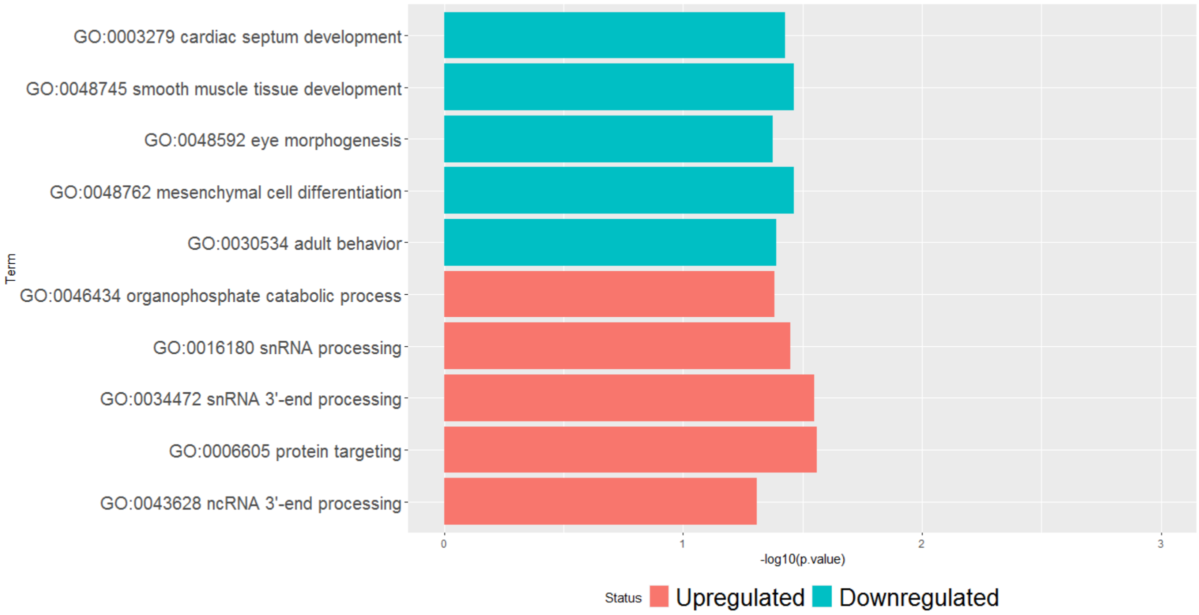


**Figure 4.18 Top 5 GO terms, in the biological process category, down and upregulated between in the intermediate-poor and intermediate-favorable identified subgroups with AML-M0-M1-M2-M4-M5.** The gene set analysis was performed in biological processes category to identify the GO terms that were differently enriched between the subgroups analyzed. In blue are represented the GO terms that are downregulated and in red are represented the GO terms that are upregulated between the intermediate-poor and the intermediate-favorable identified subgroups.

Biological processes such as cardiac septum development, smooth muscle tissue development, eye morphogenesis, mesenchymal cell differentiation, and adult behavior seemed to be downregulated in the majority of the intermediate-poor subgroups.

In contrast, organophosphate catabolic process, snRNA processing, snRNA 3’-end processing, protein targeting, and ncRNA 3’-end processing are some examples of biological gene sets that seemed to be upregulated in the intermediate-poor in comparison with the intermediate-favorable subgroups.

A bibliographic analysis was also performed to evaluate if the identified 176 genes were already referred in the literature (Annex XXI). We found that 4 identified genes (*CALHM5*, *C16orf54*, *TBC1D29P*, *CCNJL*) were never described in the PubMed literature. In addition, 69 of the 176 genes (~ 39%) had already been described in other cancer types than leukemia. Furthermore, 41 of the 176 genes (~23%) had already been described in leukemia, but not in AML related articles. Finally, 56 of the 176 genes (~32%) had been described in AML related articles.

### 4.5.2 DNA methylation as a potential prognostic biomarker in patients AML-M0-M1-M2-M4-M5 categorized in the intermediate prognostic risk group

The DNA methylation dataset with data about 485577 CpG sites from the 89 patients with AML-M0-M1-M2-M4-M5 categorized in the intermediate prognostic risk group was also analyzed by our algorithm. The first algorithm phase (data preparation for analysis) the 485577 CpG sites were reduced to 395854. In the second algorithm phase (identification of the first potential prognostic biomarkers) were selected for the next analysis step. From the third algorithm phase (confounding factor treatment) resulted CpG sites. At last, the fourth phase of the algorithm (selection criteria) identified the final 273 candidate CpG sites whose DNA methylation appeared to predict survival of AML-M0-M1-M2-M4-M5 patients categorized in the intermediate prognosis risk group (Annex XXIII and Annex XXIV). In Table 4.13 are shown the top 15 identified CpG sites with highest hazard ratio value.

**Table 4.13 List of 15 out of the 273 CpG sites identified whose DNA methylation appeared to predict survival in patients with AML-M0-M1-M2-M4-M5 categorized in the intermediate prognostic risk group.**

| CpG sites | *p-value* | Optimal cutpoint | Group1 (n) | Group2 (n) | Age test (*p-value*) | HR | gene |
|---|---|---|---|---|---|---|---|
| cg14469693 | 0.00004 | 0.8509 | 43 | 39 | 0.78403 | 3.19192 | *PRDM6* |
| cg25068253 | 0.00010 | 0.8845 | 41 | 39 | 0.05593 | 3.05123 | *RASSF10* |
| cg17755518 | 0.00039 | 0.9022 | 42 | 34 | 0.45485 | 2.97498 | *SLC44A5* |
| cg17572155 | 0.00003 | 0.7782 | 45 | 44 | 0.13191 | 2.96066 | *ADARB2* |
| cg03841832 | 0.00022 | 0.7302 | 49 | 36 | 0.06482 | 2.88363 | *SLC2A9* |
| cg14495958 | 0.00082 | 0.8522 | 46 | 33 | 0.37876 | 2.80835 | *EXD3* |
| cg01918114 | 0.00039 | 0.8505 | 44 | 37 | 0.07146 | 2.79206 | *LMF1* |
| cg13184077 | 0.00019 | 0.8994 | 42 | 39 | 0.16448 | 2.77835 | *ZNF365* |
| cg26118637 | 0.00024 | 0.8787 | 39 | 44 | 0.06856 | 2.72663 | *CNPY1* |
| cg15596913 | 0.00008 | 0.8868 | 43 | 44 | 0.15359 | 2.72333 | *ENPP7* |
| cg15861089 | 0.00064 | 0.8077 | 49 | 36 | 0.07960 | 2.69924 | *KRT86* |
| cg26208930 | 0.00070 | 0.826 | 45 | 38 | 0.13247 | 2.65737 | *TP73* |
| cg10350957 | 0.00029 | 0.8802 | 37 | 46 | 0.18076 | 2.64871 | *FAT3* |
| cg22356541 | 0.00035 | 0.8877 | 43 | 39 | 0.39014 | 2.59198 | *FGF9* |
| cg03863069 | 0.00039 | 0.9086 | 34 | 51 | 0.30620 | 2.53408 | *SH3PXD2B* |

HR, hazard ratio

In the Figure 4.19 are shown the top 4 Kaplan-Meier survival curves with the highest hazard ratio value, obtained using the methylation cutpoint of 4 identified CpG sites localized in the *PRDM6*, *RASSF10*, *SLC44A5*, and *ADARB2* genes, respectively. The survival curve obtained using the methylation cutpoint of the cg14469693 localized in the *PRDM6* gene had the highest hazard ratio value of 3.19. The hazard ratio values of the survival curves of the remaining demonstrated CpG sites localized in the *RASSF10*, *SLC44A5*, and *ADARB2* genes were 3.05, 2.97, and 2.96, respectively (Figure 4.19. b, c, d).

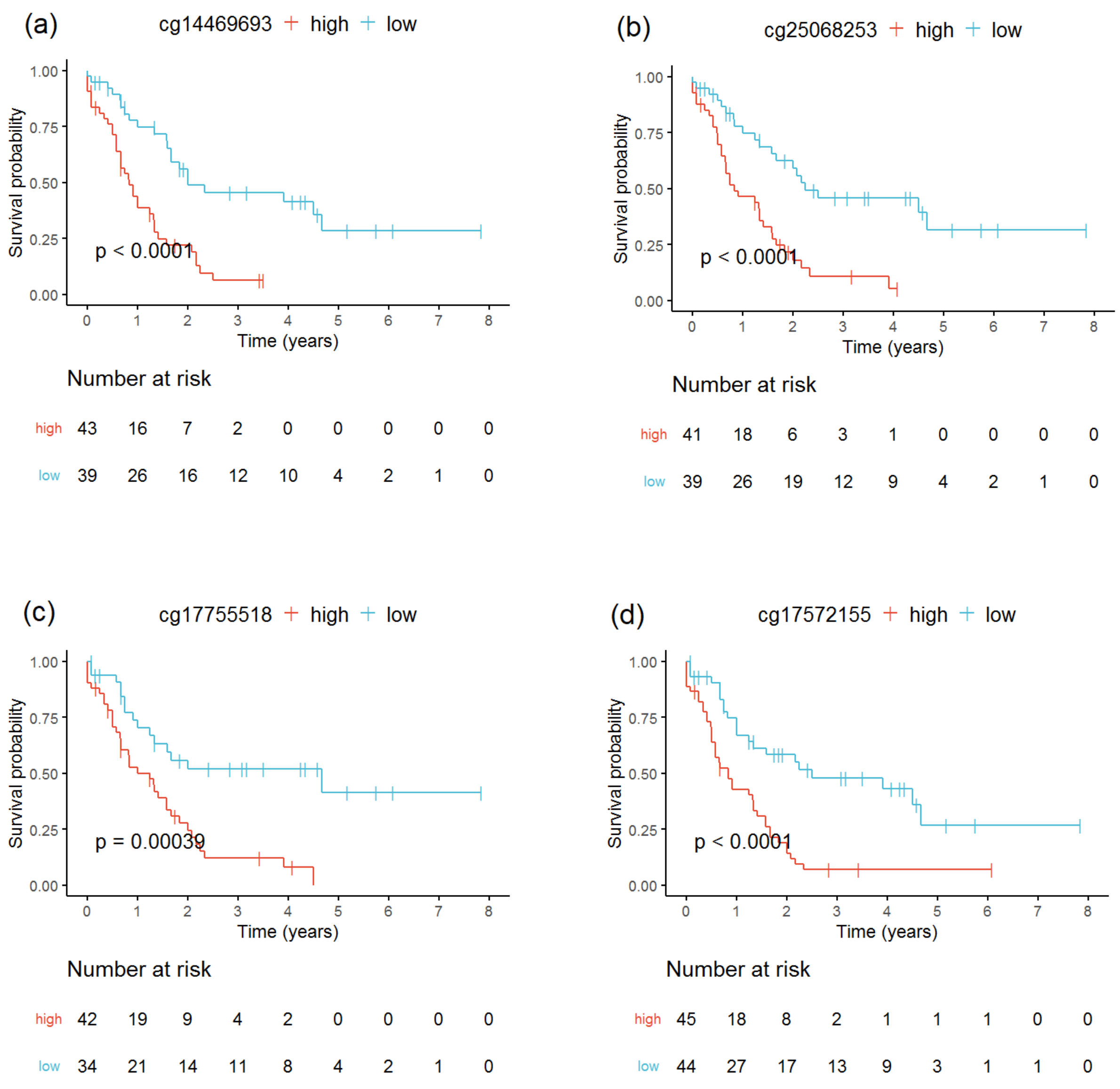


**Figure 4.19 Kaplan-Meier overall survival curves of four out of the 273 potential DNA methylation prognostic biomarkers identified for patients with AML-M0-M1-M2-M4-M5 categorized in the intermediate prognostic risk group.** Kaplan-Meier curve obtained using the methylation cutpoint of the (a) cg14469693 localized in the *PRDM6* gene (p = 0.00019, log-rank test), (b) cg25063253 localized in the *RASSF10* gene (p = 0.000097, log-rank test), (c) cg17755518 localized in the *SLC44A5* gene (p = 0.000388, log-rank test), and (d) cg17572155 localized in the *ADARB2* gene (p = 0.00034, log-rank test). The number at risk corresponds to the number of patients, at the indicated time point, that are still alive and whose follow-up continues.

Moreover, our results suggest that for each demonstrated case in the Figure 4.19, hypomethylation seemed to be related with the subgroup with better prognosis. For the remaining identified CpG sites, hypomethylation was also related with the better prognostic subgroup in 208 cases. On the other hand, hypermethylation seemed to be related with the subgroup with better prognosis in 61 cases (Annex XXIV).

Moreover, we also observed that each identified methylation cutpoint generated distinct subgroups of patients, being the subcategorization of patients dependent on the cutpoint used.

The 273 identified CpG sites with a prognostic value in the group of AML patients studied were localized in the 264 genes, which mean that were different CpG sites localized in the same gene with a predictive prognostic value. We observed that there were 9 genes with more than one identified CpG site with a potential prognostic value in the AML patients (Table 4.14).

**Table 4.14 The 9 genes with more than one CpG site with prognostic value in AML-M0-M1-M2-M4-M5 patients categorized in the intermediate prognostic risk group.**

| Gene | CpG sites | CpG sites location | Methylation status related to better prognosis |
|---|---|---|---|
| *ABLIM2* | cg16114831 | Intron | hypomethylation |
| | cg03033996 | Promoter (2-3kb) | hypomethylation |
| *ADARB2* | cg17572155 | Promoter (2-3kb) | hypomethylation |
| | cg08302300 | Intron | hypermethylation |
| | cg00615654 | Intron | hypomethylation |
| *CHN2* | cg03100044 | Promoter (<=1kb) | hypomethylation |
| | cg01617933 | Intron | hypermethylation |
| *EXD3* | cg13710542 | Promoter (<=1kb) | hypomethylation |
| | cg14495958 | 1st Intron | hypomethylation |
| *GLI2* | cg25919979 | Intron | hypomethylation |
| | cg03465652 | Intron | hypomethylation |
| *KLHL29* | cg12459514 | Intron | hypomethylation |
| | cg15736743 | Intron | hypermethylation |
| *PRICKLE2* | cg01165355 | Distal Intergenic | hypomethylation |
| | cg20584157 | Intron | hypomethylation |
| *UBE2I* | cg05246900 | Promoter (1-2kb) | hypomethylation |
| | cg02920178 | Promoter (<=1kb) | hypomethylation |

For example, the cg16114831 in an intronic region and the cg03033996 at promotor region are two distinct CpG sites that were localized in the same gene (*ABLIM2*) and that had a predictive prognostic value in the group of patients studied. The hypomethylation of both CpG sites was related with the better prognosis subgroup. Another example was the cg03100044 at promotor region and the cg01617933 at an intronic region that were localized in the *CHN2* gene. However, in this case the methylation status of the two CpG sites that were related with the subgroup with better prognosis were different. Whereas the hypomethylation of cg03100044 was related with the subgroup with better prognosis, in the of the cg01617933 is the hypermethylation.

Further, we identified the genomic regions of the 273 identified CpG's, being localized in regions such as promotor, 3'UTR, 5'UTR, exon, first intron, intron, downstream and distal intergenic (Figure 4.20 and Annex XXIV). The genomic region with more density of identified CpG sites was the promotor region, with about 48% of the identified CpG's. Moreover, the distal intergenic region was the second region with more identified CpG's, 22%, respectively.

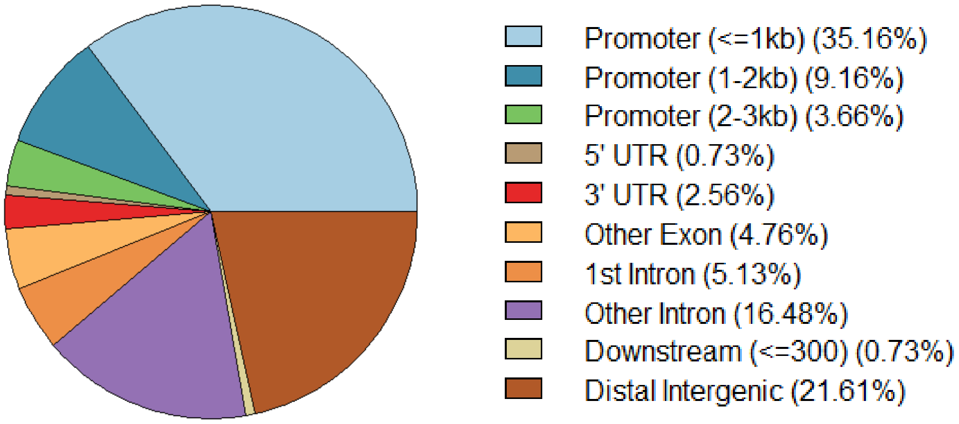


**Figure 4.20 Location of the identified CpG sites within the genes, whose methylation appeared to be potential biomarker in AML-M0-M1-M2-M4-M5 patients categorized in the intermediate prognostic risk group.** The majority of the 273 CpG sites identified by our algorithm were localized in the promotor region (~ 48%). The second genomic region with more identified CpG's was the distal intergenic region. About 17% of the CpG's were in intronic regions. The first intron, other exons, 3'UTR, 5'UTR, and downstream were the genomic regions with less identified CpG's (5.13%, 4.76%, 2.56%, 0.73%, and 0.73%, respectively). The other intron means that the CpG's were localized in other introns than the first intron. The other exon region refers exons than the first exon.

Bibliographic analysis was performed to know if the 264 identified candidate genes were already described in the literature (Annex XXIV). We noticed that 13 identified genes (*MIR769*, *ZNF846*, *KIAA2012*, *MIR4710*, *MIR4634*, *OR4K17*, *ZNF585B*, *MIR4679-2*, *LMNTD1*, *ZNF514*, *LSMEM1*, *MIR4655*, and *ITPRID1*) were never described in the PubMed literature. Also, 106 of the 264 genes (~ 40%) had already been described in other cancer types than leukemia. Moreover, 55 of the 264 genes (~21%) had already described in leukemia, but not in AML related articles. Finally, 70 of the 264 genes (~27%) had been described in AML related articles. As examples are *BSG*, *EBF3*, *MAP1LC3C*, *SLC6A5*, and *PITX1* genes.

# CHAPTER 5

# DISCUSSION

## 5.1 Potential prognostic biomarkers for AML patients with AML-M1 categorized in the intermediate prognostic risk group

The group of patients with AML-M1 categorized in the intermediate risk group was the first to be analyzed by our algorithm in order to identify potential prognostic biomarkers of both gene expression and DNA methylation that predict survival in these group of patients.

By performing the gene expression analysis, we identified 11 genes whose expression levels can be used as potential prognostic biomarkers, two of which were already cited in the AML literature. One of these genes is the Minichromosome Maintenance Complex Component 4 (*MCM4*), that codifies for one of the six subunits (MCM2 to MCM7) of the minichromosome maintenance (MCM) complex that works as a replicative helicase involved in the DNA replication process.[61] In the G1 phase of the cell cycle, the MCM complex binds to the replication origins and dissociates after the initiation of DNA replication, in order to avoid a second replication from the same origin. Moreover, this complex is involved in the unwinding of the parental DNA strands.[61] According to our results, when the patients with AML-M1 categorized in the intermediate prognostic risk group are subdivided based on the expression *MCM4* cutpoint, the subgroup with worse prognosis is characterized by low expression levels of *MCM4* gene. By searching the AML literature, we only found that the expression of *MCM4* gene is altered in aneuploid acute myeloid leukemia.[62]. Our R-based AML-literature searching function is able to detect if the gene was cited in an AML-related article, but it cannot detect if the gene has a causal relation with AML. Nonetheless, our study seems to suggest that *MCM4*'s expression levels are altered between intermediate-poor and intermediate-favorable subgroups of AML patients.

The other identified putative prognostic biomarker that was already cited in the AML literature is the Synaptosome Associated Protein 23 (*SNAP23*) gene. This gene codifies for SNAP23 protein which is part of a sub-family of SNARE proteins, known as SNAP25 protein sub-family.[63] The SNARE-complexes are involved in the fusion of membrane vesicles with the target membrane. In particular, the SNAP23 participates in driving regulated exocytosis.[63] A study reported that the *SNAP23* gene was expressed in 20 of 31 analyzed AML cases, including 4 AML cases with normal cytogenetics.[64] Our results demonstrate that high *SNAP23* expression levels is related to a worse prognosis in the patients with AML-M1 categorized in the intermediate prognostic risk group. It has been described that SNAP23 is implicated in a secretion of matrix metalloproteinases, extracellular matrix degradation and cell invasion.[65] In addition, Sun *et al*. demonstrated that the silencing of SNAP23 in ovarian cancer cells leads to diminished proliferation, impairs cell migration and invasion capacities, and inhibits apoptosis in these cells. Furthermore, the group also showed that higher levels of SNAP23 expression were associated with poor prognosis in ovarian cancer patients.[65] It is possible that the worse prognosis of the AML subgroup with high SNAP23 expression levels is also due to decreased apoptosis and higher proliferation capacities of the leukemic cells.

The generated subgroups of AML-M1 patients by the previously identified potential prognostic biomarkers of gene expression were also compared by the gene set enrichment analysis. In this step, we compared the identified subgroups with worse prognosis (intermediate-poor) with the subgroups with better prognosis (intermediate-favorable) in order to know what gene sets describing biological processes were systematically differentially expressed between them.

We observed that in most of the intermediate-poor subgroups there were some biological processes that appeared to be downregulated. Two examples are processes related to cellular response to vascular endothelial growth factor stimulus, and the vascular endothelial growth factor signaling pathway. The downregulation of these two GO terms seem to suggest that the leukemic cells of the intermediate-poor patients have a lower response to the vascular endothelial growth factor (VEGF).

It is known that AML patients shown an aberrant VEGF signaling in the bone marrow that promotes AML blast cell proliferation and survival, chemotherapy resistance, and also increased angiogenesis.[66]. It was previously reported that high levels of VEGF are associated with poor therapeutic outcome of AML patients.[66] Our results are contradictory to what is described in the literature. Therefore, we hypothesize that, although the VEGF cellular response, and consequently its signaling, seem to be downregulated in the majority of the intermediate-poor subgroups, these alterations might not have an impact in patient survival. It is possible that in the two prognostic subgroups there are alterations in gene sets that confer a better prognosis and others that confer a worse prognosis, but in the intermediate-poor subgroups there are more dysregulated gene sets that confer a poor prognosis. Furthermore, this is only an example for one enriched gene set, which cannot be observed individually, since it is part of a larger gene expression network.

In addition, we observed that the hydrogen peroxide catabolic process and the gas transport process are also downregulated in the majority of the intermediate-poor, in comparison with the intermediate-favorable subgroups. Hydrogen peroxide is a reactive oxygen species (ROS) produced through the cellular metabolism.[67] According to previous studies, AML cells show higher levels of ROS in comparison with normal leukocytes, which is a trigger for leukemogenesis. In normal conditions, the HSCs are under low ROS levels that regulate their self-renewal and proliferation capacities. However, oxidative stress caused by high ROS levels promotes HSC proliferation.[67] We theorize that in the majority of the intermediate-poor subgroups, the leukemic cells have a higher proliferation capacity than the leukemic cells of the intermediate-favorable subgroups. Furthermore, there is an association between ROS and chemotherapy resistance, that could also contribute for the poor prognosis.[67]

Gene sets related to skeletal system development also seem to be downregulated in the intermediate-poor subgroups. However, this is a large set that consists of 545 genes and is the parent term of a complex network of several child GO terms. As such, it is difficult to draw reasonable insights for why it is downregulated in the intermediate-poor subgroups.

In contrast, our analysis also identified gene sets related to biological processes that appear to be upregulated in most of the identified subgroups with worse prognosis in comparison with the more favorable ones. Positive regulation of response to biotic stimulus, regulation of innate immune response, and positive regulation of innate immune response are examples of biological processes that appear to be upregulated in most intermediate-poor AML-M1 patients. The enrichment of GO terms seem to suggest that there is more innate immune response in the intermediate-poor patients than in the intermediate-favorable patients. In addition, we also observed that regulation of natural killer cell mediated immunity also seemed to be upregulated in the intermediate-poor subgroup, which could be explained by the fact that there seems to be a general upregulation of gene sets associated with innate immune response. In contrast, however, it is possible that several gene sets related to the innate immune response were found to be upregulated merely as a consequence of the upregulation of the term related to the regulation of natural killer cell mediated immunity. It has been described that NK cells play an important antitumor role in AML.[68] Furthermore, it has been shown that NK cells are often defective in AML patients, which contributes to the immunological escape of this malignancy.[69] Malfunctioning NK cells are a predictive factor for poor prognosis and early relapse and, on the other hand, NK cell activity is positively correlated with better prognosis.[69] It is possible that the intermediate-favorable AML subgroup might have, as described in the literature, a higher NK cell activity. Therefore, downregulation of genes associated with the activation of the innate immune response could provide a selective advantage to the leukemic cells in the intermediate-favorable patients. On the other hand, the intermediate-poor subgroup could have either faulty or low-activity NK cells, so the selective advantage conferred by downregulating genes related to the activation of innate immune response would minor in the leukemic cells of these patients.

By performing the methylation analysis, we identified 130 genes as potential prognostic biomarkers in patients with AML-M1 categorized in the intermediate prognostic risk group. Of these 130 genes, 46 were already described in the AML literature. An example is the Scinderin (*SCIN*) gene, which encodes for a protein member of the actin-binding protein family.[70] Several actin-binding proteins are involved in the regulation of dynamics of actin filaments, and this is associated with cell migration, contributing to tumor cell invasion and metastasis.[70] Our results suggest that when AML-M1 patients categorized in the intermediate prognostic risk group are subcategorized based on the determined methylation cutpoint of the *SCIN* gene, the subgroup with worse prognosis displays *SCIN* hypermethylation in the promotor region. Zhang *et al.* investigated the clinical relevance of *SCIN* expression and promotor methylation in AML patients.[70] They demonstrated that AML patients had significantly lower levels of *SCIN* expression in comparison with healthy controls. The AML patients also had significantly higher levels of methylation at the *SCIN* promotor region, and this methylation was negatively correlated with *SCIN* expression. In addition, AML patients with low expression levels of *SCIN* showed lower rates of complete remission and shorter overall survival in comparison with patients with higher levels of *SCIN* expression. So, *SCIN* promotor methylation, which is associated with lower levels of *SCIN* expression, is a valuable biomarker to predict poor prognosis in AML patients.[70] This study is in concordance with our results.

Another putative prognostic biomarker identified in our methylation analysis that was already cited in the AML literature is the Major histocompatibility complex class I related - A

(*MICA*). According to our results, when the patients with AML-M1 categorized in the intermediate prognostic risk group are subcategorized based on the determined methylation cutpoint for the *MICA* gene, the subgroup with worse prognosis is related with *MICA* hypomethylation at promotor region. MICA is one of the ligands that activates the receptor natural killer group 2 member D (NKG2D) promoting the innate immune response in NK cells.[71] However, cancer cells downregulate the expression of ligands, such as MICA, in order to avoid recognition by the immune system. Baragaño Raneros et al. observed that, in AML cells, there is a hypermethylation of *MICA* in comparison with cells from healthy donors. This hypermethylation was correlated with decreased expression levels of the ligand, suggesting that, in AML there is an epigenetic silencing of MICA expression through DNA methylation, as a result of the tumor's development.[71] Nonetheless, we found that the intermediate-poor AML patients had MICA hypomethylation, which could be linked to higher transcription levels of the gene. Although MICA is usually downregulated in cancer to avoid immune detection, it has also been described tumor cells can release soluble molecules of MICA in order to evade NKG2D-mediated immune responses.[72] Release of soluble MICA by tumor cells leads to less MICA in the cell surface, which promotes a reduced susceptibility to NKG2D-mediated cytotoxicity.[72] Since we only analyze expression levels, and not protein levels, we can only speculate that the hypomethylation of *MICA* in the intermediate-poor subgroup could be linked to higher levels of *MICA*, which could be shed from the cell surface and cause less immunogenicity, and thus a worse prognosis.

## 5.2 Potential prognostic biomarkers for AML patients with AML-M2 categorized in the intermediate prognostic risk group

The group of patients with AML-M2 categorized in the intermediate risk group was the second to be analyzed by our algorithm in order to identify potential prognostic biomarkers of both gene expression and DNA methylation that could predict survival in this subgroup of patients.

By performing a gene expression analysis, we identified 58 candidate prognostic biomarker genes whose expression levels seem to be able to subdivide AML-M2 patients into two groups with distinct prognosis. Moreover, 19 of the 58 identified genes were already cited in AML literature.

The *ABCB1* gene is an example of one of the identified genes that was already cited in an AML-related article. This gene codifies for a P-glycoprotein, a member of the ATP binding cassette (ABC) transporter superfamily.[73] The ABC transporters are proteins localized in the cell membranes responsible for the translocation of solutes across the membrane, using energy generated from the ATP hydrolysis. In AML patients treated with intensive chemotherapy (combination of anthracycline and cytarabine), the high expression levels of *ABCB1* was related with lower CR rates and higher relapse rates, being a poor prognostic factor that confers worse overall- and event-free survival. Moreover, the main cause of leukemia related death is drug resistance and it has been hypothesized that ABCB1 is related to drug resistance, since these transporters export the therapeutic drugs out of the target cell. However, this association is not completely clear yet.[73] According to our results, based on the determined expression cutpoint of *ABCB1* gene, the patients with AML-M2 categorized with intermediate prognostic risk group that represent the subgroup with worse prognosis display high expression levels of the *ABCB1* gene. This result is in concordance with the findings described in the literature. It is possible that patients in the subgroup with worse prognosis, characterized by high expression levels of *ABCB1*, can develop resistance to the AML chemotherapy. As the treatment is less effective in AML patients with high expression of *ABCB1*, this subgroup of patients has lower survival rates.

The alpha-chain of the interleukin-3 receptor (IL-3RA), also known as CD123, is a subunit of the IL-3 receptor. Together with the β-subunit of the receptor, it promotes high-affinity binding to IL-3, which is primarily produced by T-lymphocytes.[74] The binding of IL-3 to its receptor stimulates hematopoietic cell's cycle progression and differentiation, and inhibits apoptosis.[74] Studies showed that the high expression of IL-3RA is present in hematological malignancies, such AML, and that it can confer a proliferative advantage to leukemic cells.[75] Furthermore, high expression levels of IL-3RA were also associated with reduced patient survival, being a poor prognostic factor.[73] Our results suggest that, when the patients with AML-M2 categorized in the intermediate prognostic risk group are subdivided based on the expression cutpoint of the *IL-3RA* gene, the subgroup with worse prognosis is related with high expression levels of this gene, which is in concordance with the literature. However, these results refer specifically to FAB-M2 AML patients, and the literature about *IL-3RA* gene expression in AML does not discriminate any subtypes of AML.

After identifying the candidate prognostic biomarkers of gene expression for intermediate-risk AML-M2 patients, we compared the subgroups of patients generated by the determined expression cutpoints. By performing a gene set enrichment analysis, we evaluated what gene sets in biological processes were systematically enriched between the subgroups with worse prognosis (intermediate-poor) and the subgroups with better prognosis (intermediate-favorable).

We identified that sprouting angiogenesis, locomotor behavior, negative regulation of supramolecular fiber organization, Rho protein signal transduction, and regulation of Rho protein signal transduction were some of the identified gene sets that appeared to be downregulated in most of the AML-M2 intermediate-poor in comparison with the intermediate-favorable subgroups.

Rho proteins are GTPases that cycle through an active GTP-bound form and an inactive GDP-bound form.[76] This family of enzymes interacts with downstream effectors that are involved in several cellular processes, such as cytoskeleton dynamics.[76] In fact, we also identified a downregulation of the biological process "negative regulation of supramolecular fiber organization", which might be related to the Rho signal transduction. However, it is important to observe that we do not know if the downregulation of the biological processes related to the Rho protein signal transduction are negative or positive regulation processes. Similarly, it is not clear if the "negative regulation of supramolecular fiber organization" refers to a polymerization or depolymerization process.

Nonetheless, Rho GTPases have been implicated in both malignant transformation and tumor development, contributing to processes like development of an inflammatory environment and induction of tumoral angiogenesis. Curiously, another downregulated biological process in the intermediate-poor subgroup is the "sprouting angiogenesis", which could also be related to the Rho protein signal transduction downregulation.

Increased angiogenesis is usually associated with an unfavorable prognosis in AML[77], but it is not possible for us to conclude that our results are in agreement with this association, since we do not know which type of angiogenic process is being downregulated.

It is also not completely clear how these downregulated biological processes are related to an unfavorable prognosis since they refer to complex networks of child biological processes with both negative and positive regulatory roles.

We identified several gene sets related to cellular respiration and ATP synthesis that were upregulated in the intermediate-poor subgroup, comparatively to the intermediate-favorable subgroup. These biological processes include mitochondrion organization, ATP synthesis coupled proton transport, energy coupled proton transport down electrochemical gradient, respiratory electron transport chain, and cellular respiration.

High proliferating cells have an increased energy demand, and high levels of ATP synthesis, especially through an aerobic pathway, consequently, cause a rise in cellular ROS.[78] It is known that ROS are key mediators in normal hematopoiesis. In fact, ROS is one of the triggers for HSC to undergo differentiation and proliferation. Although, ROS levels are maintained at relatively low levels in normal HSC, they are aberrantly increased in leukemic cells and are a known stimulator of myeloid leukemogenesis.[78]

Focusing on the DNA methylation analysis, our algorithm identified 591 genes whose DNA methylation appears to have a predictive value of survival in patients with AML-M2 categorized in the intermediate prognostic risk group. 180 of the 592 identified genes were already described in AML-related articles. Of the genes already cited in the AML literature, our algorithm identified the distal-less homeobox 4 (*DLX4*) gene as a potential prognostic biomarker of DNA methylation in patients with AML-M2 categorized in the intermediate prognostic risk group. We identified 2 CpG sites (cg10592171 and cg12224030) with a potential prognostic value localized in the promotor region of the *DLX4* gene. In both CpG sites, hypermethylation is related with the subgroup with worse prognosis. The clinical relevance of DLX4 methylation in *de novo* AML patients was investigated by Zhou *et al.*[79] This study showed that AML patients had a significant *DLX4* methylation in comparison with controls. Moreover, in comparison with the patients with unmethylated DLX4, they observed that all AML and non-M3 patients had a significant lower rate of complete remission. Furthermore, all AML, non-M3 AML, and cytogenetically normal AML cases with DLX4 methylation had a significantly shorter overall survival. As such, DLX4 methylation was considered an independent risk factor in all-AML and non-M3 AML patients, predicting a poor clinical outcome in *de novo* AML.[79] These findings are in accordance to our results, which also identified DLX4 hypermethylation as a negative prognostic factor, but only in AML-M2 patients. Zhou *et al*. argued that it is possible that a mutation in the U2AF1 gene, which encodes for a small subunit of the U2 Auxiliary Factor complex (one of the components of the spliceosome), could potentially trigger DLX4 methylation through the DNMT pathway during leukemogenesis, but this mechanism is not clear and further studies are needed.

### 5.3 Potential prognostic biomarkers for AML patients with AML-M4 categorized in the intermediate prognostic risk group

By performing the gene expression analysis, our algorithm identified 4 genes whose expression levels appear to be able to predict prognosis in the studied AML-M4 patients. The *CCDK* gene is the only identified gene that was already cited in AML-related articles. This gene a protein member of the cyclin family known as cyclin K.[80] The function of cyclin K is not yet completely understood. However, previous studies showed that the detection of this protein in non-proliferative human tissues is hard, but it is highly expressed in stem cells with rapid proliferation. In fact, there is a positive correlation between the expression of cyclin K and cellular proliferation.[80] Our results suggest that when the patients with AML-M4 categorized in the intermediate prognostic risk group are subclassified based on the determined expression cutpoint of *CCNK* gene, the subgroup with worse prognosis is characterized by high expression levels of *CCNK*. Comparing with the previous studies, maybe in AML-M4 patients with high levels of *CCNK* expression, the leukemic cells exhibit more proliferation than the leukemic cells with low expression of *CCNK*, leading to a faster progression of the disease and less survival time for the patients, which potentially explains why it seems to be a poor prognostic factor in these patients.

Our analysis identified some biological processes that seemed to be downregulated in the AML-M4 intermediate-poor patients in comparison with the intermediate-favorable subgroups, such as branching morphogenesis of an epithelial tube, detection of stimulus involved in sensory perception and detection of chemical stimulus involved in sensory perception. The first mentioned GO term is related to a decreased generation and organization of epithelial tubes, such as blood vessels. The remaining GO terms are linked to a decrease in sensory perception like pain, smell, taste. According to the literature, it has been described that sensory losses are one of the symptoms of leptomeningeal AML.[81] Leptomeningeal AML is usually diagnosed through several neurological symptoms, such as impaired vision, hearing deficits, sensory losses, or vertigo.[81] Although leptomeningeal involvement in AML is rare, it has a higher prevalence in patients with FAB-M4 and FAB-M5 subtypes, which might explain why we found such enrichment in the FAB-M4 cohort. The reach of AML to the central nervous system is strong indicator of poor overall survival, decreased disease-free survival, and a diminished rate of complete response.[81,82] Therefore, it is possible that our results indicate that the group of patients with low-prognosis have a down-regulation of sensory perception due to an involvement of the malignancy in the central nervous system.

In contrast, some biological processes were found to be upregulated in the AML-M4 intermediate-poor patients in comparison with the intermediate-favorable patients. Negative regulation of cell cycle process, RNA catabolic process, proteasomal protein catabolic process, mRNA processing, and RNA splicing are some examples of biological processes that appeared to be upregulated in the intermediate-poor subgroups. The first mentioned GO term refers to a decrease in a cellular process that is involved with cell cycle progression, which can be a process that either promotes or inhibits the cell cycle. The other GO terms seem to indicate an increase in protein turnover in the intermediate-poor patients, involving more mRNA processing, which could suggest more protein formation, and protein catabolic processes mediated by proteasome, which could suggest more protein degradation.[83] Cancer cells have an increased need for protein production and degradation. In fact, inhibition of proteasome activity in tumor cells leads to a block in cellular proliferation and an activation of apoptosis. In AML, the inhibition of proteasome activity has been explored as potential treatment and seems to show promising results.[83] Therefore, we theorized that in AML-M4 intermediate-poor subgroups there is an negative regulation of processes that inhibit the cell cycle progression, and also an increased protein turnover. These two combined, could promote the proliferation of leukemic cells, generating a more aggressive malignancy, which could explain why the subgroups where these biological processes are upregulated have worse prognosis in comparison to the other subgroups.

By performing the DNA methylation analysis, we identified 330 genes whose methylation levels were able to differentiate survival in patients with AML-M4 categorized in the intermediate prognostic risk group. One of these genes already described in AML-related articles was the Adenomatous polyposis col 2 (APC2), a tumor suppressor gene that encodes for a protein that negatively regulates beta-catenin. Our data shows that *APC2* hypomethylation in the 5'UTR region is a predictor for poor-prognosis in AML-M4 patients categorized with intermediate prognostic risk group. This gene has already been cited in the AML-literature by Xia Y. and colleagues, who described that APC2 promoter methylation levels did not seem to be affected by various chemotherapy regimens.[84] It is also possible that prognostic-predicting capabilities of the methylation of APC2 5'UTR region is not because the gene itself, but a consequence of another unknown cellular event.

Another identified gene that was already described in AML-related articles and whose DNA methylation appears to have a predictive value of survival in the AML-M4 intermediate-risk patients is the *SNRPN* gene. Based on the patient division by the identified methylation

*SNRPN* cutpoint, the subgroup with worse prognosis is related with *SNRPN* hypermethylation at the promotor region. Benetatos *et al*. studied the *SNRPN* methylation at the promotor region in 42 patients diagnosed with AML, and observed that 21 displayed hypermethylation of *SNRPN*.[85] Even though Benetatos's results seemed to suggest that abnormal methylation of *SNRPN* could be a characteristic event in AML, they did not find any association between *SNRPN* hypermethylation and the survival of AML patients.[85] Our finding contrast with Benetatos *et al.* results, possibly because we found an association of *SNRPN* hypermethylation with survival specifically in intermediate-risk AML-M4 patients, while Benetatos studied AML patients from different subclassifications and risk categories.

### 5.4 Potential prognostic biomarkers for AML patients with AML-M5 categorized in the intermediate prognostic risk group

By performing the gene expression analysis, we identified 32 genes whose expression appears to be able to predict survival in the studied intermediate-risk AML-M5 patients. Of these 32 identified genes, 5 were already described in the AML literature. An example of these five genes is the phospholipase A2 Group IVA (*PLA2G4A*), a gene that encodes an enzyme that catalyzes the hydrolysis of membrane phospholipids to release arachidonic acid and lysophospholipids.[86] This gene's survival-predicting capabilities in AML patients has already been reported by Bai and colleagues, who evaluated the prognostic value of *PLA2G4A* expression levels in non-M3/ NPM1 wildtype AML patients. They observed that the *PLA2G4A* gene is highly expressed in non-M3 AML samples in comparison to normal peripheral blood samples. Moreover, this elevated *PLA2G4A* expression was associated with a significantly shorter overall survival of AML patients. Furthermore, the group also described that *PLA2G4A* expression can possibly be an independent prognostic biomarker of OS in non-M3/NPM1 wildtype-AML patients.[86]

Our data suggests that, based on the determined expression cutpoint, *PLA2G4A* expression levels are able to subdivide the AML-M5 with intermediate-risk patients into two subgroups with significant different overall survival. In this subdivision, high expression levels of *PLA2G4A* is related to worse prognosis, which is in accordance with the results described by Bai *et al*. By promoting the release arachidonic acid and lysophospholipids, PLA2G4A may act in many signaling pathways, such as in the activation of the PI3K/Akt pathway.[86] Although PLA2G4A is likely to be implicated in several cellular processes that influence

AML development, one way to explain the inverse association with this gene's expression and M5-AML survival is that arachidonic acid can be metabolized by cyclooxygenase to synthetize prostaglandins, which have been described as key players in cell cycle progression. It has also been shown that leukotrienes, which also stem from arachidonic acid, can induce cell proliferation in several cell types and that its deregulation directly causes uncontrolled cell proliferation.[86]

Another example of gene already cited in AML is the Beta-2-Microglobulin (*B2M*) gene. Our results suggest that when the patients with AML-M5 categorized in the intermediate prognostic risk group are subdivided based on the expression *B2M* cutoff, the low expression of *B2M* is related with the subgroup of patients with a worse prognosis. Tsimberidou *et al.*, observed that in older patients (with ages of 60 or more years) with newly diagnosed AML, high serum levels of β2M were associated with poor survival.[87] Since our analysis aimed to identify potential prognostic biomarkers that could predict survival regardless of the age of the patients, this could explain why our results are not in accordance with the findings described by Tsimberidou *et al*. Furthermore, our study was specifically focused in intermediate-risk AML-M5 patients, which could also, in part, explain the difference in the results.

The subgroups of intermediate-risk AML-M5 patients, generated based on the expression cutpoint of the identified candidate gene expression biomarkers were also compared using the GSA methodology to know what gene sets and biological processes were differentially expressed between the two prognostic clusters.

Some examples of biological process-related gene sets that seemed to be downregulated in the majority of the intermediate-poor subgroups included leaflet of membrane bilayer and antigen processing and presentation of exogenous peptide antigen. The downregulation of these biological processes seems to suggest that in the intermediate-poor subgroups there is a lower presentation of exogenous antigens. Lower levels of leukemic antigens result in lower immunogenicity, in fact it has been described that cytotoxic T cells preferentially kill leukemia cells with higher expression levels of leukemic antigens like Neutrophil elastase (NE) and proteinase 3 (P3), which could explain why the subgroup with downregulation of biological processes related to exogenous antigen presentation have the worse prognosis.[88] Furthermore, it has also been shown that the levels of the antigens NE and P3 are positively correlated with remission status in AML patients.[88]

On the other hand, some biological process-related gene sets were also found to be upregulated in most of the intermediate-poor subgroups in comparison with the intermediate-favorable subgroups. Some examples include protein localization to cilium, microtubule-based protein transport, protein transport along microtubule, intraciliary transport, and cilium organization. In 2017, Singh and colleagues identified primary cilia in leukemia cells.[89] Primary cilia are microtubule-based organelles that are important for the function of signaling pathways such as the Wnt and Hedgehog pathways. Singh described that, in leukemic cells, the primary cilia often displayed aberrant morphologies, which could result in aberrant activation of the hedgehog pathway. Although the link between the former two is not clearly established in leukemic cells, it is possible that the upregulation of these biological processes in intermediate-poor patients could be related to primary cilia and/or aberrant activation of pathways like Wnt and Hedgehog.[89]

### 5.5. Potential prognostic biomarkers for patients with AML-M0-M1-M2-M4-M5 categorized in the intermediate prognostic risk group

Finally, we aimed to search for candidate prognostic biomarkers of both gene expression and DNA methylation that could predict survival of intermediate-risk AML patients, independently of its FAB classification. To achieve this, we performed our biomarker identification analysis using all intermediate-risk AML patients for which we had gene expression and DNA methylation data, which included patients classified as FAB-M0, FAB-M1, FAB-M2, FAB-M4, and FAB-M5.

The stromal cell-derived factor 1 (*SDF-1*) was one of the genes that was identified by our algorithm and one of the cited in the AML literature. The *SDF-1* gene, *CXCL12* encodes a chemokine that is expressed and produced by the bone marrow stromal cells and promotes migration and homing of HSCs and progenitor cells when it binds to their receptor (CXCR4).[90] In AML, it was already demonstrated that the binding of CXCL12 to CXCR4 promotes the development and progression of the disease. A high expression of CXCR4 was identified in AML patients with a significantly reduced survival rate and a high probability of relapse. Moreover, it was also found that the AML cells constitutively express and secrete CXCL12, which plays a role in migration and proliferation of leukemic cells. In addition, high levels of CXCR4-expressing vesicles and CXCL12 were identified in serum samples of AML patients in comparison with normal individuals.[91] Another study suggested that the

CXCL12 expression by AML cells confers survival advantage and participates in the autonomous growth to these cells.[92] Our results suggest that the subgroup of AML patients with worse prognosis are represented by lower expression levels of the *CXCL12* gene, in comparison to the intermediate-favorable AML subgroups. Our results are contradictory with the findings above described.

The *PDE7B* gene is other example of an identified candidate prognostic biomarker that was already described in the AML literature. According to our results, the intermediate-poor patients display high expression levels of *PDE7B* gene. Han and colleagues studied the effect of the expression levels of *PDE7B* in the prognostic of patients with CN-AML.[93] They observed that the patients with high expression levels of *PDE7B* gene showed a significant reduction in event-free survival and overall survival. Moreover, the *PDE7B* gene showed to be an independent risk predictor of poor prognosis in patients with CN-AML.[93] These findings seem to be in concordance with our results.

Through the gene set analysis, we identified some biological processes gene sets that appear to be differently enriched in the majority of the intermediate-poor in comparison with the intermediate-favorable subgroups. For example, we observed that the organophosphate catabolic process gene set was upregulated in the intermediate-poor subgroups. One of the main drugs used in AML therapeutics is Cytarabine.[94] This compound is converted intracellularly into the active form cytosine arabinoside triphosphate.[94] The upregulation of the organophosphate catabolic process in the intermediate-poor subgroups may indicate an increase in cytosine arabinoside triphosphate metabolism. Moreover, we also found an upregulation of the snRNA processing, snRNA 3'-end processing, and ncRNA 3'-end processing gene sets. Small nuclear RNA (snRNA) complex with proteins to form the small nuclear ribonucleoproteins (snRNPs), which are part of the spliceosome.[95] These snRNPs are necessary for the pre-mRNA splicing process. In AML, alternative RNA splicing can contribute to drug resistance. For example, it has been shown that alternative splicing can produce an enzymatically inactive deoxycytidine kinase (dCK), thus contributing to drug resistance.[95] We reasoned that the leukemic cells of the intermediate-poor patients might be drug resistant, which could explain the poor survival of these patients.

Through the application of our prognostic biomarker searching algorithm, we also identified 264 genes whose DNA methylation appears to be potentially able to predict prognosis in patients with AML-M0-M1-M2-M4-M5 categorized in the intermediate prognostic risk group. We also verified that 70 of the 264 identified genes were already cited in the AML literature.

An example of an identified candidate prognostic biomarker gene that was already cited in AML is the DEAD box polypeptide 43 (*DDX43*) gene. Our results indicate that when the patients with AML-M0-M1-M2-M4-M5 categorized in the intermediate prognostic risk group are subdivided based on the DNA methylation cutpoint of cg17188169 in the *DDX43* gene, the subgroup with worse prognosis displays *DDX43* hypermethylation in the promotor region. Lin *et al.* investigated the methylation status at the promotor region of the *DDX43* gene and its clinical relevance in patients with primary AML.[96] The authors observed that the hypomethylation of the *DDX43* gene, at the promotor region, was present in primary AML, including in patients whose AML was categorized in the intermediate prognostic risk group, and that this hypomethylation was correlated with the expression levels of the gene. Moreover, they also observed that the AML patients with hypomethylation of the *DDX43* gene in comparison with patients with methylation of the *DDX43* gene had a better overall survival. Ultimately, they concluded that hypomethylation at promotor activates the *DDX43* gene, and that such activation can be a favorable prognostic factor in AML patients.[96] These findings seem to be in concordance with our results. We hypothesize that in the intermediate-poor subgroup, the hypermethylation at the *DDX43* promotor may inhibit the expression of the *DDX43* gene, leading to an increase in proliferation of the leukemic cells.

The Carnitine palmitoyl transferase 1A (*CPT1A*) gene was also identified by our algorithm as a candidate prognostic biomarker of DNA methylation and it was also already cited in the AML literature. According to our results, hypermethylation in the promotor of the *CPT1A* gene is related with subgroup with worse prognosis. Shi *et al.* studied the expression levels of the *CPT1A* gene in samples of AML patients and its relevance to the prognosis of the AML patients.[97] Their study showed that the higher levels of *CPT1A* expression were significantly associated with poor outcomes in cytogenetically normal AML patients. This gene encodes for a protein that is a rate-limiting enzyme of fatty acid β-oxidation, a metabolic pathway where it seems to be some evidence of cancer-associated aberrant gene expression.[97]

We do not know how the promoter methylation of the *CPT1A* gene is related to its transcription in our cohort, but it could be possible that the adverse prognosis conferred by promoter hypermethylation could be linked to a deregulation of the fatty acid β-oxidation pathway.

### 5.6. Study limitations

Although our study was able to identify several candidate prognostic biomarkers the analyzed cohort, it presents several limitations such as:

- Lack of normal samples, which excludes the possibility of understanding which alterations between prognostic groups are similar to normal patterns.
- Small sample size, which reduces statistical power in our hypothesis tests.
- The analyzed datasets can be enriched in certain features (like gender, age, or clinical characteristics) which could lead to false conclusions about the intermediate-risk AML population.
- Even though we developed a searching prognostic biomarker algorithm that can be used in other datasets from different diseases, it has not undergone methodological validation.

# CHAPTER 6

# CONCLUSION

Our developed algorithm was able to identify potential prognostic biomarkers of gene expression and DNA methylation that were able to distinguish survival in patients with FAB M1, M2, M4, and M5 subtypes categorized in the intermediate prognostic risk group. Moreover, some potential biomarkers were also found for the FAB M0, M1, M2, M4, and M5 AML patients, without the subtype distinction.

For some identified candidate genes, their role in the development of AML as well as their prognostic value were described in previous studies, and for other candidate genes, their prognostic potential is still being researched. However, biomarker as well as algorithm validation are necessary to confirm the prognostic value of the identified candidate genes.

Moreover, we also identified that, although the majority of the identified potential biomarkers generate different subgroup of intermediate AML patients, most of the intermediate-poor subgroups share some gene sets that appeared to be upregulated and downregulated in comparison with the intermediate-favorable subgroups.

In summary, our data suggests that both DNA methylation and gene expression are valuable tools that can be used to stratify intermediate-risk AML patients of various FAB subtypes into subgroups with distinct overall survival. This can be useful for a better understanding and management of AML.

## Bibliography

1. Gibney, E. R. & Nolan, C. M. Epigenetics and gene expression. *Heredity (Edinb).* **105**, 4–13 (2010).

2. Antoniani, C., Romano, O. & Miccio, A. Concise Review: Epigenetic Regulation of Hematopoiesis: Biological Insights and Therapeutic Applications. *Stem Cells Transl. Med.* **6**, 2106–2114 (2017).

3. Chen, Z., Li, S., Subramaniam, S., Shyy, J. Y.-J. & Chien, S. Epigenetic Regulation: A New Frontier for Biomedical Engineers. *Annu. Rev. Biomed. Eng.* **19**, 195–219 (2017).

4. Perri, F. *et al.* Epigenetic control of gene expression: Potential implications for cancer treatment. *Crit. Rev. Oncol. Hematol.* **111**, 166–172 (2017).

5. Wouters, B. J. & Delwel, R. Epigenetics and approaches to targeted epigenetic therapy in acute myeloid leukemia. *Blood* **127**, 42–52 (2016).

6. Kouzarides, T. Chromatin Modifications and Their Function. *Cell* **128**, 693–705 (2007).

7. Kim, J. K., Samaranayake, M. & Pradhan, S. Epigenetic mechanisms in mammals. *Cell. Mol. Life Sci.* **66**, 596–612 (2009).

8. Álvarez-Errico, D., Vento-Tormo, R., Sieweke, M. & Ballestar, E. Epigenetic control of myeloid cell differentiation, identity and function. *Nat. Rev. Immunol.* **15**, 7–17 (2015).

9. Dhanoa, J. K., Sethi, R. S., Verma, R., Arora, J. S. & Mukhopadhyay, C. S. Long non-coding RNA: its evolutionary relics and biological implications in mammals: a review. *J. Anim. Sci. Technol.* **60**, 25 (2018).

10. Ozata, D. M., Gainetdinov, I., Zoch, A., O'Carroll, D. & Zamore, P. D. PIWI-interacting RNAs: small RNAs with big functions. *Nat. Rev. Genet.* **20**, 89–108 (2019).

11. Moore, L. D., Le, T. & Fan, G. DNA Methylation and Its Basic Function. *Neuropsychopharmacology* **38**, 23–38 (2013).

12. Lee, D. D. *et al.* DNA hypermethylation within TERT promoter upregulates TERT expression in cancer. *J. Clin. Invest.* **129**, 223–229 (2018).

13. Rieger, M. A. & Schroeder, T. Hematopoiesis. *Cold Spring Harb. Perspect. Biol.* **4**, a008250–a008250 (2012).

14. Hoggatt, J. & Pelus, L. M. Hematopoiesis. in *Brenner's Encyclopedia of Genetics* vol. 3 418–421 (Elsevier, 2013).

15. He, S., Nakada, D. & Morrison, S. J. Mechanisms of stem cell self-renewal. *Annu. Rev. Cell Dev. Biol.* **25**, 377–406 (2009).

16. Alonso, S. *et al.* Human bone marrow niche chemoprotection mediated by cytochrome P450 enzymes. *Oncotarget* **6**, 14905–14912 (2015).

17. Borghesi, L. Hematopoiesis in Steady-State versus Stress: Self-Renewal, Lineage Fate

Choice, and the Conversion of Danger Signals into Cytokine Signals in Hematopoietic Stem Cells. *J. Immunol.* **193**, 2053–2058 (2014).

18. Bröske, A.-M. *et al.* DNA methylation protects hematopoietic stem cell multipotency from myeloerythroid restriction. *Nat. Genet.* **41**, 1207–1215 (2009).

19. Challen, G. A. *et al.* Dnmt3a and Dnmt3b Have Overlapping and Distinct Functions in Hematopoietic Stem Cells. *Cell Stem Cell* **15**, 350–364 (2014).

20. Nakajima, H. Role of Transcription Factors in Differentiation and Reprogramming of Hematopoietic Cells. *Keio J. Med.* **60**, 47–55 (2011).

21. Virani, S., Colacino, J. A., Kim, J. H. & Rozek, L. S. Cancer Epigenetics: A Brief Review. *ILAR J.* **53**, 359–369 (2012).

22. Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* **144**, 646–674 (2011).

23. Hassanpour, S. H. & Dehghani, M. Review of cancer from perspective of molecular. *J. Cancer Res. Pract.* **4**, 127–129 (2017).

24. Bustin, S. Molecular Biology of the Cell, Sixth Edition; ISBN: 9780815344643; and Molecular Biology of the Cell, Sixth Edition, The Problems Book; ISBN 9780815344537. *Int. J. Mol. Sci.* **16**, 28123–28125 (2015).

25. Bray, F. *et al.* Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA. Cancer J. Clin.* **68**, 394–424 (2018).

26. Hu, D. & Shilatifard, A. Epigenetics of hematopoiesis and hematological malignancies. *Genes Dev.* **30**, 2021–2041 (2016).

27. KITAMURA, T. *et al.* The molecular basis of myeloid malignancies. *Proc. Japan Acad. Ser. B* **90**, 389–404 (2014).

28. Lagunas-Rangel, F. A., Chávez-Valencia, V., Gómez-Guijosa, M. Á. & Cortes-Penagos, C. Acute Myeloid Leukemia-Genetic Alterations and Their Clinical Prognosis. *Int. J. Hematol. stem cell Res.* **11**, 328–339 (2017).

29. Bullinger, L., Döhner, K. & Dohner, H. Genomics of acute myeloid leukemia diagnosis and pathways. *J. Clin. Oncol.* **35**, 934–946 (2017).

30. Short, N. J., Rytting, M. E. & Cortes, J. E. Seminar Acute myeloid leukaemia. *Lancet* **392**, 593–606 (2018).

31. Longo, D. L., Döhner, H., Weisdorf, D. J. & Bloomfield, C. D. Acute Myeloid Leukemia. *N. Engl. J. Med.* **373 VN-**, 1136–1152 (2015).

32. De Kouchkovsky, I. & Abdul-Hay, M. 'Acute myeloid leukemia: A comprehensive review and 2016 update'. *Blood Cancer J.* **6**, (2016).

33. Shallis, R. M., Wang, R., Davidoff, A., Ma, X. & Zeidan, A. M. Epidemiology of acute myeloid leukemia: Recent progress and enduring challenges. *Blood Rev.* **36**, 70–87 (2019).

34. Kelly, L. M. & Gilliland, D. G. Genetics of myeloid leukemias. *Annu. Rev. Genomics Hum. Genet.* **3**, 179–98 (2002).

35. Al-Harbi, S., Aljurf, M., Mohty, M., Almohareb, F. & Ahmed, S. O. A. An update on the molecular pathogenesis and potential therapeutic targeting of AML with t(8;21)(q22;q22.1);RUNX1-RUNX1T1. *Blood Adv.* **4**, 229–238 (2020).

36. Delaunay, J. *et al.* Prognosis of inv(16)/t(16;16) acute myeloid leukemia (AML): a survey of 110 cases from the French AML Intergroup. *Blood* **102**, 462–9 (2003).

37. Saultz, J. & Garzon, R. Acute Myeloid Leukemia: A Concise Review. *J. Clin. Med.* **5**, 33 (2016).

38. Daver, N., Schlenk, R. F., Russell, N. H. & Levis, M. J. Targeting FLT3 mutations in AML: review of current knowledge and evidence. *Leukemia* **33**, 299–312 (2019).

39. Medeiros, B. C. *et al.* Isocitrate dehydrogenase mutations in myeloid malignancies. *Leukemia* **31**, 272–281 (2017).

40. Montalban-Bravo, G. & DiNardo, C. D. The role of IDH mutations in acute myeloid leukemia. *Futur. Oncol.* **14**, 979–993 (2018).

41. Melnick, A. M. Epigenetics in AML. *Best Pract. Res. Clin. Haematol.* **23**, 463–468 (2010).

42. Yang, X., Wong, M. P. M. & Ng, R. K. Aberrant DNA Methylation in Acute Myeloid Leukemia and Its Clinical Implications. *Int. J. Mol. Sci.* **20**, 4576 (2019).

43. Bennett, J. M. *et al.* Proposals for the Classification of the Acute Leukaemias French-American-British (FAB) Co-operative Group. *Br. J. Haematol.* **33**, 451–458 (1976).

44. Arber, D. A. *et al.* The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. *Blood* **127**, 2391–2405 (2016).

45. West, H. (Jack) & Jin, J. O. Performance Status in Patients With Cancer. *JAMA Oncol.* **1**, 998 (2015).

46. Döhner, H. *et al.* Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. *Blood* **129**, 424–447 (2017).

47. Martelli, M. P., Sportoletti, P., Tiacci, E., Martelli, M. F. & Falini, B. Mutational landscape of AML with normal cytogenetics: Biological and clinical implications. *Blood Rev.* **27**, 13–22 (2013).

48. Kassim, A. A. & Savani, B. N. Hematopoietic stem cell transplantation for acute myeloid leukemia: A review. *Hematol. Oncol. Stem Cell Ther.* **10**, 245–251 (2017).

49. Bewersdorf, J. P., Shallis, R., Stahl, M. & Zeidan, A. M. Epigenetic therapy combinations in acute myeloid leukemia: what are the options? *Ther. Adv. Hematol.* **10**, 204062071881669 (2019).

50. Duchmann, M. & Itzykson, R. Clinical update on hypomethylating agents. *Int. J. Hematol.* **110**, 161–169 (2019).

51. Pleyer, L. & Greil, R. Digging deep into "dirty" drugs – modulation of the methylation machinery. *Drug Metab. Rev.* **47**, 252–279 (2015).

52. Tomczak, K., Czerwińska, P. & Wiznerowicz, M. The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. *Wspolczesna Onkol.* **1A**, A68–A77 (2015).

53. Kukurba, K. R. & Montgomery, S. B. RNA Sequencing and Analysis. *Cold Spring*

*Harb. Protoc.* **2015**, 951–69 (2015).

54. Dedeurwaerder, S. *et al.* A comprehensive overview of Infinium HumanMethylation450 data processing. *Brief. Bioinform.* **15**, 929–941 (2014).

55. Wang, Z., Wu, X. & Wang, Y. A framework for analyzing DNA methylation data from Illumina Infinium HumanMethylation450 BeadChip. *BMC Bioinformatics* **19**, 115 (2018).

56. Kwak, S. K. & Kim, J. H. Statistical data preparation: management of missing values and outliers. *Korean J. Anesthesiol.* **70**, 407 (2017).

57. Hothorn, T. & Lausen, B. On the exact distribution of maximally selected rank statistics. *Comput. Stat. Data Anal.* **43**, 121–137 (2003).

58. Li, H., Han, D., Hou, Y., Chen, H. & Chen, Z. Statistical Inference Methods for Two Crossing Survival Curves: A Comparison of Methods. *PLoS One* **10**, e0116774 (2015).

59. Qiu, P. & Sheng, J. A two-stage procedure for comparing hazard rate functions. *J. R. Stat. Soc. Ser. B (Statistical Methodol.* **70**, 071103032514002-??? (2007).

60. Piazza, R. *et al.* OncoScore: a novel, Internet-based tool to assess the oncogenic potential of genes. *Sci. Rep.* **7**, 46290 (2017).

61. Kikuchi, J. *et al.* Minichromosome maintenance (MCM) protein 4 as a marker for proliferation and its clinical and clinicopathological significance in non-small cell lung cancer. *Lung Cancer* **72**, 229–237 (2011).

62. Simonetti, G. *et al.* Aneuploid acute myeloid leukemia exhibits a signature of genomic alterations in the cell cycle and protein degradation machinery. *Cancer* **125**, 712–725 (2019).

63. Kádková, A., Radecke, J. & Sørensen, J. B. The SNAP-25 Protein Family. *Neuroscience* **420**, 50–71 (2019).

64. Bhat, R., Bhattacharyya, P. K. & Ratech, H. An Immunohistochemical Survey of SNARE Proteins Shows Distinct Patterns of Expression in Hematolymphoid Neoplasia. *Am. J. Clin. Pathol.* **145**, 604–616 (2016).

65. Sun, Q. *et al.* SNAP23 promotes the malignant process of ovarian cancer. *J. Ovarian Res.* **9**, 80 (2016).

66. Kampen, K. R., ter Elst, A. & de Bont, E. S. J. M. Vascular endothelial growth factor signaling in acute myeloid leukemia. *Cell. Mol. Life Sci.* **70**, 1307–1317 (2013).

67. Sillar, Germon, DeIuliis & Dun. The Role of Reactive Oxygen Species in Acute Myeloid Leukaemia. *Int. J. Mol. Sci.* **20**, 6003 (2019).

68. Dulphy, N. *et al.* Underground Adaptation to a Hostile Environment: Acute Myeloid Leukemia vs. Natural Killer Cells. *Front. Immunol.* **7**, (2016).

69. Lion, E., Willemen, Y., Berneman, Z. N., Van Tendeloo, V. F. I. & Smits, E. L. J. Natural killer cell immune escape in acute myeloid leukemia. *Leukemia* **26**, 2019–2026 (2012).

70. Zhang, Z.-H. *et al.* Decreased SCIN expression, associated with promoter methylation, is a valuable predictor for prognosis in acute myeloid leukemia. *Mol. Carcinog.* **57**,

735–744 (2018).

71. Baragaño Raneros, A. *et al.* Methylation of NKG2D ligands contributes to immune system evasion in acute myeloid leukemia. *Genes Immun.* **16**, 71–82 (2015).

72. Kamimura, H. *et al.* Reduced NKG2D ligand expression in hepatocellular carcinoma correlates with early recurrence. *J. Hepatol.* **56**, 381–388 (2012).

73. Boyer *et al.* Clinical Significance of ABCB1 in Acute Myeloid Leukemia: A Comprehensive Study. *Cancers (Basel).* **11**, 1323 (2019).

74. Angelova, E. *et al.* CD123 expression patterns and selective targeting with a CD123-targeted antibody-drug conjugate (IMGN632) in acute lymphoblastic leukemia. *Haematologica* **104**, 749–755 (2019).

75. Testa, U., Pelosi, E. & Frankel, A. CD 123 is a membrane biomarker and a therapeutic target in hematologic malignancies. *Biomark. Res.* **2**, 4 (2014).

76. Orgaz, J. L., Herraiz, C. & Sanz-Moreno, V. Rho GTPases modulate malignant transformation of tumor cells. *Small GTPases* **5**, e983867 (2014).

77. Najafabadi, M. M., Shamsasenjan, K. & Akbarzadehalaleh, P. Angiogenesis status in patients with acute myeloid leukemia: From diagnosis to post-hematopoietic stem cell transplantation. *Int. J. Organ Transplant. Med.* **8**, 57–67 (2017).

78. Zhou, F., Shen, Q. & Claret, F. X. Novel roles of reactive oxygen species in the pathogenesis of acute myeloid leukemia. *J. Leukoc. Biol.* **94**, 423–429 (2013).

79. Zhou, J. *et al.* DLX4 hypermethylation is a prognostically adverse indicator in de novo acute myeloid leukemia. *Tumor Biol.* **37**, 8951–8960 (2016).

80. Lei, T. *et al.* Cyclin K regulates prereplicative complex assembly to promote mammalian cell proliferation. *Nat. Commun.* **9**, 1876 (2018).

81. Patel, S. A. Acute Myeloid Leukemia Relapse Presenting as Complete Monocular Vision Loss due to Optic Nerve Involvement. *Case Rep. Hematol.* **2016**, 1–4 (2016).

82. Rozovski, U. *et al.* Incidence of and risk factors for involvement of the central nervous system in acute myeloid leukemia. *Leuk. Lymphoma* **56**, 1392–1397 (2015).

83. Csizmar, C. M., Kim, D.-H. & Sachs, Z. The role of the proteasome in AML. *Blood Cancer J.* **6**, e503–e503 (2016).

84. Xia, Y. *et al.* APC2 and CYP1B1 methylation changes in the bone marrow of acute myeloid leukemia patients during chemotherapy. *Exp. Ther. Med.* **12**, 3047–3052 (2016).

85. Benetatos, L. *et al.* CpG methylation analysis of the MEG3 and SNRPN imprinted genes in acute myeloid leukemia and myelodysplastic syndromes. *Leuk. Res.* **34**, 148–153 (2010).

86. Bai, H., Zhou, M., Zeng, M. & Han, L. PLA2G4A Is a Potential Biomarker Predicting Shorter Overall Survival in Patients with Non-M3/ NPM1 Wildtype Acute Myeloid Leukemia. *DNA Cell Biol.* **39**, 700–708 (2020).

87. Tsimberidou, A.-M. *et al.* The Prognostic Significance of Serum 2 Microglobulin Levels in Acute Myeloid Leukemia and Prognostic Scores Predicting Survival:

Analysis of 1,180 Patients. *Clin. Cancer Res.* **14**, 721–730 (2008).

88. Alatrash, G. *et al.* The Role of Antigen Cross-presentation From Leukemia Blasts on Immunity to the Leukemia-associated Antigen PR1. *J. Immunother.* **35**, 309–320 (2012).

89. Singh, M., Chaudhry, P. & Merchant, A. A. Primary cilia are present on human blood and bone marrow cells and mediate Hedgehog signaling. *Exp. Hematol.* **44**, 1181-1187.e2 (2016).

90. Peled, A. & Tavor, S. Role of CXCR4 in the Pathogenesis of Acute Myeloid Leukemia. *Theranostics* **3**, 34–39 (2013).

91. Kalinkovich, A. *et al.* Functional CXCR4-Expressing Microparticles and SDF-1 Correlate with Circulating Acute Myelogenous Leukemia Cells. *Cancer Res.* **66**, 11013–11020 (2006).

92. Kim, H.-Y. *et al.* Endogenous stromal cell-derived factor-1 (CXCL12) supports autonomous growth of acute myeloid leukemia cells. *Leuk. Res.* **37**, 566–572 (2013).

93. Cao, L. *et al.* The Prognostic Significance of PDE7B in Cytogenetically Normal Acute Myeloid Leukemia. *Sci. Rep.* **9**, 16991 (2019).

94. Karijolich, J. & Yu, Y.-T. Spliceosomal snRNA modifications and their function. *RNA Biol.* **7**, 192–204 (2010).

95. Zhou, J. & Chng, W.-J. Aberrant RNA splicing and mutations in spliceosome complex in acute myeloid leukemia. *Stem Cell Investig.* **4**, 6–6 (2017).

96. Lin, J. *et al.* DDX43 promoter is frequently hypomethylated and may predict a favorable outcome in acute myeloid leukemia. *Leuk. Res.* **38**, 601–607 (2014).

97. Shi, J. *et al.* High Expression of CPT1A Predicts Adverse Outcomes: A Potential Therapeutic Target for Acute Myeloid Leukemia. *EBioMedicine* **14**, 55–64 (2016).