



## **EVALUATING UNSUPERVISED MULTI-OMICS FACTORISATION FRAMEWORKS: MOFA2 AND GFA APPLIED TO GLIOBLASTOMA**

Master Degree Project in Bioinformatics  
Second Cycle 30 credits  
Autumn term-2025

Student: Havva Unal

Supervisor: Zelmina Lubovac, University of Skövde  
External Supervisor: Marcela Davila, University of  
Gothenburg

Examiner: Angelica Lindlöf, University of Skövde



## Abstract

Unsupervised multi-omics integration methods are increasingly used to uncover latent structure in high-dimensional biological data, yet their behaviour and interpretability can vary substantially depending on the underlying inference strategy. In this study, two unsupervised multi-omics factorisation frameworks, Multi-Omics Factor Analysis (MOFA2) and Group Factor Analysis (GFA), were systematically benchmarked using paired RNA sequencing and promoter-level DNA methylation data from glioblastoma tumours as a representative use case. The methods were compared with respect to latent factor structure, partitioning of shared versus modality-specific variation, variance distribution across factors, and biological interpretability based on functional enrichment analyses. Although both frameworks captured complementary transcriptomic and epigenetic signals, they differed markedly in how variation was organised within the latent space. MOFA2 produced a compact and strongly regularised representation with a small number of dominant factors, whereas GFA retained a more distributed latent structure that preserved weaker sources of variation. Overall, this study highlights fundamental methodological trade-offs between interpretability and completeness in unsupervised multi-omics integration. The results emphasise that method selection should be guided by analytical objectives rather than biological context alone, and demonstrate the value of comparative benchmarking for informed application of unsupervised integration frameworks.

## Contents

Abstract .....	1
Abbreviations .....	5
Introduction .....	6
Epigenetics and DNA Methylation .....	6
Transcriptome and Transcriptomics.....	7
Glioblastoma (GBM) .....	7
Multi-Omics Factor Analysis (MOFA).....	7
Group Factor Analysis (GFA).....	9
Multi-Omics Integration .....	10
Aim.....	12
Objectives.....	12
Material and methods .....	13
Data Collection and Preprocessing .....	13
RNA-seq data processing.....	13
DNA methylation preprocessing.....	13
Harmonization across omics layers.....	14
Model Initialization and ARD Behavior .....	14
Exploratory Data Analysis .....	14
Unsupervised Multi-Omics Integration using MOFA2 .....	15
Unsupervised Multi-Omics Integration using GFA.....	15
Comparative Evaluation of MOFA2 and GFA .....	16
Functional Enrichment and Biological Interpretation.....	17
Model Stability and Reproducibility Analysis.....	17
Code Availability .....	18
R packages version info .....	18
Alternative methods and motivation .....	18
Implementation.....	19
Software Environment and Project Structure.....	19
Data Acquisition and Organization.....	20

RNA-seq Preprocessing Workflow.....	20
DNA Methylation Preprocessing Workflow.....	20
Harmonization and Construction of Paired Matrices.....	21
MOFA2 Model Implementation.....	21
GFA Model Implementation.....	21
Functional Enrichment Pipeline (GSEA and ORA).....	22
Overlap and Comparative Analysis Workflow.....	22
Reproducibility and Code Management.....	23
Results.....	23
Analysis of Unpaired TCGA-GBM Dataset.....	23
Dataset Structure and Separation of Unpaired and Paired Cohorts.....	23
Differential Gene Expression Analysis for Unpaired RNA-seq Dataset.....	23
Analysis of Paired Dataset.....	26
Transcriptomic Processing and Identification of Paired RNA Samples.....	26
DNA methylation preprocessing and promoter-level matrix.....	27
MOFA2 model fitting and data preparation.....	27
Latent Factor Structure Identified by MOFA2 in RNA and DNA Methylation.....	27
Functional enrichment of MOFA2 latent factors (GSEA and ORA on gene-level weights).....	29
GFA model fitting and data preparation.....	31
Latent Factor Structure Identified by GFA in RNA and DNA Methylation.....	31
Functional enrichment of GFA latent factors (GSEA and ORA on gene-level weights).....	33
Comparative model behavior of MOFA2 and GFA on high-dimensional RNA–methylation data .	34
GO: BP-Based Functional Enrichment Framework for Cross-Model Comparison.....	35
Conclusion.....	41
Scientific contribution and novelty.....	41
Ethical considerations and societal impact.....	42
Limitations and future directions.....	42
Discussion.....	43
Comparison of multi-omics factor structure captured by MOFA2 and GFA.....	43
Functional enrichment behavior and comparison with previous studies.....	43

Benchmarking implications and methodological trade-offs .....	44
Limitations and data-related considerations.....	44
References .....	46

## **Abbreviations**

**ARD** – Automatic Relevance Determination

**BP** – Biological Process

**CpG** – Cytosine–phosphate–guanine dinucleotide

**DE** – Differential Expression

**DM** – Differential Methylation

**DMG** – Differentially Methylated Gene

**DNA** – Deoxyribonucleic Acid

**FDR** – False Discovery Rate

**GBM** – Glioblastoma

**GFA** – Group Factor Analysis

**GO** – Gene Ontology

**GO:BP** – Gene Ontology: Biological Process

**GSEA** – Gene Set Enrichment Analysis

**MCMC**– Markov Chain Monte Carlo

**METH** – Methylation

**MOFA2** – Multi-Omics Factor Analysis (version 2)

**mRNA** – Messenger RNA

**ORA** – Over-Representation Analysis

**PCA** – Principal Component Analysis

**RNA** – Ribonucleic Acid

**RNA-seq** – RNA sequencing

**TCGA** – The Cancer Genome Atlas

**TF** – Transcription Factor

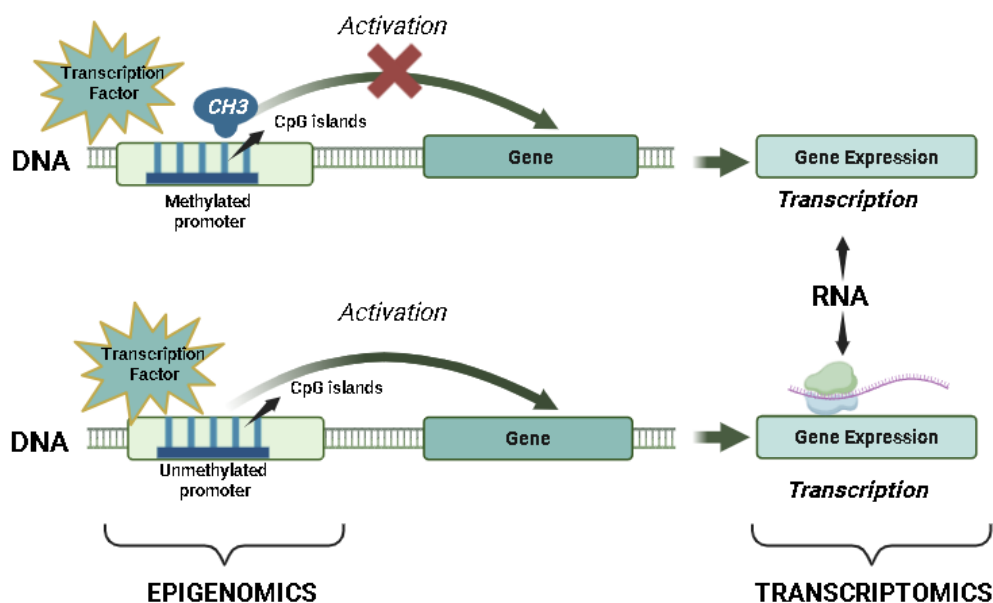
# Introduction

## Epigenetics and DNA Methylation

Epigenetics refers to heritable changes in gene expression that occur without alterations in the underlying DNA sequence. In contemporary usage, the term encompasses molecular mechanisms that regulate when and how genes are activated or silenced, including DNA methylation, histone modifications such as acetylation and phosphorylation, and other regulatory processes that influence chromatin structure and transcriptional activity. These regulatory layers play a fundamental role in normal cellular function and development; however, their disruption can contribute to severe diseases and abnormal cellular behaviour.

DNA methylation is one of the earliest discovered and most extensively studied epigenetic mechanisms, particularly in the context of human cancer. It involves the addition or removal of a methyl group ( $-CH_3$ ) at specific cytosine residues, most commonly within CpG dinucleotides. Aberrant DNA methylation patterns have been strongly associated with disease initiation and progression (Weinhold, 2006). Hypermethylation of promoter-associated CpG islands is typically linked to transcriptional repression, whereas hypomethylation is often associated with increased gene expression. Through this mechanism, DNA methylation directly influences transcription factor binding and transcriptional output, thereby forming a mechanistic link between epigenomic regulation and gene expression, as illustrated in Figure 1.

At the molecular level, DNA methylation occurs through the covalent addition of a methyl group to the 5th carbon of cytosine, resulting in the formation of 5-methylcytosine. This modification has important evolutionary and genomic consequences: spontaneous deamination of 5-methylcytosine converts it into thymine, a process that contributes to the depletion of CpG dinucleotides observed in vertebrate genomes (Bird, 1980; Wang, 1981). Despite its essential role in normal development and cellular identity, dysregulation of DNA methylation can profoundly alter transcriptional programmes and has been implicated in the initiation and progression of cancer.



**Figure 1.** Epigenomic control of gene expression. Hypermethylation of promoter-associated CpG islands prevents transcription factor binding and leads to transcriptional repression, whereas unmethylated promoters permit transcription factor binding and active gene expression. The schematic illustrates the regulatory relationship between promoter DNA methylation status and transcriptional output. Figure created with BioRender.com.

## **Transcriptome and Transcriptomics**

The transcriptome comprises the complete set of RNA molecules transcribed from the genome at a given time in a specific cell or tissue. It reflects which genes are transcriptionally active and the relative levels at which they are expressed, thereby providing a snapshot of cellular activity under defined biological conditions. Transcriptomics refers to the systematic study of the transcriptome using high-throughput technologies, most notably RNA sequencing (RNA-seq), to quantify gene expression patterns on a genome-wide scale. Transcriptomic analyses are widely applied to investigate cellular functions, disease mechanisms, and regulatory networks. Because RNA abundance can change rapidly in response to environmental cues, developmental signals, or pathological states, transcriptomic profiles provide a dynamic view of biological processes and cellular states (Wang et al., 2009).

Epigenomic mechanisms such as DNA methylation regulate gene activity without altering the underlying DNA sequence, whereas transcriptomic data capture the downstream transcriptional output of these regulatory processes. Integrating transcriptomic and epigenomic data therefore enables the investigation of how upstream regulatory modifications shape gene expression programmes, providing a more comprehensive understanding of cellular behaviour and disease-associated molecular dysregulation (Comendul et al., 2025).

## **Glioblastoma (GBM)**

Glioblastoma (GBM) is the most aggressive and the most common primary malignant brain tumour in adults. Originating from glial cells, it is characterised by rapid cellular proliferation, extensive infiltration into surrounding brain tissue, and a pronounced resistance to conventional therapies. Despite maximal treatment strategies, including surgical resection followed by radiotherapy and chemotherapy, the median survival time for patients remains approximately 15 months. This poor prognosis highlights the urgent need for a deeper molecular understanding of the disease. GBM exhibits extreme molecular heterogeneity, encompassing a wide range of genetic and epigenetic alterations that affect key biological processes such as growth factor signalling, cell-cycle regulation, metabolism, and immune responses (Stupp et al., 2005; Brennan et al., 2013). This heterogeneity is observed both between patients and within individual tumours, contributing to treatment resistance and disease recurrence. As a result, single-layer molecular analyses often fail to capture the full complexity of GBM biology.

At the molecular level, glioblastoma develops through the progressive accumulation of genetic mutations and epigenetic dysregulation that disrupt normal control of cell proliferation, apoptosis, and cellular differentiation. These alterations collectively promote uncontrolled tumour growth, enhanced invasiveness, and resistance to programmed cell death, which together underlie the aggressive clinical behaviour of GBM (Huse & Holland, 2010). Given that these processes operate across multiple regulatory layers, integrative multi-omics approaches are increasingly recognised as essential for elucidating the complex molecular architecture of GBM and for identifying biologically informed therapeutic targets.

## **Multi-Omics Factor Analysis (MOFA)**

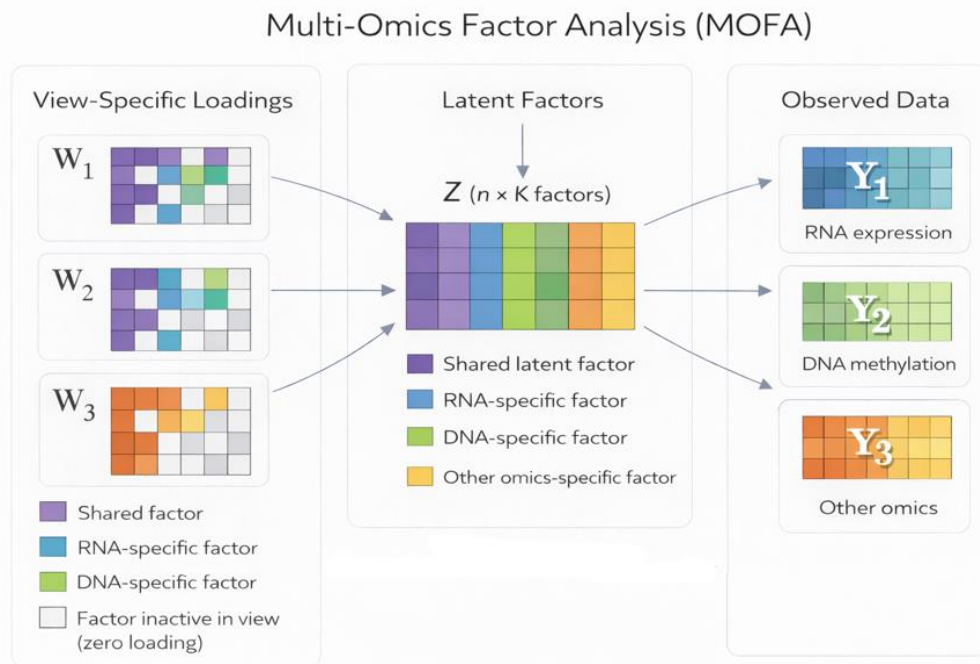
Multi-Omics Factor Analysis (MOFA2) is a probabilistic latent factor model developed for the unsupervised integration of multiple heterogeneous omics datasets (Argelaguet et al., 2018; Argelaguet et al., 2020). The method represents complex multi-omics data using a limited number of latent factors that capture the main sources of biological variation across samples. A key feature of MOFA2 is its ability to distinguish between variation that is shared across multiple omics layers and variation that is specific to a single modality, enabling the identification of both coordinated molecular patterns and modality-specific effects.

Before model fitting, each omics dataset is processed independently using modality-specific preprocessing and filtering steps and then formatted in a consistent structure, with molecular features in rows and samples in columns. MOFA2 is designed to handle differences in feature dimensionality across omics layers and can accommodate missing values, making it suitable for partially paired multi-omics datasets.

During model training, MOFA2 jointly learns two main components: a shared latent factor matrix ( $Z$ ) and a set of view-specific loading matrices  $W^{(v)}$ . The matrix  $Z$  describes how samples vary along each latent factor, while the loading matrices  $W^{(v)}$  describe how molecular features from each omics layer contribute to these factors. As illustrated in Figure 2, the generative structure of MOFA2 for each omics view  $v$  can be expressed as:

$$Y^{(v)} \approx ZW^{(v)} \quad (\text{Eq. 1})$$

where  $Y^{(v)}$  denotes the observed data matrix for omics view  $v$ . In this formulation, the same latent factors are shared across all samples through  $Z$ , while their relevance to each omics layer is determined by the corresponding loading patterns in  $W^{(v)}$ . Latent factors can therefore capture shared sources of variation when they are active in multiple omics layers, or view-specific sources of variation when their contributions are restricted to a single modality. This structure allows MOFA2 to provide an interpretable, low-dimensional representation of high-dimensional multi-omics data, facilitating the discovery of biologically meaningful associations across molecular layers while preserving modality-specific regulatory signals (Argelaguet et al., 2018; Argelaguet et al., 2020).



**Figure 2.** Multi-omics datasets, including RNA expression ( $Y_1$ ) and DNA methylation ( $Y_2$ ), are jointly modelled using a shared latent factor matrix  $Z$ . The latent factors capture both shared and view-specific sources of variation across samples, while view-specific loading matrices  $W^{(v)}$  encode the contributions of molecular features within each omics layer. Colour coding indicates whether latent factors are active in one or multiple omics views.

## Group Factor Analysis (GFA)

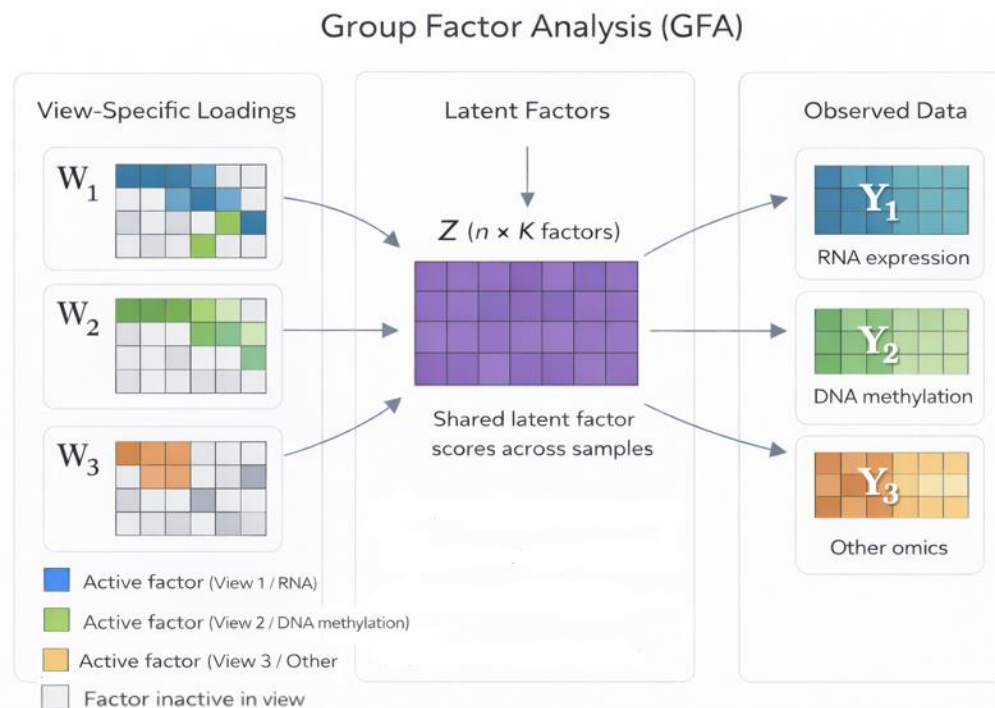
Group Factor Analysis (GFA) is a Bayesian latent variable model developed for the integration of multiple related data matrices measured on overlapping or partially missing samples (Virtanen et al., 2015). The method aims to identify latent sources of variation that capture both patterns shared across omics layers and variation that is specific to individual data modalities, while accounting for noise and heterogeneity between views. In GFA, multiple omics datasets are analysed jointly using a common latent factor representation. Each omics layer is linked to the same set of latent factors through view-specific loading patterns. For a given omics view  $v$ , the observed data matrix is approximated as:

$$Y^{(v)} \approx Z W^{(v)} \quad (\text{Eq. 2})$$

where  $Y^{(v)}$  denotes the observed data matrix for omics view  $v$ ,  $Z$  is the latent factor score matrix shared across all samples, and  $W^{(v)}$  is the corresponding view-specific loading matrix that describes how molecular features in that omics layer contribute to the latent factors.

As illustrated in Figure 3, GFA represents all latent factors uniformly at the level of  $Z$ , with view-specificity emerging exclusively from the sparsity structure of the loading matrices. Instead, this distinction emerges from the structure of the loading matrices  $W^{(v)}$ . Latent factors with non-zero loadings in multiple omics views are interpreted as shared factors, whereas factors with non-zero loadings restricted to a single view are considered view-specific. Factors with zero loadings in a given view are inactive in that modality.

This loading-driven structure allows GFA to flexibly distribute variation across a potentially large number of latent factors. By retaining both strong and weaker components, GFA is well suited for exploratory multi-omics analyses where subtle but structured biological signals across heterogeneous molecular layers are of interest (Virtanen et al., 2015).



**Figure 3.** Multi-omics datasets, including RNA expression ( $Y_1$ ) and DNA methylation ( $Y_2$ ), are jointly modelled using a shared latent factor matrix  $Z$ . In GFA, all latent factors are treated uniformly at the level of  $Z$ , while view-specificity is encoded through sparse loading matrices  $W^{(v)}$ , which determine whether a factor is active or inactive in a given omics layer. Colour coding indicates whether latent factors are active in one or multiple omics views.

## Multi-Omics Integration

Multi-omics integration aims to combine complementary molecular layers—such as gene expression, DNA methylation, proteomics, or other omics data—to provide a more comprehensive view of biological regulation than any single data type alone. Each omics layer captures a different aspect of cellular regulation. For example, transcriptomics reflects dynamic gene activity, while DNA methylation represents more stable, cell-state-dependent epigenetic control. Integrating these layers makes it possible to identify coordinated molecular programmes, distinguish shared and modality-specific sources of variation, and uncover regulatory relationships that are not apparent in single-omics analyses (Hasin et al., 2017; Hernández-Lemus & Ochoa, 2024).

In complex diseases such as glioblastoma, tumour development and progression are driven by disruptions across multiple regulatory layers, including transcriptional networks, epigenetic regulation, growth-factor signalling, and cell-cycle control. Analysing RNA expression or DNA methylation in isolation therefore provides only a partial view of tumour biology. Multi-omics approaches enable a more integrated reconstruction of disease-associated molecular pathways and can support improved biological interpretation, biomarker discovery, and patient stratification (Rosenberg et al., 2017; Sanchez-Vega et al., 2018).

To address this need, several computational frameworks have been developed for the joint analysis of multi-omics data. Among these, unsupervised latent factor models such as MOFA2 and GFA aim to identify shared and omics-specific latent components that explain structured variation across datasets. These latent factors can subsequently be interpreted using functional enrichment analyses, providing pathway-level insights into how different molecular layers interact in disease contexts (Argelaguet et al., 2018).

Although MOFA2 and GFA are both unsupervised Bayesian approaches for multi-omics integration, they differ in how they infer and organise latent structure. MOFA2 uses variational Bayesian inference combined with an Automatic Relevance Determination (ARD) prior, leading to a compact latent representation that emphasises dominant sources of variation. In contrast, GFA typically relies on Bayesian inference using Markov Chain Monte Carlo (MCMC) sampling and tends to retain a larger number of latent factors, distributing variation across a higher-dimensional latent space (Virtanen et al., 2012; Klami et al., 2015). These methodological differences motivate the systematic comparison of MOFA2 and GFA performed in this study, as summarised in Table 1.

**Table 1. Key methodological differences between MOFA2 and GFA.**

<b>Aspect</b>	<b>MOFA2</b>	<b>GFA</b>
<b>Model type</b>	Unsupervised multi-omics factor model	Unsupervised multi-omics factor model
<b>Inference approach</b>	Variational Bayesian inference with ARD	Bayesian inference via MCMC sampling
<b>Regularisation strategy</b>	Automatic Relevance Determination (ARD) prior	Conservative Bayesian shrinkage
<b>Treatment of latent dimensionality</b>	Non-informative factors are automatically suppressed	Most latent factors are retained
<b>Role of initial number of factors (K)</b>	Initial K can be set high; effective number of factors is learned during training	Initial K largely determines the size of the latent space
<b>Typical latent space</b>	Compact and low-dimensional	High-dimensional and distributed
<b>Factor strength</b>	Fewer but stronger latent factors	Many weaker but structured latent factors
<b>Interpretability</b>	High interpretability due to concentrated biological signal	Lower per-factor interpretability, but richer overall structure
<b>Sensitivity to weak signals</b>	Lower sensitivity, prioritises dominant variation	Higher sensitivity, preserves subtle variation
<b>Suitability for small sample sizes</b>	More robust due to aggressive regularisation	More sensitive to noise when sample size is limited

## Aim

The aim of this project is to benchmark two unsupervised multi-omics integration methods, MOFA2 and GFA, using matched RNA-seq and DNA methylation data from glioblastoma. Specifically, the study compares how effectively each method captures shared and omics-specific sources of variation, the stability of the inferred latent factors, and the biological interpretability of the resulting signals.

RNA-seq and DNA methylation were selected because they represent two mechanistically linked layers of gene regulation. RNA-seq reflects downstream transcriptional output, whereas promoter DNA methylation represents upstream epigenetic regulation that can activate, repress, or silence gene expression. In glioblastoma, changes in promoter methylation are frequently associated with corresponding changes in RNA expression, making these two modalities biologically coherent for joint analysis.

By integrating RNA expression and DNA methylation, the models can be evaluated on their ability to distinguish transcriptional programmes, epigenetic programmes, and coordinated regulation across both layers. All preprocessing, modelling, and evaluation steps were standardised across modalities to ensure a fair and reproducible comparison between MOFA2 and GFA.

## Objectives

The specific objectives of this project are to:

**Preprocess and harmonise** paired RNA-seq and promoter-level DNA methylation data from TCGA-GBM to construct a biologically and technically comparable multi-omics dataset suitable for integrative analysis.

**Apply and configure** MOFA2 and GFA under consistent modelling conditions, including matched latent dimensionality, to enable a fair comparison between the two frameworks.

**Characterise and compare** the latent factor structures inferred by MOFA2 and GFA, with particular focus on the partitioning of shared versus omics-specific sources of variation.

**Quantify and evaluate** variance explained by individual latent factors across transcriptomic and epigenomic layers to assess differences in model regularisation and latent space organisation.

**Assess biological interpretability** of model-derived latent factors using functional enrichment analyses, applying Gene Set Enrichment Analysis (GSEA) for RNA-derived factors and over-representation analysis (ORA) for methylation-derived factors.

**Benchmark methodological trade-offs** between MOFA2 and GFA in terms of interpretability, sensitivity to weak signals, and robustness, and relate these findings to existing multi-omics integration literature.

## Material and methods

### Data Collection and Preprocessing

RNA-seq and DNA methylation data for glioblastoma (GBM) were obtained from The Cancer Genome Atlas (TCGA) via the Genomic Data Commons (GDC) Data Portal (<https://portal.gdc.cancer.gov/>). As an initial step, RNA-seq and DNA methylation datasets were manually downloaded from the GDC web interface together with the corresponding sample sheets and barcode metadata files. These metadata files were used to identify sample types and to extract patient-level identifiers from TCGA barcodes. The manually downloaded RNA-seq dataset comprised 5 normal samples and 17 tumour samples, while the manually downloaded DNA methylation dataset comprised 140 tumour samples and 2 normal samples. At this stage, RNA-seq and DNA methylation data were handled independently, and no patient-level matching was performed across omics layers.

Because the manually downloaded RNA-seq and DNA methylation datasets could not be reliably matched at the patient level, they constituted unpaired datasets and were therefore not suitable for joint multi-omics integration. Consequently, these data were excluded from integrative modelling.

To construct a paired multi-omics dataset suitable for integration, RNA-seq data were subsequently re-downloaded programmatically within R (version 4.4.3) using the official GDC data access interfaces, together with the corresponding metadata and sample annotation files. RNA-seq data consisted of gene-level raw read counts generated using the STAR-Counts workflow (HTSeq-Counts). DNA methylation data consisted of Level-3  $\beta$ -value files generated using the Illumina HumanMethylation450K platform.

For integrative analysis, RNA-seq data were restricted to tumour samples only (Primary Tumour and Recurrent Tumour) to enable patient-level matching with the tumour DNA methylation dataset. All TCGA sample identifiers were standardised by harmonising barcodes to a consistent format, and patient-level identifiers were extracted from the TCGA barcodes. The paired multi-omics cohort was defined as the intersection of patient identifiers between the tumour-only RNA-seq dataset and the tumour DNA methylation dataset. This matching procedure yielded a final paired cohort of 84 GBM tumour samples, which constituted the input dataset for all downstream integrative analyses using MOFA2 and GFA.

### RNA-seq data processing

RNA-seq count files were merged into a gene-by-sample count matrix. Lowly expressed genes were filtered by removing features with fewer than 10 counts in at least three samples, in order to reduce noise from low-abundance transcripts. Normalisation was performed using the DESeq2 framework, including estimation of size factors to account for differences in sequencing depth across samples. For downstream multivariate analyses, variance-stabilising transformation (VST) was applied to obtain approximately homoscedastic expression values suitable for dimensionality reduction and latent factor modelling (Love et al., 2014).

Gene identifiers were mapped from ENSEMBL IDs to HGNC gene symbols using the org.Hs.eg.db annotation package. Genes that could not be mapped were retained with their original ENSEMBL identifiers to avoid unnecessary data loss. Although DESeq2 reports adjusted p-values using a default significance threshold of 0.1 for descriptive summaries, all downstream filtering, statistical analyses, and biological interpretations in this study were performed using a false discovery rate (FDR) threshold of 0.05.

### DNA methylation preprocessing

DNA methylation data were obtained as Level-3  $\beta$ -value files generated using the Illumina HumanMethylation450K platform. Individual methylation files were matched to sample metadata and

merged into a unified CpG-by-sample  $\beta$ -value matrix. CpG probes with insufficient data completeness across samples were removed, and remaining missing values were imputed on a per-sample basis using the column median. To improve statistical properties for downstream analyses,  $\beta$ -values were transformed to M-values using a logit transformation, resulting in a CpG-by-sample M-value matrix.

CpG-level M-values were subsequently summarised at the gene level using Illumina HumanMethylation450K annotation (hg19). CpG probes annotated to promoter regions (TSS200 and TSS1500) were mapped to genes, and gene-level promoter methylation values were computed by averaging M-values across promoter-associated CpGs for each gene. Promoter-level methylation data were also analysed using the limma framework in a separate exploratory analysis to assess tumour–normal differential methylation. Samples were grouped according to phenotype annotations, and linear models were fitted for each promoter-level feature followed by empirical Bayes moderation. Differential methylation was assessed using Benjamini–Hochberg adjusted p-values, and differentially methylated genes (DMGs) were defined using a false discovery rate (FDR) threshold of 0.05 and an absolute  $\log_2$  fold-change threshold of 0.5.

For integrative multi-omics analyses, tumour samples were selected based on phenotype annotations. The resulting tumour-only promoter-level methylation matrix, together with the corresponding metadata including patient identifiers derived from TCGA barcodes, was used for integration with tumour-only RNA-seq data using MOFA2 and GFA.

## **Harmonization across omics layers**

Only tumour samples present in both the RNA-seq and DNA methylation datasets were retained for integrative analysis. Sample matching was performed at the patient level, and only samples with measurements available in both modalities were included. RNA-seq expression and promoter-level DNA methylation matrices were treated as separate omics views, with each view retaining its own feature set. During model fitting, each omics layer was internally centred and scaled using the built-in scaling procedures of the respective integration frameworks to ensure comparable numerical ranges across views, in accordance with recommended practice for multi-omics factor analysis (Argelaguet et al., 2018). The resulting tumour-only RNA expression and promoter-level DNA methylation matrices constituted the input for the unsupervised multi-omics integration methods MOFA2 and GFA.

## **Model Initialization and ARD Behavior**

The number of latent factors was selected in accordance with the inference strategy and regularisation behaviour of each integration method. MOFA2 was fitted using variational Bayesian inference with an Automatic Relevance Determination (ARD) prior, which progressively downweights latent factors that do not explain consistent variation across samples. To allow the ARD mechanism to operate effectively, MOFA2 was initialised with 15 latent factors in an overcomplete setting. During model training, factors explaining negligible variance were downweighted by the ARD prior, while factors capturing meaningful signal remained active and were retained for downstream analyses (Argelaguet et al., 2018).

GFA was initialised with the same number of latent factors (15) to ensure a controlled and directly comparable latent dimensionality across models. GFA was fitted using Bayesian inference without an ARD-based pruning mechanism; consequently, all latent factors were retained during model fitting. Interpretation of GFA components was therefore based on their variance contributions across omics views and their associated functional enrichment patterns evaluated in downstream analyses.

## **Exploratory Data Analysis**

Exploratory data analysis (EDA) was performed to assess data quality and to examine major sources of variation prior to multi-omics integration. Principal component analysis (PCA) was applied separately to the transcriptomic and epigenomic datasets using tumour samples only. For RNA-seq

data, PCA was conducted on  $\log_2$ -transformed, size-factor normalised gene expression values obtained from DESeq2. For DNA methylation data, PCA was performed on promoter-level methylation profiles represented as M-values. These analyses were used to inspect global variation patterns and to identify potential outliers or technical artefacts following preprocessing. To provide a descriptive assessment of cross-omics relationships, promoter-level methylation values and gene expression levels for matched genes were examined using pairwise Pearson correlation analysis. In addition, hierarchical clustering based on Euclidean distance and Ward's linkage was applied separately to each omics layer to assess sample-level structure.

Additional quality-control summaries, including feature-wise variance distributions and sample-level metrics, were inspected to ensure comparability between datasets. Together, these exploratory analyses were used to confirm that the paired TCGA-GBM dataset was suitable for downstream unsupervised multi-omics integration using MOFA2 and GFA. All analyses were performed in R version 4.4.3.

## **Unsupervised Multi-Omics Integration using MOFA2**

MOFA2 is a probabilistic latent factor model designed to decompose multi-omics datasets into shared and view-specific components, thereby capturing major sources of variation across multiple data modalities (Argelaguet et al., 2018; Argelaguet et al., 2020). Each latent factor represents an unobserved variable that contributes to the covariance structure within and between omics layers.

In this study, MOFA2 was applied to a paired GBM RNA-seq and promoter-level DNA methylation dataset to model shared and modality-specific sources of variation. Input matrices were formatted consistently following the preprocessing and harmonisation steps described previously, and the two omics layers were treated as separate views, each retaining its own feature space, according to the generative formulation shown in Equation (1). MOFA2 was selected because it accommodates differences in feature dimensionality across views and can handle missing values within a variational Bayesian framework (Argelaguet et al., 2020).

The model was initialised in an overcomplete setting with 15 latent factors to allow the ARD prior to downweight latent components that did not explain consistent variation across samples during training (Argelaguet et al., 2018). This regularisation strategy enables MOFA2 to learn a compact latent representation without requiring manual pruning of factors.

Following model fitting, variance explained by each latent factor was quantified separately for the RNA-seq and DNA methylation views. These variance estimates were subsequently used to guide downstream analyses. For functional characterisation, factor-specific feature weights were extracted separately for RNA expression and promoter-level DNA methylation. RNA-derived feature weights were analysed using Gene Ontology Biological Process enrichment based on gene set enrichment analysis, while promoter-level methylation features were analysed using over-representation analysis in a view-specific manner.

Model outputs, including variance explained and latent factor structure, were summarised using visualisations such as bar plots and heatmaps to support downstream interpretation and comparison with results obtained from GFA. All analyses were performed in R version 4.4.3 using the MOFA2 package.

## **Unsupervised Multi-Omics Integration using GFA**

GFA is a Bayesian latent factor model for the joint analysis of multiple data matrices measured on the same set of samples. It decomposes multi-omics data into latent factors that may be shared across data modalities or specific to individual omics layers through sparsity-inducing priors on the view-specific loading patterns (Klami et al., 2014; Virtanen et al., 2015). This framework enables the modelling of both shared and modality-specific sources of variation in paired multi-omics datasets.

In this study, GFA was applied to the paired TCGA-GBM tumour cohort ( $n = 84$ ) using matched RNA-seq expression and promoter-level DNA methylation data. The same input datasets used for MOFA2 were employed to ensure methodological comparability between the two integration approaches. Prior to model fitting, the two omics layers were formatted consistently and standardised at the feature level, following the latent factor formulation described in Equation (2). Sample matching was performed at the patient level using harmonised identifiers, and only samples present in both omics layers were included.

The GFA model was initialised with 15 latent factors in an overcomplete setting, reflecting common practice for high-dimensional multi-omics data where automatic pruning of latent components is not applied. Model fitting was performed using Bayesian inference based on Markov Chain Monte Carlo sampling, as implemented in the GFA framework.

Following model fitting, variance explained by each latent factor was quantified separately for the RNA-seq and DNA methylation views. These variance estimates were used to guide downstream analyses. Latent factors were categorised as shared or view-specific based on whether non-zero loadings were observed in one or both omics layers. For functional characterisation, factor-specific feature weights were extracted separately for RNA expression and promoter-level DNA methylation and analysed using Gene Ontology Biological Process enrichment in a view-specific manner.

Model outputs, including variance explained and latent factor structure, were summarised using visualisations such as bar plots and heatmaps to support downstream interpretation and comparison with MOFA2 results. All analyses were performed in R version 4.4.3 using the GFA package.

## **Comparative Evaluation of MOFA2 and GFA**

To enable a fair and reproducible comparison between the two unsupervised multi-omics integration frameworks, MOFA2 and GFA, a systematic evaluation was performed using both statistical and biological criteria. The comparison focused on how each model partitioned variation into shared versus view-specific components, the compactness of the latent representations, and the biological interpretability of the inferred factors.

For both models, variance explained was quantified at the factor level and separately for the RNA-seq and DNA methylation views. In MOFA2, variance explained values were obtained directly from the model outputs. For GFA, an equivalent reconstruction-based approach was applied, in which factor scores and loading matrices were used to reconstruct factor-specific signal contributions within each omics layer. This ensured that variance explained was computed in a manner directly comparable between the two frameworks. Based on these estimates, factors were classified as RNA-specific, methylation-specific, or shared according to their relative contributions across views.

Model complexity and interpretability were assessed by comparing the number of latent factors, the distribution of variance explained across factors, and the sparsity structure of the corresponding loading matrices. To evaluate cross-model consistency, pairwise Pearson correlations were computed between MOFA2 and GFA factor scores, with particular emphasis on factors explaining the largest proportions of variance. High correlations were interpreted as evidence of convergence on similar underlying sources of biological variation, whereas low correlations indicated model-specific latent structure.

Biological concordance between MOFA2 and GFA was further assessed through functional enrichment analyses based exclusively on Gene Ontology Biological Process (GO:BP) terms (Ashburner et al., 2000; Gene Ontology Consortium, 2021). RNA-derived factors were analysed using Gene Set Enrichment Analysis (GSEA), while promoter-level DNA methylation factors were analysed using over-representation analysis (ORA). Enrichment analyses were performed in a view-aware manner, and overlap and divergence in enriched GO:BP terms were examined using factor-level

comparisons as well as model-level aggregation. Set-intersection analyses were used to quantify shared and model-specific biological processes captured by each framework.

Comparative results—including variance explained profiles, factor classifications, cross-model correlations, and overlap of enriched GO:BP terms—were summarised using barplots, heatmaps, correlation plots, and UpSet diagrams. Together, these analyses provided a structured and transparent framework for benchmarking MOFA2 and GFA on the paired TCGA-GBM multi-omics dataset.

## **Functional Enrichment and Biological Interpretation**

Functional enrichment analyses were performed to support the biological interpretation of molecular signals captured by MOFA2 and GFA across transcriptomic and epigenetic layers. Analysis strategies were selected according to the statistical properties of each omics layer and the structure of the respective integration frameworks. For RNA expression data, Gene Set Enrichment Analysis (GSEA) was applied using ranked gene weights derived from model outputs, thereby leveraging the full continuous distribution of gene-level contributions without applying arbitrary thresholds (Subramanian et al., 2005). For promoter-level DNA methylation data, over-representation analysis (ORA) was used based on sets of promoter-associated genes showing the strongest model-derived contributions (Boyle et al., 2004; Khatri et al., 2012). All enrichment analyses were performed in R using the clusterProfiler package, which provides a unified framework for both GSEA and ORA with Gene Ontology annotations (Yu et al., 2012).

Enrichment analyses were initially performed using model-derived latent representations; however, downstream biological interpretation focused primarily on aggregated, model-level enrichment patterns rather than individual latent factors. This approach was adopted to enable a robust and interpretable comparison between MOFA2 and GFA, given differences in latent dimensionality, sparsity structure, and factor relevance across the two frameworks.

To obtain a single representative functional profile per model and omics layer, significantly enriched Gene Ontology Biological Process (GO:BP) terms were aggregated across all latent factors within each framework (Ashburner et al., 2000; Gene Ontology Consortium, 2021). Only terms passing false discovery rate correction (adjusted p-value  $\leq 0.05$ ) were retained. When the same GO term appeared multiple times across different latent factors, the most significant instance was selected based on the lowest adjusted p-value; in cases of ties, the term with the highest absolute effect size was retained.

Aggregated enrichment results were subsequently used for global functional comparison between MOFA2 and GFA across RNA and DNA methylation layers. For structured interpretation, enriched GO:BP terms were grouped into broader functional domains, including RNA processing and gene expression regulation, chromatin organisation and epigenetic regulation, immune-related processes, developmental and signalling pathways, and cell-cycle control. This domain-level organisation facilitated coherent biological interpretation and cross-model comparison.

## **Model Stability and Reproducibility Analysis**

Model stability was assessed to evaluate the robustness of the latent structures inferred by MOFA2 and GFA with respect to random initialisation. For both frameworks, stability analysis was performed by refitting models multiple times using different random seeds and comparing variance explained profiles and dominant latent factors across runs.

For MOFA2, stability assessment focused on the reproducibility of latent factors retained following application of the ARD prior. Across repeated model fits, variance explained by retained factors was compared for the RNA-seq and DNA methylation views to assess consistency across runs. For GFA, which does not apply ARD-based factor pruning and retains all latent components during model fitting, stability assessment focused on factors explaining the largest proportions of variance. Variance

contributions and loading structures of high-variance factors were compared across repeated runs, while low-variance factors were not considered further in stability evaluation.

Overall, model stability was assessed based on the consistency of variance explained and the reproducibility of dominant latent factors across repeated fits. Formal resampling or cross-validation procedures were not applied, given the high dimensionality of the data and the limited sample size of the paired TCGA-GBM cohort.

## Code Availability

All scripts used for data preprocessing, multi-omics integration, functional enrichment analyses, and cross-model comparisons were implemented in R and are publicly available in the following GitHub repository:

### GitHub repository:

<https://github.com/Hunal33/gbm-multiomics-thesis>

## R packages version info

All analyses were performed in R using packages from CRAN and Bioconductor. The following R packages and versions were used throughout the analysis: data.table (v1.17.8), dplyr (v1.1.4), tidyr (v1.3.1), tibble (v3.2.1), readr (v2.1.5), stringr (v1.5.2), purrr (v1.1.0), ggplot2 (v4.0.0), ggrepel (v0.9.6), pheatmap (v1.0.13), DESeq2 (v1.46.0), limma (v3.62.2), edgeR (v4.4.2), MOFA2 (v1.16.0), GFA (v1.0.5), clusterProfiler (v4.14.6), enrichplot (v1.26.6), AnnotationDbi (v1.68.0), org.Hs.eg.db (v3.20.0), matrixStats (v1.5.0), and BiocManager (v1.30.26).

## Alternative methods and motivation

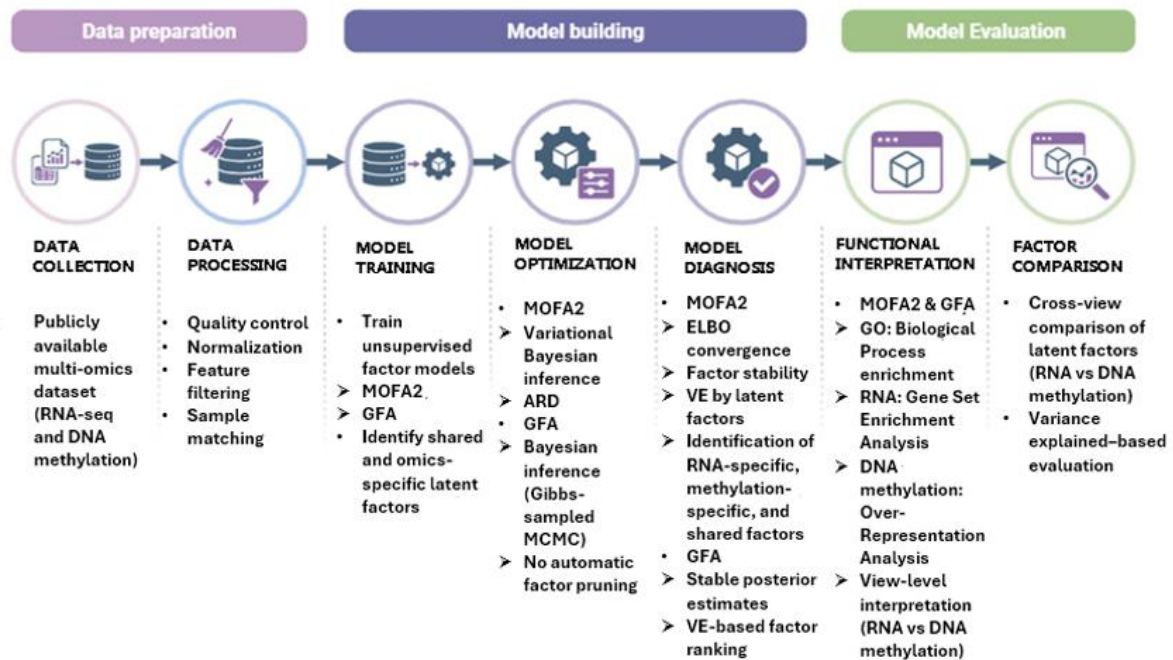
Several alternative approaches for multi-omics data integration have been proposed in the literature, including iClusterPlus (Mo et al., 2013), Similarity Network Fusion (SNF; Wang et al., 2014), and Joint and Individual Variation Explained (JIVE; Lock et al., 2013). These methods were considered during the initial planning phase of the project but were not selected for the final benchmarking workflow due to methodological and practical considerations.

iClusterPlus supports both supervised and unsupervised integration through penalised regression but typically requires predefined class labels for optimal performance and has limited scalability for high-dimensional omics datasets with large numbers of features. Similarity Network Fusion integrates multiple omics layers by constructing and combining sample similarity networks. While effective for sample clustering, SNF does not provide latent factor representations or feature-level loadings, limiting its suitability for factor-based biological interpretation (Vahabi & Michailidis, 2022). JIVE decomposes datasets into joint and individual components but relies on linearity assumptions and is less flexible in handling partially paired data, which is a common characteristic of large cancer genomics cohorts (Lock et al., 2013).

In contrast, MOFA2 and GFA were selected because they provide probabilistic, factor-based frameworks for modelling both shared and modality-specific sources of variation in an unsupervised manner (Argelaguet et al., 2018; Klami et al., 2015). Both methods are designed for high-dimensional settings, can accommodate heterogeneous omics data types, and produce interpretable latent factors that can be linked to biological pathways through functional enrichment analysis.

The overall analytical workflow adopted in this study is summarised in Figure 4, outlining the main steps from data collection and preprocessing to model fitting, diagnostic assessment, and factor-level biological interpretation. Focusing on MOFA2 and GFA enables a systematic comparison of two complementary Bayesian integration frameworks under consistent data and preprocessing conditions.

# Workflow of Multi-omics integration



**Figure 4.** Schematic overview of the multi-omics integration workflow applied to the paired TCGA-GBM dataset. The figure illustrates the main analytical steps, including modality-specific data preprocessing, unsupervised model training using MOFA2 and GFA, model diagnostics and evaluation, and factor-level functional interpretation through enrichment analyses. The workflow highlights the sequential relationship between data preparation, model fitting, and biological interpretation.

## Implementation

### Software Environment and Project Structure

All computational analyses were performed in R version 4.4.3 using Bioconductor version 3.20. This environment provided access to current genomic annotation resources and established implementations of high-dimensional statistical methods. Key packages included DESeq2 for RNA-seq normalisation and expression-based preprocessing, MOFA2 and GFA implementations for unsupervised multi-omics integration, and clusterProfiler for functional enrichment analyses based on GSEA and ORA. Gene identifier harmonisation and annotation were conducted using org.Hs.eg.db in combination with AnnotationDbi.

The analysis workflow was organised using a structured and modular project directory to support reproducibility and transparency. A RAW directory contained all files obtained from the GDC, including RNA-seq count files, DNA methylation  $\beta$ -value matrices, sample sheets, and metadata. A PREPROCESSING directory contained scripts for RNA-seq filtering and normalisation, gene identifier annotation, and promoter-level DNA methylation summarisation.

Separate MOFA and GFA directories were used to store model-specific scripts, trained model objects, factor loading matrices, sample-level factor scores, variance-explained summaries, and diagnostic outputs. Functional enrichment results were organised in an ENRICHMENT directory, which

contained GSEA and ORA outputs for each latent factor and omics layer. All visual outputs, including PCA plots, volcano plots, heatmaps, bar plots, and UpSet diagrams, were saved in a dedicated FIGURES directory.

This modular organisation allowed each stage of the analysis—from raw data import and preprocessing to integrative modelling and functional interpretation—to be executed and inspected independently. At the same time, it ensured traceability of inputs, parameters, and outputs across the workflow, supporting reproducibility and systematic comparison of results throughout the study.

## Data Acquisition and Organization

RNA-seq gene-level count data generated using the STAR–Counts workflow and DNA methylation Level-3  $\beta$ -value files for the TCGA GBM cohort were obtained from GDC Data Portal using the official GDC data access tools. Corresponding sample sheets and metadata files were retrieved alongside the molecular data to ensure consistent annotation, traceability, and downstream harmonisation.

Sample identifiers were standardised to a consistent TCGA barcode format to enable reliable cross-omics matching. Patient-level identifiers were extracted from TCGA barcodes and used for direct matching between RNA-seq and DNA methylation datasets. Only tumour samples were considered for downstream analyses, and patient-level matching across omics layers was applied to define a paired cohort for integrative modelling. Samples lacking complete cross-omics coverage were excluded to ensure valid joint inference in subsequent multi-omics analyses.

## RNA-seq Preprocessing Workflow

RNA-seq preprocessing was performed by merging raw STAR–Counts output files into a single gene-by-sample count matrix. To reduce noise from extremely lowly expressed features, genes with fewer than ten counts in at least three samples were removed prior to downstream analyses.

Gene identifiers were standardised by removing ENSEMBL version suffixes, and ENSEMBL gene IDs were mapped to HGNC gene symbols using the org.Hs.eg.db annotation database. Genes that could not be mapped to HGNC symbols were retained with their original ENSEMBL identifiers to avoid unnecessary data loss.

Following filtering and annotation, size-factor normalisation was performed using the DESeq2 package. Variance-stabilising transformation (VST) was applied for exploratory data analysis, including principal component analysis, to obtain approximately homoscedastic expression values suitable for dimensionality reduction. For downstream multi-omics integration, RNA-seq inputs were prepared according to the requirements of each integration framework, as described in the harmonisation step.

## DNA Methylation Preprocessing Workflow

DNA methylation preprocessing was performed by merging Illumina HumanMethylation450K Level-3  $\beta$ -value files into a unified CpG-by-sample matrix. Probe-level quality control was applied to remove CpG probes with insufficient data completeness across samples. Remaining missing values were imputed on a per-sample basis using the column median, a commonly used pragmatic approach for high-dimensional methylation data following probe-level filtering (Chen et al., 2013).

Following filtering and imputation,  $\beta$ -values were transformed to M-values using a logit transformation to improve statistical properties for downstream multivariate analyses. CpG probe annotation was performed using Illumina HumanMethylation450K annotation resources provided through Bioconductor.

To enable integration with gene-level RNA-seq data, CpG-level M-values were summarised at the promoter level. CpG probes annotated to promoter regions (TSS200 and TSS1500) were mapped to genes, and promoter-level methylation values were computed by averaging M-values across promoter-associated CpGs for each gene. This resulted in a gene-by-sample promoter methylation matrix, facilitating alignment with transcriptomic data during downstream integrative analyses.

## **Harmonization and Construction of Paired Matrices**

To enable joint modelling across omics layers, RNA-seq and promoter-level DNA methylation matrices were restricted to tumour samples present in both datasets. Sample matching was performed at the patient level using harmonised TCGA identifiers, and only samples with corresponding measurements in both modalities were retained. This ensured that all integrative analyses were conducted on paired observations derived from the same individuals.

Following sample matching, the two omics layers were aligned at the gene level. Promoter-level methylation features were retained only for genes represented in the RNA-seq expression matrix, resulting in gene-matched input matrices across modalities. This step established direct gene-level correspondence between the transcriptomic and epigenomic views while preserving modality-specific feature definitions.

Prior to model fitting, input matrices were prepared according to the requirements of each integration framework. For GFA, input matrices were explicitly centred and scaled on a per-feature basis to ensure comparable numerical ranges across omics layers. For MOFA2, centring and scaling were handled internally during model training, and no additional external scaling was applied.

These harmonisation procedures yielded a paired multi-omics dataset suitable for downstream MOFA2 and GFA analyses, ensuring consistent sample alignment and numerical comparability across omics layers.

## **MOFA2 Model Implementation**

MOFA2 was applied to the paired RNA-seq and promoter-level DNA methylation matrices to model shared and modality-specific sources of variation across glioblastoma samples. Input data were provided in a samples-by-features format using the harmonised paired dataset described above.

The model was initialised with 15 latent factors to provide sufficient capacity for capturing major sources of variation across omics layers. Model training was performed using variational Bayesian inference with an ARD prior, which downweights latent factors that do not explain consistent variation across samples. Model optimisation proceeded until convergence, as assessed by stabilisation of the variational objective.

Following model fitting, sample-level latent factor scores and view-specific feature loading matrices were extracted for each omics layer. Variance explained was computed separately for the RNA-seq and DNA methylation views using model-derived estimates. Latent factors with negligible contributions to variance across both modalities were not prioritised for downstream biological interpretation.

## **GFA Model Implementation**

GFA was applied to the paired RNA-seq and promoter-level DNA methylation dataset using the same harmonised input matrices as those used for MOFA2. For consistency across integration frameworks, both models were initialised with the same number of latent factors (15), ensuring a directly comparable latent dimensionality. The paired matrices were provided as separate views corresponding to transcriptomic and epigenomic measurements.

GFA was fitted using Bayesian inference based on Markov Chain Monte Carlo (MCMC) sampling, estimating posterior distributions for latent factor scores and view-specific loading matrices for each omics layer. In contrast to MOFA2, GFA did not apply an Automatic Relevance Determination mechanism during training; consequently, all latent factors were retained throughout model fitting.

Following model fitting, variance explained was computed for each latent factor and separately for the RNA-seq and DNA methylation views. These variance estimates were used to guide downstream analyses, without applying factor pruning during model fitting.

## **Functional Enrichment Pipeline (GSEA and ORA)**

Functional enrichment analyses were performed to support the biological interpretation of latent factors inferred by MOFA2 and GFA. Because RNA expression data and promoter-level DNA methylation data yield factor weight structures with different statistical properties, enrichment strategies were selected in a view-appropriate manner.

For RNA-derived factor weights, genes were ranked according to their continuous loading values. GSEA was performed using the Gene Ontology Biological Process ontology as implemented in the clusterProfiler package. This ranked-list approach was applied without imposing arbitrary cut-offs, allowing the full distribution of RNA factor loadings to be used as input.

For methylation-derived factors, CpG probes were mapped to promoter-associated genes using Illumina HumanMethylation450K annotation resources. For each factor, promoter-level gene sets were constructed by selecting genes associated with CpGs exhibiting the largest absolute factor loadings. Functional enrichment of these discrete gene sets was assessed using ORA with GO:BP terms, using the set of promoter-associated genes as background.

For both GSEA and ORA, multiple testing correction was applied using the Benjamini–Hochberg false discovery rate procedure, with a significance threshold of  $FDR \leq 0.05$ . Enrichment outputs were retained at the factor and omics-view levels for downstream comparative analyses between MOFA2 and GFA.

## **Overlap and Comparative Analysis Workflow**

To support systematic comparison of biological outputs generated by MOFA2 and GFA, enriched GO:BP terms were aggregated at the model level rather than analysed exclusively on a factor-by-factor basis. For each model, enrichment results derived from all latent factors were combined into unified GO:BP term sets. This approach enabled comparison of overall biological coverage independently of factor number, sparsity patterns, or differences in model structure.

Enrichment results were aggregated separately for MOFA2 and GFA within each omics layer. RNA-seq-derived and DNA methylation-derived enrichment outputs were processed independently to reflect differences in data characteristics and enrichment procedures, with GSEA applied to RNA-derived factors and ORA applied to methylation-derived factors. This separation ensured that enrichment outputs obtained under different statistical assumptions were not combined during comparison. Functional overlap between MOFA2 and GFA was assessed by intersecting the aggregated GO:BP term sets derived for each omics layer. These intersections were visualised using UpSet diagrams generated with the ComplexUpset framework. UpSet plots were used solely as a visualisation approach to summarise shared and model-specific GO:BP term sets and did not alter the underlying enrichment results. Separate UpSet plots were generated for RNA-seq-based and DNA methylation-based enrichment outputs.

This workflow was defined to enable structured comparison of enrichment outputs across models with respect to overall biological coverage and cross-omics consistency.

## **Reproducibility and Code Management**

All analyses were implemented using modular R scripts, with each script corresponding to a distinct stage of the analytical workflow, including data preprocessing, MOFA2 model fitting, GFA model fitting, functional enrichment analyses, and cross-model comparisons. This modular structure allowed individual analysis steps to be executed and inspected independently, while preserving a clear and structured dependency between different stages of the pipeline.

Scripts were designed to enable re-execution of the complete analysis workflow starting from raw data files obtained from the GDC Data Portal, using standardised input formats and explicitly defined analysis parameters. This ensured that all preprocessing steps, model configurations, and downstream analyses could be reproduced in a consistent and transparent manner.

To support reproducibility, random initialisation was controlled for stochastic components where applicable, including MOFA2 variational optimisation and GFA model fitting. Key parameters, file paths, and analysis options were explicitly specified within each script to ensure consistent behaviour across repeated runs and to facilitate inspection and debugging.

Together, these implementation practices were adopted to ensure transparent execution, traceable analytical workflows, and reproducible generation of all results reported for MOFA2 and GFA.

## **Results**

### **Analysis of Unpaired TCGA-GBM Dataset**

#### **Dataset Structure and Separation of Unpaired and Paired Cohorts**

The TCGA-GBM project provides multiple molecular data types; however, these data layers were not generated uniformly for the same set of individuals. As a result, the datasets analysed in this study comprised two distinct cohorts serving different analytical purposes: an unpaired cohort used for single-omics analyses and a paired cohort used for integrative multi-omics modelling.

The unpaired RNA-seq STAR-Counts dataset consisted of 22 samples, including 17 glioblastoma tumour samples and 5 normal brain controls. These RNA-seq samples did not overlap with the available DNA methylation data at the patient level and therefore could not support joint multi-omics integration. Consequently, this RNA-seq cohort was analysed independently to characterise transcriptomic differences between tumour and normal samples.

The corresponding DNA methylation dataset also lacked overlapping sample identifiers with the unpaired RNA-seq cohort. Because transcriptomic and methylation measurements could not be linked to the same individuals, integrative modelling was not performed for the unpaired dataset.

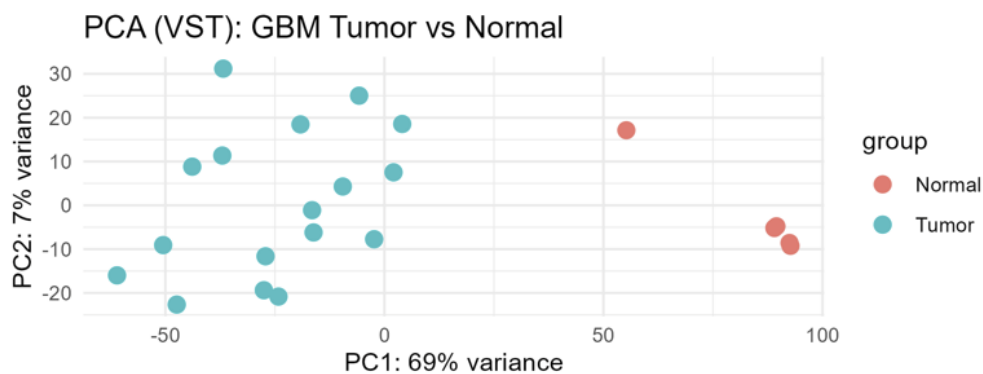
In contrast, a separate cohort of 84 glioblastoma tumour samples contained both RNA-seq and DNA methylation measurements derived from the same individuals. This paired dataset enabled sample-level integration across omics layers and formed the basis for all subsequent integrative analyses presented in the following sections.

#### **Differential Gene Expression Analysis for Unpaired RNA-seq Dataset**

Differential gene expression analysis was performed on the unpaired TCGA-GBM RNA-seq dataset to characterise transcriptional differences between glioblastoma tumours and normal brain tissue. After preprocessing and filtering, 26,582 genes across 22 samples (17 tumour and 5 normal brain tissues) were retained for statistical testing using DESeq2.

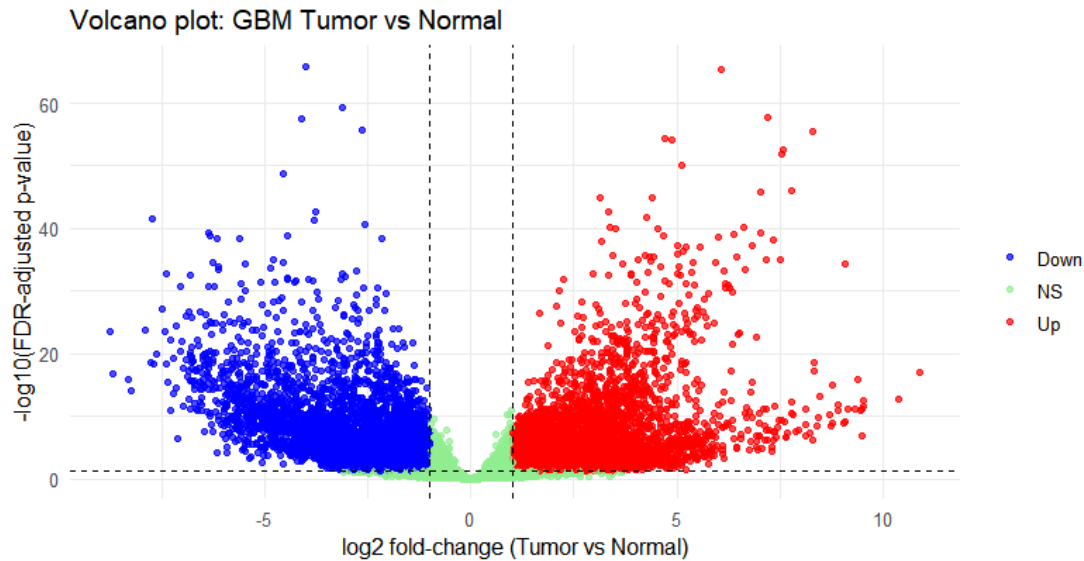
At a FDR threshold of  $< 0.05$ , a total of 13,080 genes were identified as significantly differentially expressed between tumour and normal samples. Of these, 6,938 genes showed higher expression in tumours ( $\log_2$  fold-change  $> 0$ ), while 6,142 genes showed lower expression in tumours ( $\log_2$  fold-change  $< 0$ ), indicating widespread transcriptional dysregulation associated with glioblastoma.

To examine sample-level structure, variance-stabilised expression values were used for principal component analysis (PCA). PCA revealed a clear separation between tumour and normal samples along the first principal component (Figure 5), indicating that the tumour–normal contrast represents the dominant source of transcriptional variation in the unpaired dataset. Normal samples formed a compact cluster, whereas tumour samples showed greater dispersion, reflecting transcriptional heterogeneity among glioblastoma samples. No strong outliers were observed among tumour samples.



**Figure 5.** Principal component analysis (PCA) of variance-stabilising transformed (VST) RNA-seq expression values from the unpaired TCGA-GBM cohort. Each point represents one sample and is coloured by sample type. The first principal component (PC1) explains 69% of the total variance and separates tumour samples (green) from normal samples (orange), while the second principal component (PC2) explains an additional 7% of the variance.

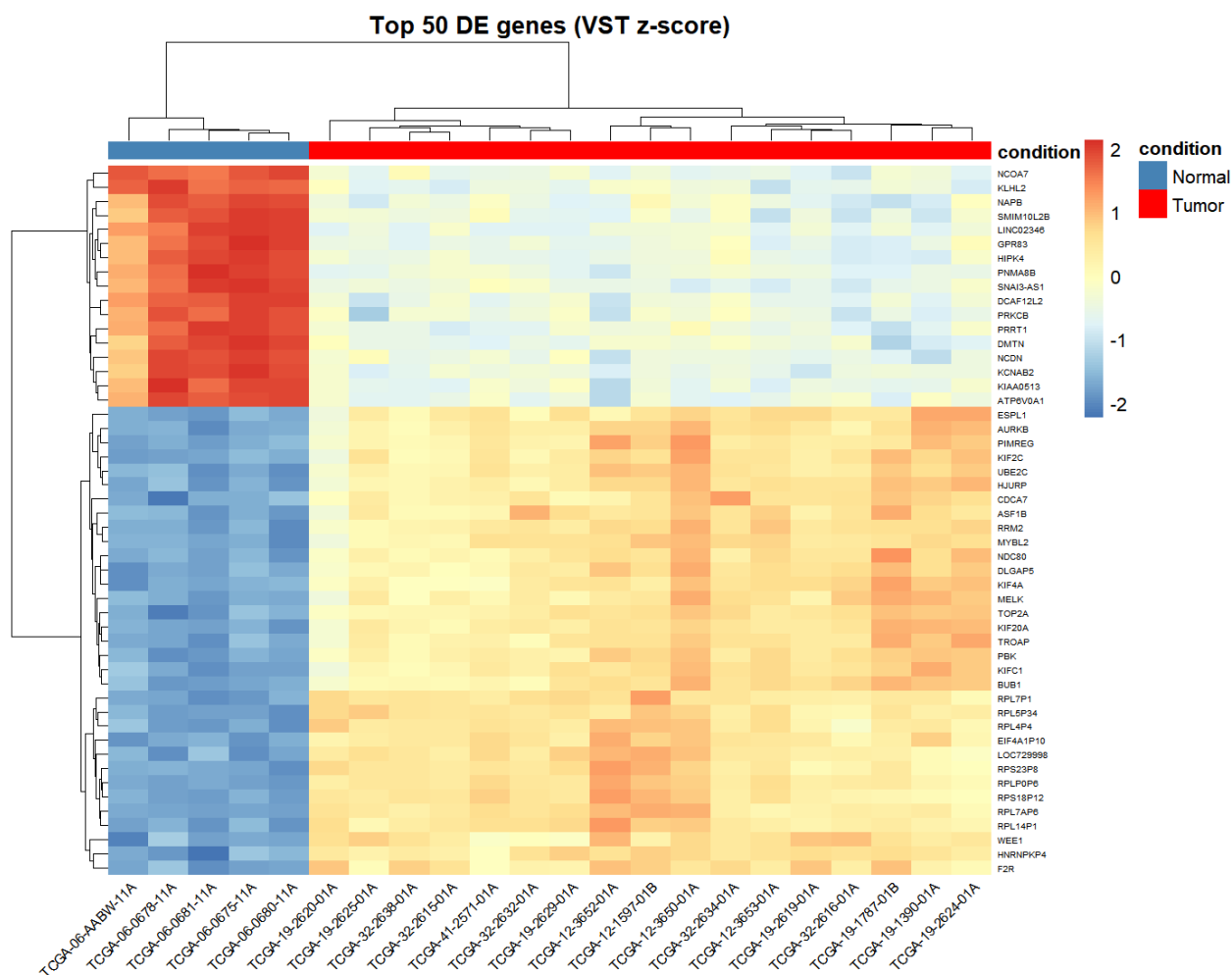
The volcano plot (Figure 6) illustrates the relationship between effect size and statistical significance for differential gene expression, with each point representing a single gene plotted by  $\log_2$  fold change and  $-\log_{10}$  adjusted p-value. Genes with positive  $\log_2$  fold changes are up-regulated in tumour samples, whereas genes with negative  $\log_2$  fold changes are down-regulated relative to normal brain tissue. The broad and approximately symmetric distribution of significantly differentially expressed genes in both directions highlights the extensive and bidirectional transcriptional reprogramming observed in the unpaired RNA-seq dataset.



**Figure 6.** Volcano plot of differential gene expression between GBM tumour and normal brain samples in the unpaired RNA-seq dataset. Each point represents one gene, plotted by  $\log_2$  fold-change and  $-\log_{10}$  false discovery rate (FDR)-adjusted p-value. Genes with positive  $\log_2$  fold-change are up-regulated in tumours (red), whereas genes with negative  $\log_2$  fold-change are down-regulated (blue). Genes that do not reach statistical significance ( $FDR \geq 0.05$ ) are shown in green.

A heatmap of the top 50 differentially expressed genes (Figure 7) provides a focused view of transcriptional differences between tumour and normal samples. Unsupervised hierarchical clustering largely separates tumour and normal tissues into two groups, indicating that a relatively small subset of strongly differentially expressed genes is sufficient to discriminate between the two conditions. Genes up-regulated in tumour samples display coordinated expression patterns across the tumour group, whereas a distinct set of genes shows consistently lower expression in tumours compared with normal brain tissue. The clear separation between these expression patterns highlights structured and coherent transcriptional differences that are primarily driven by condition-level effects rather than isolated sample-specific variation.

Together, these analyses indicate that the unpaired TCGA-GBM RNA-seq dataset captures strong and structured transcriptional differences between tumour and normal brain tissue. The DESeq2 workflow yielded stable estimates and consistent expression patterns across samples, providing a transcriptomic reference that contextualises and supports the interpretation of the tumour-only multi-omics integration analyses presented in subsequent sections.



**Figure 7.** Heatmap of variance-stabilised and z-scored RNA-seq expression values for the top 50 differentially expressed genes ranked by adjusted p-value in the unpaired TCGA-GBM cohort. Rows represent genes and columns represent samples (5 normal and 17 tumour). Hierarchical clustering was applied to both genes and samples, and the heatmap illustrates the separation between tumour and normal samples based on their expression profiles.

Together, these analyses show that the unpaired TCGA-GBM RNA-seq dataset captures strong and structured transcriptional differences between tumour and normal brain tissue. The DESeq2 workflow produced stable estimates and consistent patterns across samples, providing a transcriptomic reference that contextualises the tumour-only multi-omics integration analyses presented in subsequent sections.

## Analysis of Paired Dataset

### Transcriptomic Processing and Identification of Paired RNA Samples

For integrative multi-omics analysis, only tumour samples were considered, reflecting the availability of tumour-only DNA methylation profiles in the TCGA-GBM cohort. RNA-seq data were initially processed to obtain a clean transcriptomic dataset suitable for downstream integration.

Tumour RNA-seq data comprised 386 samples corresponding to 288 unique patients, while the tumour DNA methylation dataset consisted of 140 samples derived from 140 unique patients. Patient-level identifiers extracted from TCGA barcodes were used to assess overlap between the two omics layers.

This matching procedure identified 84 individuals with both RNA-seq and DNA methylation measurements, defining the paired tumour cohort used for all subsequent integrative analyses.

Restricting integration to this paired cohort ensured that transcriptomic and epigenomic profiles originated from the same individuals, enabling direct modelling of shared and modality-specific sources of variation using MOFA2 and GFA.

## **DNA Methylation Preprocessing and Promoter-level Matrix**

DNA methylation data from the TCGA-GBM cohort were summarised at the promoter level to enable integration with gene-level RNA-seq data. CpG probes annotated to promoter-proximal regions (TSS200 and TSS1500) were mapped to genes and aggregated to obtain gene-level promoter methylation estimates. For each gene, promoter-level methylation values were computed by averaging methylation signals across all associated CpG probes, resulting in a single promoter-level value per gene. Following gene-level summarisation, the promoter-level methylation matrix was restricted to tumour samples with corresponding RNA-seq measurements. Gene identifiers were harmonised across transcriptomic and epigenomic datasets to ensure direct correspondence between features. This harmonisation yielded aligned RNA-seq and promoter-level DNA methylation matrices derived from the same set of 84 tumour samples. The resulting paired RNA expression and promoter methylation datasets formed the input for all subsequent integrative multi-omics analyses using MOFA2 and GFA, enabling investigation of shared and modality-specific molecular variation across regulatory layers.

## **MOFA2 Model Fitting and Data Preparation**

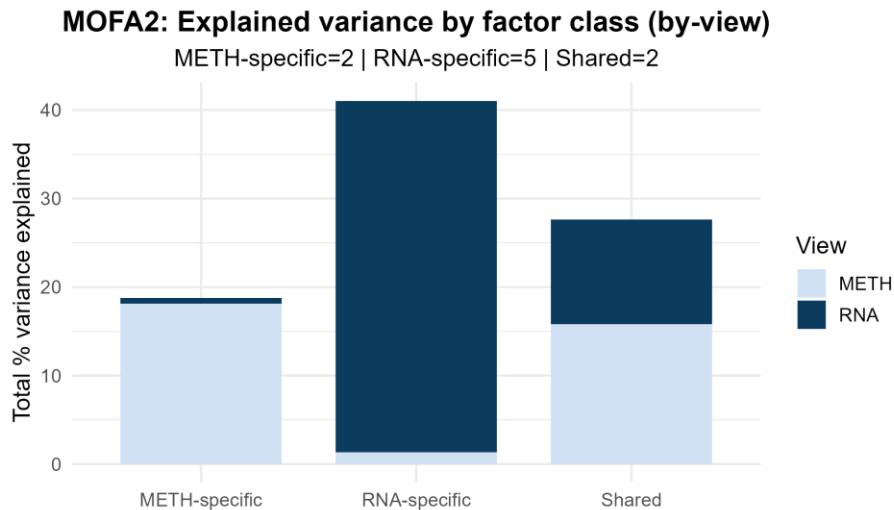
Paired glioblastoma tumour samples with both RNA-seq and DNA methylation measurements were used for MOFA2 modelling. After harmonisation of sample identifiers across omics layers, the paired cohort comprised 84 tumour samples with complete data available in both modalities. For integration, RNA-seq and DNA methylation data were restricted to a shared gene space to ensure direct comparability across views. The final RNA-seq input matrix consisted of 15,201 gene-level expression features, while the promoter-level DNA methylation matrix contained 15,201 corresponding promoter-associated features, represented as M-values. Both matrices were provided to MOFA2 in a feature-by-sample format with matched sample ordering.

MOFA2 was initialised with 15 latent factors and trained until convergence, as indicated by stabilisation of the Evidence Lower Bound (ELBO). Following model training and automatic pruning of non-informative components, 9 latent factors remained active in the fitted model. These factors constitute the latent structure learned by MOFA2 from the paired RNA-seq and DNA methylation data and were retained for downstream inspection and analysis.

## **Latent Factor Structure Identified by MOFA2 in RNA and DNA Methylation**

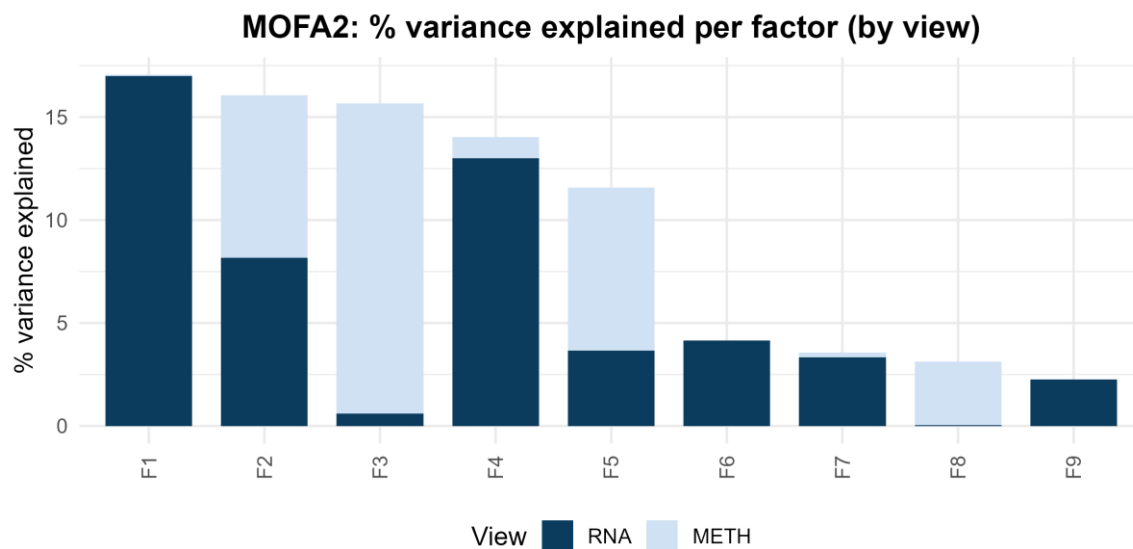
Variance explained (VE) was quantified for each MOFA2 latent factor to assess how variability was distributed across RNA expression and promoter-level DNA methylation. Factors were classified using an 80/20 variance contribution criterion across views: factors were defined as RNA-specific when the RNA share of VE was  $\geq 0.80$ , methylation-specific when the RNA share was  $\leq 0.20$ , and shared otherwise. Factors with total VE below 2% were defined as weak; however, no weak factors were identified in the paired GBM dataset.

Using this classification, five RNA-specific factors, two methylation-specific factors, and two shared factors were identified. RNA-specific factors captured the majority of variance within the transcriptomic layer, whereas methylation-specific factors explained variance restricted to promoter-level DNA methylation (Figure 8). The shared factors contributed variance to both RNA expression and DNA methylation, indicating a limited but detectable degree of coordinated cross-omics structure in the paired GBM cohort.



**Figure 8.** Variance explained by RNA expression and promoter-level DNA methylation across MOFA2 latent factor classes in the paired TCGA-GBM cohort. Bars represent the percentage of variance explained within each factor class, separated by omics layer. Factor classes distinguish RNA-dominant, methylation-dominant, and shared latent components.

MOFA2 factors showed a clear view-specific structure in terms of variance explained. RNA-specific factors accounted for the largest proportion of explained variance and were driven predominantly by transcriptomic variation, whereas methylation-specific factors explained variance largely restricted to the promoter-level DNA methylation view. Shared factors contributed detectable variance to both omics layers, indicating limited but measurable cross-omics coupling within the paired GBM cohort. The relative contribution of RNA expression and DNA methylation to each individual factor is summarised in Figure 9.



**Figure 9:** Percentage of variance explained by individual MOFA2 latent factors, separated into RNA expression and promoter-level DNA methylation components in the paired TCGA-GBM cohort. Factors were classified as RNA-specific when the proportion of variance explained by RNA expression was  $\geq 0.80$ , as methylation-specific when  $\leq 0.20$ , and as shared otherwise. Latent factors explaining less than 2% of the total variance were classified as weak.

Together, these results indicate that transcriptomic variation represents the dominant source of latent structure in the paired GBM dataset, while promoter-level DNA methylation contributes additional, partially independent signals across the remaining nine active factors.

### **Functional enrichment of MOFA2 latent factors (GSEA and ORA on gene-level weights)**

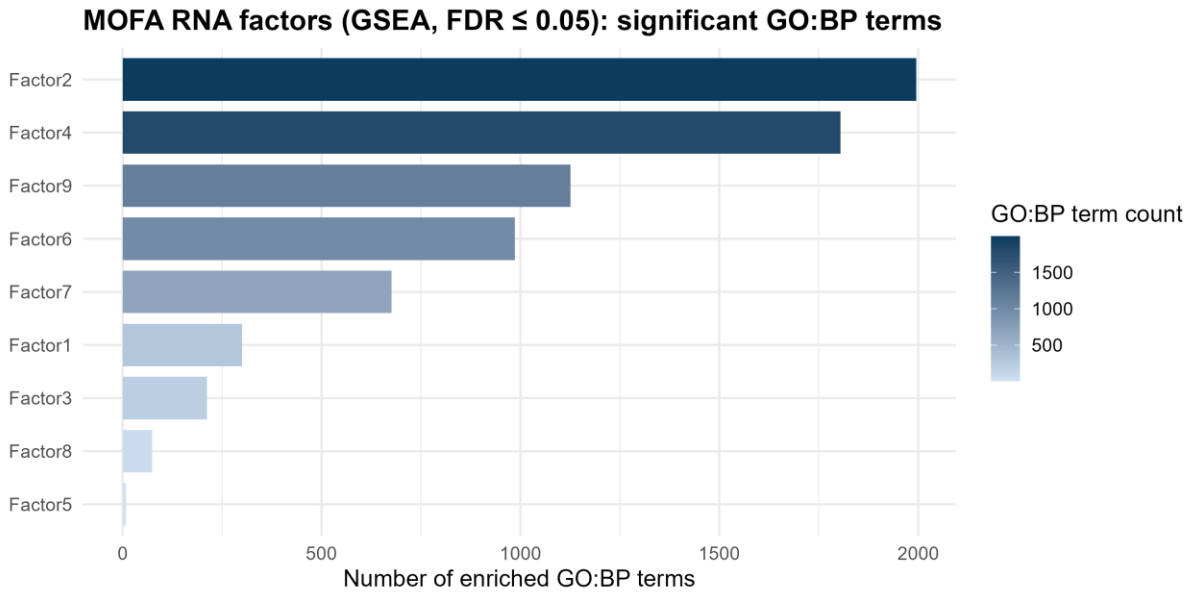
Functional enrichment analysis was performed to interpret the biological processes associated with latent factors inferred by MOFA2 from paired RNA expression and promoter-level DNA methylation data. Gene-level factor weights obtained from each omics view were used as input for enrichment analyses. Two complementary strategies—GSEA and ORA—were applied, selected according to the statistical structure of the factor weight distributions in each data modality.

For RNA-derived MOFA2 factors, enrichment analysis was conducted using GSEA. RNA factor loadings form continuous, signed weight distributions across a large number of genes, which are well suited for ranked-list-based enrichment approaches. GSEA evaluates whether predefined gene sets are systematically enriched toward the extremes of a ranked gene list without relying on an arbitrary significance threshold for individual genes. This allows the detection of coordinated transcriptional programmes driven by modest but consistent effects across many genes. Genes were ranked according to their signed RNA factor loadings, and enrichment was assessed using GO:BP terms with FDR control at  $\leq 0.05$ .

In contrast, promoter-level DNA methylation factors were analysed using ORA. Methylation-derived factor weights were summarised at the gene level and interpreted primarily through subsets of genes with the strongest absolute loadings. Compared to RNA expression, methylation signals tend to be sparser and more localised, often dominated by a limited number of genes showing strong effects. Under these conditions, ORA provides a more appropriate framework by testing whether strongly weighted genes are over-represented in specific biological processes relative to a defined background gene universe. ORA was therefore applied to top-ranked promoter-associated genes for each methylation factor using GO:BP terms, with  $FDR \leq 0.05$ .

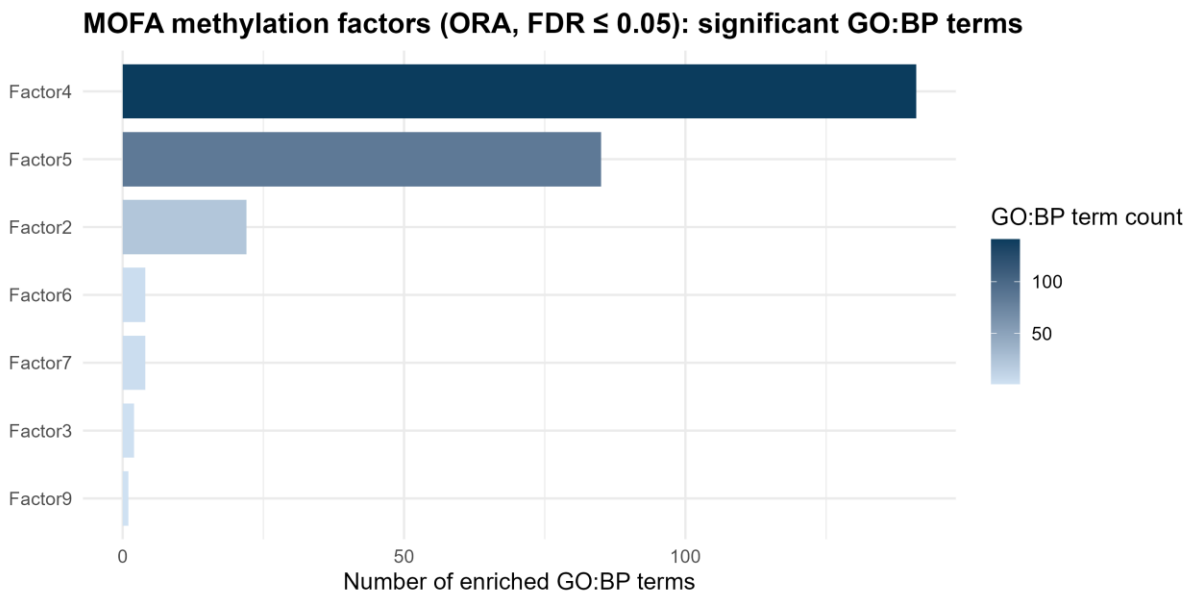
Prior to enrichment analysis, gene identifiers were harmonised by mapping gene symbols to ENTREZ identifiers using standard annotation resources. Only successfully mapped genes were retained for downstream analyses. RNA-derived factors showed a small proportion of unmapped genes, primarily due to deprecated symbols or pseudogene annotations, whereas promoter-level methylation features exhibited near-complete mappability, reflecting the structured annotation of array-based CpG probes.

RNA-derived MOFA2 factors exhibited substantial heterogeneity in their functional enrichment profiles. As shown in Figure 10, several RNA factors were associated with a large number of significantly enriched GO:BP terms, indicating that these latent components capture broad transcriptional programmes. In particular, two RNA factors (F2 and F4) showed the strongest enrichment signals, each associated with more than one thousand GO:BP terms passing the FDR threshold. Other RNA factors displayed more moderate enrichment, while a subset yielded relatively few significant terms, reflecting variability in the functional breadth of RNA-associated latent factors.



**Figure 10.** Number of significantly enriched GO:BP terms identified by GSEA for each MOFA2 RNA latent factor in the paired TCGA-GBM cohort. Bars indicate the count of GO:BP terms passing a false discovery rate (FDR) threshold of 0.05, with factors ordered by decreasing number of enriched terms.

In contrast, enrichment profiles for methylation-derived MOFA2 factors were markedly narrower. As summarised in Figure 11, only a subset of methylation factors showed significant GO:BP enrichment under ORA, and the total number of enriched terms per factor was substantially lower than that observed for RNA-derived factors. This pattern suggests that promoter-level methylation factors capture more focused, gene-restricted biological signals rather than broad pathway-level programmes, consistent with the regulatory characteristics of DNA methylation.



**Figure 11.** Bar plot showing the number of significantly enriched GO:BP terms identified by ORA for each DNA methylation-derived MOFA2 latent factor in the paired TCGA-GBM cohort (FDR ≤ 0.05). Factors are ordered by decreasing number of enriched GO:BP terms.

Taken together, these results demonstrate that MOFA2 captures distinct functional architectures across transcriptomic and epigenomic layers. RNA-derived latent factors are associated with broad, coordinated biological processes detectable through ranked-list enrichment, whereas methylation-derived factors reflect more localised regulatory signals identifiable through over-representation testing. The combined use of GSEA and ORA therefore provides a statistically appropriate and modality-aware framework for functional interpretation of MOFA2 latent factors in multi-omics integration analyses.

## **GFA Model Fitting and Data Preparation**

GFA was applied to paired RNA-seq and promoter-level DNA methylation data from 84 TCGA-GBM tumour samples. The same preprocessed RNA expression and DNA methylation matrices used in the MOFA2 analysis were employed here, ensuring direct comparability between the two multi-omics integration frameworks. Sample identifiers were harmonised across omics layers, yielding a fully paired dataset with one-to-one correspondence between transcriptomic and epigenomic profiles for all samples.

Prior to model fitting, RNA expression and DNA methylation matrices were independently centred and scaled at the feature level. Each omics layer was supplied to the GFA model as a separate view in a samples-by-features format ( $84 \times 15,201$  features per view). No additional feature filtering was applied at this stage, allowing the model to operate on the full high-dimensional input data.

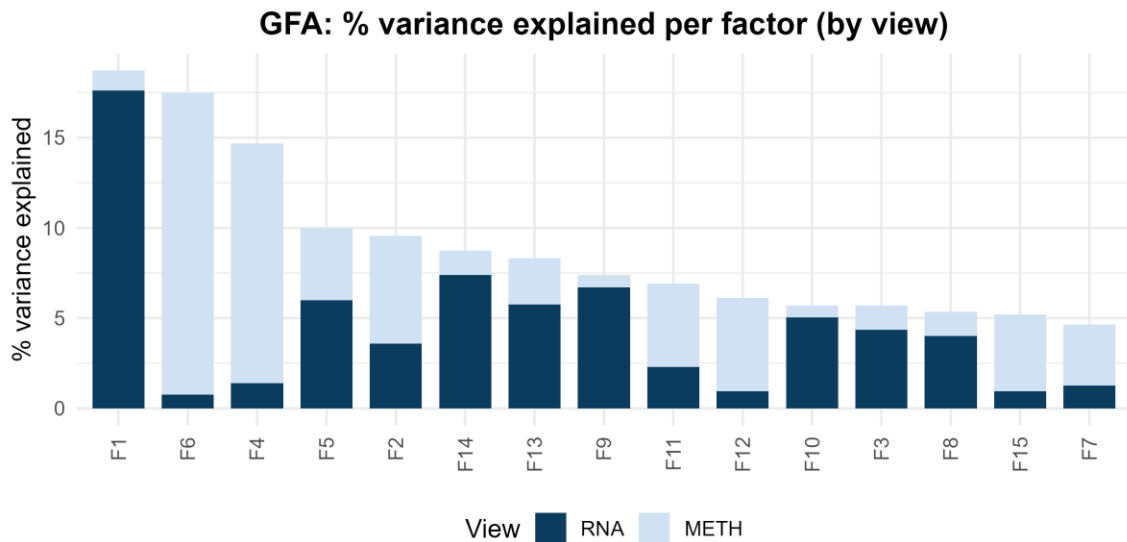
The GFA model was fitted using a fixed number of latent factors ( $K = 15$ ), matching the dimensionality specified in the MOFA2 analysis. Model fitting resulted in a decomposition of the paired RNA expression and DNA methylation data into a shared sample-level latent factor matrix and view-specific loading matrices for each omics layer.

All 15 inferred latent factors were retained for downstream analyses. These factors were subsequently used to quantify variance explained across omics layers, classify factors into shared and omics-specific components, assess cross-omics relationships at the factor level, and perform functional enrichment analyses based on factor-specific molecular signatures.

## **Latent Factor Structure Identified by GFA in RNA and DNA Methylation**

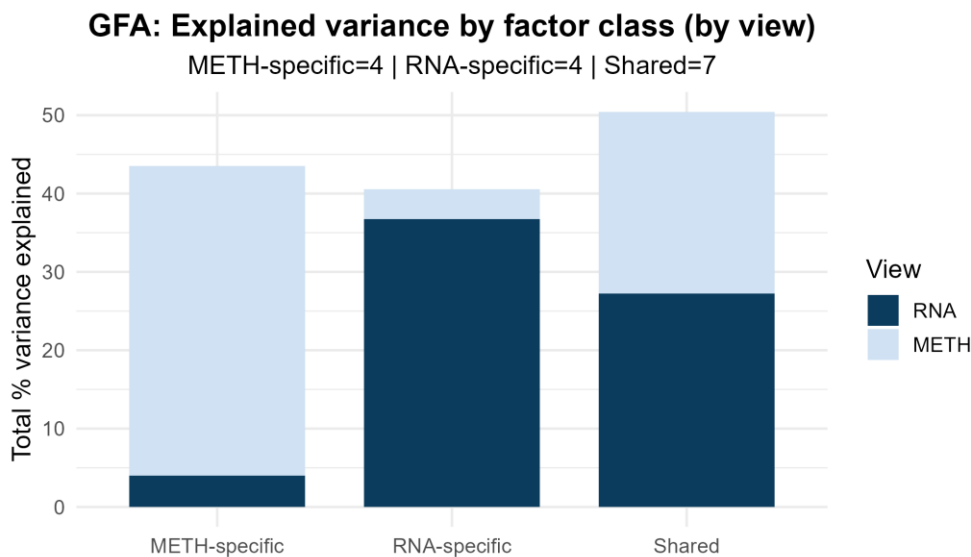
GFA was applied to paired RNA expression and promoter-level DNA methylation data from 84 TCGA glioblastoma samples to characterise the latent structure underlying joint and view-specific sources of molecular variation. The model was fitted with 15 latent factors, which were ranked according to the total proportion of variance explained across both omics layers.

The variance explained by individual latent factors showed marked heterogeneity across RNA expression and DNA methylation views (Figure 12). Several high-ranking factors exhibited pronounced modality dominance, with some factors explaining the majority of their variance through DNA methylation, while others were primarily driven by RNA expression. In contrast, a subset of factors displayed more balanced contributions from both omics layers, indicating shared latent components capturing coordinated cross-omics variation. Ordering factors by decreasing total variance explained revealed a gradual decline in explanatory power across the top 15 factors, without a sharp cutoff separating dominant and weaker components.



**Figure 12.** Stacked bar plot showing the percentage of variance explained by the top 15 GFA latent factors in RNA expression (dark blue) and DNA methylation (light blue) in the paired TCGA-GBM cohort. Factors are ordered by decreasing total variance explained across both omics layers.

To provide a higher-level summary of the inferred latent structure, factors were grouped according to the distribution of their explained variance across the two omics layers, resulting in methylation-specific, RNA-specific, and shared factor classes (Figure 13). RNA-specific factors contributed predominantly through the transcriptomic view, with only minor contributions from DNA methylation. Conversely, methylation-specific factors explained variance mainly within the methylation view, with limited contribution from RNA expression. Shared factors accounted for a substantial fraction of the total explained variance and received contributions from both RNA and methylation, suggesting partially coordinated signals across the two molecular layers within the selected top factors.



**Figure 13.** Total percentage of variance explained by RNA-specific, methylation-specific, and shared GFA latent factors across RNA expression and DNA methylation views, considering the top 15 factors. Factor classes are defined based on the relative contribution of each omics layer to the latent factors.

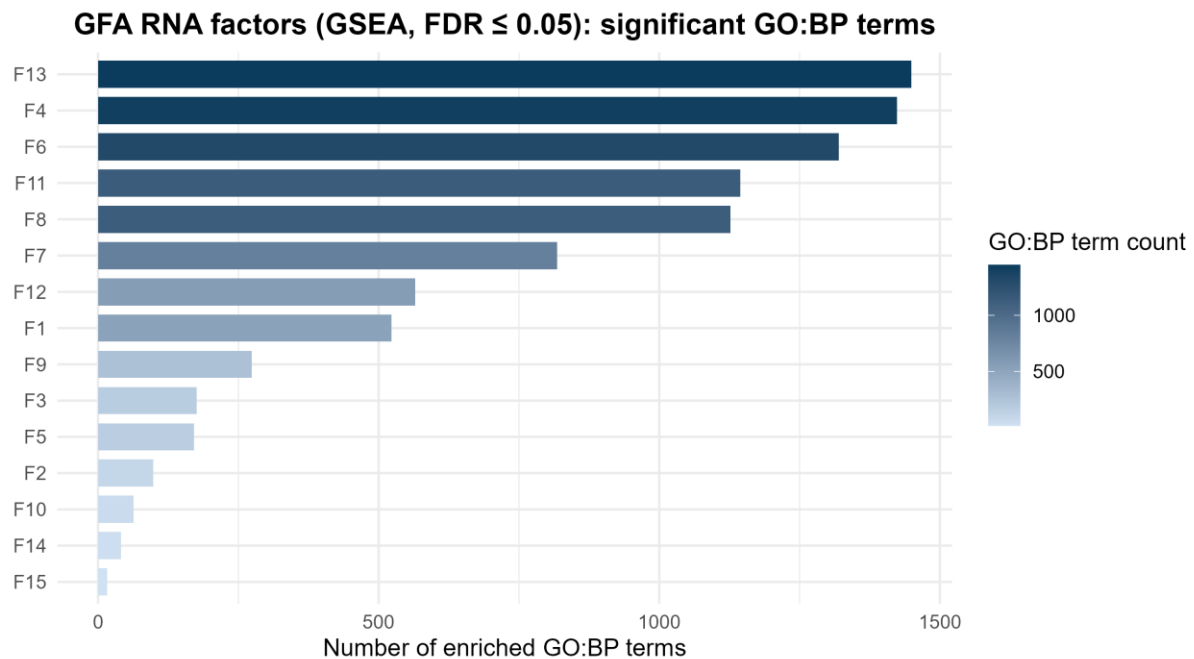
Overall, these results indicate that the latent structure captured by GFA in this dataset comprises a combination of strongly view-specific factors and a substantial set of shared components. This balance suggests that glioblastoma-associated molecular variation includes both modality-specific signals and coordinated cross-omics patterns that are jointly captured by the GFA model. Notably, although factors were classified as RNA-specific or methylation-specific based on dominance of explained variance, small residual variance contributions from the non-dominant omics layer were still observed. This reflects the shared latent score structure of GFA and indicates weak cross-omics coupling rather than strict modality exclusivity (Figure 13).

### Functional Enrichment of GFA Latent Factors (GSEA and ORA on Gene-level Weights)

Functional enrichment analysis was performed to interpret the biological signals captured by the latent factors identified by GFA, using gene-level weight vectors derived from RNA expression and promoter-level DNA methylation views. As in the MOFA2 analysis, two complementary enrichment strategies—GSEA and ORA—were applied in a view-specific manner.

For RNA-derived GFA factors, enrichment was assessed using GSEA. Genes were ranked according to their signed RNA loading values for each GFA factor, and enrichment was evaluated against GO:BP terms with FDR control at  $\leq 0.05$ . All RNA factors exhibiting non-zero signal were included without manual factor selection, allowing enrichment results to reflect the latent structure inferred by the model.

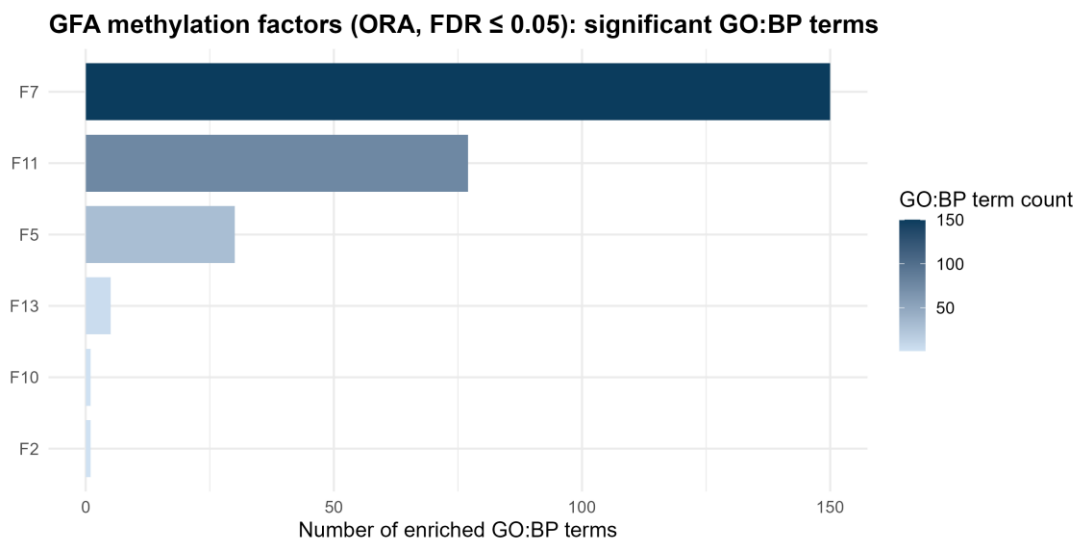
As shown in Figure 14, RNA-derived GFA factors exhibited marked heterogeneity in their functional enrichment profiles. Several factors were associated with a large number of significantly enriched GO:BP terms, indicating broad transcriptional programmes captured by these latent components. In particular, factors F5 and F12 showed the strongest enrichment signals, each associated with more than 1,400 significant GO:BP terms at  $FDR \leq 0.05$ . Additional RNA factors (e.g. F7, F13, F11, and F9) displayed moderate enrichment, while a subset of factors yielded relatively few significant terms. This variability suggests that different RNA-specific GFA factors capture transcriptional programmes of differing biological breadth.



**Figure 14.** Number of significantly enriched GO:BP terms identified by GSEA ( $FDR \leq 0.05$ ) for each RNA-derived GFA latent factor in the paired TCGA-GBM cohort. Factors are ordered by decreasing number of enriched terms.

In contrast, promoter-level DNA methylation factors were analysed using ORA. Methylation-derived factor loadings were interpreted through subsets of genes with the largest absolute promoter-level weights. ORA was applied to test whether these genes were over-represented in specific GO:BP terms relative to the background gene universe, with  $FDR \leq 0.05$ .

The enrichment profiles for methylation-derived GFA factors were substantially narrower than those observed for RNA-derived factors. As shown in Figure 15, only a small subset of methylation factors showed significant GO:BP enrichment. Specifically, factors F15 and F5 exhibited clear enrichment signals, with approximately 85 and 70 significantly enriched GO:BP terms, respectively. Other methylation factors showed very limited or no significant enrichment under the applied thresholds. This pattern indicates that GFA methylation factors capture more focused, gene-restricted regulatory signals rather than broad pathway-level programmes.



**Figure 15.** Bar plot showing the number of significantly enriched GO:BP terms identified by ORA for each DNA methylation-derived MOFA2 latent factor in the paired TCGA-GBM cohort ( $FDR \leq 0.05$ ). Factors are ordered by decreasing number of enriched GO:BP terms.

Together, these results demonstrate that GFA captures distinct functional architectures across transcriptomic and epigenomic layers. RNA-derived latent factors are associated with broad, coordinated biological processes that are effectively detected using ranked-list enrichment approaches such as GSEA, whereas methylation-derived factors reflect more localised regulatory effects that are appropriately characterised using ORA. The combined use of GSEA and ORA therefore provides a consistent and statistically justified framework for functional interpretation of GFA latent factors across heterogeneous omics modalities.

### Comparative Model Behavior of MOFA2 and GFA on High-dimensional RNA–methylation Data

Both MOFA2 and GFA were applied to paired RNA expression and promoter-level DNA methylation data from TCGA-GBM, together comprising more than 80,000 molecular features. Despite the high dimensionality of the dataset, model fitting was stable for both methods, and no numerical or convergence problems were observed. MOFA2 was initialised with twenty latent factors in an over-complete setting. During model training, the ARD prior gradually reduced the contribution of factors that did not explain consistent variation across samples. As a result, only a subset of factors remained

active after convergence. One factor explaining less than 2% of the total variance across both omics layers was classified as weak and excluded from further biological interpretation. The remaining MOFA2 factors captured the major sources of variation in the data, including RNA-specific, methylation-specific, and shared components, and these factors were used for all downstream analyses.

In contrast, GFA was fitted using Bayesian MCMC inference and does not include an automatic mechanism to remove weak factors during model fitting. Therefore, all latent factors specified at initialisation ( $K = 15$ ) were retained in the final model. The relevance of individual GFA factors was assessed after model fitting based on the proportion of variance explained in each omics view. Downstream analyses focused on factors that explained substantial variance, while components with negligible contributions across both RNA expression and DNA methylation were retained in the model but not prioritised for biological interpretation.

Overall, both models converged reliably on the high-dimensional paired dataset and captured meaningful structure across RNA expression and DNA methylation layers. However, the two frameworks differed in how latent structure was represented and interpreted. MOFA2 resulted in a more compact set of active factors through ARD-based suppression during training, whereas GFA retained the full set of latent factors specified at initialisation. These differences reflect distinct modelling strategies and influence how latent components are prioritised and interpreted in downstream analyses.

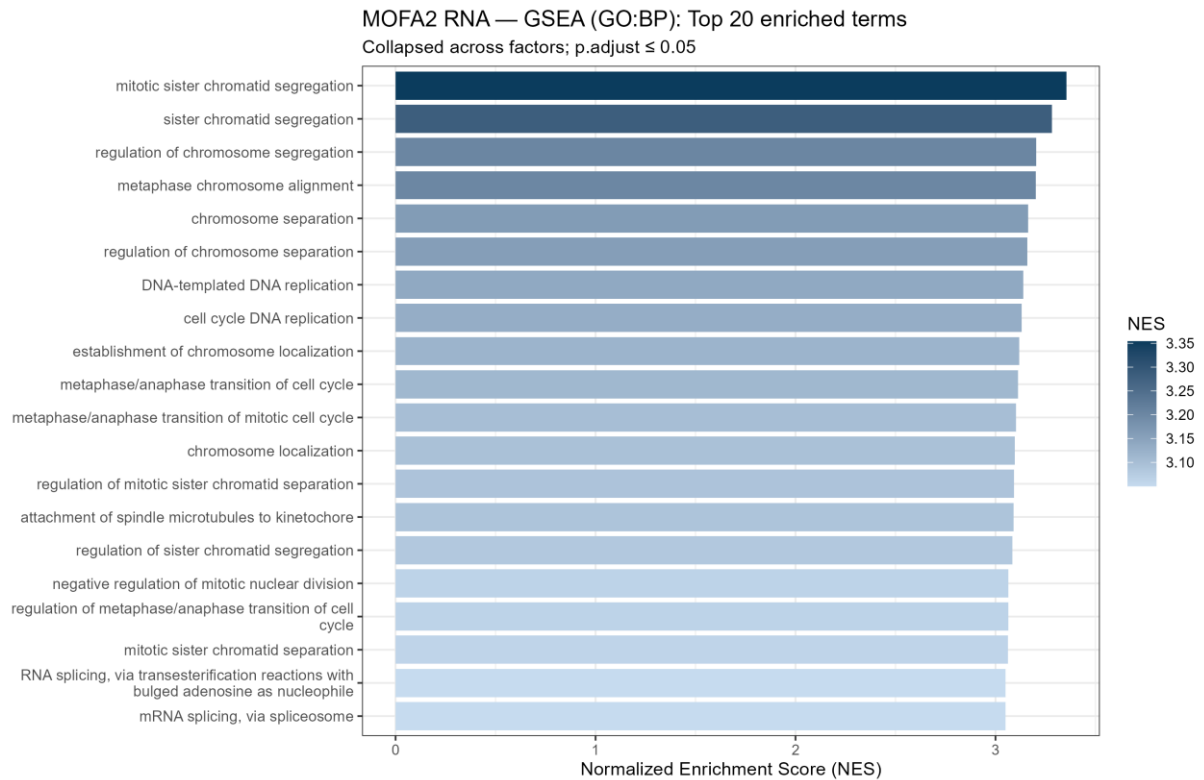
## **GO: BP-Based Functional Enrichment Framework for Cross-Model Comparison**

To obtain a global overview of the biological processes captured by the RNA layer, GSEA results were summarised across all latent factors inferred by MOFA2 and GFA. For each method, RNA GSEA results from all factors were combined, and enrichment was evaluated at the level of GO:BP terms.

Only significantly enriched terms were retained using an adjusted p-value threshold of  $p_{\text{adjust}} \leq 0.05$ . Because the same GO term may appear in multiple factors, factor-specific results were collapsed by selecting, for each GO term and method, the entry with the lowest adjusted p-value; in cases of ties, the term with the highest absolute normalised enrichment score (NES) was retained. This procedure yielded a single, representative enrichment profile per integration method, independent of direct factor-level correspondence.

The top 20 enriched GO:BP terms were then selected separately for MOFA2 and GFA based on statistical significance and NES and visualised as bar plots (Figures 16–17). Across the aggregated RNA GSEA results, the enrichment profiles derived from MOFA2 exhibited a high degree of internal consistency, with closely related GO:BP terms repeatedly capturing overlapping aspects of cell cycle progression and mitotic regulation. The dominance of these processes suggests that a substantial fraction of transcriptomic variation captured by MOFA2 is organised along proliferative axes at the global model level. Importantly, the absence of strongly negatively enriched terms among the top-ranked categories indicates that the summarised RNA enrichment profile of MOFA2 is characterised by a coherent directionality rather than opposing regulatory trends.

In the MOFA2 RNA analysis (Figure 16), the most strongly enriched GO:BP terms were dominated by cell cycle-related processes, including mitotic sister chromatid segregation, chromosome segregation, DNA replication, and cell cycle DNA replication. All top-ranked terms showed positive enrichment scores, indicating a coherent activation of proliferative and mitotic programmes across the RNA data when summarised globally.

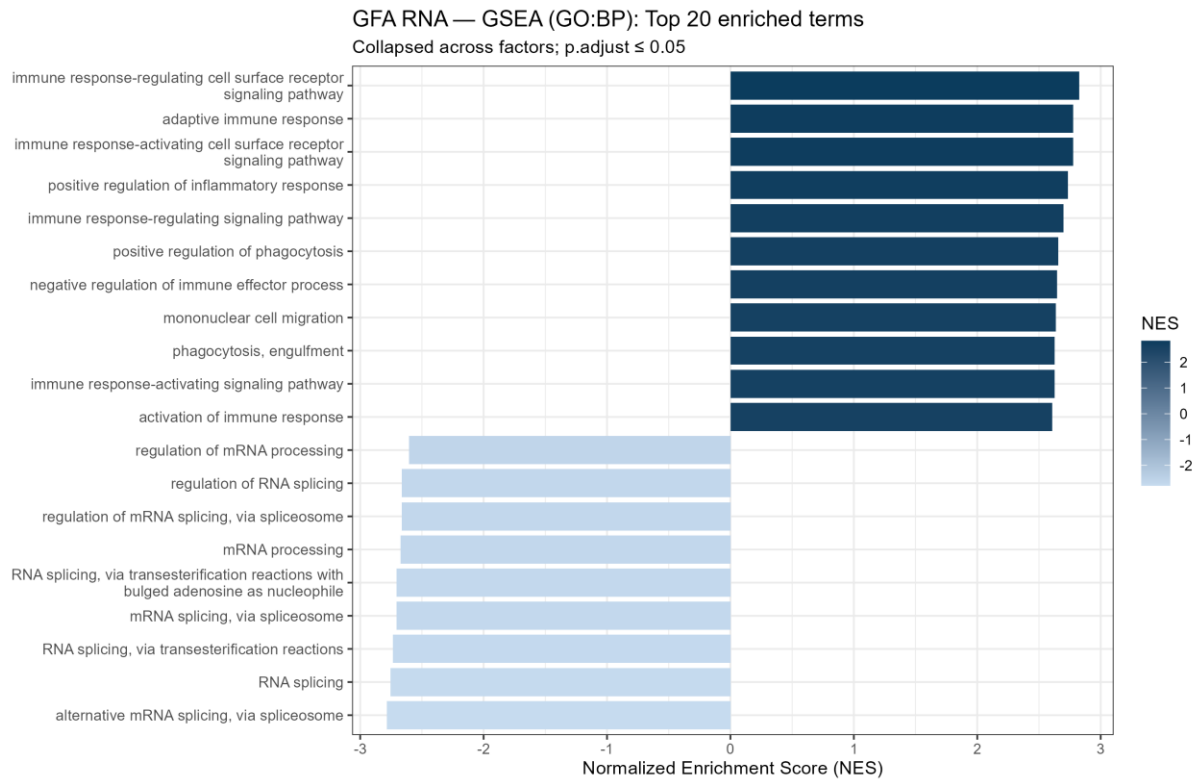


**Figure 16.** Top 20 enriched GO:BP terms identified by GSEA in the RNA layer after aggregating results across all MOFA2 latent factors ( $FDR \leq 0.05$ ).

In contrast, the GFA RNA analysis (Figure 17) revealed a markedly different enrichment profile when results were aggregated across latent factors. The top positively enriched GO:BP terms were predominantly associated with immune-related processes, including immune response–regulating cell surface receptor signalling pathway, adaptive immune response, activation of immune response, and phagocytosis. These terms indicate a strong contribution of immune and inflammatory signalling pathways to the global RNA signal captured by GFA.

Notably, alongside these positively enriched immune-related categories, several RNA metabolism and processing–related processes exhibited negative normalised enrichment scores. These included regulation of RNA splicing, mRNA processing, and RNA splicing via spliceosome, suggesting an opposing regulatory trend within the same aggregated RNA enrichment profile. The simultaneous presence of positively and negatively enriched biological processes highlights a more heterogeneous global RNA signal structure in GFA compared to MOFA2.

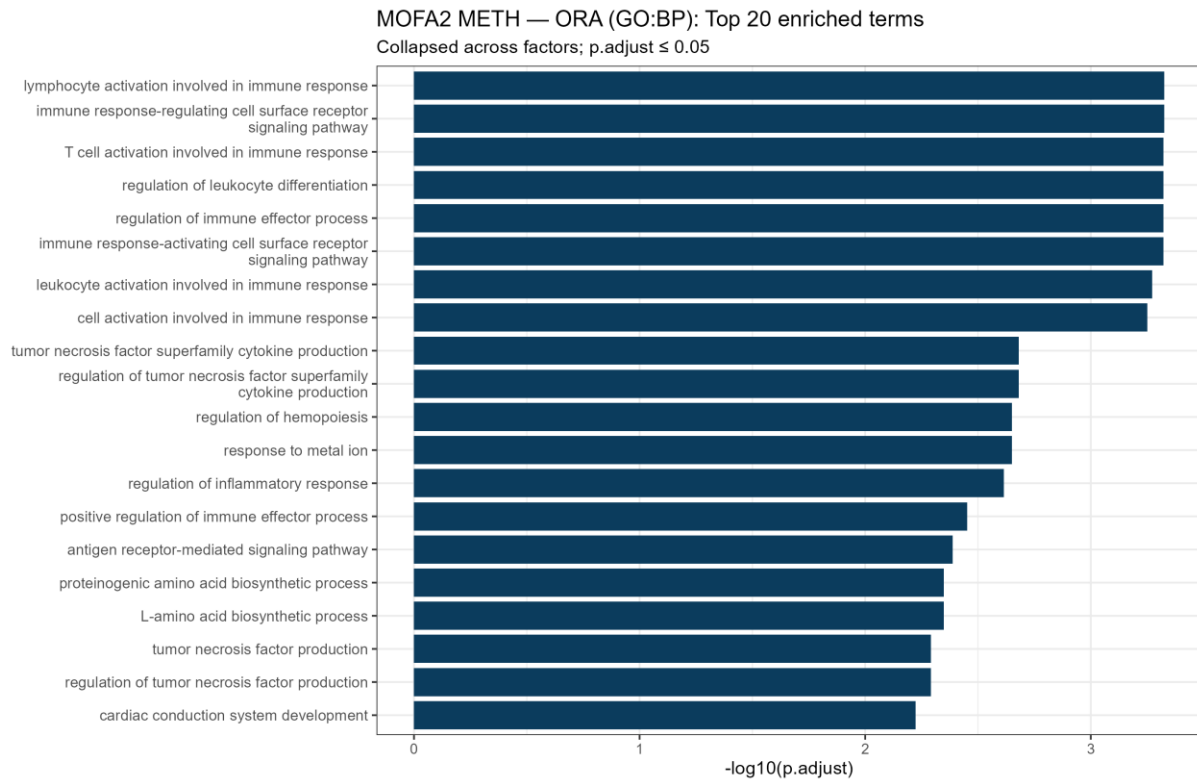
Taken together, the RNA-based GSEA results demonstrate that MOFA2 and GFA capture distinct dominant biological themes when enrichment results are summarised across all latent factors. While MOFA2 emphasises cell cycle– and mitosis-related transcriptional programmes with consistently positive enrichment, GFA highlights immune-associated signalling pathways together with inverse regulation of RNA processing mechanisms. This divergence indicates that the two integration frameworks organise transcriptomic variation into fundamentally different latent structures, despite being applied to the same RNA expression data.



**Figure 17.** Top 20 enriched GO:BP terms identified by GSEA in the RNA layer after aggregating results across all GFA latent factors ( $FDR \leq 0.05$ ).

To characterise biological processes associated with DNA methylation-driven signals, ORA was performed on gene sets derived from MOFA2 and GFA integration outputs. ORA results from all latent factors were aggregated to obtain a global view of enriched GO:BP terms for each method. Only significantly enriched terms were retained using an adjusted p-value threshold of  $p.adjust \leq 0.05$ , and results were collapsed across factors by selecting the most significant occurrence of each GO term. The top 20 enriched terms for each method were visualised (Figures 18–19).

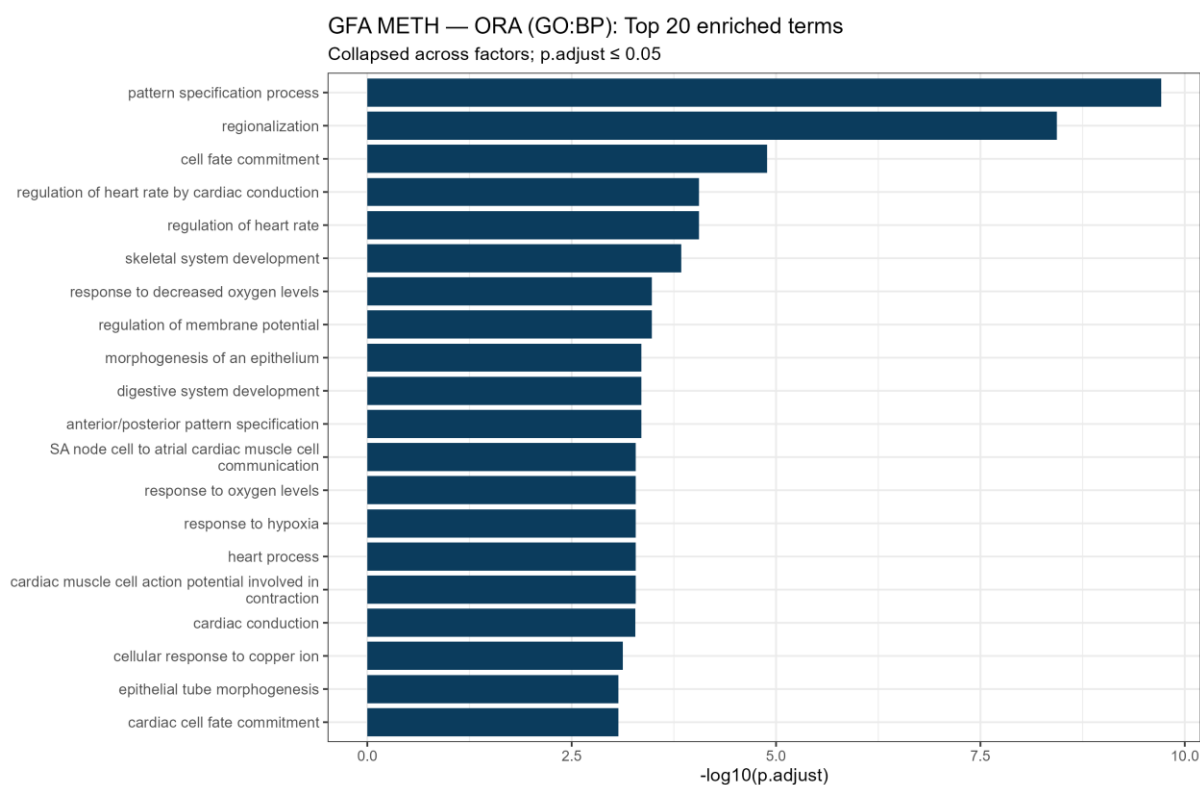
In the MOFA2 DNA methylation analysis (Figure 18), enriched GO:BP terms were predominantly related to immune-associated processes. Highly significant terms included lymphocyte activation involved in immune response, immune response-regulating cell surface receptor signalling pathway, T cell activation, and regulation of immune effector process. Additional enriched categories involved inflammatory responses, cytokine production, and tumour necrosis factor-related pathways. These results indicate that, when aggregated across factors, MOFA2 captures DNA methylation signatures linked to immune regulation and inflammatory signalling.



**Figure 18.** GO:BP terms enriched in the DNA methylation layer of MOFA2 identified by ORA. Bars represent enrichment significance expressed as  $-\log_{10}(\text{FDR-adjusted } p\text{-values})$ .

In contrast, the GFA DNA methylation analysis (Figure 19) revealed a distinct enrichment profile dominated by developmental and physiological processes. The most significant GO:BP terms included pattern specification process, regionalization, cell fate commitment, and multiple processes related to cardiac function, such as regulation of heart rate, cardiac conduction, and cardiac muscle cell action potential. Enrichment of hypoxia- and oxygen-related response terms was also observed. This pattern suggests that GFA highlights DNA methylation signals associated with developmental patterning and tissue-specific physiological regulation rather than immune-related processes.

Together, these ORA results demonstrate that MOFA2 and GFA prioritise different biological themes in the DNA methylation layer when results are summarised globally. While MOFA2 emphasises immune and inflammatory processes, GFA captures methylation patterns associated with developmental regulation and organ-specific functional pathways. These differences reflect method-specific representations of DNA methylation-associated variation in the dataset.



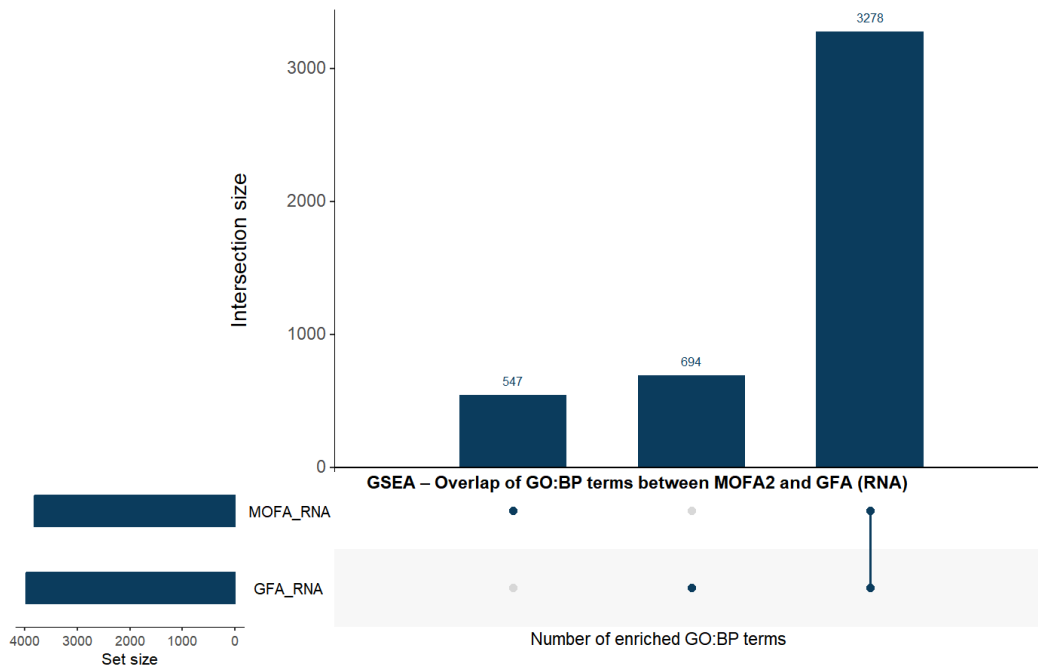
**Figure 19.** Top 20 enriched GO:BP terms identified by ORA in the DNA methylation layer after aggregating results across all GFA latent factors ( $FDR \leq 0.05$ ). Bars represent enrichment significance expressed as  $-\log_{10}(FDR\text{-adjusted } p\text{-values})$ .

To assess the concordance between MOFA2 and GFA at the level of functional interpretation, overlap analyses were performed on GO:BP terms identified by RNA-based GSEA and DNA methylation-based ORA. For both omics layers, enriched GO:BP terms passing FDR correction (adjusted  $p \leq 0.05$ ) were aggregated across all factors for each model prior to comparison.

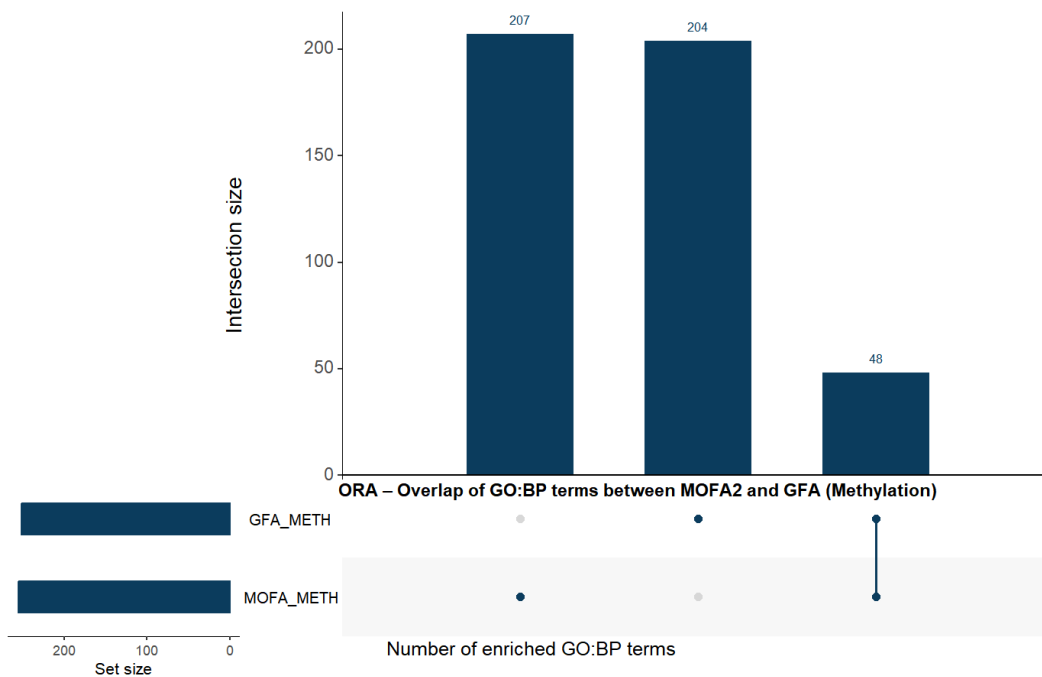
For RNA-based GSEA results (Figure 20), a substantial overlap in enriched GO:BP terms was observed between MOFA2 and GFA. A total of 3,278 GO:BP terms were shared between the two methods, indicating strong agreement in the biological processes captured at the transcriptomic level. In addition to this shared component, MOFA2 identified 547 GO:BP terms that were not detected by GFA, while GFA uniquely captured 694 terms. Despite these model-specific contributions, the dominance of the shared intersection suggests that both factorisation approaches converge on a largely consistent functional signal when applied to RNA expression data.

In contrast, overlap analysis of DNA methylation-based ORA results revealed a markedly different pattern (Figure 21). The number of shared GO:BP terms between MOFA2 and GFA was limited to 48, while a large proportion of enriched terms were model-specific. MOFA2 uniquely identified 207 GO:BP terms, whereas GFA captured 204 distinct terms not shared with MOFA2. This reduced overlap highlights a stronger divergence between the two methods at the level of promoter-associated methylation signals, suggesting that MOFA2 and GFA structure and prioritise epigenetic variation in fundamentally different ways.

Taken together, these results demonstrate that MOFA2 and GFA show high functional concordance for RNA-based enrichment analyses but diverge substantially when applied to DNA methylation-based ORA. This contrast underscores the greater methodological sensitivity of methylation-derived functional interpretation to model assumptions and factor structure, compared to transcriptomic enrichment results.



**Figure 20.** UpSet plot summarising the overlap and model-specific GO:BP terms identified by RNA-based GSEA in MOFA2 and GFA. Bars represent the number of enriched GO:BP terms unique to each model or shared between them after FDR correction ( $FDR \leq 0.05$ ).



**Figure 21.** UpSet plot summarising the overlap and model-specific GO:BP terms identified by DNA methylation-based ORA in MOFA2 and GFA. Bars represent the number of enriched GO:BP terms unique to each model or shared between them after FDR correction ( $FDR \leq 0.05$ ).

## Conclusion

This study compared two unsupervised multi-omics integration methods, MOFA2 and GFA, using paired RNA-seq and promoter-level DNA methylation data from glioblastoma tumours. The primary aim was to assess how each method captures modality-specific and shared biological variation in high-dimensional multi-omics data. Both approaches supported a consistent overall conclusion: transcriptomic and promoter-level DNA methylation variation in glioblastoma largely represent complementary and non-redundant sources of information, with limited evidence for strongly coupled cross-omics latent structure, although weak cross-layer contributions were observed.

MOFA2 produced a compact and strongly regularised latent representation through variational Bayesian inference combined with ARD prior. As a result, only a small number of latent factors remained active, leading to a clear separation between RNA-driven, methylation-driven, and shared sources of variation. This behaviour was reflected in downstream functional analyses. RNA-driven MOFA2 factors were associated with a large number of significantly enriched GO Biological Process terms in gene set enrichment analysis, indicating coherent and dominant transcriptional programmes. In contrast, methylation-driven MOFA2 factors showed more focused but biologically interpretable enrichment profiles. Overall, MOFA2 concentrated biological signal into a limited number of well-defined latent factors, supporting straightforward biological interpretation.

GFA exhibited a different behaviour. Due to its more conservative shrinkage strategy and the absence of aggressive factor suppression during model fitting, variation was distributed across a larger number of latent factors. Functional enrichment analyses reflected this structure: RNA-associated GFA factors generally showed fewer significantly enriched GO:BP terms per factor, while methylation-associated factors displayed narrower and more selective enrichment patterns. Variance decomposition further indicated that methylation-related components accounted for the largest proportion of explained variance, reflecting both the structure of the methylation data and the conservative regularisation strategy employed by GFA. Although individual GFA factors were less sharply defined than those identified by MOFA2, the model preserved weaker patterns that may be suppressed under stronger regularisation.

Taken together, these results highlight a clear methodological trade-off between the two approaches. MOFA2 prioritises parsimony and interpretability by concentrating biological signal into a small number of dominant latent factors, whereas GFA prioritises completeness by retaining a broader and more distributed latent structure. Importantly, the two methods did not lead to conflicting biological interpretations; instead, they converged on the same underlying biological themes while representing them at different levels of resolution.

Overall, this study demonstrates that the choice of a multi-omics integration method should be guided by the specific analytical objective. For analyses aiming to identify a small number of robust, pathway-rich biological signals, MOFA2 is well suited. In contrast, for more exploratory analyses that seek to preserve subtle and distributed sources of variation across molecular layers, GFA offers distinct advantages. Applying both methods in parallel can therefore provide a more comprehensive and reliable characterisation of multi-omics variation in glioblastoma.

## Scientific Contribution and Novelty

The primary contribution of this project is methodological rather than biological. This study provides a controlled and systematic benchmarking of two unsupervised multi-omics integration frameworks, MOFA2 and GFA, applied to the same paired RNA-seq and promoter-level DNA methylation dataset from glioblastoma. By standardising data preprocessing, model inputs, and evaluation criteria, the analysis isolates differences arising from the modelling approaches themselves rather than from technical or dataset-specific factors.

A central methodological contribution of this work is the explicit comparison of how distinct inference strategies and regularisation mechanisms shape the latent factor space. In particular, the ARD prior employed by MOFA2 actively suppresses latent factors that do not explain stable variance across samples, resulting in a compact and parsimonious representation. From a statistical perspective, this behaviour increases model stability and reduces the risk of overfitting, thereby enhancing interpretability. However, it also implies a reduced sensitivity to weaker sources of variation that may be biologically meaningful but explain limited variance. In contrast, GFA retains a broader set of active factors under a fixed latent dimensionality, capturing more distributed and low-amplitude signals at the cost of increased interpretational complexity.

While MOFA2 and GFA have previously been applied separately, this work offers a direct comparative perspective that clarifies how regularisation and inference choices influence latent factor dimensionality, stability, and downstream interpretability. The novelty of the study therefore lies in its evaluative design and in providing evidence-based guidance for method selection and interpretation in unsupervised multi-omics analyses.

## **Ethical Considerations and Societal Impact**

This study is based exclusively on publicly available, de-identified datasets from the TCGA and does not involve new data collection or direct interaction with human subjects. As such, it does not raise ethical concerns related to patient consent, privacy, or data ownership beyond those already addressed by the original data providers.

From a broader perspective, the societal and ethical relevance of this work lies in its implications for the responsible interpretation of multi-omics analyses in biomedical research. Unsupervised integration methods are increasingly used to derive biological hypotheses and, in some contexts, to inform translational or clinical research. This study demonstrates that methodological choices—such as inference strategy and regularisation—can substantially influence the inferred latent structure and downstream biological interpretation, even when applied to the same dataset. Overlooking these methodological effects may lead to overconfident or misleading conclusions regarding shared molecular mechanisms.

By highlighting the strengths and limitations of different modelling assumptions, this work contributes to more transparent, cautious, and reproducible multi-omics analyses. In the long term, such methodological awareness is essential to ensure that data-driven insights are interpreted responsibly, particularly in disease contexts where premature or overstated conclusions could influence downstream research priorities or clinical expectations.

## **Limitations and Future Directions**

This study focused on benchmarking unsupervised multi-omics integration methods using paired bulk RNA-seq and promoter-level DNA methylation data, which entails several limitations. First, the analysis was restricted to two molecular layers, limiting the ability to assess how post-transcriptional and post-translational regulation aligns with the transcriptional and epigenetic variation captured by the inferred latent factors. A natural extension would therefore be the inclusion of additional omics layers, such as proteomics or phosphoproteomics, to evaluate whether the compact or distributed factor structures observed for MOFA2 and GFA remain consistent across regulatory levels.

A further limitation relates to the unsupervised nature of the analysis. While latent factors were interpreted based on variance structure and functional enrichment, no direct external validation was performed. Future work could assess the robustness and generalisability of the inferred factors by projecting them onto independent glioblastoma or brain tumour cohorts, or by evaluating their association with external clinical and molecular annotations, such as tumour subtype, treatment response, or patient survival. Such analyses would help distinguish stable, biologically meaningful patterns from dataset-specific or method-induced structure.

From a methodological perspective, this study highlights that regularisation choices, such as the use of Automatic Relevance Determination in MOFA2, fundamentally shape the latent factor space. Future research could therefore explore sensitivity analyses that systematically vary regularisation strength or prior assumptions to better characterise the trade-off between model stability, sparsity, and sensitivity to weak signals. In addition, combining unsupervised integration with supervised approaches within a unified framework could provide complementary perspectives, balancing hypothesis discovery with predictive relevance.

Finally, extending the benchmarking framework to single-cell or spatial multi-omics data represents an important future direction. Such data would enable the investigation of cell-type-specific regulation and intratumoural heterogeneity, which are not accessible in bulk analyses but are central to understanding glioblastoma biology.

## Discussion

### Comparison of Multi-Omics Factor Structure Captured by MOFA2 and GFA

In this study, MOFA2 and GFA were systematically compared using paired RNA-seq and promoter-level DNA methylation data from TCGA-GBM tumours. Although both methods aim to uncover latent structure in multi-omics data, they rely on fundamentally different Bayesian inference strategies. These differences strongly influenced the dimensionality of the latent space, the degree of regularisation, and the interpretability of the inferred factors (Argelaguet et al., 2018; Virtanen et al., 2015).

MOFA2 applies variational Bayesian inference combined with an ARD prior, which actively downweights latent factors that do not explain stable variation across samples. As expected from the design of the method, this resulted in a compact latent representation in which only a limited number of factors remained active after model convergence. These active factors showed a clear separation into RNA-dominant, methylation-dominant, and shared components, which facilitated biological interpretation and structured downstream analyses. Similar behaviour has been reported in previous MOFA-based studies, where ARD was shown to facilitate interpretability through increased parsimony by concentrating biological signal into a small number of robust latent axes (Argelaguet et al., 2018; Meng et al., 2021).

In contrast, GFA relies on Bayesian inference implemented via Gibbs-sampled MCMC and applies more conservative shrinkage during model fitting (Virtanen et al., 2012; Klami et al., 2015). As a result, latent variation was distributed across a larger number of factors rather than concentrated into a small set of dominant components. This behaviour is consistent with previous descriptions of GFA in high-dimensional settings, where weak but structured sources of variation are preserved rather than suppressed (Virtanen et al., 2015). In the present analysis, GFA captured a broader spectrum of RNA-specific, methylation-specific, and shared factors, but individual factors were less sharply defined than those inferred by MOFA2.

Overall, both methods recovered comparable high-level structure in terms of the relative independence of transcriptomic and promoter-level methylation variation (Sanchez-Vega et al., 2018). MOFA2 favoured a compressed and highly regularised representation, whereas GFA maintained a more diffuse and higher-dimensional latent space.

### Functional Enrichment Behavior and Comparison with Previous Studies

Functional enrichment analyses further highlighted the methodological differences between MOFA2 and GFA. RNA-driven MOFA2 factors were associated with large numbers of enriched GO Biological Process terms, reflecting broad and coordinated transcriptional programmes. This behaviour is well documented for RNA-seq data, where strong global correlation structure and high variance often lead

to enrichment of many overlapping biological processes in GSEA-based analyses (Subramanian et al., 2005; Wang et al., 2009). While such breadth can complicate fine-grained pathway interpretation, it also indicates that MOFA2 effectively concentrates dominant transcriptional signal into a small number of interpretable latent factors.

In contrast, methylation-driven MOFA2 factors exhibited narrower and more focused enrichment profiles. This pattern is consistent with the regulatory role of promoter-level DNA methylation, which typically influences gene expression in a more targeted and context-dependent manner (Bird, 2002; Jones, 2012). The observed differences between RNA- and methylation-derived enrichment profiles therefore reflect fundamental biological distinctions between the two omics layers rather than methodological artefacts.

GFA showed a different enrichment pattern. Only a subset of GFA factors yielded significant GO:BP enrichment, and enrichment signals were generally weaker and more factor-specific. This behaviour reflects the distribution of variation across many latent dimensions, which reduces pathway-level signal concentration within individual factors. Similar observations have been reported in recent comparative studies, where integration methods with larger latent spaces were shown to preserve subtle biological variation at the cost of reduced pathway interpretability (Cantini et al., 2019; Huang et al., 2022). In this context, the limited number of enriched GFA factors observed here represents an inherent methodological trade-off rather than a lack of biological signal.

## **Benchmarking Implications and Methodological Trade-offs**

From a benchmarking perspective, the results indicate that MOFA2 and GFA provide complementary rather than conflicting views of multi-omics structure. Both methods converged on the same overarching biological interpretation: transcriptomic and promoter-level DNA methylation variation in glioblastoma largely represent non-redundant sources of information, with limited evidence for strongly coupled cross-omic latent structure. However, the two frameworks differ substantially in how this structure is prioritised and represented (Brennan et al., 2013; Sanchez-Vega et al., 2018).

MOFA2 prioritises parsimony and interpretability by aggressively suppressing weak latent components during training, making it well suited for analyses aimed at identifying a small number of robust, pathway-rich factors. GFA, in contrast, prioritises completeness and sensitivity to weaker signals, retaining a richer and more distributed latent structure that captures diffuse variation across omics layers. These differences arise directly from the underlying inference strategies of the two models and align with conclusions from recent benchmarking studies, which emphasise that no single integration framework is universally optimal (Hernández-Lemus & Ochoa, 2024).

## **Limitations and Data-related Considerations**

A key limitation of this study relates to data availability and cohort overlap within the TCGA-GBM resource. After systematic harmonisation of TCGA identifiers, it became clear that RNA-seq and DNA methylation datasets did not overlap at the patient level outside a tumour-only cohort. As a result, integrative analyses were restricted to tumour samples with matched RNA-seq and promoter-level DNA methylation measurements.

Although this restriction reduced the number of samples available for integration, it was necessary to ensure biological validity. Integrating unmatched molecular profiles would confound shared and modality-specific signals and undermine interpretation of cross-omic structure. By limiting analyses to paired tumour samples, the inferred latent factors reflect biologically corresponding measurements (Argelaguet et al., 2018).

Additional limitations arise from both methodological and biological factors. The absence of normal samples in the paired dataset precluded modelling tumour–normal contrasts within the integrative framework. Furthermore, the use of bulk RNA-seq and promoter-level DNA methylation data does not

capture cell-type-specific effects, distal regulatory elements, or intratumoural heterogeneity, all of which are known to be important in glioblastoma (Huse & Holland, 2010; Patel et al., 2014).

Despite these limitations, the consistency of findings across two fundamentally different integration frameworks supports the robustness of the main conclusions. The convergence of biological interpretations between MOFA2 and GFA suggests that the inferred latent structure reflects genuine biological variation rather than method-specific artefacts. Together, these results highlight both the constraints imposed by current data availability and the value of applying complementary multi-omics integration approaches when studying complex diseases such as glioblastoma.

## References

- Argelaguet, R., Arnol, D., Bredikhin, D., Deloro, Y., Velten, B., Marioni, J. C., & Stegle, O. (2020). MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome biology*, 21(1), 111. <https://doi.org/10.1186/s13059-020-02015-1>
- Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., Buettner, F., Huber, W., & Stegle, O. (2018). Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular systems biology*, 14(6), e8124. <https://doi.org/10.15252/msb.20178124>
- Aryee, M. J., Jaffe, A. E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A. P., Hansen, K. D., & Irizarry, R. A. (2014). Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics (Oxford, England)*, 30(10), 1363–1369. <https://doi.org/10.1093/bioinformatics/btu049>
- Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, 25(1), 25–29. <https://doi.org/10.1038/75556>
- Bird, A. P. (1980). DNA methylation and the frequency of CpG in animal DNA. *Nucleic acids research*, 8(7), 1499–1504. <https://doi.org/10.1093/nar/8.7.1499>
- Boyle, E. I., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J. M., & Sherlock, G. (2004). GO::TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics (Oxford, England)*, 20(18), 3710–3715. <https://doi.org/10.1093/bioinformatics/bth456>
- Brennan, C. W., Verhaak, R. G. W., McKenna, A., Campos, B., Nounshmehr, H., Salama, S. R., Zheng, S., Chakravarty, D., Sanborn, J. Z., Berman, S. H., Bejan, A. R., et al. (2013). The somatic genomic landscape of glioblastoma. *Cell*, 155(2), 462–477. <https://doi.org/10.1016/j.cell.2013.09.034>
- Chen, Y. A., Lemire, M., Choufani, S., Butcher, D. T., Grafodatskaya, D., Zanke, B. W., Gallinger, S., Hudson, T. J., & Weksberg, R. (2013). Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics*, 8(2), 203–209. <https://doi.org/10.4161/epi.23470>
- Comendul, A., Ruf-Zamojski, F., Ford, C. T., Agarwal, P., Zaslavsky, E., Nudelman, G., Hariharan, M., Rubenstein, A., Pincas, H., Nair, V. D., Michaelas, A. M., Fremont-Smith, P. D., Ricke, D. O., Sealfon, S. C., Woods, C. W., Claypool, K. T., & Jaimes, R. III. (2025). *Comprehensive guide for epigenetics and transcriptomics data quality control*. *iScience Protocols*, 6(1), 103607. <https://doi.org/10.1016/j.xpro.2025.103607>
- Deaton, A. M., & Bird, A. (2011). CpG islands and the regulation of transcription. *Genes & development*, 25(10), 1010–1022. <https://doi.org/10.1101/gad.2037511>
- Du, P., Zhang, X., Huang, C. C., Jafari, N., Kibbe, W. A., Hou, L., & Lin, S. M. (2010). Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC bioinformatics*, 11, 587. <https://doi.org/10.1186/1471-2105-11-587>
- Ehrlich, M., & Wang, R. Y. (1981). 5-Methylcytosine in eukaryotic DNA. *Science (New York, N.Y.)*, 212(4501), 1350–1357. <https://doi.org/10.1126/science.6262918>

Emam, M., Tarek, A., Soudy, M. et al. Comparative evaluation of multiomics integration tools for the study of prediabetes: insights into the earliest stages of type 2 diabetes mellitus. *Netw Model Anal Health Inform Bioinforma* 13, 8 (2024). <https://doi.org/10.1007/s13721-024-00442-9>

EXPLAINED (JIVE) FOR INTEGRATED ANALYSIS OF MULTIPLE DATA TYPES. *The annals of applied statistics*, 7(1), 523–542. <https://doi.org/10.1214/12-AOAS597>

Fortin, JP., Labbe, A., Lemire, M. et al. Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome Biol* 15, 503 (2014). <https://doi.org/10.1186/s13059-014-0503-2>

Gene Ontology Consortium (2021). The Gene Ontology resource: enriching a GOld mine. *Nucleic acids research*, 49(D1), D325–D334. <https://doi.org/10.1093/nar/gkaa1113>

Hasin, Y., Seldin, M., & Lusis, A. (2017). Multi-omics approaches to disease. *Genome biology*, 18(1), 83. <https://doi.org/10.1186/s13059-017-1215-1>

Hernández-Lemus, E., & Ochoa, S. (2024). Methods for multi-omic data integration in cancer research. *Frontiers in genetics*, 15, 1425456. <https://doi.org/10.3389/fgene.2024.1425456>

Hoadley, K. A., Yau, C., Hinoue, T., Wolf, D. M., Lazar, A. J., Drill, E., Shen, R., Taylor, A. M., Cherniack, A. D., Thorsson, V., Akbani, R., Bowlby, R., Wong, C. K., Wiznerowicz, M., Sanchez-Vega, F., Robertson, A. G., Schneider, B. G., Lawrence, M. S., Noushmehr, H., Malta, T. M., ... Laird, P. W. (2018). Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell*, 173(2), 291–304.e6. <https://doi.org/10.1016/j.cell.2018.03.022>

Huse, J. T., & Holland, E. C. (2010). Genetic alterations and pathogenesis of glioblastoma. *Annual Review of Medicine*, 61, 397–406. <https://doi.org/10.1146/annurev.med.050108.151650>

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 374(2065), 20150202. <https://doi.org/10.1098/rsta.2015.0202>

Jones P. A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature reviews. Genetics*, 13(7), 484–492. <https://doi.org/10.1038/nrg3230>

Khatri, P., Sirota, M., & Butte, A. J. (2012). Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS computational biology*, 8(2), e1002375. <https://doi.org/10.1371/journal.pcbi.1002375>

Klami, A., Virtanen, S., Leppäaho, E., & Kaski, S. (2014). Group Factor Analysis [Preprint, accepted version]. Imperial College London, Spiral Institutional Repository. <https://spiral.imperial.ac.uk/server/api/core/bitstreams/0c0e996a-81c0-4a68-a106-5272471e46de/content>

Klami, A., Virtanen, S., Leppäaho, E., & Kaski, S. (2015). Group factor analysis. *IEEE Transactions on Neural Networks and Learning Systems*, 26(9), 2136–2147. <https://doi.org/10.1109/TNNLS.2014.2376974>

Lehne, B., Drong, A. W., Loh, M., Zhang, W., Scott, W. R., Tan, S. T., Afzal, U., Scott, J., Jarvelin, M. R., Elliott, P., McCarthy, M. I., Kooner, J. S., & Chambers, J. C. (2015). A coherent approach for analysis of the Illumina HumanMethylation450 BeadChip improves data quality and performance in epigenome-wide association studies. *Genome biology*, 16(1), 37. <https://doi.org/10.1186/s13059-015-0600-x>

- Lock, E. F., Hoadley, K. A., Marron, J. S., & Nobel, A. B. (2013). JOINT AND INDIVIDUAL VARIATION EXPLAINED (JIVE) FOR INTEGRATED ANALYSIS OF MULTIPLE DATA TYPES. *The annals of applied statistics*, 7(1), 523–542. <https://doi.org/10.1214/12-AOAS597>
- Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12), 550. <https://doi.org/10.1186/s13059-014-0550-8>
- Mo, Q., Wang, S., Seshan, V. E., Olshen, A. B., Schultz, N., Sander, C., Powers, R. S., Ladanyi, M., & Shen, R. (2013). Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proceedings of the National Academy of Sciences of the United States of America*, 110(11), 4245–4250. <https://doi.org/10.1073/pnas.1208949110>
- Pagès, H., Carlson, M., Falcon, S., & Li, N. (2024). AnnotationDbi: Manipulation of SQLite-based annotations in Bioconductor. <https://bioconductor.org/packages/AnnotationDbi>
- R. P. McDonald, “Three common factor models for groups of variables,” *Psychometrika*, vol. 35, no. 1, pp. 111–128, 1970.
- Rohart F, Gautier B, Singh A, Lê Cao K-A (2017) mixOmics: An R package for ‘omics feature selection and multiple data integration. *PLoS Comput Biol* 13(11): e1005752. <https://doi.org/10.1371/journal.pcbi.1005752>
- Rosenberg, S., Verreault, M., Schmitt, C., Guegan, J., Guehenec, J., Levasseur, C., Marie, Y., Bielle, F., Mokhtari, K., Hoang-Xuan, K., Ligon, K., Sanson, M., Delattre, J. Y., & Idbaih, A. (2017). Multi-omics analysis of primary glioblastoma cell lines shows recapitulation of pivotal molecular features of parental tumors. *Neuro-oncology*, 19(2), 219–228. <https://doi.org/10.1093/neuonc/now160>
- Sanchez-Vega, F., Mina, M., Armenia, J., Chatila, W. K., Luna, A., La, K. C., Dimitriadoy, S., Liu, D. L., Kantheti, H. S., Saghafinia, S., Chakravarty, D., Daian, F., Gao, Q., Bailey, M. H., Liang, W. W., Foltz, S. M., Shmulevich, I., Ding, L., Heins, Z., Ochoa, A., ... Schultz, N. (2018). Oncogenic Signaling Pathways in The Cancer Genome Atlas. *Cell*, 173(2), 321–337.e10. <https://doi.org/10.1016/j.cell.2018.03.035>
- Stupp, R., Mason, W. P., van den Bent, M. J., Weller, M., Fisher, B., Taphoorn, M. J. B., ... Mirimanoff, R. O. (2005). Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *New England Journal of Medicine*, 352(10), 987–996. <https://doi.org/10.1056/NEJMoa043330>
- Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., & Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43), 15545–15550. <https://doi.org/10.1073/pnas.0506580102>
- Teschendorff, A. E., & Relton, C. L. (2018). Statistical and integrative system-level analysis of DNA methylation data. *Nature reviews. Genetics*, 19(3), 129–147. <https://doi.org/10.1038/nrg.2017.86>
- Vahabi, N., & Michailidis, G. (2022). Unsupervised multi-omics data integration methods: A comprehensive review. *Frontiers in Genetics*, 13, 854752. <https://doi.org/10.3389/fgene.2022.854752>
- Verhaak, R. G., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M. D., Miller, C. R., Ding, L., Golub, T., Mesirov, J. P., Alexe, G., Lawrence, M., O'Kelly, M., Tamayo, P., Weir, B. A., Gabriel, S., Winckler, W., Gupta, S., Jakkula, L., Feiler, H. S., ... Cancer Genome Atlas Research Network

(2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer cell*, 17(1), 98–110. <https://doi.org/10.1016/j.ccr.2009.12.020>

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics*, 10(1), 57–63. <https://doi.org/10.1038/nrg2484>

Weinhold B. (2006). Epigenetics: the science of change. *Environmental health perspectives*, 114(3), A160–A167. <https://doi.org/10.1289/ehp.114-a160>

Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS: A Journal of Integrative Biology*. 2012;16(5):284-287. doi:[10.1089/omi.2011.0118](https://doi.org/10.1089/omi.2011.0118)