universität wien

MASTERARBEIT | MASTER'S THESIS

Titel | Title

Evaluation of Methods for the Identification of Gene Co-Expression Networks in Malignant Pleural Mesothelioma

verfasst von | submitted by

Dipl.-Ing. Markus Riedl BSc

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2026

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 875
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Bioinformatik
Betreut von | Supervisor:	Univ.-Prof. Dipl.-Phys. Dr. Ivo Hofacker

Acknowledgements

I would like to express my sincere gratitude to Ivo Hofacker, Thomas Mohr and Walter Berger for their excellent supervision and continuous support throughout this Master's thesis. I greatly appreciated their valuable input and constructive feedback whenever I consulted them.

I also wish to thank all members of the Berger and Heffeter groups at the Institute of Cancer Research of the Medical University of Vienna for providing a warm, welcoming and inspiring work environment.

My deepest thanks go to my girlfriend Eva for her unwavering support, patience and encouragement throughout my studies. Finally, I am profoundly grateful to my family, and especially my parents. Without their constant support and encouragement, this accomplishment would not have been possible.

Abstract

Malignant Pleural Mesothelioma (MPM) is a rare but aggressive cancer arising from the mesothelial cells lining the pleural cavity. It is characterised by poor prognosis, resistance to conventional therapies, and a complex molecular landscape that remains incompletely understood. To better elucidate the transcriptional patterns underlying MPM, this thesis applies gene co-expression network analysis as a means to identify biologically meaningful gene modules associated with clinical traits. Two module detection methods—Independent Component Analysis (ICA) and Weighted Gene Correlation Network Analysis (WGCNA)—were systematically evaluated.

Following parameter optimisation using known regulatory modules and biological pathways, both methods identified gene clusters with significant correlations with clinical traits. However, WGCNA outperformed ICA, producing modules with higher module membership, greater biological homogeneity, and more specific functional enrichment, as demonstrated through analysis of Gene Ontology (GO) terms and WikiPathways. Notably, significant functional enrichments within gene modules included associations with the Hippo pathway, cell cycle control and DNA replication as well as the extracellular matrix.

These findings establish gene co-expression network analysis as a valuable approach for uncovering key transcriptional patterns in MPM. More specifically, they highlight WGCNA as a robust and biologically interpretable method for gene co-expression analysis in MPM, while underscoring the importance of systematic parameter assessment in machine learning-based network analysis methods.

Zusammenfassung

Das maligne Pleuralmesotheliom (MPM) ist eine seltene, aber aggressive Krebsart, welche durch schlechte Prognose und Resistenzen gegenüber herkömmlichen Therapien charakterisiert ist. Um die den Transkriptionsmustern von MPM zugrunde liegenden Mechanismen besser zu verstehen, wird in dieser Masterarbeit die Gen-Koexpressionsnetzwerkanalyse eingesetzt, um biologisch relevante Gen-Module zu identifizieren, die mit klinischen Merkmalen assoziiert sind. Zwei Methoden zur Identifikation von Modulen—ICA und WGCNA—wurden systematisch untersucht.

Nach einer Parameteroptimierung anhand bekannter regulatorischer Module und biologischer Signalwege konnten mit beiden Methoden Gen-Cluster mit signifikanten Korrelationen zu klinischen Merkmalen identifiziert werden. WGCNA erzielte hierbei bessere Ergebnisse als ICA in der Modulzugehörigkeit der einzelnen Gene, in der biologischen Homogenität sowie in der spezifischeren funktionellen Charakterisierung der Module in Genontologie und WikiPathways Gensets. Besonders hervorzuheben sind funktionelle Assoziationen von Gen-Modulen mit dem Hippo-Signalweg, der Zellzykluskontrolle und DNA-Replikation sowie der extrazellulären Matrix.

Diese Ergebnisse unterstreichen den Wert der Gen-Koexpressionsnetzwerkanalyse als wertvollen Ansatz zur Aufdeckung zentraler Transkriptionsmuster bei MPM. Insbesondere wird WGCNA als robuste und biologisch interpretierbare Methode zur Gen-Koexpressionsnetzwerkanalyse bei MPM hervorgehoben. Darüber hinaus unterstreicht die Arbeit die Bedeutung einer systematischen Parameterbewertung bei maschinellen Lernverfahren für Netzwerkanalysen.

Contents

List of Figures

List of Tables

Abbreviations

BHI Biological Homogeneity Index

EMT Epithelial-to-Mesenchymal Transition

GO Gene Ontology

ICA Independent Component Analysis

MPM Malignant Pleural Mesothelioma

ORA Over-Representation Analysis

ROS Reactive Oxygen Species

WGCNA Weighted Gene Correlation Network Analysis

1 Introduction

1.1 Cancer

Tumours are cell masses which show aberrant and excessive growth and proliferation caused by genetic and epigenetic alterations. Benign tumours remain *in situ* and often do not pose any health threat to the affected person. Malignant tumours are termed cancer and gain the capability to invade adjacent tissue and migrate to other parts of the body in a process called metastasis. Cancer is a major global health challenge, causing almost 18 % of all deaths in 2019, making it the second largest cause of death after cardiovascular diseases [1].

Cancers can be classified according to the tissue or cell type from which they arise. The most common forms of cancer are carcinomas, accounting for approximately 80 % of all cancers and occurring from epithelial cells. More rarely occurring are sarcomas, which develop from connective tissue or muscle cells. Lymphomas and leukaemias, on the other hand, derive from hematopoietic and lymphatic tissue [2].

The morbidity and mortality of different types of cancer can vary. While breast cancer has the highest incidence, with 11.7 % of all cancer localisations, it is associated with only 6.9 % of all cancer-related deaths. Lung cancer, on the other hand, has an incidence of 11.4 % and mortality of 18.0 %, which makes it the deadliest cancer by absolute numbers [3].

1.1.1 Hallmarks of Cancer

Cancer cells distinguish themselves from normal cells in a variety of characteristics that were summarised and described by Douglas Hanahan and Robert A. Weinberg in the Hallmarks of Cancer, consisting of 10 core features that are regularly found in cancers of all types.

The original set of six hallmarks was published in 2000 [4], followed by the addition of two hallmarks as well as two enabling characteristics in 2011 [5].

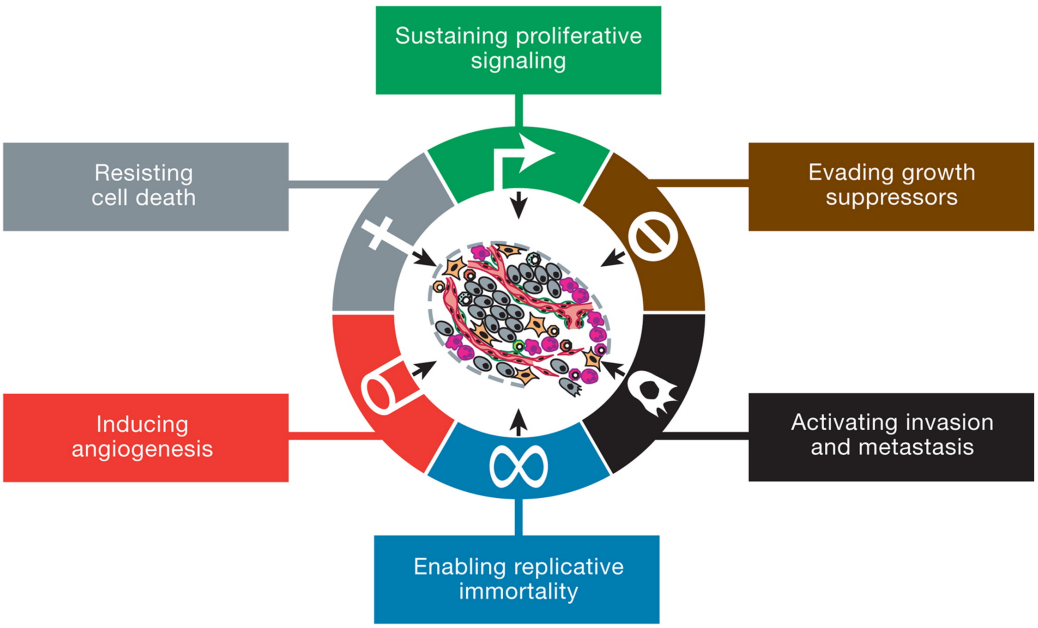

Figure 1.1.: The six Hallmarks of Cancer as initially defined by Hanahan and Weinberg in 2000 [4].

Sustained growth signaling

One key characteristic of cancer is the increased proliferation, which is, in large part, caused by sustained growth signalling. Cancer cells are often provided with elevated levels of growth-promoting ligands by autocrine signalling [6] or by stimulating cells in the tumour-associated stroma [7]. Additionally, cancer growth can be deregulated by mutations occurring in genes that encode for proteins that are part of pathways downstream of the growth factors and their receptors. Such mutation events often affect key members of signalling pathways such as B-Raf, PI3K, PTEN or the Ras protein family [8–10].

Evading growth suppressors

Apart from the reinforcement of growth signalling, cancer cells also frequently evade growth-suppression from important tumour suppressor genes such as *TP53* and *RB1*. The protein encoded by the latter, retinoblastoma-associated protein (pRb), is a crucial gatekeeper for the progression of the cell cycle, particularly for the G_1/S phase transition. In a hypophos-

phorylated state, the protein blocks E2F transcription factors that govern the expression of proteins associated with G_1/S transition [11]. Loss or deregulated phosphorylation of pRb leads to the inactivation of the G_1 checkpoint and potentially unobstructed tumour growth [12].

Resisting cell death

Programmed cell death, apoptosis, is a fate shared by all normal cells and serves as a barrier for cancer development. The highly organised and regulated programme consists of regulators that react to internal and external signals, as well as effectors. A balance of pro- and antiapoptotic agents, commonly described as the "apoptotic trigger", is largely made up of members of the Bcl-2 protein family [13]. One of the proapoptotic members, BAX2, can be upregulated by p53, the protein encoded by *TP53*, as part of the DNA damage response. Cancer cells often develop mechanisms to bypass these death signals, such as by overexpressing antiapoptotic proteins (like Bcl-2), downregulating pro-apoptotic factors (like BAX), or mutating key regulators such as *TP53*. This resistance to cell death allows malignant cells to survive longer than normal, accumulate additional mutations, and become more aggressive and resistant to treatment, ultimately contributing to tumour growth and therapy failure.

Enabling replicative immortality

Normal cells are limited in their replicative potential. The Hayflick limit describes the natural tendency to enter senescence, the depletion of growth, after approximately 50 doublings of a cell population [14]. The telomeres, repetitive terminal segments of linear chromosomes, are centrally involved in the enforcement of this limit. These structures undergo shortening upon every replicative event due to the unidirectional nature of the DNA polymerase, which causes the so-called end replication problem. The shortened telomere structures are recognised as DNA damage and trigger the appropriate response, including cell cycle arrest. Cancer cells often exhibit the hallmark of replicative immortality, in which this mechanism fails due to the activity of telomerase This specialized reverse transcriptase adds repeat segments to the ends of telomeric DNA [15].

Angiogenesis

The formation of new blood vessels includes the generation of endothelial cells, their assembly during vasculogenesis and the sprouting of new vessels in angiogenesis. This process is active

throughout embryogenesis and becomes largely quiescent afterwards [16]. Activation of angiogenesis occurs only transiently in adults in the context of wound healing and the female reproductive cycle. Tumour progression, however, often involves the constitutive activation of angiogenesis in order to provide cancer cells with the necessary nutrients and oxygen. The angiogenic switch is regulated by activating and inhibiting proteins such as vascular endothelial growth factor A (VEGFA) and thrombospondin 1 (TBHS1) [17].

Invasion and metastasis

Cell-to-cell and cell-to-extracellular-matrix adhesion is provided by cell adhesion molecules (CAMs). In most cancers, the function of the CAM and tumour suppressor E-cadherin (CDH1) is lost, thereby paving the way for migration of cancer cells to distant parts of the body and the formation of metastases. Metastasis is preceded by local invasion of the tumour's surrounding tissue and the intravasation of cells into blood and lymphatic vessels. After transit to other parts of the body, the cancer cells may escape those vessels again and settle in the parenchyma of distant tissue. The Epithelial-to-Mesenchymal Transition (EMT) is a critical event leading to this hallmark. In this reversible process, epithelial cells lose their polarity as well as adhesion and exhibit a migratory, mesenchymal phenotype.

Genome Instability and Mutation

The first enabling characteristic described by Hanahan and Weinberg describes the elevated mutation rates and genomic instability. Tumour progression is accompanied by the successive accumulation of genetic changes in the form of mutations. While normal cells possess a very low mutation rate, cancer cells, on the other hand, often exhibit enhanced rates of mutation [18]. They are often more susceptible to mutational agents and have dysfunctional DNA maintenance and damage repair systems [19]. Most notable and already mentioned previously is the central role of TP53 in this regard. Another role in tumour progression is played by heritable epigenetic alterations by histone modification and DNA methylation that modulate the accessibility to, and expression of, protein-coding segments of the genome [20].

Tumour-promoting inflammation

The role of the immune system in cancer is ambiguous. It is involved in both anti-tumour response as well as in the promotion of tumour progression and the acquisition of hallmark capabilities by inflammation. Various immune cells are part of the tumour micro environment

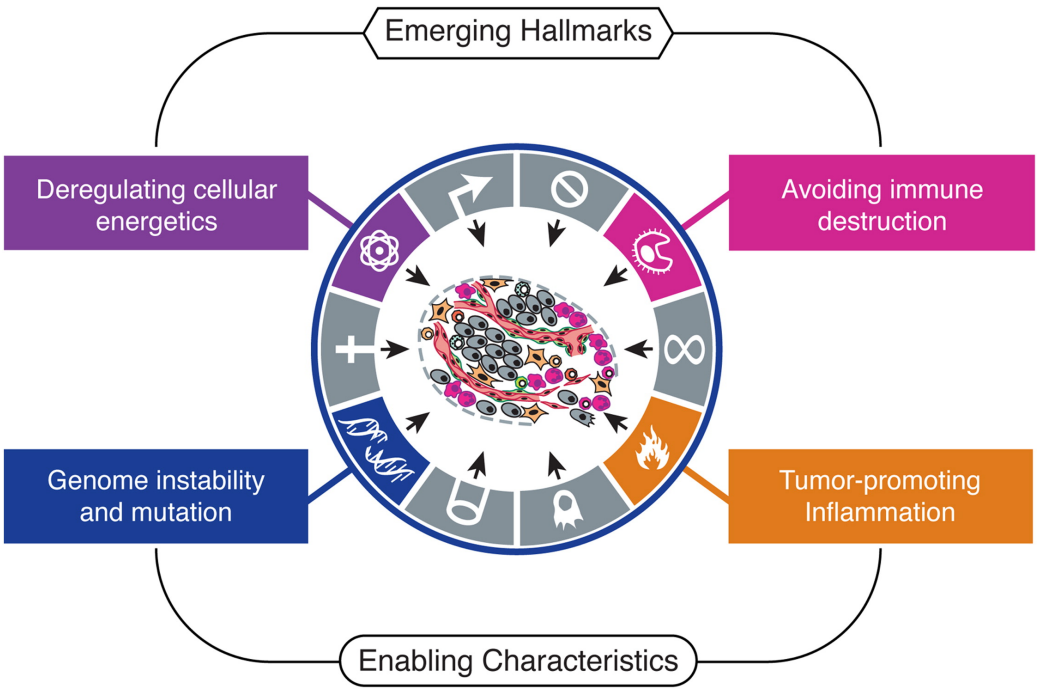

Figure 1.2.: Emerging Hallmarks and Enabling Characteristics as defined by Hanahan and Weinberg in 2011 [5].

and release a variety of molecules such as growth factors, survival factors, proangiogenic agents and inductive signals. These may further reinforce as hallmark capabilities such as growth signalling, evading cell death, angiogenesis, invasion and metastasis described before. Additionally, mutagenic agents, particularly Reactive Oxygen Species (ROS), can accelerate tumour progression.

Reprogramming energy metabolism

The observation that cancer cells often shift their metabolism to only anaerobic glycolysis has already been made almost 100 years ago by Otto Warburg [21]. Cancer cells need to compensate for the significantly lower efficiency of glycolysis and do this by increasing glucose uptake via overexpression of glucose transporters such as GLUT1.

One driver for this metabolic shift is hypoxia, the undersupply of oxygen in the tumour tissue. The response to these circumstances includes the pleiotropic upregulation of glucose transporters [22]. Interestingly, this upregulation and the suppression of oxidative phos-

phorylation can also be linked to the activation of oncogenes such as *RAS* and *MYC* [23, 24].

Evading immune destruction

The immune system fulfils an important protective mechanism against tumours by recognising presented tumour antigens and subsequently eliminating the cell. However, this only affects immunologically vulnerable cancer cells and thereby augments selective pressure towards reduced immunogenicity. Eventually, cancer cells may exert influence on the immune system's anti-tumoural activity by suppressing immunological response or by loss of antigen expression [25]. One such way by which tumour cells modulate the immune system is by producing transforming growth factor beta (TGF-β), aiding the conversion of naïve $CD4^+$ T cells to regulatory T cells, which are actively immunosuppressive.

1.2 Malignant Pleural Mesothelioma

The mesothelium is composed of a slim layer of mesoderm-derived mesothelial cells and covers several body cavities such as the thoracic cavity with the pleura, the abdominal cavity with the peritoneum and the pericardial cavity with the pericardium. Mesothelial cells are specialised in secreting a lubricating fluid, providing a frictionless layer to enable movement of adjacent organs. They are furthermore involved in the transport of solutes and cells across serous membranes, presentation of antigens and inflammation [26]. A peculiarity of these cells is the expression of both mesenchymal and epithelial intermediate filaments. Mesothelial cells also show capabilities of transdifferentiation, comparable to changes seen in EMT, with implications in malignant diseases [27, 28].

Malignant mesothelioma is a rare, aggressive cancer that arises from the mesothelium, most often originating in the pleura. Mesothelial cells retain pluripotency and can therefore give rise to different histologies. Malignant Pleural Mesothelioma (MPM) is distinguished into three histological subtypes: epithelioid, sarcomatoid and biphasic [29]. Epithelioid mesothelioma cells can be recognised with a roundish morphology, while the sarcomatoid subtype is oval or spindle-like, as seen in Figure 1.3. The biphasic subtype is a mix of the two, whereas the proportions might differ from case to case. The median survival across all histological types is 12 months [30]. However, prognosis is closely linked to histological subtype, with median

survival times of approximately of 14, 10 or 4 for the epithelioid, biphasic and sarcomatoid forms, respectively [31].

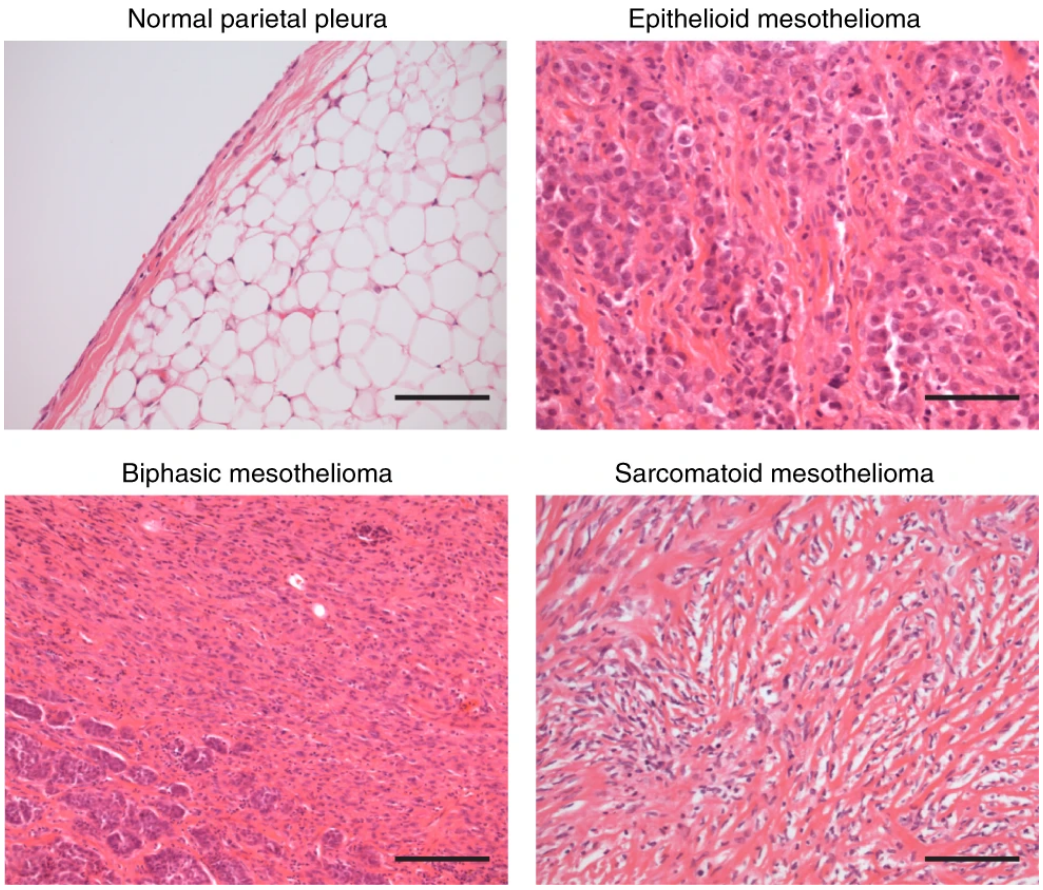

Figure 1.3.: Normal pleura and the three histological subtypes of malignant mesothelioma. Source: [32].

The rarity and histological heterogeneity of the disease make it difficult to correctly diagnose and distinguish mesothelioma from metastases of other cancers or benign mesothelial hyperplasia. Patients with MPM most commonly seek medical attention with symptoms such as shortness of breath, dry cough and chest pain [33]. Many of these symptoms are related to pleural effusion, leading to investigations with X-ray and computer tomography (CT) as well as cytological evaluation of the fluid, including immunohistochemistry. A biopsy from the thoracic cavity is finally needed to confirm the diagnosis. Early diagnosis is critical for mesothelioma since delayed diagnosis will inevitably lead to tumour progression [33].

1.2.1 Risk factors and carcinogenesis

The majority of cases of malignant pleural mesothelioma are linked to occupational asbestos exposure. Additionally, several lines of evidence suggest that susceptibility to this cancer is heritable through germline mutations of certain genes, most notably *BAP1*.

Exposure to asbestos

Asbestos is a term describing a group of fibrous silicate minerals that were used commercially. Most countries have banned the use of asbestos, while in others it remains in use.

Early hypotheses proposed that asbestos fibres physically interfere with the mitotic spindle, leading to extensive chromosomal damage and alterations [34]. Later, however, it was shown that affected cells were not immortalised in this process but rather died within a few days [35]. Meanwhile, numerous studies indicated that carcinogenesis is primarily attributable to chronic inflammation caused by the deposition and long-term presence of asbestos fibres [36]. Phagocytosis of these mineral fibres by macrophages cause the release of cytokines like Interleukin (IL)-1β and tumour necrosis factor (TNF)-α, as well as ROS, all of which have been linked to asbestos-induced carcinogenesis [35, 37, 38]. Cell necrosis caused by asbestos also leads to the release of high-mobility group protein B1 (HMGB1), which is a damage-associated molecular pattern (DAMP) and further reinforces the inflammatory response [39].

Mutation of *BAP1*

BRCA1 associated protein-1 (BAP1) is a deubiquitylase that is involved in the regulation of cell pathways, including cell cycle control, cell differentiation, cell death and DNA damage response. Germline mutations in the *BAP1* gene are associated to markedly increased incidences of several cancers, including mesothelioma [40] and melanocytic tumours [41].

The critical role of BAP1 as tumour suppressor gene is reflected in some of its molecular interactions. BAP1 modulates the stability of ITPR3, a channel protein releasing Ca^{2+} ions from the endoplasmic reticulum into the cytoplasm, from where it ultimately gets into the mitochondria. There, Ca^{2+} is involved in oxidative phosphorylation and, at higher levels, apoptosis [42]. Apart from the role of *BAP1* mutations in the reinforcement of the Warburg effect and apoptosis, they also impair ferroptosis—another possible mode of programmed cell death besides apoptosis.

The detection of *BAP1* germline mutations can help with an early response to malignant mesothelioma and lead to some of the very few cases of patients who remain tumour-free for 10 years postsurgically [43]. Thus, screening of *BAP1* and other genes involved in DNA repair and tumour suppression is important in families with a history of mesothelioma.

1.3 Analysis of gene co-expression

Gene co-expression network analysis is an essential analysis method for genome-wide gene expression data. It aims to find and define groups of genes with similar expression profiles. Such genes tend to be regulated by common transcription factors and related in terms of function. Gene co-expression network analysis therefore, allows for extensive insight and discovery of gene functions and transcriptional regulatory programmes.

Gene co-expression analysis is a widely used approach to determine potential functional roles of poorly annotated genes by applying the so-called guilt-by-association principle. However, such associations should be interpreted with caution since there is likely only a small core of approximately $< 20\,\%$ of genes relevant for a module's main biological functions [44].

Transcriptional regulation is highly dynamic and often specific to context [45]. Differential co-expression network analysis enables the examination of changes in co-expression patterns under varying conditions such as disease states [46], tissue types [47], and developmental stages [48]. Differences in co-expression are assessed by the presence or absence of the modules, their change in structure and by rearrangements of genes between modules [49].

The identification of gene modules is a central task in any co-expression network analysis and most commonly accomplished by clustering. However, several approaches were developed and are implemented in a variety of algorithms.

1.3.1 Weighted Correlation Network Analysis

The weighted correlation network analysis, also known as Weighted Gene Correlation Network Analysis (WGCNA), by Langfelder and Horvath [50] is the most widely applied gene co-expression network analysis method and considered state-of-the-art for this type of ana-

lysis [49]. It constructs an undirected, weighted network of all genes and expresses the strength of connection by a measure of their correlation of gene expression [51].

A network in WGCNA is specified as an $n \times n$ adjacency matrix for n genes, with entries a_{ij} in $[0, 1]$ denoting the connection strength or correlation measure between genes.

Firstly, in WGCNA, a co-expression similarity measure s_{ij} is computed and will by default use the absolute value of the correlation between genes i and j:

$$s_{ij} = \left|\mathrm{cor}(x_i, x_j)\right| \tag{1.1}$$

Here, x_i and x_j denote the gene expressions of i and j, respectively. This unsigned co-expression similarity measure obscures the distinction between normal and reciprocal correlations in gene expression. Alternatively, the signed co-expression similarity transforms the correlation value in order to consider the sign of the correlation:

$$s_{ij} = 0.5 + 0.5 \cdot \mathrm{cor}(x_i, x_j) \tag{1.2}$$

Scale-free networks

Most biological networks can be modelled accurately with a scale-free topology [52, 53]. The distribution of degrees in such a network follows a power law:

$$P(k) \sim k^{-\gamma}$$

In gene regulatory networks, this implies that a small subset of genes—most likely transcription factors—is highly connected, whereas the majority of genes exhibits low connectivity.

A scale-free network shows properties of robustness against random errors, however, high vulnerability against coordinated, targeted attacks [54]. This property makes such gene co-expression networks highly interesting for identifying hub genes involved in particular diseases that might ultimately lead to new potential drug targets.

In WGCNA, the adjacency a_{ij} is then given by computing

$$a_{ij} = {s_{ij}}^{\beta} \tag{1.3}$$

with β being chosen appropriately to converge towards scale-free properties. Figure 1.4 illustrates signed and unsigned co-expression similarity measures along with their corresponding powers.

Connectivity and Topological Overlap

While the scale-free topology accurately models the degree distribution, it fails to adequately capture the modularity of biological networks, such as gene co-expression networks. The topological overlap is a measures of pairwise connectedness among nodes of a network and was found to be useful by better reflecting this modularity [53]. The Topological Overlap Matrix (TOM) with entries t_{ij} can be computed by

$$t_{ij} = \frac{\sum_{u \notin \{i,j\}} a_{iu} a_{uj} + a_{ij}}{\min\{k_i, k_j\} + 1 - a_{ij}}$$

where k_i is the connectivity of the respective genes and is defined by

$$k_i = \sum_{u \neq i} a_{iu}$$

This produces a nonnegative, symmetric similarity measure, which can be converted into a dissimilarity measure d^t_{ij} via

$$d^t_{ij} = 1 - t_{ij}$$

This dissimilarity or distance measure is subsequently used to cluster the genes using agglomerative average-linkage hierarchical clustering. From the resulting dendrogram, gene modules can be defined by using a constant-height cutoff value or one of two dynamic tree cutoff algorithms provided by the R package Dynamic Tree Cut [55].

Module eigengenes

After clustering, the first principal component of a module's gene expressions—termed the module eigengene (ME)—represents the weighted average expression profile of the module [56]. This provides a concise representation of the module, which can be correlated with each gene's expression profile to yield a continuous measure of module membership. Additionally, module eigengenes can be correlated with external data, such as clinical variables.

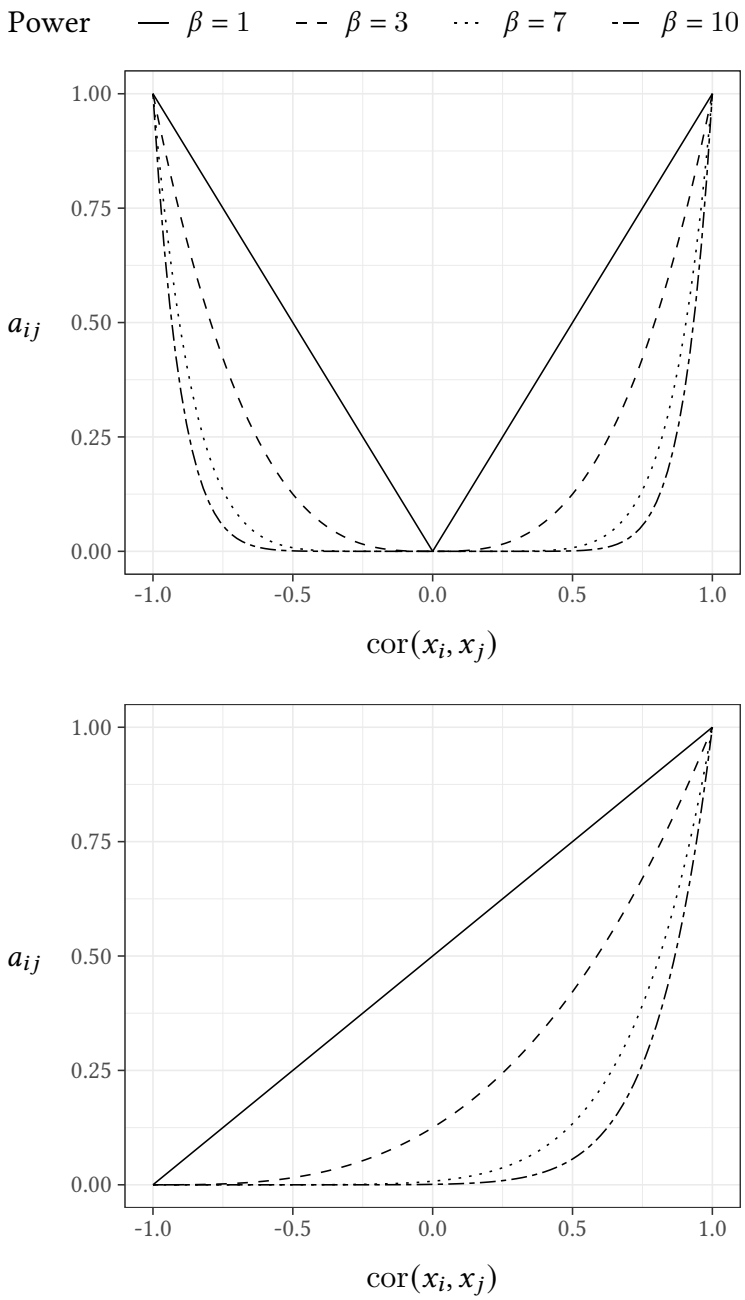

Figure 1.4.: Unsigned (top) and signed (bottom) correlations as co-expression similarity measures, corresponding to equations 1.1 and 1.2, respectively. Dashed lines show the effect of different values for β in equation 1.3.

1.3.2 Independent Component Analysis

Independent Component Analysis (ICA) is a method that decomposes an observed signal into n components by maximising their statistical independence. The assumptions underlying ICA is the statistical independence of the source signal as well as its non-normal origin. A fundamental concept applied in this method is the central limit theorem, which states that linear combinations of independent non-Gaussian sources are more Gaussian than their original variables. Several measures of non-Gaussianity are possible, such as kurtosis, negentropy and mutual information.

The observed expression of the j-th gene in the i-th sample x_{ij} is assumed to be a linear combination of n components:

$$x_{ij} = a_{i1}s_{1j} + a_{i2}s_{2j} + \cdots + a_{in}s_{nj}$$

or, in matrix form, as illustrated in figure 1.5. The matrix A, commonly referred to as mixing matrix, can be interpreted as the activation level of each component in the samples or individuals. The matrix S is the source matrix and describes the affiliation of genes to components [57]. The values for the source matrix can be estimated by seeking the unmixing matrix $\mathrm{W} \in \mathbb{R}^{n \times i}$ and maximizing the non-gaussianity of

$$\mathrm{S} = \mathrm{W} \cdot \mathrm{X}$$

An important property of ICA is that these affiliations of genes to components—and, by extension, to co-expression modules—are not exclusive; the sets of genes may overlap. In contrast, many other co-expression network analysis methods such as WGCNA, partition genes into mutually exclusive sets. Allowing overlap in co-expression modules can enhance the biological relevance of the analysis, as evidence suggests that gene regulation is highly combinatorial and context-specific [58].

ICA finds broad application in biomedical signal processing, most notably for electroencephalogram (EEG) and magnetic resonance tomography (MRT).

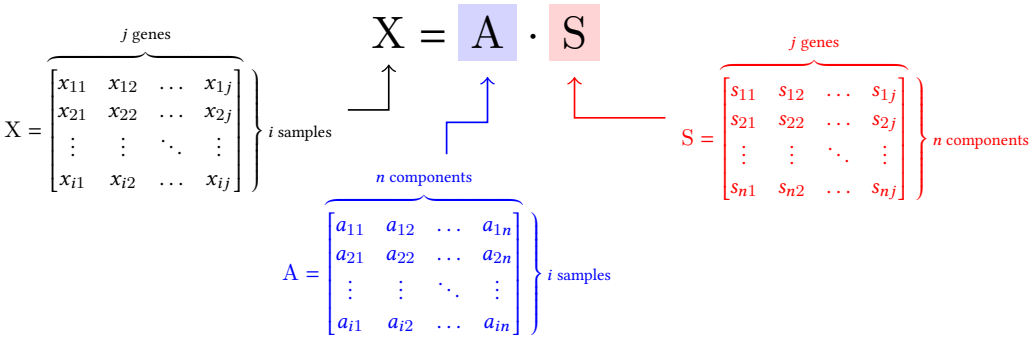

Figure 1.5.: Matrix form of the decomposition of the expression matrix X into the source signal S and the mixing matrix A in ICA.

2 Materials & Methods

2.1 Overview

An overview of the workflow can be seen in Figure 2.1. All analyses were carried out using the statistical software R version 4.1.1.

2.2 Known modules

2.2.1 Regulatory circuits

The 'regulatory circuits' data set by Marbach et al. [59] contains a comprehensive set of gene-regulatory networks for 394 human cell types and tissues. In essence, the data set contains weighted directed networks with edges from regulators to respective gene targets. The regulatory circuit of mesothelioma was loaded into R, and known modules were defined as the sets of gene targets of each regulator with an edge weight greater than a cutoff value ($a_{ij} > \text{cutoff}$). The cutoff was used as a parameter and varied:

$$\texttt{cutoff} \in \{0.1, 0.2, 0.5, 0.01, 0.02, 0.05, 0.001, 0.002, 0.005\}$$

Additionally, one set of known modules was constructed using no cutoff. Known modules were filtered to contain ≥ 5 genes. Subsequently, pairs of modules with Jaccard index of > 0.8 were iteratively merged until no highly similar module pairs remained.

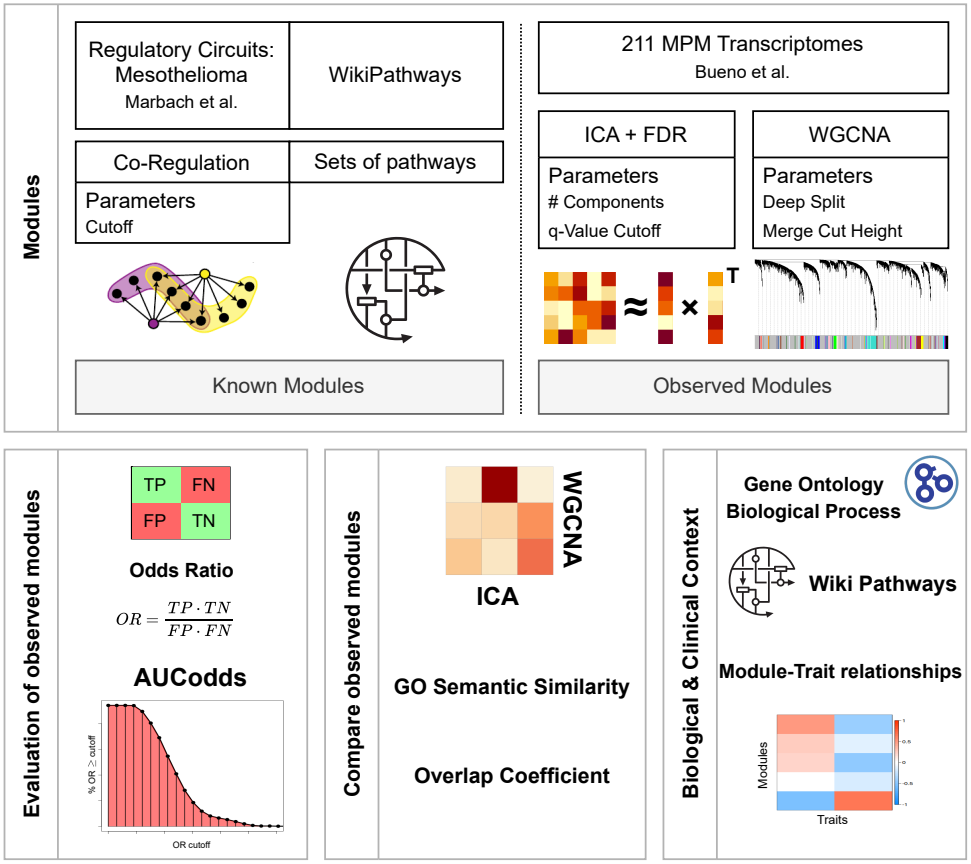

Figure 2.1.: Overview of the data and analyses conducted. Observed modules were assessed against known reference modules from two sources using the aucodds approach. Modules identified through ICA and WGCNA were compared across multiple characteristics. Finally, the biological and clinical relevance of the findings was interpreted and discussed.

2.2.2 WikiPathways

The WikiPathways *Homo sapiens* data set (version 20211010) was downloaded using the R package `rWikiPathways` v1.14.0 [60]. Known modules from WikiPathways were defined as the gene sets of each pathway in the data set. Similarly to known modules from the regulatory circuits, the modules were filtered to contain at least five genes, and similar modules were merged based on a Jaccard index of > 0.8, ultimately yielding 653 modules.

2.3 Observed modules

2.3.1 BUENO data set

The BUENO data set contains RNA-sequencing data from 211 mesothelioma patients. A detailed description of the preparation and sequencing of the samples can be obtained from the study by Raphael Bueno et al. [61]. In brief, samples were obtained from patients undergoing extirpative surgery for MPM and sequenced on an Illumina HiSeq 2500 sequencer with multiplexed libraries and 2 × 75 bp paired-end reads. The data was obtained from the European Genome-phenome Archive (EGA) [62] with ID EGAD00001001915 on July 26$^{\text{th}}$ 2019.

2.3.2 Preprocessing

Quantification of transcript abundances was done using the pseudo-aligner kallisto [63] and an index of the human reference genome GRCh38.p12. Transcript abundances were imported into R and mapped to genes using the package `tximport`. Genes with less than 10 counts across all samples were filtered, and variance stabilising transformation, available from the `DESeq2` v1.26.0 R package, was applied to the raw counts, yielding 25 624 genes in 211 samples.

2.3.3 WGCNA

Variance-stabilised transformed data of the BUENO data set was used, and the network was constructed using the Pearson correlation and a signed co-expression similarity measure as seen in equation 1.2. For convergence towards scale-free topology properties, the `pickSoftThreshold` function from the WGCNA package v1.70.3 [50] was used, and the power was determined to be $\beta = 7$. Subsequently, the package's `blockwiseModules` function constructed modules using the topological overlap dissimilarity measure, and modules were detected using the Dynamic Tree Cut algorithm [55]. Parameters for module detection and module merging were varied and are listed in Table 2.1. The minimum module size was set to 30, and the reassign threshold to 10^{-6}.

Table 2.1.: Parameters for module detection by Dynamic Tree Cut algorithm as well as settings for module merging in WGCNA.

Parameter and values	Description
`deepSplit` $\in \{1, 2, 3, 4\}$	Sensitvity of module splitting.
`detectCutHeight` $\in \{0.990, 0.995, 0.998\}$	Cut height for module detection.
`mergeCutHeight` $\in \{0.15, 0.20, 0.25\}$	Cut height for module merging.

2.3.4 Independent Component Analysis

For ICA, the variation standardised transformed expression matrix of the BUENO data set was used. Expression values were preprocessed by subtraction from the mean and division by the standard deviation to achieve a zero-mean ($\bar{X} = 0$) and a unit variance ($s^2 = 1$). ICA was performed using the R implementation of the FastICA algorithm [64] (R package `fastICA` v1.2.3 [65]). The number of components was varied according to Table 2.2. Modules were detected from the source matrix S using a False Discovery Rate (FDR) estimation as proposed by Rotival et al. [57]. The R package `fdrtool` v1.2.17 [66] was used with the cutoff method "`fndr`". Any genes with a q-value below a certain cutoff value ($q < \text{cutoff}$) were considered module members. The q-value cutoff was varied as a parameters as seen in Table 2.2.

Table 2.2.: Parameters for the number of components and module detection in ICA.

Parameter and values	Description
`nComponents` $\in \{10, 30, 50, \ldots, 170, 190, 210\}$	Number of indenpendent components.
`cutoff` $\in \{10^{-1}, 10^{-2}, \ldots, 10^{-7}, 10^{-8}\}$	Cutoff of q-values.

2.4 Gene Ontology Term Enrichment Analysis

Over-Representation Analysis (ORA) of Gene Ontology (GO) biological process terms was conducted for known and observed modules using the `enrichGO` function of the R package `clusterProfiler` v4.2.2 [67]. Annotation data for *Homo sapiens* was used from the R package `org.Hs.eg.db` v3.13.0 [68]. The 'universe', the set of all genes, used as background for the ORA, was comprised of all genes in the respective data set from which the known or observed modules were defined from. In the case of the known modules from the mesothelioma regu-

latory circuits data, only target genes were used. Multiple testing correction was conducted using the Benjamini-Hochberg method [69], and p and q-value cutoffs were each set to 0.05.

Resulting GO terms were simplified to reduce redundancy using the `simplify` function of `clusterProfiler`, with cutoff set to 0.7 for the Wang measure and using the minimum of the adjusted p-value.

2.5 Evaluation of known modules

The sets of known modules that were derived from the regulatory circuits data set by varying the q-value cutoff were evaluated using two biologically relevant measures: biological homogeneity index and GO semantic similarity. For each set of known modules, a set of randomised modules with an identical module size distribution was constructed by sampling genes from the regulatory circuits or WikiPathways data set, respectively.

2.5.1 Biological Homogeneity Index

Each set of known modules was evaluated using the Biological Homogeneity Index (BHI) [70] according to

$$\mathrm{BHI} = \frac{1}{k}\sum_{j=1}^{k}\frac{1}{n_j(n_j-1)}\sum_{x\neq y\in C_j} I(\mathcal{B}_x, \mathcal{B}_y)$$

with k being the number of known modules, n_j the number of annotated genes in module j, C_j the set of GO-annotated genes in module j, and $\mathcal{B}_x$ denotes the set of GO terms annotated to gene x. The indicator function I returns 1 if any of the GO terms of genes x and y match:

$$I\left(\mathcal{B}_x, \mathcal{B}_y\right) = \begin{cases} 1 & |\mathcal{B}_x \cap \mathcal{B}_y| \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

2.5.2 Gene Ontology Semantic Similarity

Co-expressed genes are likely involved in similar biological processes. This assumption is used as basis for the guilt-by-association principle in gene co-expression network analysis. Here, this assumption is used to evaluate known modules by assessing the semantic similarity of

enriched biological process gene ontology terms in known modules. The R package `GOSemSim` v2.20.0 was used to calculate the semantic similarity using the relevance method, proposed by Schlicker et al. [71]. The measure is based on the information content defined as

$$IC(t) = -\log(p(t))$$

where $p(t)$ is the frequency of a gene ontology term t relative to the entire gene ontology graph. The relevance method is then given by

$$\mathrm{sim}_{rel}(t_1, t_2) = \frac{2\, IC(LCA(t_1, t_2))\, (1 - p(LCA(t_1, t_2)))}{IC(t_1) + IC(t_2)}$$

with $LCA(t_1, t_2)$ being the least common ancestor of terms t_1 and t_2 in the gene ontology graph.

The semantic similarity of two genes is computed by using the combination method Best-Match Average (BMA) for all GO term semantic similarities, defined as

$$\mathrm{sim}(g_1, g_2) = \frac{\sum_{i=1}^{m} \max_{1 \le j \le n} \mathrm{sim}_{rel}(t_i, t_j) + \sum_{j=1}^{n} \max_{1 \le i \le m} \mathrm{sim}_{rel}(t_i, t_j)}{m + n}$$

2.6 Evaluation of observed modules

Observed modules derived from ICA and WGCNA were evaluated using the known modules from regulatory circuits and WikiPathways. For each pair of observed and known modules, a 2×2 contingency table was constructed consisting of the number of genes that are in both, in either one, or in neither of the respective known and observed modules. Genes that were not represented in any of the known modules were not considered in the contingency table.

A Fisher's exact test was conducted for this contingency table using the R function `fisher.test` with the alternative hypothesis set to '`greater`'. Multiple testing correction was conducted using the R function `p.adjust` with the Benjamini-Hochberg method. Significant pairs of known and observed modules were defined at a significance level of $\alpha = 0.01$.

From this set of significant module pairs, only the pairs with the maximal odds ratio for each known module were used for further evaluation by the aucodds score. The aucodds

score is defined as the area under the curve formed by the fraction of module pairs with an odds ratio greater than or equal to a threshold in the range from 1 to 1000. The area under the curve was approximated using the trapezoid rule and scaled by normalisation into the range $[0, 1]$.

Based on this aucodds score, one set of observed modules from each method was chosen and used for further evaluation.

2.7 Comparison of observed modules

Observed modules were compared to identify similarities and overlaps between modules derived from ICA and WGCNA. First, the gene ontology semantic similarity was computed for each gene module, as described in subsection 2.5.2. Additionally, the overlap coefficient was calculated as

$$\text{overlap}(X, Y) = \frac{|X \cap Y|}{\min(|X|, |Y|)}$$

where X and Y denote the sets of genes in the respective observed modules. The overlap coefficient was chosen over the more commonly used Jaccard index, defined as

$$J(X, Y) = \frac{|X \cap Y|}{|X \cup Y|}$$

because it captures cases in which one module is a subset of another, making this measure more interpretable when module sizes differ substantially.

Finally, the module eigengenes of the observed modules from ICA and WGCNA were correlated with each other.

2.8 Correlation to clinical parameters

The module eigengenes of the observed modules were correlated to clinical parameters using the Pearson correlation coefficient. Module eigengenes are defined as the first principal component of the module's gene expression matrix. Clinical parameters from the Bueno cohort that were used include the overall survival in days, the histological subtype in the form of the percentage of sarcomatoid cells in the tumour, and the expression-based consensus

clusters by encoding the discrete variables numerically. The epithelioid cluster was encoded numerically as 0, the biphasic-E cluster as 0.33, the biphasic-S as 0.66 and the sarcomatoid cluster as 1.

2.9 Biological context of observed modules

Modules of co-expressed genes were put into biological context by conducting GO term enrichment analysis as described in section 2.4. Additionally, an ORA of WikiPathways was conducted similarly by using the `enrichWP` function of the `clusterProfiler` R package v4.2.2 [67]. The organism parameter was set to '*Homo sapiens*'. Other parameters were set as in the ORA of GO terms.

3 Results

3.1 Evaluation of known modules

The known modules from the regulatory circuits dataset were constructed by varying the edge weight cutoff. Additionally, one set of known modules was defined as the pathways found in the WikiPathways dataset, ultimately yielding 11 sets of known modules. The distributions of the number of genes in the modules can be seen in Figure 3.1.

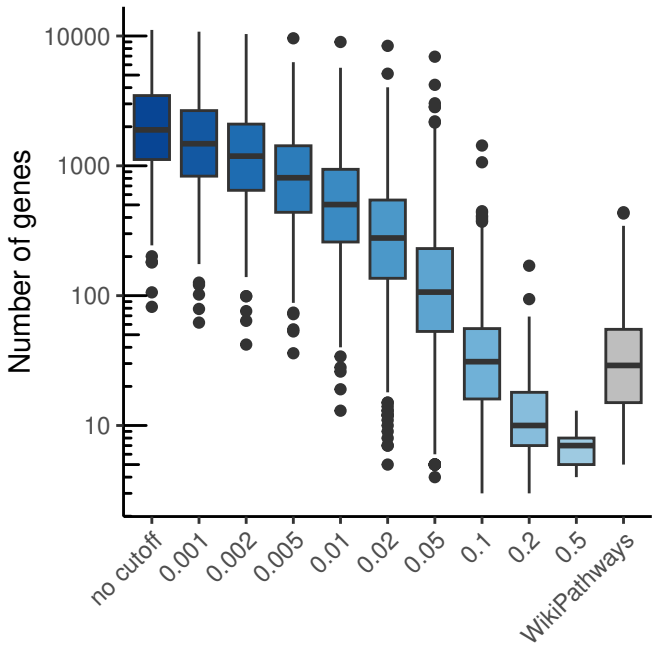

Figure 3.1.: Distributions of the number of genes in known modules derived from WikiPathways (grey) as well as from different cutoffs of regulatory circuits (shades of blue).

The sizes of the known modules from regulatory circuits decrease by increasing the cutoff parameter, from a median of 1892 genes per module when no cutoff is applied to only a median of 7 genes per module with the most stringent cutoff of 0.5. The WikiPathways-derived set of known module has a median of 29 genes per module.

The BHI is a measure describing the fraction of gene pairs in a set of genes with common biological functions. Figure 3.2a shows the calculated BHI of each set of known modules compared to a randomised set of modules with an identical module size distribution. More stringent cutoffs improve the biological homogeneity of known modules from the regulatory circuits dataset compared to the random modules. The fold change in mean BHI for the known modules with no cutoff was only 1.09 while the fold change with a cutoff of 0.2 was 2.46. The known modules from WikiPathways had a fold change of 7.54. The most stringent cutoff of 0.5 showed a lower biological homogeneity than the random modules (fold change 0.78).

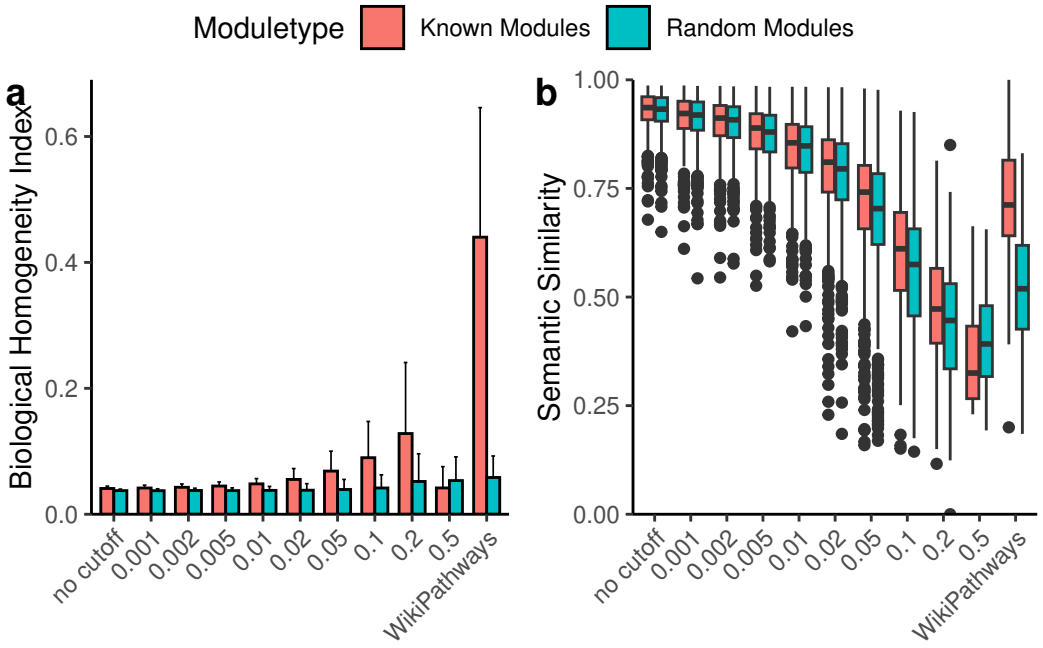

Figure 3.2.: Evaluations of known modules (■) compared to randomized modules with the same module size distribution (■). Evaluation measures are (**a**) the biological homogeneity index and (**b**) the GO semantic similarity. Known modules definitions include different cutoffs as parameter for known modules derived from regulatory circuits as well as WikiPathways.

The semantic similarity of GO biological process terms was analysed within each known and randomised module. The measure did not provide large differences between the known and random modules to be reliably used for evaluation. Out of all regulatory circuits-derived known modules, only the cutoffs 0.05, 0.1 and 0.2 were significantly different from the random modules (Mann-Whitney U test; $p = 2.62 \times 10^{-5}$, 9.04×10^{-7}, 1.51×10^{-5} for 0.05, 0.1 and 0.2, respectively). The GO semantic similarity of biological processes was significantly higher in observed modules from WikiPathways.

For further analyses, the known modules set at a cutoff of 0.05 was used. This set of known modules showed significant improvements in BHI over the set of random modules. Furthermore, the number of genes in the modules (median: 106.5; mean: 235.1) is in a realistic range for the number of targets of a transcription factor [72, 73].

3.2 Evaluation of observed modules

Observed modules were evaluated using an odds ratio-based approach against the sets of known modules. The aucodds measure is the area under the curve of the fraction of odds ratios greater than or equal to a threshold from 1 to 1000. All decisions were made considering both sets of known modules, from regulatory circuits and WikiPathways. Subsequent analyses were conducted using only the set of observed modules derived from the optimal parameters of each method, ICA and WGCNA.

3.2.1 ICA parameters

Results from the evaluation of ICA parameters are shown in Figure 3.3. First, the number of components was evaluated for observed modules from ICA. The optimal number of components was chosen in a process akin to the elbow method commonly used in cluster analysis to avoid over-fitting. As indicated by the vertical line in Figure 3.3a, 90 components were chosen.

The second parameter in ICA was the cutoff for the q-value used to define the members of each of the 90 components or modules. The optimal cutoff was chosen by balancing more stringent q-value cutoff values and a high aucodds scores for both sets of known modules. The chosen cutoff of 10^{-4} is indicated in Figure 3.3b by a vertical line.

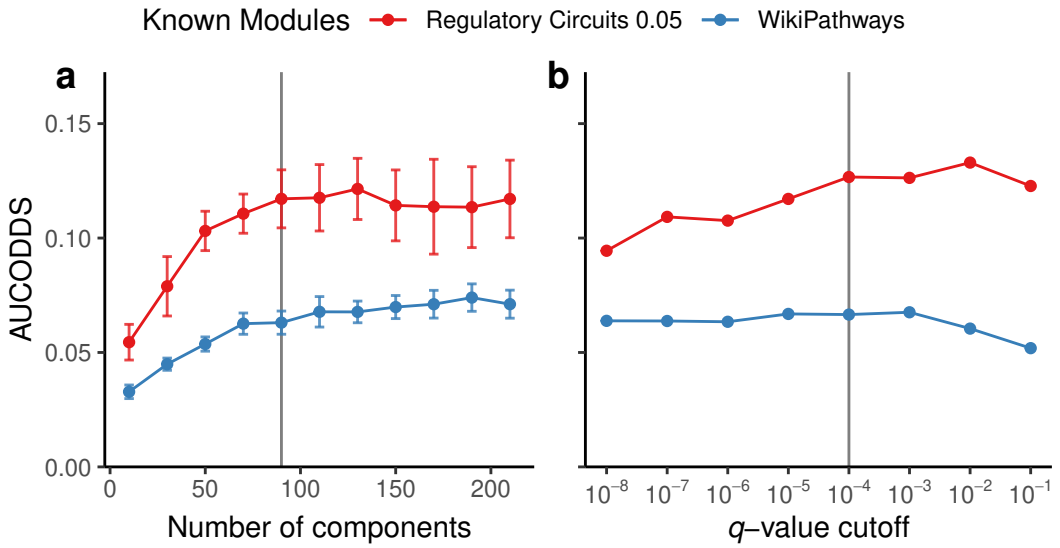

Figure 3.3.: Aucodds scores of parameter settings for the (**a**) number of components and (**b**) q-value cutoff in ICA. The q-value cutoff is shown for 90 components. Colours distinguish between two sets of known modules from regulatory circuits with cutoff of 0.05 (●) and WikiPathways (●).

3.2.2 WGCNA parameters

Three parameters influencing the module detection using topology-based dynamic tree cutting were varied in WGCNA. Figure 3.4 shows the aucodds value for each parameter setting. The largest range of variation is caused by changing the `deepSplit` parameter, corresponding to the sensitvity of the module detection algorithm. The `mergeCutHeight` parameter had only a marginal impact on the score, whereas the `detectCutHeight` parameter had none. WGCNA has been shown to be relatively robust and insensitive to parameter tuning [74]. Parameters were chosen as shown by the vertical lines in Figure 3.4. The `deepSplit` parameter of three was chosen based on the high aucodds score in both known module datasets. For `mergeCutHeight` and `detectCutHeight`, the middle values 0.02 and 0.995, respectively, were chosen. Figure 3.5 shows the dendrogram of the hierarchical clustering along with the module detection results for different values of the `deepSplit` parameter. Adjusting this parameter allows for controlling the number and the sizes of the modules. Higher values for `deepSplit` generate more modules with fewer genes, whereas lower values produce fewer modules containing more genes.

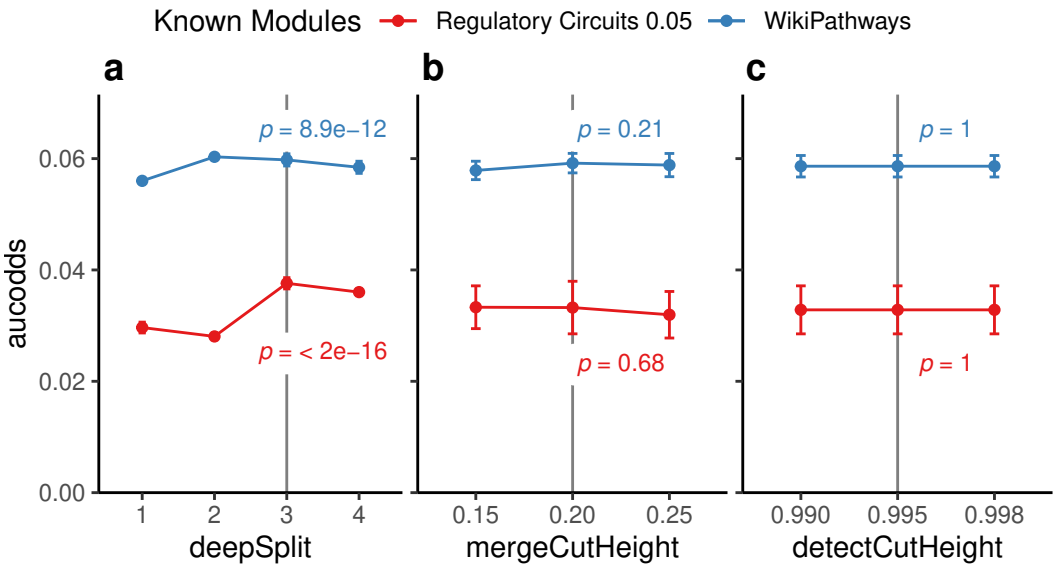

Figure 3.4.: Performance of parameter settings (**a**) `deepSplit`, (**b**) `mergeCutHeight` and (**c**) `detectCutHeight` for the module detection in WGCNA using the aucodds score for known modules from regulatory circuits with cutoff 0.05 (●) and WikiPathways (●). *p*-values were calculated using one-way ANOVA.

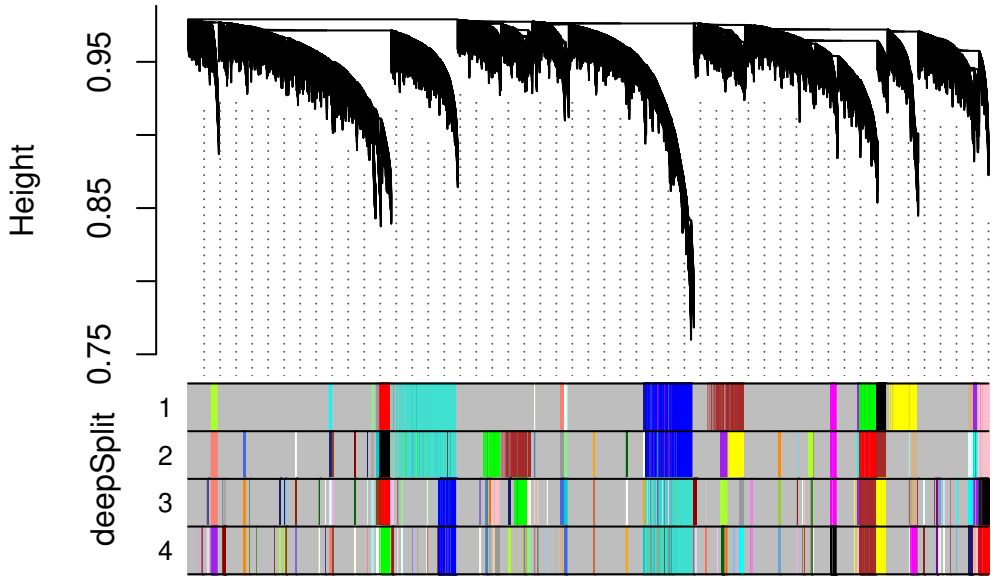

Figure 3.5.: WGCNA dendrogram and gene modules for different values of the `deepSplit` parameter.

3.3 Comparison of observed modules

3.3.1 Characteristics of observed modules

ICA and WGCNA resulted in different average sizes of modules (Figure 3.6a). ICA modules were generally much larger than WGCNA modules, with the median size of ICA modules being approximately six times larger than WGCNA's. The sizes of the module had similar variation. The standard deviation of the number of genes in modules of ICA was 199.9 genes, compared to 188.8 genes in WGCNA.

The module membership is regularly used in gene co-expression analysis for characterising the obtained modules. It is calculated by correlating the expression profiles of individual genes to that of the module's eigengene. It is thereby a direct measure of coherence in a module's gene expressions. Module membership for all genes to their respective modules revealed that genes in WGCNA modules are much stronger correlated to their module's expression profile than the genes in ICA (Figure 3.6b). By this measure, ICA is greatly outperformed by WGCNA and fails to cluster distinct groups of co-expressed genes very well.

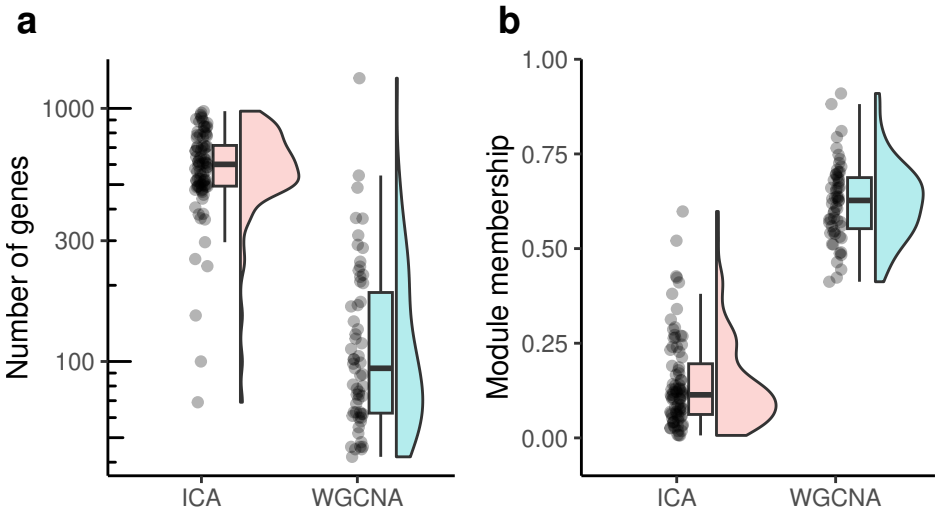

Figure 3.6.: Basic characteristics of observed modules. (**a**) Distribution of the number of genes in ICA- and WGCNA-derived observed modules. (**b**) Distribution of the module memberships in ICA and WGCNA.

Two key characteristics which distinguish ICA from WGCNA are that the former allows for module overlap (i.e., that one gene can belong with more than one module) and that it is exhaustive in assigning genes to modules (i.e., each gene is associated to a module). Observed modules from ICA had a very narrow range of overlap coefficients (Figure 3.7). On average, 73 genes were shared between pairs of modules. Every module pair shared at least one gene, while the greatest overlap was seen with 263 genes shared (minimum and maximum overlap coefficients 0.01 and 0.32, respectively).

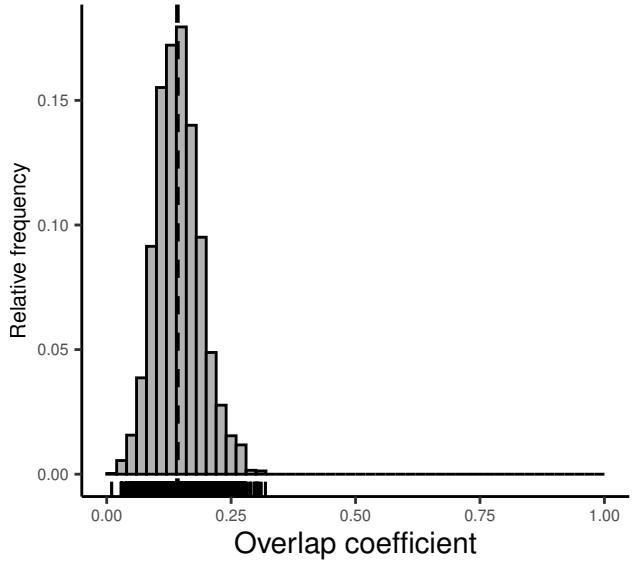

Figure 3.7.: Pairwise overlap coefficient between observed modules from ICA.

In WGCNA, the 'grey' module is a proxy for genes that are not associated with any module. Out of 25 624 genes, 16 405 were assigned to the grey module (64.02 %), leaving only 9219 (35.98 %) in gene co-expression modules.

3.3.2 Comparison of ICA and WGCNA modules

The observed modules that were identified using ICA and WGCNA were compared a) by the overlap coefficient, describing the overlap of genes between modules, b) by their GO semantic similarity, describing overlap in terms of biological function and c) by correlating the module's eigengene expressions.

The vast majority of modules between ICA and WGCNA do not overlap in their member genes, with a median overlap coefficient of 0.02 (Figure 3.8a). Only 12 module pairs of a total of 5251 have an overlap coefficient > 0.75. The semantic similarity of biological process terms between observed modules is far less skewed and has a median of 0.31 (Figure 3.8b). 28 module pairs show a similarity > 0.75. Semantic similarity of biological process terms between observed modules could not be computed for modules without any enriched GO terms. This is the case for 25 modules of WGCNA, while only 6 modules in ICA did not have any significantly enriched GO terms. The module pair ICA 44 and WGCNA tan (12) showed high overlap and functional similarity and was analysed by functional and pathway term enrichment in section 3.5.

The expression profiles of module eigengenes were compared by pairwise correlation between observed modules of ICA and WGCNA (Figure 3.9). A large fraction of 46 ICA modules are similarly and highly positively correlated to the blue WGCNA module (2) and negatively to the purple WGCNA module (10). Consequently, these ICA modules are also similarly correlated with each other.

Next, it was investigated whether the two co-expression clustering methods yield similar results in terms of the quality of the functional enrichment. For this, two measures were used that assess different qualities of enriched GO terms. First, the GO term depth score uses information from the ontology structure for each enriched term to quantify its specificity. Terms that are lower in the ontology tree are more specific descriptions of the biological associations of genes in a module. In general, WGCNA generates more specific GO terms than ICA (Figure 3.10a; One-sided Wilcoxon Rank Sum test $p = 8.14 \times 10^{-17}$). The second measure that was used to compare the two methods is the BHI, which was already used for assessing known modules. Also here, WGCNA produces modules that are more homogeneous in their enriched biological process terms than ICA (Figure 3.10b; One-sided Wilcoxon Rank Sum test $p = 0.0029$).

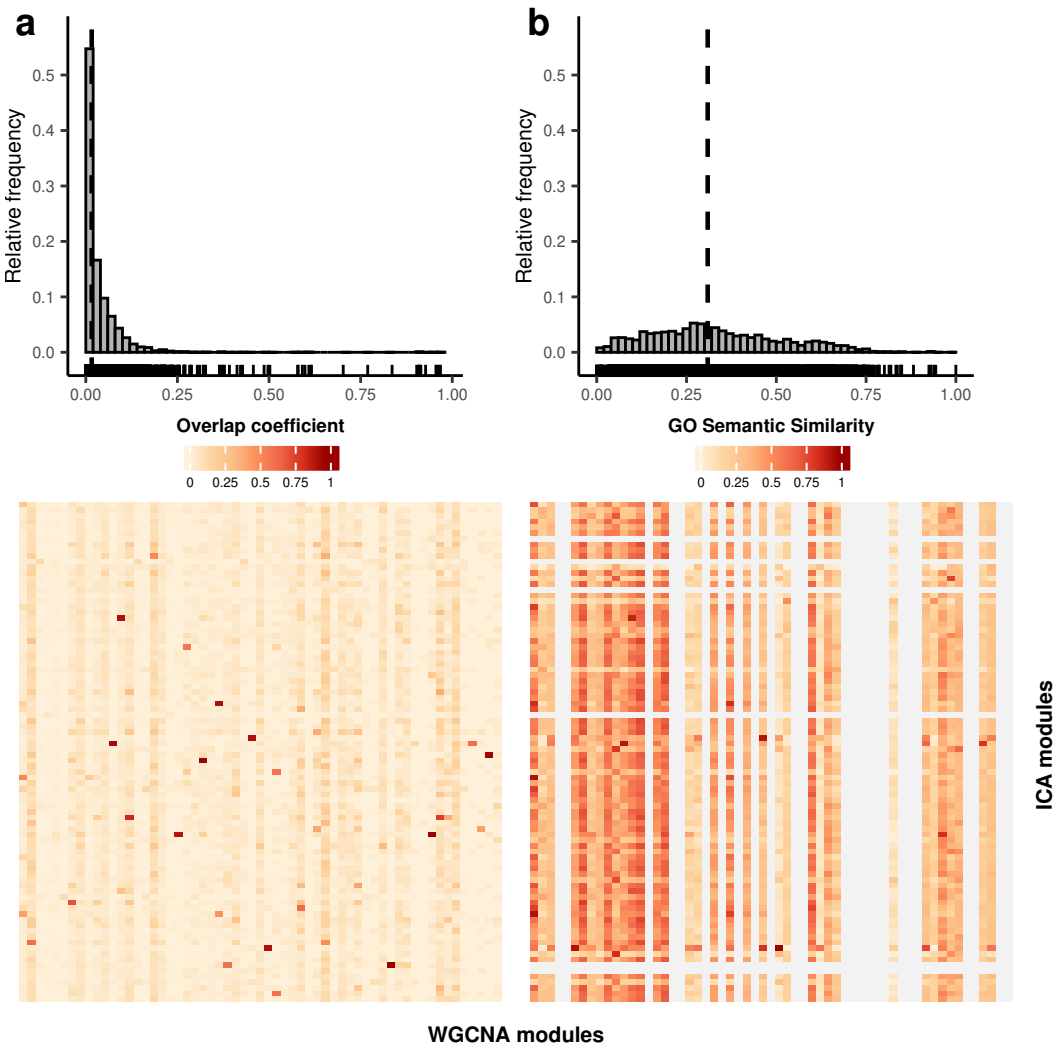

Figure 3.8.: Relative frequency distribution (top) and heatmap (bottom) of (**a**) the overlap coefficient and (**b**) the GO semantic similarity between each pair of observed modules from WGCNA and ICA. Grey rows and columns indicate modules of ICA and WGCNA with no enriched GO terms.

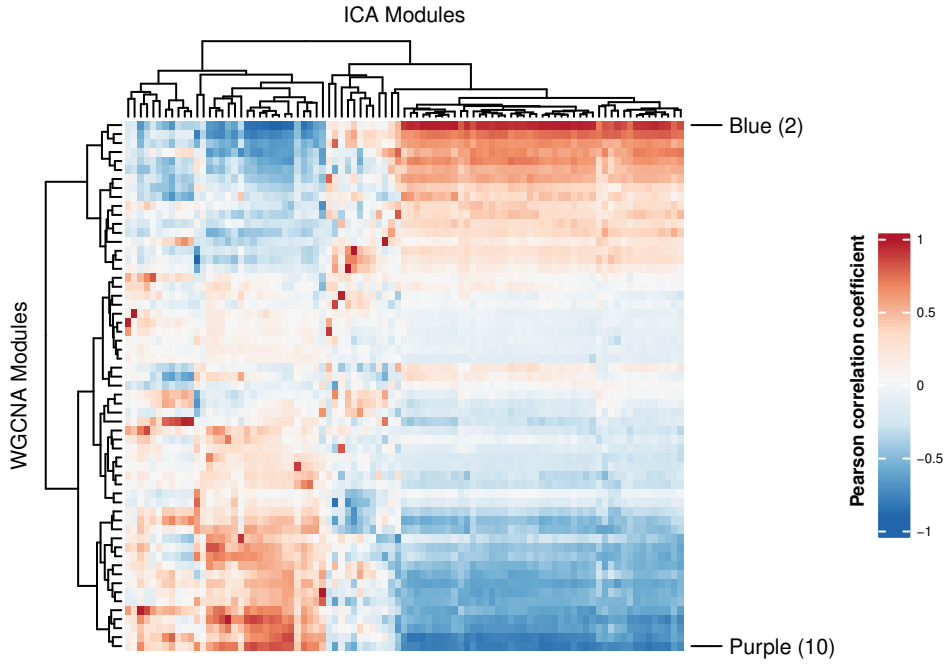

Figure 3.9.: Pearson correlation of module eigengene expression between ICA and WGCNA modules. Clustering of modules by complete-linkage agglomerative hierarchical clustering and the euclidean distance measure.

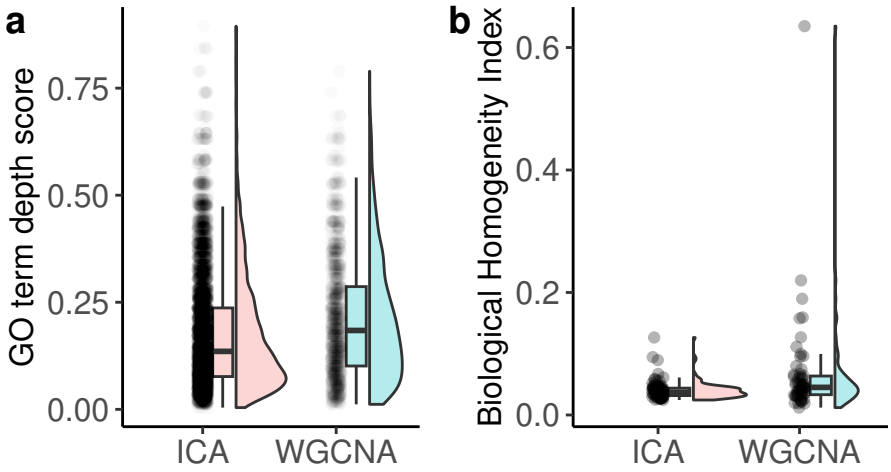

Figure 3.10.: Distributions of (**a**) GO term specificity and (**b**) BHI of ICA and WGCNA observed modules

3.4 Module–trait relationship

Expression profiles of the eigengenes of observed modules were correlated to the sample's clinical parameters overall survival and histology as well as the expression-based clusters according to Bueno et al. [61] and are shown in Figure 3.11.

Pearson correlations with overall survival were generally very low in both ICA and WGCNA, with maximum absolute correlations being 0.274 and 0.311, respectively. Higher correlations were observed with epithelioid and sarcomatoid histologies and expression-based clusters. MPM patients with sarcomatoid histology have a significantly worse prognosis than patients diagnosed with epithelioid histology [31].

The resulting correlations were partitioned into three clusters using k-means clustering, resulting in three groups of co-expressed modules. The clusters differ most noticeably by the correlations with histology and consensus cluster. A large fraction of ICA modules show a strong negative correlation to the sarcomatoid histology and consensus cluster, with 46 modules in the blue cluster shown in Figure 3.11a. Twenty-one modules of the red cluster in ICA are positively correlated with the sarcomatoid histology and consensus cluster. Analogous, in WGCNA, 17 belong to the blue, 26 to the red cluster of co-expressed modules (Figure 3.11b). Notably, the WGCNA modules purple (10) and blue (12), which were also shown to correlate strongly with a large portion of ICA modules, are most strongly correlated to consensus clusters, in opposite manner. Red and blue clusters of gene modules were investigated further using functional enrichment analysis in subsection 3.5.1.

3.5 Functional enrichment analysis

In-depth analysis of the biological context of the gene modules was carried out using functional enrichment analyses of GO biological process terms and WikiPathways using over-representation analysis for modules of interest from ICA and WGCNA. Hereinafter, selected functional enrichment results from WGCNA and ICA modules are presented based on their relevance to clinical traits.

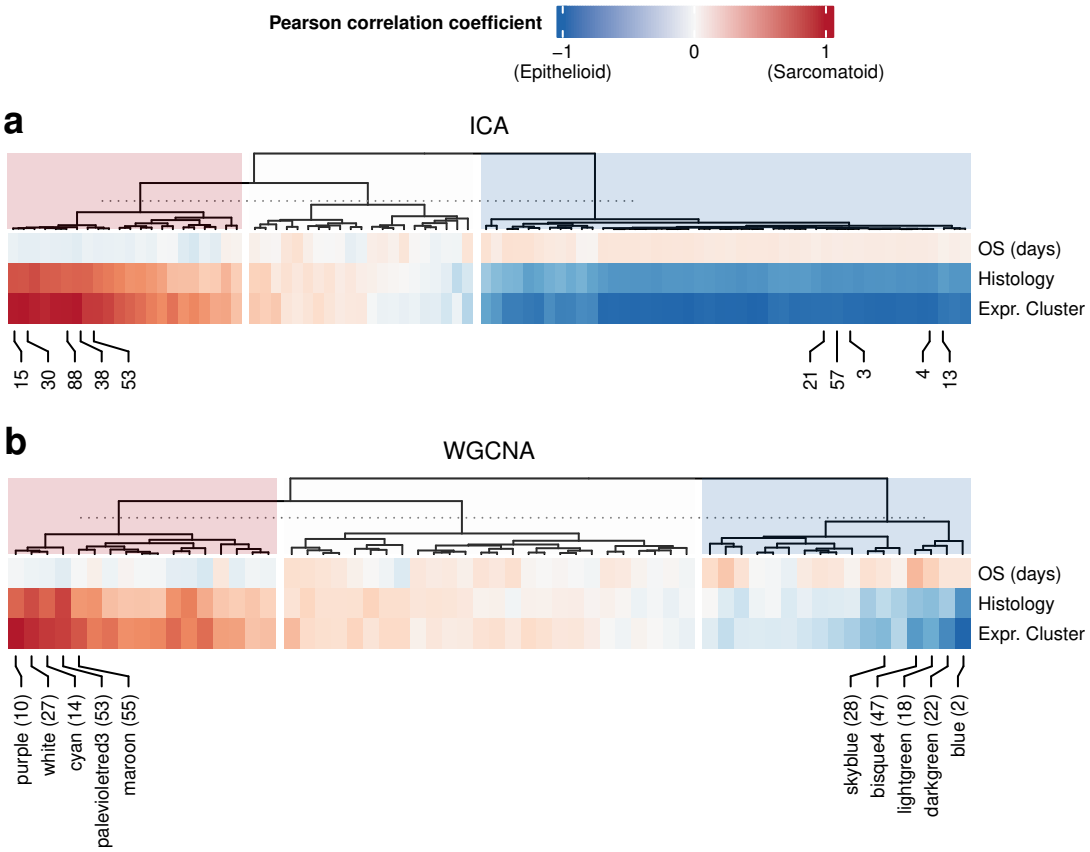

Figure 3.11.: Pearson correlation coefficient between expression profiles of eigengenes from observed modules for (**a**) ICA and (**b**) WGCNA, against the clinical parameters overall survival, sarcomatoid histology and the sarcotmatoid expression-based consensus cluster as defined by Bueno et al. [61]. Dendrograms show the complete-linkage hierarchical clustering with the euclidean distance measure. The correlations were clustered further by k-means clustering, yielding red, blue and white clusters as indicated in the dendrogram. Labels indicate modules with the highest correlation in red and the lowest correlations in blue clusters.

3.5.1 Correlation clusters

Correlations of module eigengenes to clinical parameters yielded clusters of modules in ICA and WGCNA. From each of the clusters, five modules (indicated by annotation in Figure 3.11a) were compared by functional enrichment. For the red clusters, the five maximally positively correlated modules were chosen, and for the blue clusters, the five maximally negatively correlated.

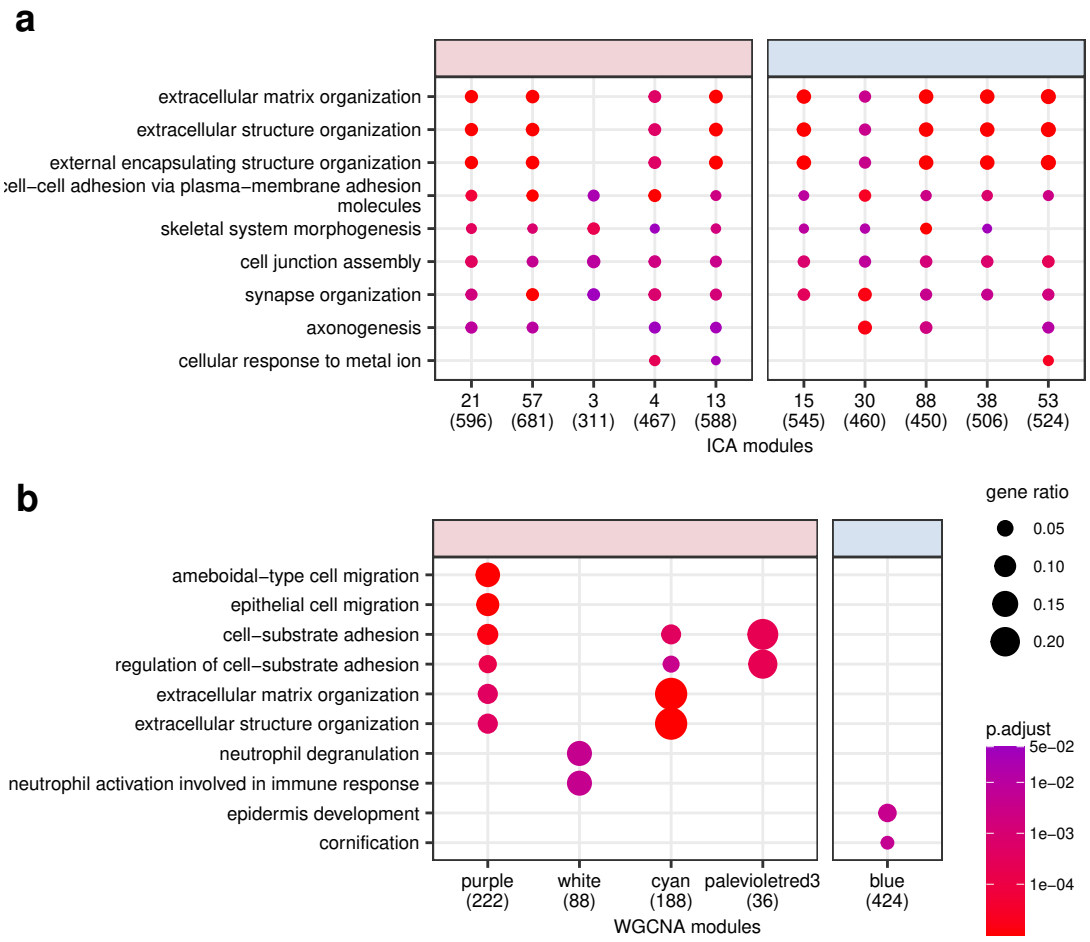

Figure 3.12.: Enriched GO terms of top-five (a) ICA and (b) WGCNA modules in red and blue correlation groups. Only modules with significantly enriched GO terms are shown. In parentheses, the number of annotated genes used in the ORA is shown.

Modules of the red and blue clusters of ICA, correlating with either histological phenotype, are enriched in similar biological processes involving the organisation of extracellular

structures, cell-cell adhesion and cell junction assembly (Figure 3.12a). All ten modules had significantly enriched biological process terms. ICA modules generally have only a low fraction of genes annotated to each GO term.

In WGCNA, only four modules of the red cluster and one of the blue cluster had any enriched GO terms (Figure 3.12b). Again, extracellular structure and matrix organisation were identified among enriched terms. Additionally, cell-substrate adhesion was significantly enriched in three out of four modules correlating with the sarcomatoid histology.

3.5.2 WGCNA module 10

The purple (10th) module of WGCNA, containing 237 genes, showed high correlation to the sarcomatoid expression-based cluster (Pearson correlation: 0.73). Enriched biological process terms and pathways are shown in Figure 3.13. Genes enriched in the module show biological functions primarily associated with cell migration as well as to cell adhesion. Genes that belong to this module include *PDGFR*β, *TGFR*β1, *LOX*, *LOXL2*. Among the genes in this module are furthermore several α and β subunits of integrins *ITGA1*, *ITGA2*, *ITGAV*, *ITGB1*, and *ITGB3*.

Some of the module members involve genes that are specifically associated with the WikiPathways 'Malignant Pleural Mesothelioma' as well as to Hippo-Merlin and TGF-β signalling pathways. The TGF-β signalling pathways are involved in numerous cellular processes, including cell growth, differentiation, migration and apoptosis. Alongside the gene *TGFBR1*, encoding for the membrane-bound receptor, is also the gene *GREM1* associated with this WikiPathway term, which acts as an inhibitor in the signalling pathway.

3.5.3 Highly overlapping modules ICA 44 and WGCNA 12

While most module pairs of ICA and WGCNA-derived modules do not show a high overlap, there are exceptions, such as the tan module (12) from WGCNA and the ICA module 44. The overlap coefficient between these modules is 0.93, and the GO semantic similarity is 0.92. The smaller WGCNA module is a subset of the larger ICA module (Figure 3.14a), and due to the large size difference between the modules, the Jaccard index between them is only 0.28.

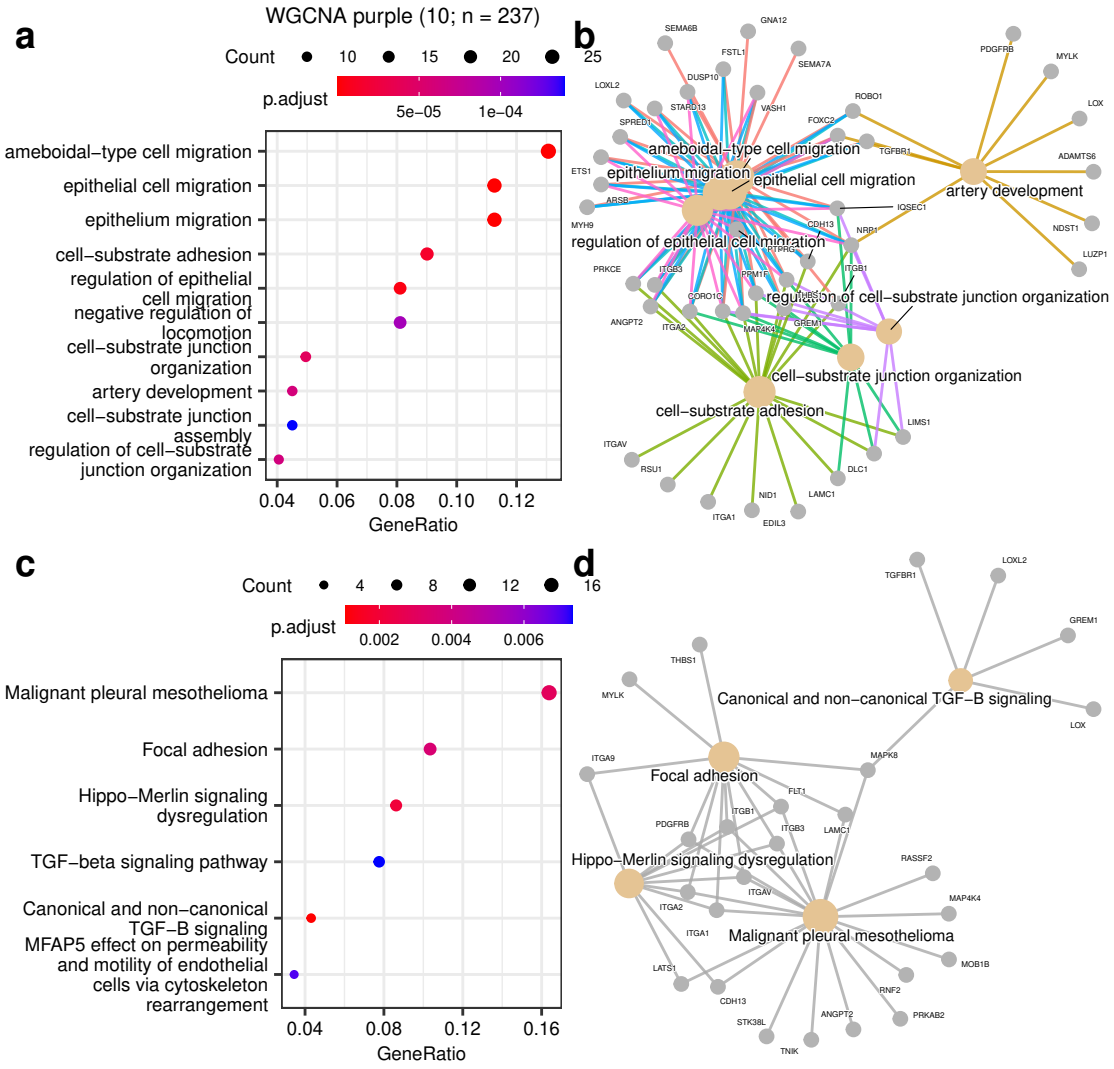

Figure 3.13.: Over-representation analysis of (a-b) GO terms and (c-d) WikiPathways. Dotplots depict the gene ratio and Benjamini-Hochberg corrected p-value of (a) GO terms and (b) WikiPathways. Concept-network plots show (b) GO terms and (d) WikiPathways with associated genes and their relationship.

The expression profiles of both modules are correlated with lower overall survival (Pearson correlation coefficient -0.27 and -0.29 for ICA and WGCNA, respectively) as well as to the sarcomatoid histology and expression cluster (histology: 0.38 and 0.35; consensus cluster: 0.59 and 0.53).

The modules are functionally involved in biological processes and pathways associated to cell cycle control, mitosis and DNA replication (Figure 3.14b,c). The gene-concept networks for the biological process GO terms and WikiPathwys are shown in Figure S2 and Figure S3, respectively. Terms associated with WGCNA modules showed higher gene ratios than those from ICA modules, suggesting that WGCNA modules are biologically more coherent.

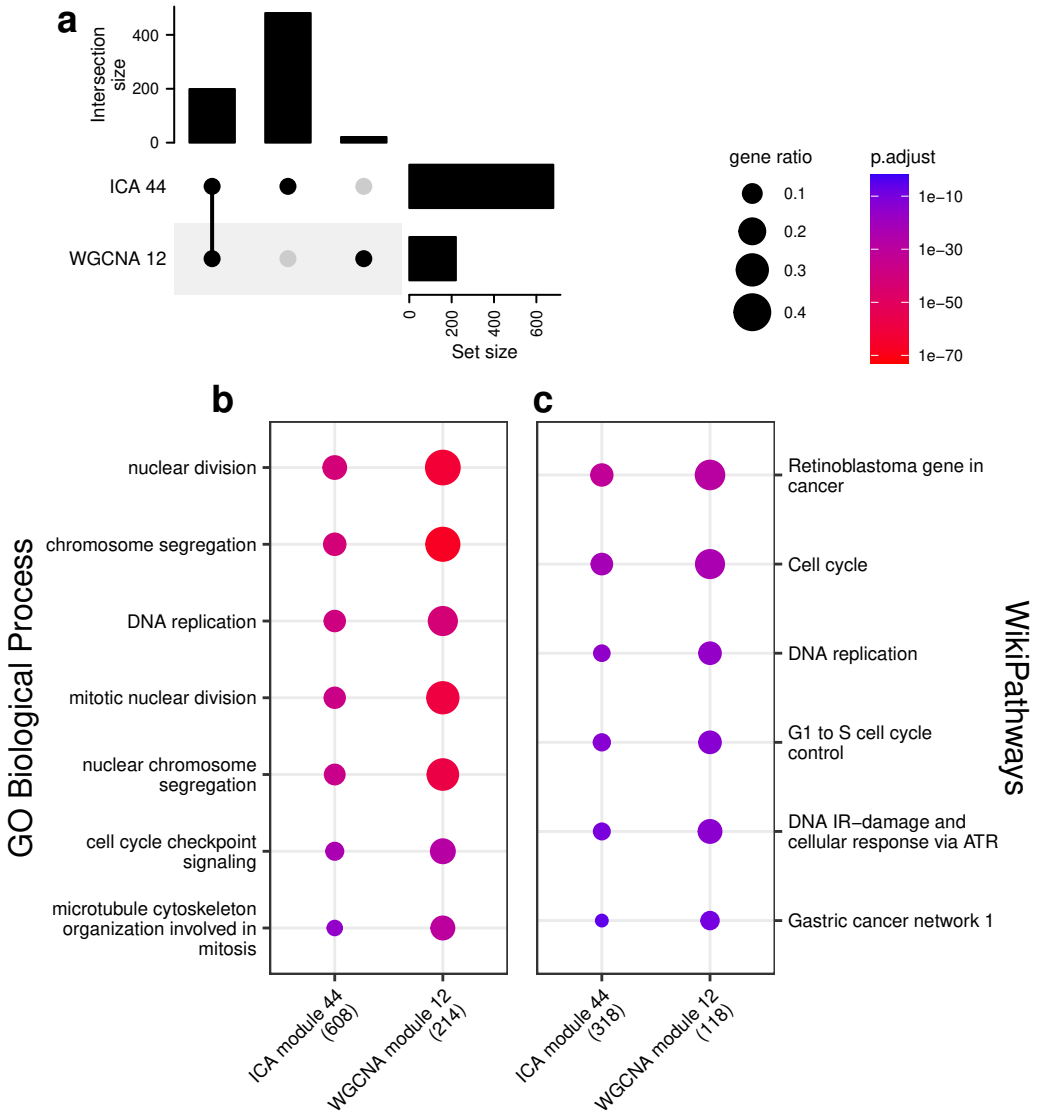

Figure 3.14.: Comparison of functional enrichment between ICA module 44 and WGCNA module 12. **(a)** Upset plot showing the overlap between the two observed modules. **(b)** GO biological process terms and **(c)** WikiPathways enriched in ICA module 44 and WGCNA module 12. The number of annotated genes in each module is shown in parentheses.

4 Discussion

Malignant pleural mesothelioma is a highly aggressive cancer with a generally bad prognosis for affected patients. Bueno et al. [61] have performed a comprehensive characterisation of a large cohort of MPM patients and defined distinct molecular subtypes using RNA-seq. Here, this large dataset was used for the analysis of co-expressed genes in MPM, as well as the evaluation of two methods for this purpose. Compared to previous studies that compare gene co-expression network analysis methods, such as Saelens et al. [74], a larger focus was laid on the biological interpretability of the results.

Co-expression network analysis methods identify groups of genes with similar expression profiles, called modules. Several of these methods were previously evaluated by Saelens et al. [74] using a dataset of tissue-specific regulatory circuits published by Marbach et al. [59]. This dataset was also used in this thesis in order to define a known set of modules for the evaluation and optimisation of ICA and WGCNA parameters. However, the definition of known modules also required the choice of a parameter, namely the cutoff for the edge weights in the directed network. In order to choose a reasonable value for this parameter, the module size and GO biological process terms were considered; the latter in comparison with randomised gene modules. Measures for these GO terms rely on the assumption that target genes of transcription factors are involved in the same biological processes and that one transcription factor only regulates one process. However, this assumption may be flawed, as many transcription factors regulate genes across multiple biological processes [75]. This has been demonstrated in *Escherichia coli*; however, whether this holds true in eukaryotes—and in humans specifically—remains unknown. Also, the number of genes affected by a transcription factor can vary by a wide range, up to thousands of genes [73]. Furthermore, the optimal number of genes in a module is dependent on the biological question under investigation and the purpose of the application of co-expression network analysis methods. These aspects make

the judgement of the cutoff parameter based on the module size challenging. Additionally, to the regulatory circuits dataset, the WikiPathways dataset [60] was added as well. This dataset contains sets of genes involved in the same processes and pathways rather than genes regulated by the same transcription factor. The assumption that is made here is that co-expressed genes are oftentimes involved in the same biological processes and pathways.

The parameters of ICA and WGCNA were optimised using the aucodds approach described by Salelens et al. [74], applied to the known modules from the regulatory circuits and WikiPathways datasets. In ICA, the number of modules or components must be specified explicitly, in contrast to WGCNA, and this was the first parameter to be optimised. Saelens et al. explored a wide range of values, from 50 to 600 components, using a large step size, whereas in this study the range was limited to 10–210 components with a smaller step size. There were no major improvements in aucodds at higher numbers of components, and the number of components used by Saelens et al. was not reported. The second parameter was the q-value cutoff for the definition of module member genes. Also, this parameter was not reported by Saelens et al., but the same value range was used here. For WGCNA, Saelens et al. reported that only the `mergeCutHeight` showed a significant impact on the performance of the method. Here, however, it was found that `mergeCutHeight` had a much smaller impact on the module detection in WGCNA than the `deepSplit` parameter. This parameter controls the sensitivity for the tree cut algorithm and thereby directly determines the total number of detected modules (as shown in Figure 3.5) and, in consequence, the module sizes and memberships. The power, for convergence towards scale-free topology properties, was not varied in this thesis but was set by the `pickSoftThreshold` function of WGCNA, since this is the recommended procedure by the package authors. WGCNA has several other parameters that might significantly influence the module detection performance, some of which are not factors of the method itself, but for example, the stringency of filtering out lowly expressed genes during preprocessing of the data. Generally, it was found here as well as by Saelens et al. that WGCNA is a highly robust method with regard to its module detection parameters.

Generally, ICA produced larger modules with lower gene module memberships than WGCNA, indicating poorer performance in defining co-expressed gene modules. This may be due to suboptimal parameters choices, such as selecting too many or too few components or an inadequate q-value cutoff. Alternative methods for determining optimal ICA parameters

have been described in the literature and may yield improved gene module memberships [57, 76]. However, evaluating such approaches was beyond the scope of this work.

Among important properties of gene co-expression network analysis methods are their ability to allow for overlap in modules. In the two methods evaluated by this thesis, ICA allows for the overlap of gene modules while WGCNA forms mutually exclusive sets of genes. However, WGCNA introduces the concept of module eigengene as a representation of a module's gene expression profile, which can be correlated to individual gene expression patterns [77]. This allows for quasi-fuzzy clustering of genes based on these correlations.

Results from the BHI and GO term depth score analysis of observed modules suggest that WGCNA modules are biologically more homogeneous and specific than ICA's, which makes WGCNA biologically more interpretable. Furthermore, comparisons of enriched GO terms show that the gene ratios in ICA are generally lower than in WGCNA (Figure 3.12 and 3.14).

Module eigengenes were used to correlate the expression profile to clinical features of patients. This analysis revealed that modules could be clustered into three clusters based on their correlation to these clinical traits. Genes that correlate with the sarcomatoid histology and lower overall survival were functionally enriched in biological processes such as extracellular matrix and structure organisation. A previous large-scale molecular characterisation of epithelial and sarcomatoid mesothelioma showed significant differences in mean expression and methylation of genes involved in related pathways and processes [78].

In their comparative analysis, Saelens et al. [74] conducted only limited biological and functional investigations of the module detection methods. Specifically, they employed the F-aucodds method to assess the enrichment of observed modules with functional gene sets. However, ORA of GO terms and pathways is widely regarded as a standard approach for evaluating the biological relevance of gene sets, offering more insightful and interpretable results. Therefore, in this thesis, the biological and functional interpretation of the observed modules was performed using this established method.

The purple WGCNA module was correlated with the sarcomatoid histology and expression-based cluster. Most notably, this module was enriched in cell migration and cell adhesion biological process terms and pathways. Among the genes in this module are members of the lysyl oxidase gene family. The altered expression of these genes, commonly found in MPM, is linked to poorer overall survival in patients [79]. Drugs that target lysyl oxidases are currently

in active development, and clinical trials show promising results as a therapeutic target in MPM. Furthermore, enrichment of genes relating to the Malignant pleural mesothelioma and Hippo-Merlin signalling dysregulation WikiPathways demonstrate the biological relevance of this module. The Hippo signalling pathway, mainly playing a role in organ size control and tissue homeostasis, is commonly dysregulated in MPM. Central to this pathway is the tumour suppressor gene *NF2* encoding the protein Merlin. *NF2* is amongst the genes with the highest number of genomic alterations in MPM, mutated in almost one-fifth of patients in the Bueno cohort [61]. Another key member in this signalling cascade is the kinase and tumour suppressor LATS1, which inhibits the transcriptional co-activators YAP1 and TAZ. MOB1, which is found co-expressed with LATS1 in the purple module, is needed in the Hippo pathway as an activator of LATS1/2. These and other genes found within the purple module are highly relevant for the disease occurrence and progression in MPM. Currently, the Hippo pathway is considered highly relevant for diagnostics as well as therapeutic options [80].

Only a small number of modules from WGCNA and ICA have a considerable overlap. Module 12 from WGCNA is an almost complete subset of ICA's module 44, with an overlap coefficient of 0.93. Both modules are correlated to lower overall survival, as well as the sarcomatoid histology and expression-based cluster. Among the genes found in these overlapping modules are members of the minichromosome maintenance protein complex (MCM). This complex is also known as the DNA helicase complex and is therefore central to DNA replication. Numerous works have examined the connection between MCM proteins and various types of cancer [81–83]. Two other genes found in both modules are several members of the E2F family of transcription factors, such as *E2F1*, *E2F2*, *E2F7*, and *E2F8*. This family of proteins is critical in the cell cycle regulation and contains both activators of transcription as well as repressors. Expression of E2Fs and, in consequence, E2F targets is linked to poor prognosis in several cancers. However, their role in cancer was also shown to be ambiguous and specific to context. Reports describe *E2F8* to have both tumour suppressive as well as oncogenic potential [84].

4.1 Conclusion & Outlook

In this thesis, clusters of co-expressed genes in MPM were identified using two different module detection methods: ICA and WGCNA. After parameter optimisation based on known modules from regulatory networks and biological pathways, both methods produced gene modules significantly correlated with clinical traits. However, the modules derived from WGCNA exhibited higher module membership values, thereby outperforming ICA in defining co-expressed gene modules. Additionally, WGCNA modules demonstrated greater biological homogeneity and specificity in functional enrichment analyses, enhancing their biological interpretability. In conclusion, WGCNA outperformed ICA and proved to be a robust method for generating biologically meaningful insights.

Gene co-expression network analysis provides valuable insights into key transcriptional patterns in malignant diseases such as cancer. However, selecting optimal parameters for such machine learning methods remains a challenging task. Future research should focus on systematically assessing and characterising these parameters using objective evaluation metrics.

Bibliography

[1] Rafael Lozano et al. “Global and Regional Mortality from 235 Causes of Death for 20 Age Groups in 1990 and 2010: A Systematic Analysis for the Global Burden of Disease Study 2010”. In: *The Lancet* 380.9859 (15th Dec. 2012), pp. 2095–2128. ISSN: 0140-6736. DOI: 10.1016/S0140-6736(12)61728-0. URL: https://www.sciencedirect.com/science/article/pii/S0140673612617280 (visited on 07/07/2025).

[2] James W. Vardiman et al. “The 2008 Revision of the World Health Organization (WHO) Classification of Myeloid Neoplasms and Acute Leukemia: Rationale and Important Changes”. In: *Blood* 114.5 (30th July 2009), pp. 937–951. ISSN: 0006-4971. DOI: 10.1182/blood-2009-03-209262. URL: https://doi.org/10.1182/blood-2009-03-209262 (visited on 12/07/2021).

[3] Hyuna Sung et al. “Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries”. In: *CA: A Cancer Journal for Clinicians* 71.3 (2021), pp. 209–249. ISSN: 1542-4863. DOI: 10.3322/caac.21660. URL: https://onlinelibrary.wiley.com/doi/abs/10.3322/caac.21660 (visited on 18/11/2021).

[4] Douglas Hanahan and Robert A. Weinberg. “The Hallmarks of Cancer”. In: *Cell* 100.1 (7th Jan. 2000), pp. 57–70. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/S0092-8674(00)81683-9. pmid: 10647931. URL: https://www.cell.com/cell/abstract/S0092-8674(00)81683-9 (visited on 08/06/2021).

[5] Douglas Hanahan and Robert A. Weinberg. “Hallmarks of Cancer: The Next Generation”. In: *Cell* 144 (Mar. 2011), pp. 646–674. ISSN: 00928674. DOI: 10.1016/j.cell.2011.02.013. URL: http://linkinghub.elsevier.com/retrieve/pii/S0092867411001279 (visited on 02/11/2011).

[6] Sergei Grivennikov and Michael Karin. "Autocrine IL-6 Signaling: A Key Event in Tumorigenesis?" In: *Cancer Cell* 13.1 (8th Jan. 2008), pp. 7–9. ISSN: 1535-6108. DOI: 10.1016/j.ccr.2007.12.020. URL: https://www.sciencedirect.com/science/article/pii/S1535610807003790 (visited on 16/11/2021).

[7] Nikki Cheng et al. "Transforming Growth Factor-β Signaling–Deficient Fibroblasts Enhance Hepatocyte Growth Factor Signaling in Mammary Carcinoma Cells to Promote Scattering and Invasion". In: *Molecular Cancer Research* 6.10 (1st Oct. 2008), pp. 1521–1533. ISSN: 1541-7786, 1557-3125. DOI: 10.1158/1541-7786.MCR-07-2203. pmid: 18922968. URL: https://mcr.aacrjournals.org/content/6/10/1521 (visited on 16/11/2021).

[8] M. A. Davies and Y. Samuels. "Analysis of the Genome to Personalize Therapy for Melanoma". In: *Oncogene* 29.41 (41 Oct. 2010), pp. 5545–5555. ISSN: 1476-5594. DOI: 10.1038/onc.2010.323. URL: https://www.nature.com/articles/onc2010323 (visited on 16/11/2021).

[9] T. L. Yuan and L. C. Cantley. "PI3K Pathway Alterations in Cancer: Variations on a Theme". In: *Oncogene* 27.41 (41 Sept. 2008), pp. 5497–5510. ISSN: 1476-5594. DOI: 10.1038/onc.2008.245. URL: https://www.nature.com/articles/onc2008245 (visited on 16/11/2021).

[10] Julian Downward. "Targeting RAS Signalling Pathways in Cancer Therapy". In: *Nature Reviews Cancer* 3.1 (1 Jan. 2003), pp. 11–22. ISSN: 1474-1768. DOI: 10.1038/nrc969. URL: https://www.nature.com/articles/nrc969 (visited on 17/11/2021).

[11] R. A. Weinberg. "The Retinoblastoma Protein and Cell Cycle Control". In: *Cell* 81.3 (5th May 1995), pp. 323–330. ISSN: 0092-8674. DOI: 10.1016/0092-8674(95)90385-2. pmid: 7736585.

[12] Deborah L. Burkhart and Julien Sage. "Cellular Mechanisms of Tumour Suppression by the Retinoblastoma Gene". In: *Nature reviews. Cancer* 8.9 (Sept. 2008), pp. 671–682. ISSN: 1474-175X. DOI: 10.1038/nrc2399. pmid: 18650841. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6996492/ (visited on 20/12/2021).

[13] J. M. Adams and S. Cory. "The Bcl-2 Apoptotic Switch in Cancer Development and Therapy". In: *Oncogene* 26.9 (9 Feb. 2007), pp. 1324–1337. ISSN: 1476-5594. DOI: 10.1038/

sj.onc.1210220. URL: https://www.nature.com/articles/1210220 (visited on 19/11/2021).

[14] Jerry W. Shay and Woodring E. Wright. "Hayflick, His Limit, and Cellular Ageing". In: *Nature Reviews Molecular Cell Biology* 1.1 (1 Oct. 2000), pp. 72–76. ISSN: 1471-0080. DOI: 10.1038/35036093. URL: https://www.nature.com/articles/35036093 (visited on 16/11/2021).

[15] Maria A. Blasco. "Telomeres and Human Disease: Ageing, Cancer and Beyond". In: *Nature Reviews. Genetics* 6.8 (Aug. 2005), pp. 611–622. ISSN: 1471-0056. DOI: 10.1038/nrg1656. pmid: 16136653.

[16] Werner Risau. "Mechanisms of Angiogenesis". In: *Nature* 386.6626 (6626 Apr. 1997), pp. 671–674. ISSN: 1476-4687. DOI: 10.1038/386671a0. URL: https://www.nature.com/articles/386671a0 (visited on 24/11/2021).

[17] Douglas Hanahan and Judah Folkman. "Patterns and Emerging Mechanisms of the Angiogenic Switch during Tumorigenesis". In: *Cell* 86.3 (9th Aug. 1996), pp. 353–364. ISSN: 0092-8674. DOI: 10.1016/S0092-8674(00)80108-7. URL: https://www.sciencedirect.com/science/article/pii/S0092867400801087 (visited on 23/11/2021).

[18] Lawrence A. Loeb. "Mutator Phenotype May Be Required for Multistage Carcinogenesis". In: *Cancer Research* 51.12 (15th June 1991), pp. 3075–3079. ISSN: 0008-5472, 1538-7445. pmid: 2039987. URL: https://cancerres.aacrjournals.org/content/51/12/3075 (visited on 02/12/2021).

[19] Stephen P. Jackson and Jiri Bartek. "The DNA-damage Response in Human Biology and Disease". In: *Nature* 461.7267 (7267 Oct. 2009), pp. 1071–1078. ISSN: 1476-4687. DOI: 10.1038/nature08467. URL: https://www.nature.com/articles/nature08467 (visited on 02/12/2021).

[20] María Berdasco and Manel Esteller. "Aberrant Epigenetic Landscape in Cancer: How Cellular Identity Goes Awry". In: *Developmental Cell* 19.5 (16th Nov. 2010), pp. 698–711. ISSN: 1878-1551. DOI: 10.1016/j.devcel.2010.10.005. pmid: 21074720.

[21] Otto Warburg. "Investigations from the Kaiser Wilhelm Institute for Biology, Berlin-Dahlem". In: *London, Constable & Co. Ltd* (1930).

[22] Ralph J. DeBerardinis et al. "The Biology of Cancer: Metabolic Reprogramming Fuels Cell Growth and Proliferation". In: *Cell Metabolism* 7.1 (1st Jan. 2008), pp. 11–20. ISSN: 1550-4131. DOI: 10.1016/j.cmet.2007.10.002. pmid: 18177721. URL: https://www.cell.com/cell-metabolism/abstract/S1550-4131(07)00295-1 (visited on 03/12/2021).

[23] Chi V Dang and Gregg L Semenza. "Oncogenic Alterations of Metabolism". In: (1999), p. 5.

[24] Arvind Ramanathan, Connie Wang and Stuart L. Schreiber. "Perturbational Profiling of a Cell-Line Model of Tumorigenesis by Using Metabolic Measurements". In: *Proceedings of the National Academy of Sciences of the United States of America* 102.17 (26th Apr. 2005), pp. 5992–5997. ISSN: 0027-8424. DOI: 10.1073/pnas.0502267102. pmid: 15840712. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1087961/ (visited on 03/12/2021).

[25] Dass S. Vinay et al. "Immune Evasion in Cancer: Mechanistic Basis and Therapeutic Strategies". In: *Seminars in Cancer Biology.* A Broad-Spectrum Integrative Design for Cancer Prevention and Therapy 35 (1st Dec. 2015), S185–S198. ISSN: 1044-579X. DOI: 10.1016/j.semcancer.2015.03.004. URL: https://www.sciencedirect.com/science/article/pii/S1044579X1500019X (visited on 10/12/2021).

[26] Steven E Mutsaers. "The Mesothelial Cell". In: *The International Journal of Biochemistry & Cell Biology* 36.1 (Jan. 2004), pp. 9–16. ISSN: 13572725. DOI: 10.1016/S1357-2725(03)00242-5. URL: https://linkinghub.elsevier.com/retrieve/pii/S1357272503002425 (visited on 15/07/2021).

[27] Stefano Turini et al. "Epithelial to Mesenchymal Transition in Human Mesothelial Cells Exposed to Asbestos Fibers: Role of TGF-β as Mediator of Malignant Mesothelioma Development or Metastasis via EMT Event". In: *International Journal of Molecular Sciences* 20.1 (3rd Jan. 2019), E150. ISSN: 1422-0067. DOI: 10.3390/ijms20010150. pmid: 30609805.

[28] Alexandra Schramm et al. "Prognostic Significance of Epithelial-Mesenchymal Transition in Malignant Pleural Mesothelioma". In: *European Journal of Cardio-Thoracic Surgery: Official Journal of the European Association for Cardio-Thoracic Surgery* 37.3 (Mar. 2010), pp. 566–572. ISSN: 1873-734X. DOI: 10.1016/j.ejcts.2009.08.027. pmid: 19781955.

[29] Aliya N. Husain et al. "Guidelines for Pathologic Diagnosis of Malignant Mesothelioma: 2012 Update of the Consensus Statement from the International Mesothelioma Interest Group". In: *Archives of Pathology & Laboratory Medicine* 137.5 (28th Aug. 2012), pp. 647–667. ISSN: 0003-9985. DOI: 10.5858/arpa.2012-0214-OA. URL: https://doi.org/10.5858/arpa.2012-0214-OA (visited on 03/01/2022).

[30] Volker Neumann et al. "Malignant Pleural Mesothelioma: Incidence, Etiology, Diagnosis, Treatment, and Occupational Health". In: *Deutsches Arzteblatt International* 110.18 (May 2013), pp. 319–326. ISSN: 1866-0452. DOI: 10.3238/arztebl.2013.0319. pmid: 23720698.

[31] Robert Ryan Meyerhoff et al. "Impact of Mesothelioma Histologic Subtype on Outcomes in the Surveillance, Epidemiology, and End Results Database". In: *The Journal of Surgical Research* 196.1 (1st June 2015), pp. 23–32. ISSN: 1095-8673. DOI: 10.1016/j.jss.2015.01.043. pmid: 25791825.

[32] Joanna Obacz et al. "Biological Basis for Novel Mesothelioma Therapies". In: *British Journal of Cancer* 125.8 (8 Oct. 2021), pp. 1039–1055. ISSN: 1532-1827. DOI: 10.1038/s41416-021-01462-2. URL: https://www.nature.com/articles/s41416-021-01462-2 (visited on 31/12/2021).

[33] Michele Carbone et al. "Mesothelioma: Scientific Clues for Prevention, Diagnosis, and Therapy". In: *CA: A Cancer Journal for Clinicians* 69.5 (2019), pp. 402–429. ISSN: 1542-4863. DOI: 10.3322/caac.21572. URL: https://acsjournals.onlinelibrary.wiley.com/doi/abs/10.3322/caac.21572 (visited on 14/03/2021).

[34] K. Olofsson and J. Mark. "Specificity of Asbestos-Induced Chromosomal Aberrations in Short-Term Cultured Human Mesothelial Cells". In: *Cancer Genetics and Cytogenetics* 41.1 (Aug. 1989), pp. 33–39. ISSN: 0165-4608. DOI: 10.1016/0165-4608(89)90105-2. pmid: 2766251.

[35] Haining Yang et al. "TNF-alpha Inhibits Asbestos-Induced Cytotoxicity via a NF-kappaB-dependent Pathway, a Possible Mechanism for Asbestos-Induced Oncogenesis". In: *Proceedings of the National Academy of Sciences of the United States of America* 103.27 (5th July 2006), pp. 10397–10402. ISSN: 0027-8424. DOI: 10.1073/pnas.0604008103. pmid: 16798876.

[36] Michele Carbone and Haining Yang. "Molecular Pathways: Targeting Mechanisms of Asbestos and Erionite Carcinogenesis in Mesothelioma". In: *Clinical Cancer Research:*

An Official Journal of the American Association for Cancer Research 18.3 (1st Feb. 2012), pp. 598–604. ISSN: 1557-3265. DOI: 10.1158/1078-0432.CCR-11-2259. pmid: 22065079.

[37] An Xu et al. "Genotoxic Mechanisms of Asbestos Fibers: Role of Extranuclear Targets". In: *Chemical Research in Toxicology* 20.5 (May 2007), pp. 724–733. ISSN: 0893-228X. DOI: 10.1021/tx600364d. pmid: 17447795.

[38] Yaohe Wang et al. "Interleukin-1β and Tumour Necrosis Factor-α Promote the Transformation of Human Immortalised Mesothelial Cells by Erionite". In: *International Journal of Oncology* 25.1 (1st July 2004), pp. 173–178. ISSN: 1019-6439. DOI: 10.3892/ijo.25.1.173. URL: https://www.spandidos-publications.com/10.3892/ijo.25.1.173 (visited on 22/12/2021).

[39] Haining Yang et al. "Programmed Necrosis Induced by Asbestos in Human Mesothelial Cells Causes High-Mobility Group Box 1 Protein Release and Resultant Inflammation". In: *Proceedings of the National Academy of Sciences of the United States of America* 107.28 (13th July 2010), pp. 12611–12616. ISSN: 1091-6490. DOI: 10.1073/pnas.1006542107. pmid: 20616036.

[40] Joseph R. Testa et al. "Germline BAP1 Mutations Predispose to Malignant Mesothelioma". In: *Nature genetics* 43.10 (28th Aug. 2011), pp. 1022–1025. ISSN: 1061-4036. DOI: 10.1038/ng.912. pmid: 21874000. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3184199/ (visited on 10/11/2021).

[41] Thomas Wiesner et al. "Germline Mutations in BAP1 Predispose to Melanocytic Tumors". In: *Nature genetics* 43.10 (28th Aug. 2011), pp. 1018–1021. ISSN: 1061-4036. DOI: 10.1038/ng.910. pmid: 21874003. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3328403/ (visited on 10/11/2021).

[42] Yilei Zhang et al. "BAP1 Links Metabolic Regulation of Ferroptosis to Tumour Suppression". In: *Nature Cell Biology* 20.10 (Oct. 2018), pp. 1181–1192. ISSN: 1476-4679. DOI: 10.1038/s41556-018-0178-0. pmid: 30202049.

[43] Sandra Pastorino et al. "A Subset of Mesotheliomas With Improved Survival Occurring in Carriers of BAP1 and Other Germline Mutations". In: *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* (30th Oct. 2018), JCO2018790352. ISSN: 1527-7755. DOI: 10.1200/JCO.2018.79.0352. pmid: 30376426.

[44] Jesse Gillis and Paul Pavlidis. ""Guilt by Association" Is the Exception Rather Than the Rule in Gene Networks". In: *PLoS Computational Biology* 8.3 (29th Mar. 2012), e1002444. ISSN: 1553-734X. DOI: 10.1371/journal.pcbi.1002444. pmid: 22479173. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3315453/ (visited on 11/01/2022).

[45] Shane Neph et al. "Circuitry and Dynamics of Human Transcription Factor Regulatory Networks". In: *Cell* 150.6 (14th Sept. 2012), pp. 1274–1286. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2012.04.040. pmid: 22959076. URL: https://www.cell.com/cell/abstract/S0092-8674(12)00639-3 (visited on 04/01/2022).

[46] David Amar, Hershel Safer and Ron Shamir. "Dissection of Regulatory Networks That Are Altered in Disease via Differential Co-expression". In: *PLOS Computational Biology* 9.3 (7th Mar. 2013), e1002955. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1002955. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002955 (visited on 07/04/2021).

[47] Emma Pierson et al. "Sharing and Specificity of Co-expression Networks across 35 Human Tissues". In: *PLOS Computational Biology* 11.5 (13th May 2015), e1004220. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1004220. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004220 (visited on 04/01/2022).

[48] Zhigang Xue et al. "Genetic Programs in Human and Mouse Early Embryos Revealed by Single-Cell RNA Sequencing". In: *Nature* 500.7464 (7464 Aug. 2013), pp. 593–597. ISSN: 1476-4687. DOI: 10.1038/nature12364. URL: https://www.nature.com/articles/nature12364 (visited on 04/01/2022).

[49] Sipko van Dam et al. "Gene Co-Expression Analysis for Functional Classification and Gene–Disease Predictions". In: *Briefings in Bioinformatics* 19.4 (20th July 2018), pp. 575–592. ISSN: 1477-4054. DOI: 10.1093/bib/bbw139. URL: https://doi.org/10.1093/bib/bbw139 (visited on 01/04/2021).

[50] Peter Langfelder and Steve Horvath. "WGCNA: An R Package for Weighted Correlation Network Analysis". In: *BMC Bioinformatics* 9.1 (29th Dec. 2008), p. 559. ISSN: 1471-2105. DOI: 10.1186/1471-2105-9-559. URL: https://doi.org/10.1186/1471-2105-9-559 (visited on 02/04/2021).

[51] Wei Zhao et al. "Weighted Gene Coexpression Network Analysis: State of the Art". In: *Journal of Biopharmaceutical Statistics* 20.2 (19th Mar. 2010), pp. 281–300. ISSN: 1054-3406. DOI: 10.1080/10543400903572753. pmid: 20309759. URL: https://doi.org/10.1080/10543400903572753 (visited on 06/04/2021).

[52] Wilberforce Zachary Ouma, Katja Pogacar and Erich Grotewold. "Topological and Statistical Analyses of Gene Regulatory Networks Reveal Unifying yet Quantitatively Different Emergent Properties". In: *PLoS Computational Biology* 14.4 (30th Apr. 2018), e1006098. ISSN: 1553-734X. DOI: 10.1371/journal.pcbi.1006098. pmid: 29708965. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5945062/ (visited on 05/01/2022).

[53] E. Ravasz et al. "Hierarchical Organization of Modularity in Metabolic Networks". In: *Science* (30th Aug. 2002). DOI: 10.1126/science.1073374. URL: https://www.science.org/doi/abs/10.1126/science.1073374 (visited on 07/01/2022).

[54] Réka Albert, Hawoong Jeong and Albert-László Barabási. "Error and Attack Tolerance of Complex Networks". In: *Nature* 406.6794 (6794 July 2000), pp. 378–382. ISSN: 1476-4687. DOI: 10.1038/35019019. URL: https://www.nature.com/articles/35019019 (visited on 07/01/2022).

[55] Peter Langfelder, Bin Zhang and Steve Horvath. "Defining Clusters from a Hierarchical Cluster Tree: The Dynamic Tree Cut Package for R". In: *Bioinformatics* 24.5 (1st Mar. 2008), pp. 719–720. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btm563. URL: https://doi.org/10.1093/bioinformatics/btm563 (visited on 10/01/2022).

[56] Steve Horvath and Jun Dong. "Geometric Interpretation of Gene Coexpression Network Analysis". In: *PLOS Computational Biology* 4.8 (15th Aug. 2008), e1000117. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1000117. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000117 (visited on 11/01/2022).

[57] Maxime Rotival et al. "Integrating Genome-Wide Genetic Variations and Monocyte Expression Data Reveals Trans-Regulated Gene Modules in Humans". In: *PLOS Genetics* 7.12 (1st Dec. 2011), e1002367. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1002367. URL: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1002367 (visited on 12/01/2022).

[58] Mark B. Gerstein et al. "Architecture of the Human Regulatory Network Derived from ENCODE Data". In: *Nature* 489.7414 (7414 Sept. 2012), pp. 91–100. ISSN: 1476-4687.

DOI: 10.1038/nature11245. URL: https://www.nature.com/articles/nature11245 (visited on 17/01/2022).

[59] Daniel Marbach et al. "Tissue-Specific Regulatory Circuits Reveal Variable Modular Perturbations across Complex Diseases". In: *Nature Methods* 13.4 (4 Apr. 2016), pp. 366–370. ISSN: 1548-7105. DOI: 10.1038/nmeth.3799. URL: https://www.nature.com/articles/nmeth.3799 (visited on 20/04/2021).

[60] Denise N. Slenter et al. "WikiPathways: A Multifaceted Pathway Database Bridging Metabolomics to Other Omics Research". In: *Nucleic Acids Research* 46.D1 (4th Jan. 2018), pp. D661–D667. ISSN: 1362-4962. DOI: 10.1093/nar/gkx1064. pmid: 29136241.

[61] Raphael Bueno et al. "Comprehensive Genomic Analysis of Malignant Pleural Mesothelioma Identifies Recurrent Mutations, Gene Fusions and Splicing Alterations". In: *Nature Genetics* 48.4 (Apr. 2016), pp. 407–416. ISSN: 1061-4036, 1546-1718. DOI: 10.1038/ng.3520. URL: http://www.nature.com/articles/ng.3520 (visited on 08/04/2021).

[62] Mallory Ann Freeberg et al. "The European Genome-phenome Archive in 2021". In: *Nucleic Acids Research* 50.D1 (7th Jan. 2022), pp. D980–D987. ISSN: 0305-1048. DOI: 10.1093/nar/gkab1059. URL: https://doi.org/10.1093/nar/gkab1059 (visited on 28/01/2022).

[63] Nicolas L. Bray et al. "Near-Optimal Probabilistic RNA-seq Quantification". In: *Nature Biotechnology* 34.5 (5 May 2016), pp. 525–527. ISSN: 1546-1696. DOI: 10.1038/nbt.3519. URL: https://www.nature.com/articles/nbt.3519 (visited on 08/03/2022).

[64] A. Hyvarinen. "Fast and Robust Fixed-Point Algorithms for Independent Component Analysis". In: *IEEE Transactions on Neural Networks* 10.3 (May 1999), pp. 626–634. ISSN: 1941-0093. DOI: 10.1109/72.761722.

[65] J L Marchini, C Heaton and B D Ripley. *fastICA: FastICA Algorithms to Perform ICA and Projection Pursuit*. manual. 2021. URL: https://CRAN.R-project.org/package=fastICA.

[66] Bernd Klaus and Korbinian Strimmer. *Fdrtool: Estimation of (Local) False Discovery Rates and Higher Criticism*. manual. 2021. URL: https://CRAN.R-project.org/package=fdrtool.

[67] Tianzhi Wu et al. "clusterProfiler 4.0: A Universal Enrichment Tool for Interpreting Omics Data". In: *The Innovation* 2.3 (28th Aug. 2021), p. 100141. ISSN: 2666-6758. DOI: 10.1016/j.xinn.2021.100141. URL: https://www.sciencedirect.com/science/article/pii/S2666675821000667 (visited on 19/01/2022).

[68] Marc Carlson. *Org.Hs.Eg.Db: Genome Wide Annotation for Human.* manual. 2021.

[69] Yoav Benjamini and Yosef Hochberg. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 57.1 (1995), pp. 289–300. ISSN: 0035-9246. JSTOR: 2346101. URL: https://www.jstor.org/stable/2346101 (visited on 21/01/2022).

[70] Susmita Datta and Somnath Datta. "Methods for Evaluating Clustering Algorithms for Gene Expression Data Using a Reference Set of Functional Classes". In: *BMC Bioinformatics* 7.1 (31st Aug. 2006), p. 397. ISSN: 1471-2105. DOI: 10.1186/1471-2105-7-397. URL: https://doi.org/10.1186/1471-2105-7-397 (visited on 19/04/2021).

[71] Andreas Schlicker et al. "A New Measure for Functional Similarity of Gene Products Based on Gene Ontology". In: *BMC Bioinformatics* 7.1 (15th June 2006), p. 302. ISSN: 1471-2105. DOI: 10.1186/1471-2105-7-302. URL: https://doi.org/10.1186/1471-2105-7-302 (visited on 07/04/2022).

[72] Mark Koudritsky and Eytan Domany. "Positional Distribution of Human Transcription Factor Binding Sites". In: *Nucleic Acids Research* 36.21 (1st Dec. 2008), pp. 6795–6805. ISSN: 0305-1048. DOI: 10.1093/nar/gkn752. URL: https://doi.org/10.1093/nar/gkn752 (visited on 14/12/2023).

[73] Alexei A. Sharov, Yuhki Nakatake and Weidong Wang. "Atlas of Regulated Target Genes of Transcription Factors (ART-TF) in Human ES Cells". In: *BMC Bioinformatics* 23.1 (16th Sept. 2022), p. 377. ISSN: 1471-2105. DOI: 10.1186/s12859-022-04924-3. URL: https://doi.org/10.1186/s12859-022-04924-3 (visited on 29/08/2024).

[74] Wouter Saelens, Robrecht Cannoodt and Yvan Saeys. "A Comprehensive Evaluation of Module Detection Methods for Gene Expression Data". In: *NATURE COMMUNICATIONS* (2018), p. 12.

[75] Daniela Ledezma-Tejeida et al. "Limits to a Classic Paradigm: Most Transcription Factors in E. Coli Regulate Genes Involved in Multiple Biological Processes". In: *Nucleic Acids*

Research 47.13 (26th July 2019), pp. 6656–6667. ISSN: 0305-1048. DOI: 10.1093/nar/gkz525. URL: https://doi.org/10.1093/nar/gkz525 (visited on 29/08/2024).

[76] Ulykbek Kairov et al. "Determining the Optimal Number of Independent Components for Reproducible Transcriptomic Data Analysis". In: *BMC Genomics* 18.1 (11th Sept. 2017), p. 712. ISSN: 1471-2164. DOI: 10.1186/s12864-017-4112-9. URL: https://doi.org/10.1186/s12864-017-4112-9 (visited on 13/01/2022).

[77] Peter Langfelder and Steve Horvath. "Eigengene Networks for Studying the Relationships between Co-Expression Modules". In: *BMC Systems Biology* 1.1 (21st Nov. 2007), p. 54. ISSN: 1752-0509. DOI: 10.1186/1752-0509-1-54. URL: https://doi.org/10.1186/1752-0509-1-54 (visited on 03/07/2025).

[78] Yuna Blum et al. "Dissecting Heterogeneity in Malignant Pleural Mesothelioma through Histo-Molecular Gradients for Clinical Applications". In: *Nature Communications* 10.1 (22nd Mar. 2019), p. 1333. ISSN: 2041-1723. DOI: 10.1038/s41467-019-09307-6. URL: https://www.nature.com/articles/s41467-019-09307-6 (visited on 27/06/2025).

[79] Lara Perryman and Steven G. Gray. "Fibrosis in Mesothelioma: Potential Role of Lysyl Oxidases". In: *Cancers* 14.4 (15th Feb. 2022), p. 981. ISSN: 2072-6694. DOI: 10.3390/cancers14040981. pmid: 35205728. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8870010/ (visited on 06/02/2025).

[80] Yoshitaka Sekido. "Targeting the Hippo Pathway Is a New Potential Therapeutic Modality for Malignant Mesothelioma". In: *Cancers* 10.4 (22nd Mar. 2018), p. 90. ISSN: 2072-6694. DOI: 10.3390/cancers10040090. pmid: 29565815. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5923345/ (visited on 03/07/2025).

[81] Si Yu et al. "MCMs in Cancer: Prognostic Potential and Mechanisms". In: *Analytical Cellular Pathology (Amsterdam)* 2020 (3rd Feb. 2020), p. 3750294. ISSN: 2210-7177. DOI: 10.1155/2020/3750294. pmid: 32089988. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7023756/ (visited on 06/07/2023).

[82] Yaoqi Sun, Zhongping Cheng and Shupeng Liu. "MCM2 in Human Cancer: Functions, Mechanisms, and Clinical Significance". In: *Molecular Medicine* 28.1 (27th Oct. 2022), p. 128. ISSN: 1528-3658. DOI: 10.1186/s10020-022-00555-9. URL: https://doi.org/10.1186/s10020-022-00555-9 (visited on 06/07/2023).

[83] Yukun Li et al. “Systemic Analysis of the DNA Replication Regulator MCM Complex in Ovarian Cancer and Its Prognostic Value”. In: *Frontiers in Oncology* 11 (9th June 2021), p. 681261. ISSN: 2234-943X. DOI: 10.3389/fonc.2021.681261. pmid: 34178669. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8220296/ (visited on 06/07/2023).

[84] Lindsey N. Kent and Gustavo Leone. “The Broken Cycle: E2F Dysfunction in Cancer”. In: *Nature Reviews Cancer* 19.6 (June 2019), pp. 326–338. ISSN: 1474-175X, 1474-1768. DOI: 10.1038/s41568-019-0143-7. URL: https://www.nature.com/articles/s41568-019-0143-7 (visited on 06/07/2023).

A Supplementary Figures

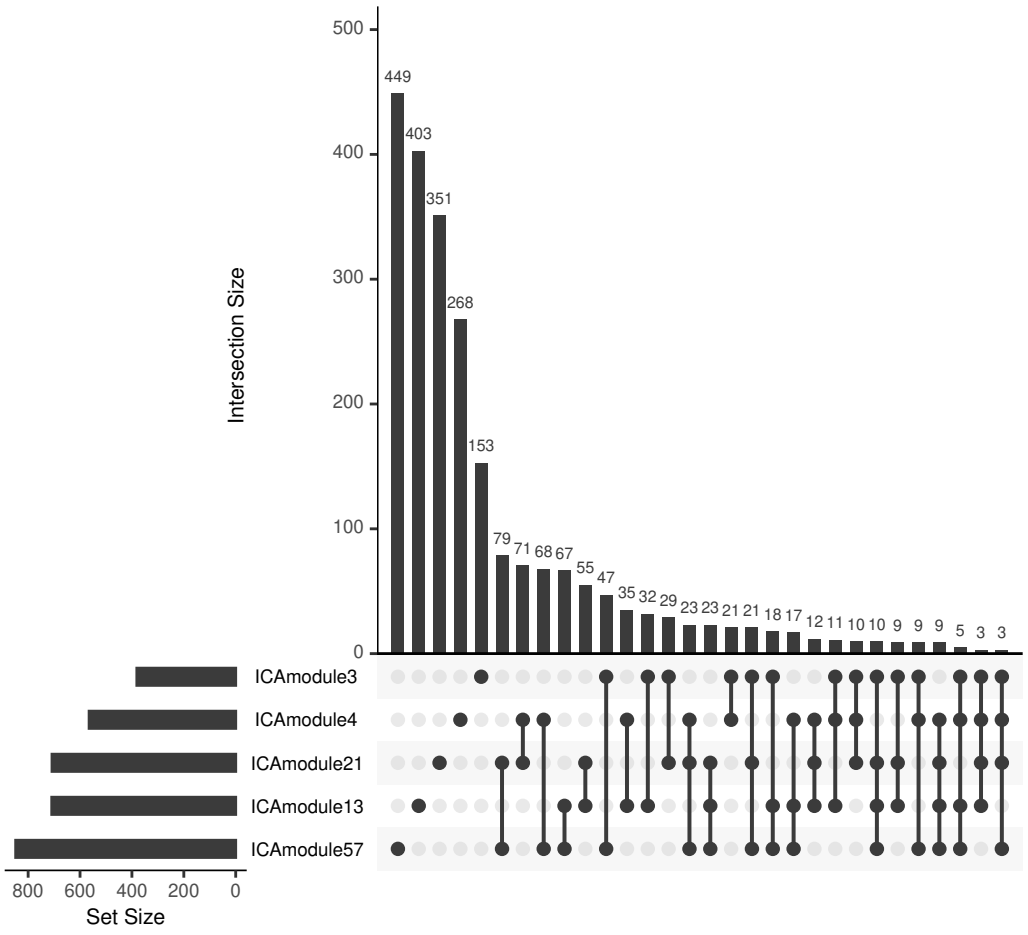

Figure S1.: Upset plot of ICA modules with similar expression patterns showing their overlap of member genes.

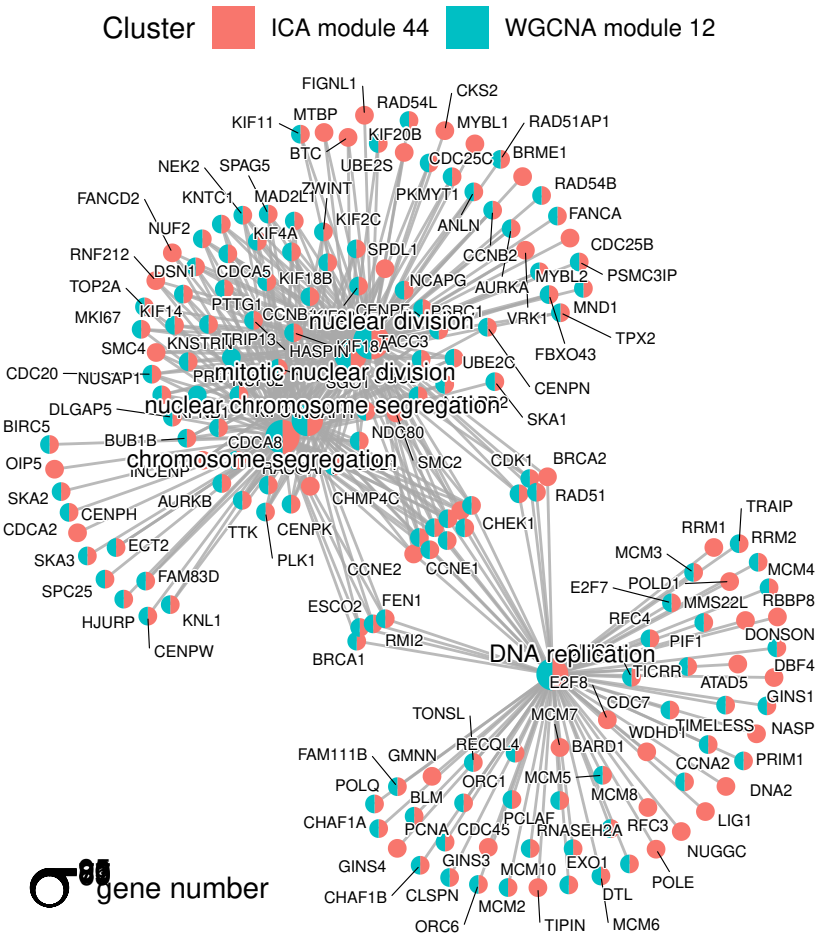

Figure S2.: Gene-concept network with biological process GO terms of overlapping genes between ICA module 44 and WGCNA module 12 (tan).

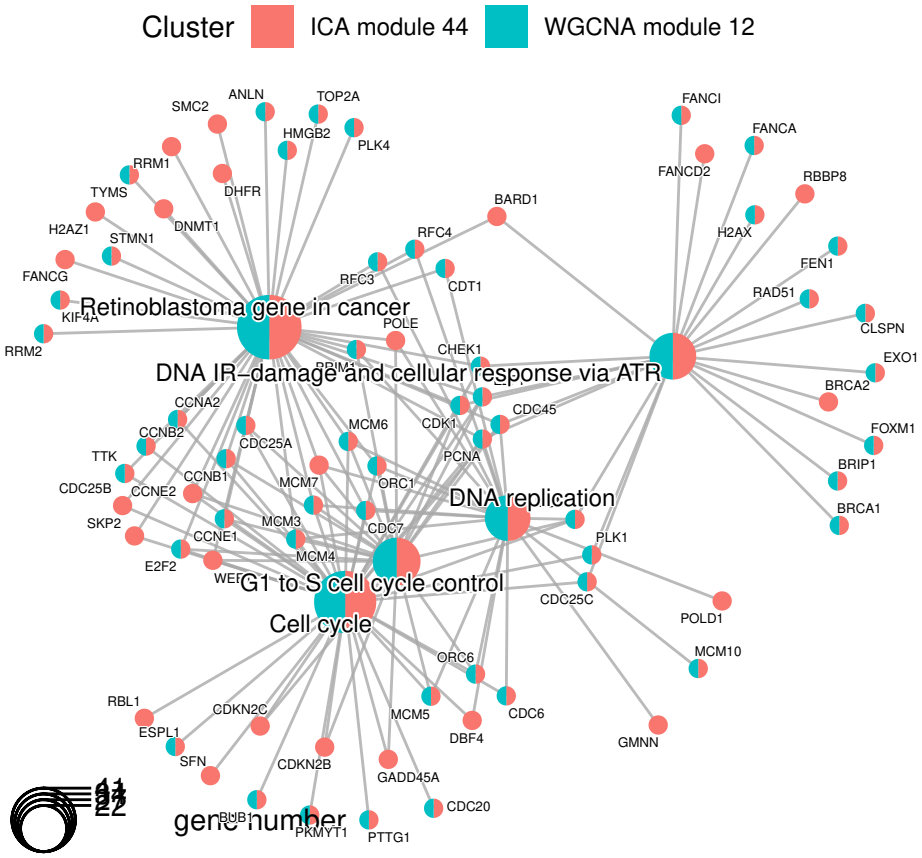

Figure S3.: Gene-concept network with WikiPathway terms of overlapping genes between ICA module 44 and WGCNA module 12 (tan).