IBIMA Publishing
Communications of International Proceedings
https://ibimapublishing.com/p-articles/47AI/2026/4713326/
Vol. 2026 (12), Article ID 4713326

Mathematical Modeling of Epigenetic Regulation: Correlating DNA Methylation and Gene Expression in Breast Cancer Analysis*

Hubert BŁASZKIEWICZ and Robert WASZKOWSKI

Military University of Technology, Faculty of Cybernetics
Institute of Computer and Information Systems
Kaliskiego St. 2, 00-908 Warsaw, Poland

Correspondence should be addressed to: Hubert BŁASZKIEWICZ, u66516@student.wat.edu.pl

*Presented at the 47th IBIMA International Conference, 29-30 June 2026, Madrid, Spain

Abstract

This study explores the computational integration of multi-omics data to analyze the regulatory mechanisms of gene expression in cancer. Specifically, it focuses on the correlation between DNA methylation in promoter regions and transcriptomic expression levels. A mathematical model is developed to define patients as data vectors containing paired high-dimensional methylation and expression values. Using this model, a formal aggregation function is defined to reduce noise from individual CpG probes, establishing a metric for regulatory strength. The model is validated using a case study of the *BRCA1* gene in the TCGA-BRCA cohort ($n = 498$). The analysis reveals a statistically significant negative correlation ($r = -0.32$, $p < 0.001$), confirming the hypothesis of epigenetic silencing. This framework provides a standardized approach for bioinformatics pipelines aiming to identify epigenetically regulated tumor suppressor genes.

Keywords: bioinformatics, data science, mathematical modeling, API integration, DNA methylation, gene expression, TCGA.

Introduction

Breast cancer constitutes a heterogeneous group of neoplasms whose etiology is inextricably linked to the accumulation of both genetic errors and epigenetic alterations. A critical role in the carcinogenesis process is played by the dysregulation of tumor suppressor genes, among which *BRCA1* (Breast Cancer 1) is paramount. This gene is responsible for DNA repair via homologous recombination, and the loss of its function inevitably leads to genomic instability (Koboldt et al., 2012).

While genetic mutations are a well-documented cause of loss of function, recent studies emphasize the role of DNA methylation—a mechanism regulating transcriptional activity without altering the nucleotide sequence (Bird, 2002). Methylation involves the covalent attachment of a methyl group (-CH_3) to cytosine, typically within CpG dinucleotides. Two key methylation states are critical for pathogenesis. First, hypermethylation of promoter regions involves an increase in methylation levels that hinders the binding of transcription factors, ultimately leading to gene silencing and reduced expression. Conversely, global hypomethylation refers to a decrease in methylation levels across the rest of the genome, a condition first described decades ago that may lead to the activation of oncogenes and chromosomal instability (Feinberg and Vogelstein, 1983; Esteller, 2008; Laird, 2003).

Cite this Article as: Hubert BŁASZKIEWICZ and Robert WASZKOWSKI Vol. 2026 (12) "Mathematical Modeling of Epigenetic Regulation: Correlating DNA Methylation and Gene Expression in Breast Cancer Analysis" Communications of International Proceedings, Vol. 2026 (12), Article ID 4713326, https://doi.org/10.5171/2026.4713326

In the domain of bioinformatics and data science, modeling approaches often focus on statistical testing. However, a formal mathematical description of the data structures—treating patients as vectors of multi-omics features—allows for more precise algorithmic definitions. This study presents a functional model of patient data, enabling the definition of aggregation and correlation functions to evaluate epigenetic silencing algorithmically.

Related Work

The relationship between epigenetics and cancer has been extensively studied, providing a theoretical foundation for our mathematical modeling. The conceptualization of cancer not merely as a genetic disease, but as a complex genomic and epigenomic disruption, was famously articulated in the updated hallmarks of cancer (Hanahan and Weinberg, 2011). Esteller (2008) provided a foundational overview, highlighting that promoter hypermethylation is a hallmark of tumor suppressor silencing, a concept extensively expanded upon in subsequent literature covering the broader epigenetic landscape (Sharma et al., 2010; Kulis and Esteller, 2010). His work established that epigenetic changes are as significant as genetic mutations, proving that genetics and epigenetics are inextricably linked in pathogenesis (You and Jones, 2012). Building on this, Jones and Baylin (2007) introduced the concept of the "epigenetic progenitor," suggesting that epigenetic alterations might precede genetic mutations in cancer development, a perspective further reinforced by extensive reviews on the cancer epigenome (Baylin and Jones, 2011; Shen and Laird, 2013; Dawson and Kouzarides, 2012).

In the specific context of breast cancer and the *BRCA1* gene, the phenomenon of epigenetic silencing is well-documented. Early studies by Rice et al. (2000) and Catteau and Morris (2002) identified that aberrant methylation of the *BRCA1* CpG island promoter directly correlates with decreased mRNA levels in sporadic breast tumors. This epigenetic downregulation creates a "BRCAness" phenotype, mirroring the clinical presentation of familial cases driven by germline mutations (Stefansson et al., 2011; Herman and Baylin, 2003; Robertson, 2005). Koboldt et al. (2012) utilized comprehensive molecular profiling to classify tumor subtypes within The Cancer Genome Atlas (TCGA) project, demonstrating the undeniable utility of multi-omics approaches in identifying therapeutic targets (Weinstein et al., 2013; Tomczak et al., 2015). This was further refined by pan-cancer analyses and specific profiling of invasive breast cancer subtypes, highlighting the profound diversity of epigenetic patterns across tumor tissues (Hoadley et al., 2014; Ciriello et al., 2015).

Quantifying the precise impact of methylation on expression requires robust computational models. The advent of high-throughput technologies has spurred the development of advanced statistical and bioinformatic frameworks to integrate disparate data types (Teschendorff and Relton, 2018; Wang et al., 2014; Huang et al., 2017). Packages such as TCGAbiolinks have democratized access to paired patient data, allowing researchers to build automated pipelines for multi-omics integration (Colaprico et al., 2016; Silva et al., 2016). Furthermore, the translation of biological signals into formalized mathematical vectors—similar to methods used in signal processing and quantitative data aggregation—enables rigorous algorithmic definitions (Pery and Waszkowski, 2024; Gevaert et al., 2013; Kristensen et al., 2014). This study builds upon these molecular and computational foundations by proposing a strictly mathematical approach to evaluating regulatory strength.

Mathematical Model

To rigorously analyze the dependency between the epigenetic state and transcriptomic output, we propose a formal mathematical model. Let the set of all patients in the cohort be denoted as P.

Definition of the Patient Vector

We define a patient p_i as an ordered pair of high-dimensional vectors representing their molecular profile (see Eq. 1):

$$p_i = (M_i, E_i), \quad p_i \in P, i \in \{1, \dots, n\}$$

Here, n represents the total number of patients in the studied cohort, which in this study, after filtering for clinical data completeness, is equal to 498. The variable M_i denotes the methylation profile vector, while E_i represents the gene expression scalar.

Methylation Vector Space

The methylation data for the *BRCA1* promoter is derived from the Illumina Human Methylation 450k platform. We define the methylation vector M_i for the i-th patient as a vector of Beta-values (see Eq. 2 and 3):

$$M_i = [\beta_{i,1}, \beta_{i,2}, \dots, \beta_{i,k}]$$

$$\beta_{i,j} \in [0,1], \quad k \in \mathbb{N}$$

The component $\beta_{i,j}$ represents the methylation level of the j-th CpG probe located in the promoter region. This functional region is defined as the set of genomic coordinates annotated as TSS1500, TSS200, 5'UTR, or 1stExon. The variable k signifies the number of identified probes in this region, which is 48 in our study.

Since analyzing individual probes may introduce single-probe bias (noise), we define an aggregation function $\mathcal{F}_{agg}$ that maps the vector M_i to a global methylation index μ_i (see Eq. 4):

$$\mu_i = \mathcal{F}_{agg}(M_i) = \frac{1}{k}\sum_{j=1}^{k} \beta_{i,j}$$

This mean methylation level μ_i serves as the primary independent variable in our correlation model.

Expression Normalization Function

The transcriptomic data, obtained via RNA-Seq (Illumina HiSeq), is initially represented as raw counts. To enable comparative analysis between samples of different sequencing depths, we define the normalization function $\mathcal{F}_{norm}$ based on the Counts Per Million (CPM) method (see Eq. 5):

$$\epsilon_i = \mathcal{F}_{norm}(c_i) = \log_2\left(\frac{c_i}{\sum C_i} \cdot 10^6 + 1\right)$$

In this formulation, c_i represents the raw count for the *BRCA1* gene, and $\sum C_i$ is the total library size for patient p_i. The logarithmic transformation is systematically applied to stabilize the variance of the expression data, making it highly suitable for linear correlation analysis.

Correlation Metric

The research hypothesis H_1 posits a negative dependence between μ and ϵ. This is verified using the Pearson correlation coefficient r, defined within our model as (see Eq. 6):

$$r(\mu, \epsilon) = \frac{\sum_{i=1}^{n}(\mu_i - \bar{\mu})(\epsilon_i - \bar{\epsilon})}{\sqrt{\sum_{i=1}^{n}(\mu_i - \bar{\mu})^2}\sqrt{\sum_{i=1}^{n}(\epsilon_i - \bar{\epsilon})^2}}$$

where $\bar{\mu}$ and $\bar{\epsilon}$ are the respective arithmetic means of the methylation and expression vectors.

Methodology and Implementation

From a data science perspective, the integration of multi-omics data presents significant challenges related to high dimensionality, heterogeneous data structures, and computational complexity (Huber et al., 2015; Gentleman et al., 2004). To address these issues, we designed an automated bioinformatic pipeline in the R programming environment (R Core Team, 2023). The system architecture involves API communication, data transformation, and matrix operations utilizing specialized packages such as TCGAbiolinks and SummarizedExperiment.

Data Acquisition via REST API

The raw data were dynamically retrieved from the NCI Genomic Data Commons (GDC) database. Instead of manual dataset downloads, we implemented an automated data querying mechanism using the GDC REST API

(Grossman et al., 2016). The query targeted the TCGA-BRCA project, strictly filtering for "Primary Tumor" samples.

The code below demonstrates the programmatic approach to querying and downloading the high-dimensional DNA methylation arrays. The data were parsed directly into a SummarizedExperiment object, which efficiently handles complex multi-assay data structures in memory.

```
# Formulating the API query for DNA methylation data
query_met <- GDCquery(
  project = "TCGA-BRCA",
  data.category = "DNA Methylation",
  data.type = "Methylation Beta Value",
  platform = "Illumina Human Methylation 450",
  sample.type = "Primary Tumor"
)

# Downloading and preparing the data structure
GDCdownload(query_met, method = "api", files.per.chunk = 10)
met_se <- GDCprepare(query_met, summarizedExperiment = TRUE)

# Extraction of the high-dimensional methylation matrix
betas <- as.matrix(assay(met_se, "Beta_value"))
```

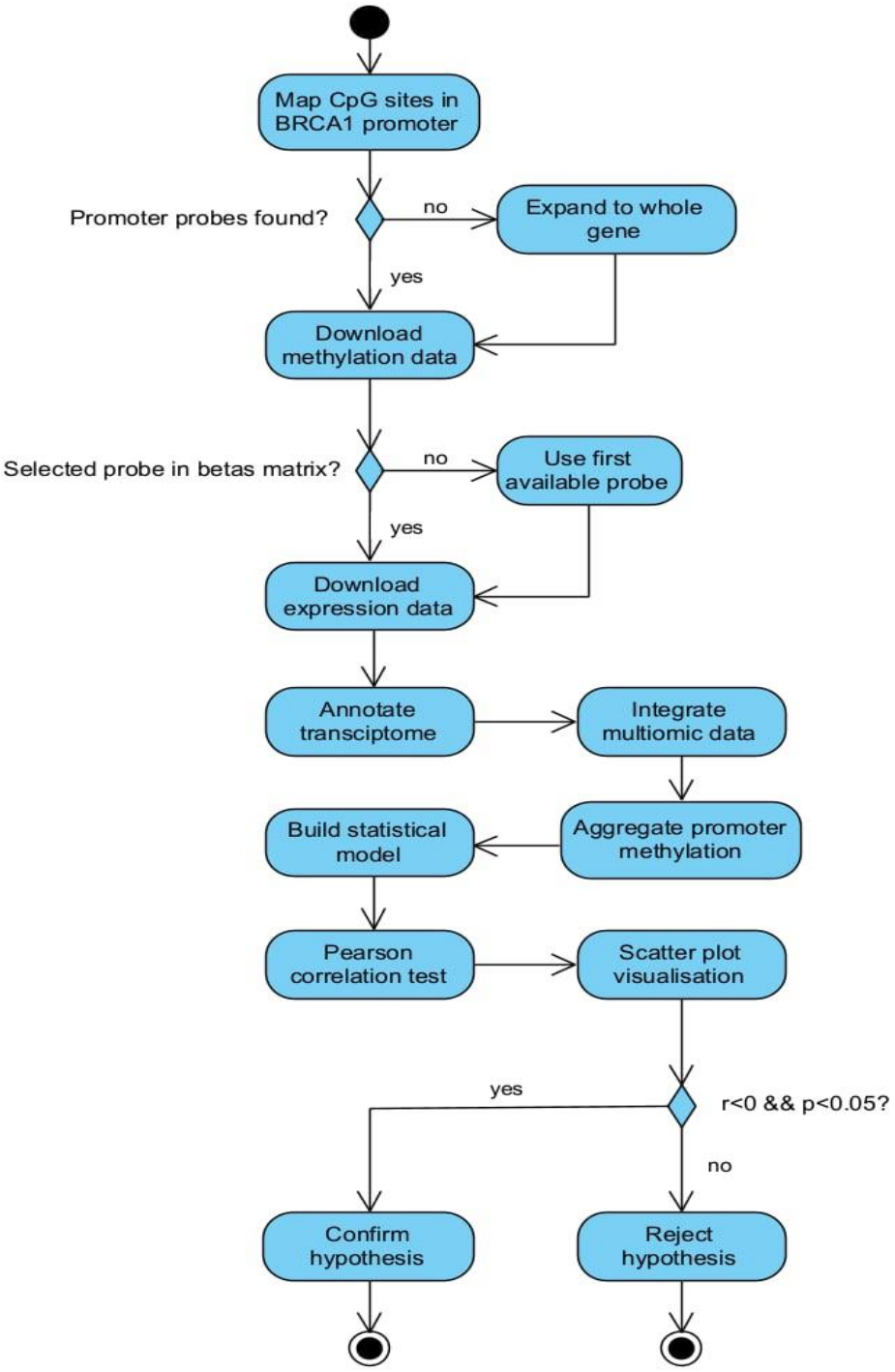

Figure 1 - UML Activity Diagram illustrating the automated data science pipeline for multi-omics data acquisition, transformation, and statistical integration.

Data Transformation and Matrix Integration

The implementation followed a structured computational pipeline to ensure data integrity across different omics layers. The process consisted of several algorithmic stages:

1. **Feature Mapping and Filtering:** Identification of all epigenetic probes mapped to the *BRCA1* gene body using genomic annotations (IlluminaHumanMethylation450kanno), followed by array subsetting to isolate strictly promoter-associated probes (TSS1500, TSS200, 5'UTR, 1stExon).

2. **Transcriptomic Normalization:** Raw RNA-Seq read counts were extracted and normalized to adjust for varying sequencing depths across the cohort. This was achieved through matrix operations computing the library sizes and transforming the data into a $\log_2(\text{CPM} + 1)$ feature space.

3. **Sample Intersection and Integration:** The core challenge of multi-omics integration is ensuring that different molecular layers accurately map to the same entity. We implemented a string manipulation algorithm to truncate sample barcodes and computed the set intersection to fuse the methylation and expression matrices (see the code below).

```
# Uniformizing sample identifiers to the patient level (16 characters)
colnames(expr_mat_symbol) <- substr(colnames(expr_mat_symbol), 1, 16)
colnames(betas) <- substr(colnames(betas), 1, 16)

# Identifying patients with complete multi-omics profiles (set intersection)
common_samples <- intersect(colnames(expr_mat_symbol), colnames(betas))

# Subsetting high-dimensional matrices for the integration phase
expr_final <- expr_mat_symbol[, common_samples, drop=FALSE]
met_final <- betas[, common_samples, drop=FALSE]
```

Statistical Computing

With the data structures harmonized, the final computational step involved the execution of the aggregation function $\mathcal{F}_{agg}$ over the promoter subsets. The column means of the subsetted methylation matrix were calculated (colMeans) to derive the global methylation index μ_i for each patient. Finally, a Pearson correlation algorithm was applied to the integrated data frame to test the statistical dependency between the independent (μ) and dependent (ϵ) variables.

Results and Analysis

The study analyzed paired data from $n = 498$ breast cancer patients. The application of the mathematical model yielded the following statistical parameters.

Statistical Correlation

The calculated Pearson correlation coefficient was $r = -0.32$. The associated p-value was $p = 2.36 \times 10^{-23}$, which is significantly lower than the standard alpha level of 0.05.

This result indicates a moderate negative linear relationship. As the mean methylation of the *BRCA1* promoter increases, the expression of the gene decreases. This provides strong statistical evidence to reject the null hypothesis and accept the alternative hypothesis: hypermethylation of the promoter region is a mechanism that silences *BRCA1* expression in breast cancer.

Visual Analysis

To visually inspect the relationship, a scatter plot was generated using the ggplot2 library (Wickham, 2016). The code used for visualization is presented below:

```
p <- ggplot(df, aes(x = Metylacja, y = Ekspresja)) +
  geom_point(alpha = 0.6, size = 2.5, color = "#C62828") +
  geom_smooth(method = "lm", se = TRUE, color = "#1565C0",
              fill = "#BBDEFB", alpha = 0.3, linewidth = 1) +
  labs(
    title = title_text,
    subtitle = sprintf("Pearson r = %.3f | p = %s | n = %d",
                       kt$estimate,
```

```
            format.pval(kt$p.value, digits = 3),
            n_points),
  x = "Mean Promoter Methylation (Beta-value)",
  y = paste0("mRNA Expression Level ", gene_symbol, " (log2 CPM)")
) +
theme_minimal(base_size = 14)
```

Figure 2 presents the resulting scatter plot. The blue regression line illustrates the negative trend. The spread of data points suggests that while methylation is a significant regulator, it is not the sole determinant of expression levels.

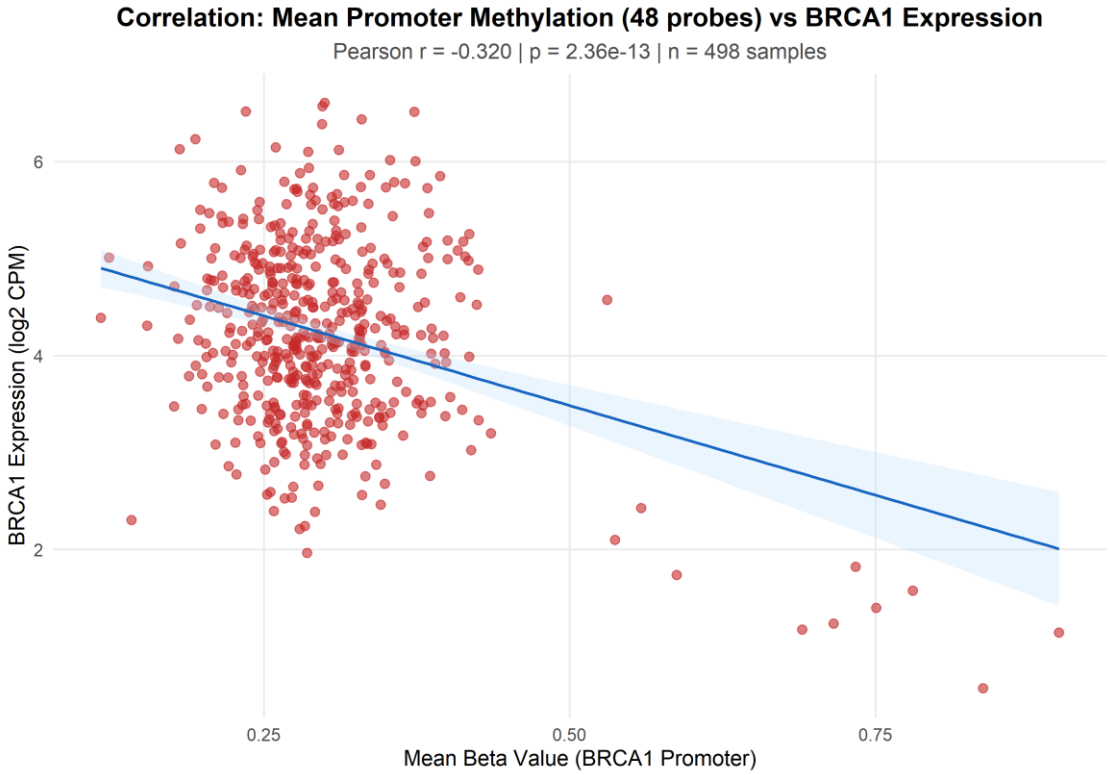

Figure 2 - Scatter plot presenting the correlation between mean promoter methylation (aggregation of 48 CpG probes) and BRCA1 expression (log2 CPM). Pearson r=-0.32, n=498.

Discussion and Conclusions

This study introduced a formal mathematical framework for analyzing multi-omics patient data. By vectorizing the methylation profiles and defining standardized normalization metrics for expression, we successfully quantified the regulatory impact of epigenetics on gene transcription.

The obtained correlation of $r = -0.32$ confirms the alternative hypothesis that DNA methylation acts as a silencer for *BRCA1* in breast cancer. However, the fact that the correlation is moderate and not close to -1.0 is biologically significant. It implies the existence of a highly multifaceted regulatory environment. As established in the literature, loss of heterozygosity (LOH), structural somatic mutations, histone modifications, and post-transcriptional miRNA regulation concurrently contribute to the final transcript abundance (Jones and Baylin, 2007; Hanahan and Weinberg, 2011). Methylation is a potent, but not exclusive, determinant of expression.

Furthermore, the choice of a simple arithmetic mean for the aggregation function $\mathcal{F}_{agg}$ inherently assumes that every CpG site within the promoter contributes equally to the silencing mechanism. This constitutes a limitation of the current model, as biological reality dictates that methylation at specific transcription factor binding sites may exert a disproportionately large effect compared to neighboring loci, and hypermethylation at CpG island shores often correlates differently than at core promoters (Irizarry et al., 2009). Future iterations of this

mathematical model should incorporate weighted aggregation functions, potentially optimized via machine learning algorithms, to better capture the spatial complexity of methylation patterns.

Ultimately, the application of this computational model extends well beyond *BRCA1*. The formalized vectors establish a standardized, reproducible data science pipeline that can be universally applied to evaluate the epigenetic regulation of any tumor suppressor gene across different cancer cohorts. Identifying such epigenetically silenced targets holds profound clinical promise. Because DNA methylation is an enzymatically reversible process, precise quantitative modeling provides crucial insights for the development of targeted epigenetic therapies and robust non-invasive biomarkers in personalized oncology.

Acknowledgment

This work was financed by Military University of Technology under research project UGB 531-000091-W500-22.

References

- Baylin, S. B., & Jones, P. A. (2011). A decade of exploring the cancer epigenome. *Nature Reviews Cancer*, 11(10), 726-734.
- Bird, A. (2002). DNA methylation patterns and epigenetic memory. *Genes & Development*, 16(1), 6-21.
- Catteau, A., & Morris, J. R. (2002). BRCA1 methylation: a significant role in tumour develop-ment? *Seminars in Cancer Biology*, 12(5), 359-371.
- Ciriello, G., et al. (2015). Comprehensive molecular portraits of invasive lobular breast cancer.
- *Cell*, 163(2), 506-519.
- Colaprico, A., et al. (2016). TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Research*, 44(8), e71.
- Dawson, M. A., & Kouzarides, T. (2012). Cancer epigenetics: from mechanism to therapy. *Cell*, 150(1), 12-27.
- Esteller, M. (2008). Epigenetics in cancer. *New England Journal of Medicine*, 358(11), 1148-1159.
- Feinberg, A. P., & Vogelstein, B. (1983). Hypomethylation distinguishes genes of some human cancers from their normal counterparts. *Nature*, 301(5895), 89-92.
- Gentleman, R. C., et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5(10), R80.
- Gevaert, O., et al. (2013). Glioblastoma multiforme: exploratory radiogenomics analysis by using quantitative image features. *Radiology*, 267(2), 568-577.
- Grossman, R. L., et al. (2016). Toward a shared vision for cancer genomic data. *New England Journal of Medicine*, 375(12), 1109-1112.
- Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell*, 144(5), 646-674.
- Herman, J. G., & Baylin, S. B. (2003). Gene silencing in cancer in association with promoter hypermethylation. *New England Journal of Medicine*, 349(21), 2042-2054.
- Hoadley, K. A., et al. (2014). Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell*, 158(4), 929-944.
- Huang, S., et al. (2017). Multi-omics integration to identify prognostic biomarkers. *Briefings in Bioinformatics*, 18(4), 549-560.
- Huber, W., et al. (2015). Orchestrating high-throughput genomic analysis with Bioconductor.
- *Nature Methods*, 12(2), 115-121.
- Irizarry, R. A., et al. (2009). The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nature Genetics*, 41(2), 178-186.
- Jones, P. A., & Baylin, S. B. (2007). The epigenomics of cancer. *Cell*, 128(4), 683-692.

- Koboldt, D. C., et al. (2012). Comprehensive molecular portraits of human breast tumours.
- *Nature*, 490(7418), 61-70.
- Kristensen, V. N., et al. (2014). Principles and methods of integrative genomic analyses in cancer. *Nature Reviews Cancer*, 14(5), 299-313.
- Kulis, M., & Esteller, M. (2010). DNA methylation and cancer. *Advances in Genetics*, 70, 27-56.
- Laird, P. W. (2003). The power and the promise of DNA methylation markers. *Nature Reviews Cancer*, 3(4), 253-266.
- Pery, M., Waszkowski, R. (2024). Mathematical modeling in color difference analysis in consecutive video frames for steganography. *Proceedings of the 44th IBIMA Conference*.
- R Core Team. (2023). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria.
- Rice, J. C., et al. (2000). Aberrant methylation of the BRCA1 CpG island promoter is associated with decreased BRCA1 mRNA in sporadic breast cancer cells. *Oncogene*, 19(25), 3396-3403.
- Robertson, K. D. (2005). DNA methylation and human disease. *Nature Reviews Genetics*, 6(8), 597-610.
- Sharma, S., Kelly, T. K., & Jones, P. A. (2010). Epigenetics in cancer. *Carcinogenesis*, 31(1), 27-36.
- Shen, H., & Laird, P. W. (2013). Interplay between the cancer genome and epigenome. *Cell*, 153(1), 38-55.
- Silva, T. C., et al. (2016). TCGA Workflow: Analyze cancer genomics and epigenomics data using Bioconductor packages. *F1000Research*, 5.