Maturitätsarbeit an der Kantonsschule Zürich Nord

Immunregulation von T-Zellen durch Vitamin D3

Eine Netzwerkanalyse

Sofia Surace
M6e

Zürich, Dezember 2025

Betreuer: Patrick Aschwanden

Abstract

Vor dem Hintergrund potenzieller Therapieansätze für **Rheumatoide Arthritis** untersucht diese Arbeit, ob Vitamin D3 das Verhalten von Immunzellen gezielt beeinflussen kann. Im Mittelpunkt steht die Frage, ob dendritische Zellen nach einer Vitamin-D3-Behandlung die Genregulation humaner $CD4^+$-T-Zellen messbar verändern.

Zu diesem Zweck wurden Vitamin-D3-behandelte tolerogene dendritische Zellen (VitD3-tolDCs) eingesetzt und deren Einfluss auf die Genexpression der T-Zellen analysiert. Zur Konstruktion der Gen-Koexpressionsnetzwerke aus RNA-Seq-Daten (GEO: GSE128816) wurde der Algorithmus ***bc3net*** verwendet (100 Bootstraps, Estimator: Spearman). Anschliessend wurden Unterschiede zwischen Behandlungs- und Kontrollgruppe mit ***DiffCoEx*** identifiziert. Wegen fehlender permutationsbasierter Signifikanztests sind die Ergebnisse als explorativ zu interpretieren. Dennoch generieren sie klare Hypothesen, die experimentell validiert werden sollten.

Die Ergebnisse zeigen spezifisch veränderte Gen-Module, unter anderem im VEGF-Signalweg, in der T-Zell-Rezeptor-Bindung sowie in Transkriptionskomplexen. Dies deutet auf eine **antiinflammatorische Modulation** der T-Zellen hin und stützt die Hypothese, dass Vitamin D3 über dendritische Zellen regulierend in die Immunantwort eingreifen kann.

Inhaltsverzeichnis

0 Vorwort

Mein Interesse an der Medizin, insbesondere an Autoimmunerkrankungen, ist persönlich motiviert. Die Rheumatoide Arthritis meiner Mutter und die Debatte um komplementäre Therapieansätze rückten Vitamin D3 in meinen Fokus. Diese Arbeit entsprang dem Wunsch, dessen vieldiskutierte, aber oft unklare Wirkung wissenschaftlich fundiert zu untersuchen.

Da mir im Rahmen einer Maturaarbeit die Mittel für experimentelle Laborforschung fehlten, wählte ich einen bioinformatischen Ansatz. Ziel war es, meine Programmierkenntnisse zu nutzen, um die potenziellen Effekte von Vitamin D3 auf Genexpression zu modellieren und zu analysieren.

Die Einarbeitung in die komplexe Methodik der Netzwerkanalyse war herausfordernd. Umso dankbarer bin ich für die wertvolle Unterstützung, die ich erhalten habe:

Ich möchte mich bei Prof. Dr. Dr. Caroline Ospelt bedanken, die mir einen wissenschaftlich fundierten Weg für die Analyse aufzeigte, sowie bei Prof. Dr. Amedeo Caflisch für den Einblick in verschiedene Ansätze der biologischen Simulation.

Ein ganz besonderer Dank gebührt Dr. Izaskun Mallona Gonzalez. Sie war meine wichtigste Stütze während der anspruchsvollen Erstellung und Analyse der Netzwerke. Ihre Bereitschaft, die Stärken und Schwächen der Methodik offen zu diskutieren, und unsere intellektuell anregenden Gespräche waren für das Gelingen dieser Arbeit von unschätzbarem Wert.

1 Einleitung

Das menschliche Immunsystem muss pathogenbedingte Bedrohungen neutralisieren und gleichzeitig körpereigene Strukturen tolerieren. Eine Störung dieses Gleichgewichts kann zu Autoimmunerkrankungen wie Multipler Sklerose oder Rheumatoider Arthritis führen. Die Erforschung von zellulären und molekularen Mechanismen, die **immunologische Toleranz** fördern, ist daher von zentraler Bedeutung.

Dendritische Zellen (DCs) steuern die Aktivierung und Differenzierung von T-Zellen und können entzündungsfördernde oder -hemmende Signale vermitteln. Unter Einfluss von Vitamin D3 (VitD3) lassen sich **tolerogene dendritische Zellen** (tolDCs) erzeugen, die regulatorische T-Zellen induzieren und somit entzündungshemmende Prozesse fördern. Obwohl die **immunsuppressiven** Effekte von VitD3 vielfach beschrieben wurden, werden die zugrunde liegenden Signalwege zwischen tolDCs und naiven T-Helferzellen ($CD4^+$-T-Zellen) noch unzureichend verstanden.

Moderne Transkriptom- und Netzwerkmethoden ermöglichen es, Genregulationsmechanismen systematisch zu analysieren. Koexpressionsnetzwerke erlauben Rückschlüsse auf **funktionell verwandte Gene**, während differenzielle Koexpressionsanalysen Veränderungen zwischen experimentellen Bedingungen sichtbar machen. Dadurch können modulare Strukturen identifiziert werden, die auf **zentrale Signalwege** oder regulatorische Knotenpunkte hinweisen.

Da aus Rechenzeitgründen keine permutationsbasierte Signifikanzprüfung durchgeführt werden konnte, sind die Ergebnisse als **explorativ** zu betrachten. Dennoch soll die Analyse dazu beitragen, Hypothesen über immunmodulatorische Prozesse und potenzielle Zielgene zu generieren, die in zukünftigen Studien experimentell überprüft werden können.

Diese Arbeit verfolgt drei Ziele:

1. Erstellung eines Genkoexpressionsnetzwerks für T-Zellen, die mit VitD3-behandelten tolDCs (VitD3-tolDCs) oder mit reifen immunogenen dendritischen Zellen (mDCs) ko-kultiviert wurden (*bc3net)*.
2. Identifikation von Modulen, die ihre Koexpression auf ähnliche Art ändern (*DiffCoEx)*.
3. Funktionelle Einordnung der Module im Kontext immunologischer Toleranzmechanismen (ORA).

Kapitel 2 beschreibt die biologischen und methodischen Grundlagen. Kapitel 3 erläutert Datensatz und Analyseverfahren. Kapitel 4 präsentiert die Ergebnisse, Kapitel 5 diskutiert sie im Forschungskontext und benennt Limitationen. Kapitel 6 fasst die wichtigsten Erkenntnisse zusammen und skizziert weitere Forschungsansätze.

2 Theoretische Grundlagen

2.1 Immunologie

2.1.1 Grundlagen des Immunsystems

Das Immunsystem ist das biologische Abwehrsystem des Körpers, das ihn vor schädlichen Einflüssen wie krankheitserregenden Mikroorganismen (Pathogenen) und abnormalen körpereigenen Zellen, beispielsweise Tumorzellen, schützt [1], [2]. Seine Hauptaufgabe ist es, die Unversehrtheit des Organismus zu gewährleisten. Man unterscheidet grundsätzlich zwischen zwei eng miteinander verknüpften Armen des Immunsystems: dem angeborenen und dem erworbenen (oder adaptiven) Immunsystem.

- **Das angeborene Immunsystem** agiert als erste Verteidigungslinie. Es reagiert sofort und unspezifisch[1] auf eine Bedrohung, kann aber keine langanhaltende Immunität aufbauen.
- **Das adaptive Immunsystem** reagiert langsamer, dafür aber hochspezifisch auf bestimmte Erreger. Eine seiner wichtigsten Eigenschaften ist die Fähigkeit, ein immunologisches Gedächtnis zu bilden, was bei einem erneuten Kontakt mit demselben Pathogen eine schnellere und stärkere Abwehrreaktion ermöglicht.

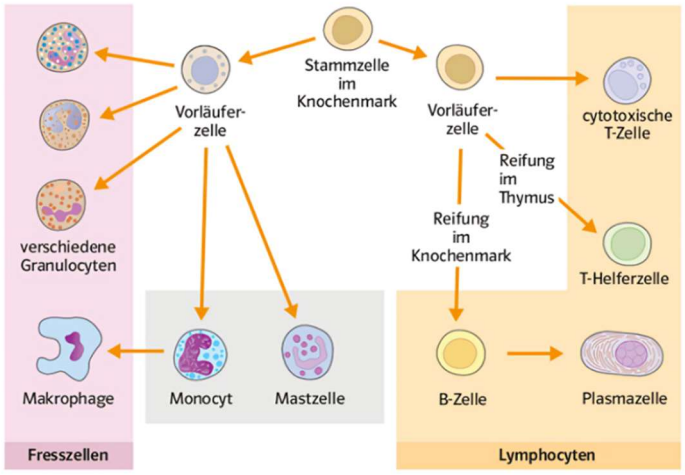

Abbildung 1: Stammbaum der Leukozyten [2]
Aus Stammzellen im Knochenmark können sich entweder lymphatisch Lymphozyten oder myeloisch verschiedene Fresszellen entwickeln.

Alle Zellen, die an diesen Abwehrprozessen beteiligt sind, haben einen gemeinsamen Ursprung: die blutbildenden (hämatopoetischen) Stammzellen im Knochenmark. Diese Stammzellen entwickeln sich entlang zweier Hauptlinien weiter (Abb. 1): der myeloischen und der lymphatischen Linie. Aus der **lymphatischen Vorläuferzelle** entstehen unter anderem die B- und T-Lymphozyten (B- und T-Zellen), welche die zentralen Akteure des adaptiven Immunsystems sind. Aus der **myeloischen Vorläuferzelle** gehen unter anderem Monozyten, Granulozyten und dendritische Zellen (Fresszellen) hervor, die wichtige Funktionen im angeborenen Immunsystem übernehmen.

2.1.2 Dendritische Zellen: die Dirigenten der Immunantwort

Eine besondere Rolle im Immunsystem nehmen die sogenannten **antigenpräsentierenden Zellen (APCs)** ein. Zu ihnen gehören Makrophagen, B-Lymphozyten und insbesondere die dendritischen Zellen. Sie fungieren als Brücke zwischen der angeborenen und der adaptiven Immunantwort. Ihre Aufgabe ist es, Pathogene aufzunehmen, sie in kleinere Stücke (**Antigene**) zu

[1] In diesem Kontext bedeutet unspezifisch, dass es gegen jede erkannte Bedrohung vorgeht und sich z.B. nicht auf ein bestimmten Erreger spezialisiert hat.

zerlegen und diese Antigene den Zellen des adaptiven Immunsystems – vor allem den T-Zellen – zu präsentieren.

Dendritische Zellen (DCs) gelten als die effektivsten APCs. Man findet sie zunächst in einem unreifen Zustand in verschiedenen Geweben des Körpers, wo sie auf potenzielle Gefahren, wie beispielsweise Tumorzellen oder Pathogene, lauern. Treffen diese **unreifen dendritischen Zellen (iDCs)** auf Gefahrensignale, wie sie von Bakterien oder geschädigten Körperzellen ausgehen, beginnen sie zu reifen.

Dieser Reifungsprozess führt zu zwei grundlegend unterschiedlichen Typen von dendritischen Zellen, die die nachfolgende Immunreaktion in gegensätzliche Richtungen lenken (Abb. 2):

- **Reife, immunogene dendritische Zellen (mDCs)**: Dies ist der «Standardweg» bei einer Infektion. Die reifende Zelle exprimiert vermehrt Moleküle an ihrer Oberfläche, die für die Aktivierung von T-Zellen notwendig sind (z.B. CD80, CD86, MHC-II). Sie wandert in die nächstgelegenen Lymphknoten und schüttet **pro-inflammatorische** (entzündungsfördernde) Botenstoffe wie das Zytokin Interleukin-12 (IL-12) aus. Dort aktivieren diese Zytokine naive (inaktive) T-Zellen und lösen so eine starke, auf Angriff ausgerichtete Immunantwort aus.
- **Tolerogene dendritische Zellen (tolDCs)**: Unter dem Einfluss bestimmter Substanzen entwickeln sich aus unreifen DCs keine immunogene, sondern tolerogene Zellen. Zu diesen Substanzen gehören unter anderem **VitD3** [3], Vitamin A und das Zytokin TGF-β. Tolerogene DCs zeichnen sich dadurch aus, dass sie nur wenige der aktivierenden Oberflächenmoleküle tragen und stattdessen **anti-inflammatorische** (entzündungshemmende) Zytokine wie Interleukin-10 (IL-10) produzieren. Ihre Hauptfunktion besteht nicht darin, eine Abwehrreaktion auszulösen, sondern diese gezielt zu unterdrücken. Dadurch verhindern sie, dass das Immunsystem in gesundem Gewebe unnötig oder gegen harmlose körpereigene Strukturen reagiert (**periphere Toleranz**).

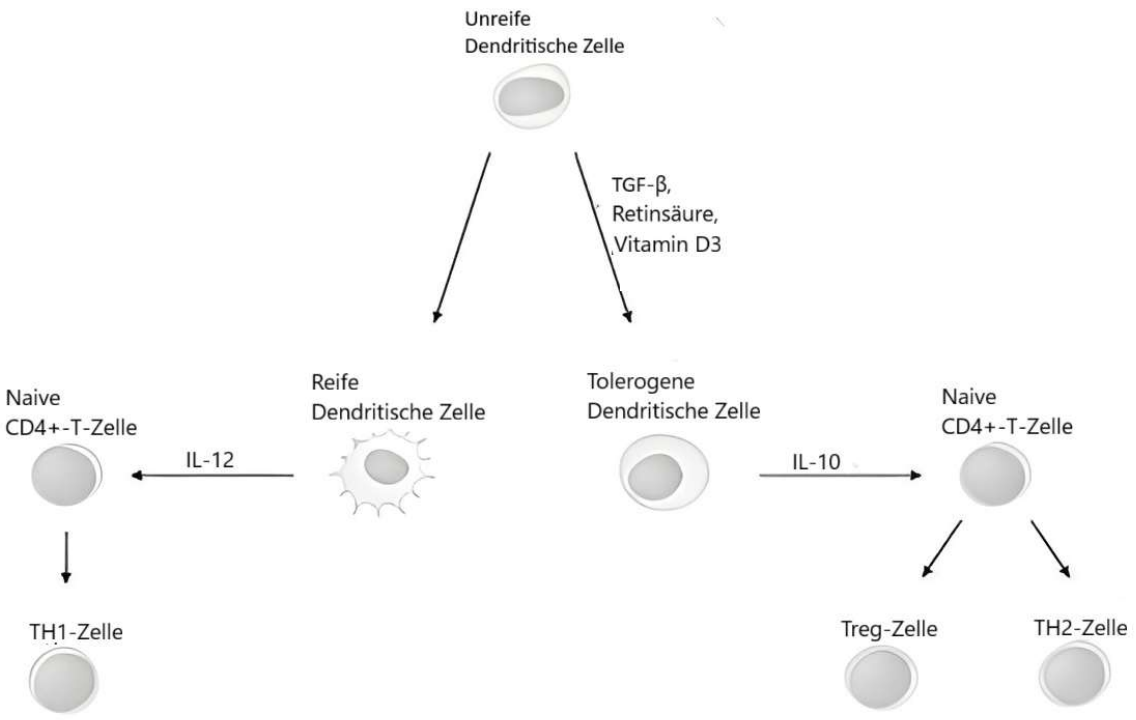

Abbildung 2: Einfluss dendritischer Zellen auf die T-Zell-Differenzierung [1], abgeänderte Abbildung
Reife DCs sezernieren u.a. das pro-inflammatorische Zytokin IL-12, welches die Differenzierung von inflammatorisch wirkenden TH1-Zellen begünstigt. Durch den Einfluss von TGF-β, Retinsäure oder VitD3 können aus unreifen DCs tolerogene DCs entstehen, die vor allem über die Sekretion des antiinflammatorischen IL-10 die Differenzierung von T_{reg}- sowie TH2-Zellen induzieren.

2.1.3 T-Zellen und die Balance zwischen Angriff und Toleranz

Naive T-Helferzellen ($CD4^+$-T-Zellen) sind die zentralen Schaltstellen der adaptiven Immunantwort. Je nachdem, welche Signale sie von einer dendritischen Zelle erhalten, entwickeln sie sich in unterschiedliche Subpopulationen mit spezialisierten Aufgaben. Für das Verständnis von Autoimmunerkrankungen und deren potenzieller Behandlung sind vor allem zwei Entwicklungspfade von Bedeutung:

Der pro-inflammatorische Pfad: Präsentiert eine mDC ein Antigen und schüttet dabei IL-12 aus, differenziert die naive T-Zelle zu einer **T-Helferzelle vom Typ 1 (TH1)**. TH1-Zellen koordinieren eine aggressive Immunantwort, indem sie selbst **entzündungsfördernde Zytokine** wie Interferon-γ (IFN-γ) und Tumornekrosefaktor-α (TNF-α) produzieren, die zur Eliminierung von Pathogenen oder infizierten Zellen beitragen. Bei Autoimmunerkrankungen wie der Rheumatoiden Arthritis ist dieser Pfad überaktiv und richtet sich fälschlicherweise gegen körpereigenes Gewebe.

Der anti-inflammatorische Pfad: Präsentiert hingegen eine tolDC ein Antigen und schüttet dabei IL-10 aus, wird die Entwicklung der naiven T-Zelle in eine gänzlich andere Richtung gelenkt. Aus ihr entsteht eine **regulatorische T-Zelle (T_{reg})**. T_{reg}-Zellen spielen eine zentrale Rolle bei der Unterdrückung fehlgeleiteter Immunreaktionen und verhindern damit **Autoimmunität**, also Angriffe des Immunsystems auf körpereigenes Gewebe. Sie sichern diese **Selbsttoleranz**, indem

sie selbst die anti-inflammatorischen Zytokine IL-10 und TGF-β freisetzen und die Aktivität anderer Immunzellen, einschliesslich der TH1-Zellen, dämpfen.

2.1.4 Autoimmunerkrankungen und Rheumatoide Arthritis

Während ein ausbalanciertes Immunsystem zwischen schädlichen und körpereigenen Strukturen unterscheiden kann, ist diese Fähigkeit bei **Autoimmunerkrankungen** gestört. Die Rheumatoide Arthritis (RA) ist ein Beispiel für eine solche **immunvermittelte entzündliche Erkrankung** (IMID): Hier verschiebt sich das im vorigen Kapitel beschriebene Gleichgewicht von einer kontrollierten, toleranten Immunantwort hin zu einem selbsterhaltenden, pro-inflammatorischen Zustand, der sich fälschlicherweise **gegen den eigenen Körper** richtet.

RA ist eine systemische Autoimmunerkrankung, die sich vor allem durch die **Entzündung der Gelenkinnenhaut** (Synovium) auszeichnet. Dies führt zu den charakteristischen Symptomen wie Schmerzen, Schwellungen und Steifheit in den Gelenken, insbesondere in Händen, Füssen und Knien. Unbehandelt schreitet die Entzündung fort und führt zur **unumkehrbaren Zerstörung** von Knorpel und Knochen, was in Deformitäten und starker funktioneller Beeinträchtigung resultiert. Obwohl die genauen Auslöser der RA noch nicht vollständig geklärt sind, geht man von einem Zusammenspiel aus genetischer Veranlagung und umweltbedingten Faktoren aus.

2.1.5 Pathologische Prozesse im Gelenk eines RA-Patienten

Im gesunden Zustand ist das **Synovium** eine dünne Zellschicht, die das Gelenk auskleidet, den Knorpel mit Nährstoffen versorgt und Gelenkschmiere[2] produziert [4]. Bei der Rheumatoiden Arthritis verändert sich diese Struktur durch folgende Prozesse dramatisch (Abb. 3) [5]:

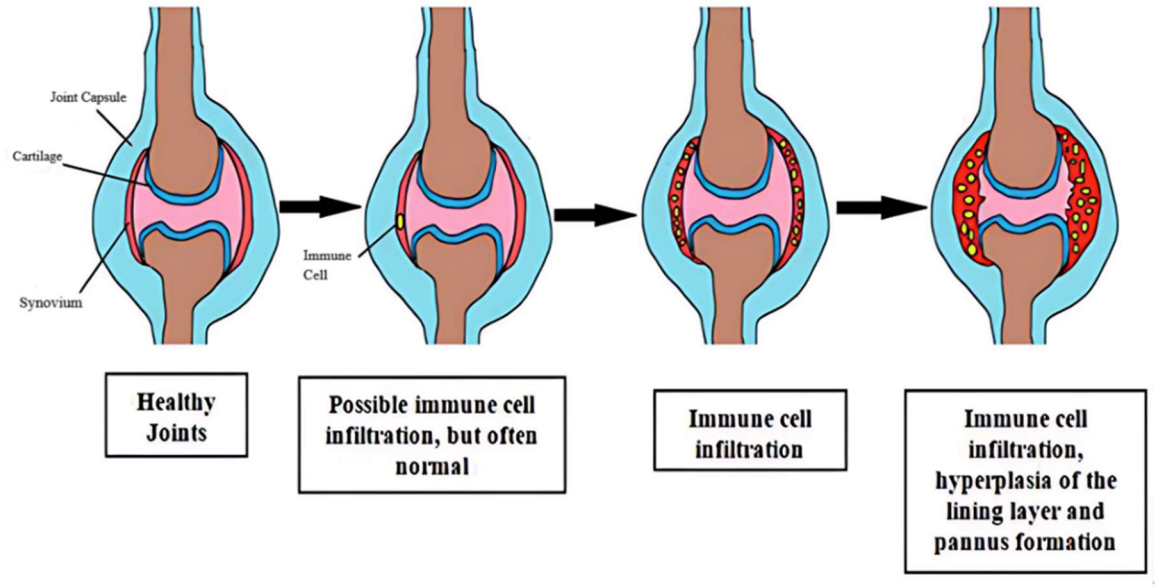

Abbildung 3: Pathophysiologie von Rheumatoider Arthritis [5], abgeänderte Abbildung
RA wird durch eine Immunzell-Infiltration eine Hyperplasie der Gelenkinnenhaut ausgelöst und es bildet sich ein Pannus.

Infiltration von Immunzellen: Das Synovium wird von einer grossen Anzahl an Immunzellen, darunter T-Zellen, B-Zellen und Makrophagen, überschwemmt. Zur Erhaltung der

[2] Gelenkschmiere ist eine Flüssigkeit, welche der Verringerung des Reibungswiderstandes dient.

vergrösserten inflammatorischen Zellmasse werden neue Blutgefässe gebildet (**Angiogenese**). Diese versorgen die Immunzellen mit Sauerstoff und Nährstoffen und ermöglichen so das Fortschreiten der Entzündung [6], [7]. Dabei wirkt vor allem der vaskuläre endotheliale Wachstumsfaktor (vascular endothelial growth factor, VEGF) mit. VEGF-**Serumkonzentrationen** [3] sind bei RA-Patienten erhöht und korrelieren direkt mit der **Krankheitsaktivität** [7].

Pannusbildung: Durch die chronische Entzündung verdickt sich die Gelenkinnenhaut stark (**Hyperplasie**) und beginnt, tumorartig in den Gelenkknorpel und den angrenzenden Knochen hineinzuwachsen. Dieses wuchernde, aggressive Gewebe wird als **Pannus** bezeichnet.

Gewebezerstörung: Die im Pannus aktiven Zellen setzen durch die Freisetzung von **pro-inflammatorischen Zytokinen** wie TNF-α, IL-1 und IL-6 eine Kaskade zerstörerischer Prozesse in Gang. Beispielsweise regen diese Zytokine Knorpelzellen (Chondrozyten) und Bindegewebszellen (synoviale Fibroblasten) [8] an, Enzyme freizusetzen, die den **Knorpel abbauen** (z.B. Kollagenasen). Gleichzeitig fördern sie die Aktivierung von Osteoklasten, spezialisierten Zellen, die für den **Abbau von Knochengewebe** verantwortlich sind.

2.2 Genetik

2.2.1 Was ist ein Gen?

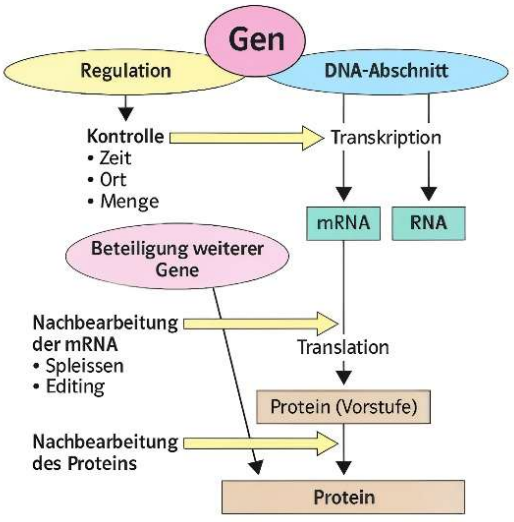

Abbildung 4: Vom Gen zum Genprodukt [9]
Die Herstellung von Proteinen folgt einem mehrstufigen Prozess. Bei jedem Schritt sind mehrere epigenetische Faktoren beteiligt, welche das Produkt verändern können.

In der Desoxyribonukleinsäure (DNA) sind alle Erbinformationen eines Organismus gespeichert. Bestimmte Abschnitte der DNA werden **Gene** genannt [9], [10]. Sie dienen als «Bauanleitungen» für Proteine, welche die Grundlage biologischer Systeme bilden. Die Bildung eines Genprodukts wird Expression genannt und erfolgt über mehrere Schritte. Zunächst wird das Gen in einem Prozess namens **Transkription** «abgelesen»: Dabei wird mithilfe des **RNA-Polymerase**-Enzyms eine Arbeitskopie des Gens in Form von messenger-RNA (mRNA) erstellt. Anschliessend wird die mRNA in der **Translation** als Vorlage für die Herstellung von Proteinen genutzt.

Die Expression eines Gens geschieht nicht zufällig, sondern wird durch ein komplexes Zusammenspiel von **regulatorischen Elementen** auf der DNA und speziellen Proteinen, den **Transkriptionsfaktoren**, exakt gesteuert (Abb. 4). Diese Steuerung ist die Basis für genbasierte Netzwerke, da die Expression eines Gens von innerlichen und äusserlichen Faktoren beeinflusst wird. Für eine Entzündung beispielsweise werden mehr pro-inflammatorische Zytokine (z.B. TNF-α) benötigt. Die dafür kodierenden Gene sind also stärker exprimiert [11].

[3] Mit „Serumkonzentration“ wird die Konzentration im Blutserum gemeint.

Da die Quantifizierung der Proteinkonzentrationen oft schwierig ist, misst man in der Forschung jedoch zumeist die mRNA – die bereits transkribierte, jedoch noch nicht translatierte Protein-Vorstufe [12], [13]. Die Gesamtheit aller RNA-Transkripte in einer Zelle zu einem bestimmten Zeitpunkt wird als **Transkriptom** bezeichnet und gibt einen Einblick in die zelluläre Aktivität [14].

2.2.2 Messung der Genexpression

Um genregulatorische Netzwerke zu analysieren, muss die Aktivität von Tausenden von Genen gleichzeitig gemessen werden [13], [14]. Die Daten werden typischerweise mit einer der folgenden Methoden erzeugt:

DNA-Microarrays: Bei dieser Methode werden kurze, bekannte DNA-Sonden[4] [15] auf einem Chip fixiert. Die Menge der gebundenen RNA an einer bestimmten Sonde ist proportional zur Expression des entsprechenden Gens und wird über ein Fluoreszenzsignal gemessen.

RNA-Sequencing (RNA-Seq): Diese Methode sequenziert die RNA-Moleküle einer Probe direkt und zählt sie. Sie bietet eine höhere Genauigkeit und einen grösseren dynamischen Bereich als Microarrays und kann auch bisher unbekannte Transkripte entdecken.

Beide Technologien liefern riesige Datensätze, die zeigen, welche Gene in einer Zelle unter bestimmten Bedingungen «an-» oder «ausgeschaltet»

sind. Die wahre Schwierigkeit liegt in der Analyse und Interpretation dieser Daten.

2.3 Bioinformatik: Gennetzwerke

2.3.1 Was ist ein Gennetzwerk?

Gennetzwerke sind eine Art, die Expression von Genen zu quantifizieren und zu vergleichen. Einige Begriffe aus der Graphentheorie sind in diesem Kontext relevant (Tab. 1) [13]:

Tabelle 1: Graphentheorie in der Bioinformatik

Begriff	**Anwendung**
Knoten	Ein einzelnes Gen.
Kante	Die mathematische Verwandtschaft zwischen zwei Genen.
Topologie	Die Struktur eines Netzwerks.
Nachbarn	Gene, welche durch eine Kante verbunden sind.
Grad eines Knotens	Die Gesamtanzahl der Nachbarn.
Dichte	Verhältnis der vorhandenen Kanten zu den möglichen Kanten in einem Netzwerk [16].
Durchmesser	Längster direkter Weg zwischen zwei Knoten [16].
Modul / Cluster	Eine Gruppe von verwandten Genen.

[4] Eine DNA-Sonde ist ein DNA-Molekül, das eine komplementäre Basensequenz zum gesuchten Gen aufweist.

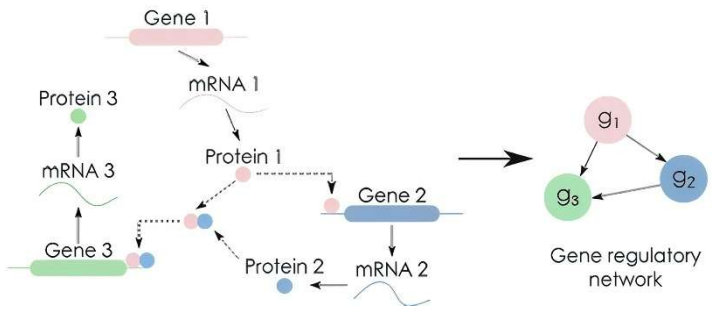

Abbildung 5: Schematische Darstellung eines genregulatorischen Netzwerks [13]
Ein komplexes biophysisches Modell beschreibt die Interaktion zwischen drei Genen: direkte Regulation (Gen 2 durch Gen 1) sowie kombinatorische Regulation durch eine Komplexbildung (Gen 3 durch die Gene 1 und 2). Das gerichtete Netzwerk stellt die abstrahierte Struktur des Systems dar.

Da meistens mehrere Gene an der Herstellung eines Proteins und immer eine Vielzahl von Genen in komplexeren biologischen Prozessen beteiligt ist, lassen sich aus den einzelnen Strukturen biologisch relevante Netzwerke erstellen (Abb. 5).

Die Kanten können sich dabei nur durch ihre Anwesenheit auszeichnen (Abb. 6a), gerichtet sein (Abb. 6b) oder eine Gewichtung haben (Abb. 6c). Kombinationen sind möglich. Gewichtete Kanten stellen meistens die statistische Ähnlichkeit im Expressionsmuster (**Koexpression**) als Verbindungsstärke zwischen zwei Genen dar.

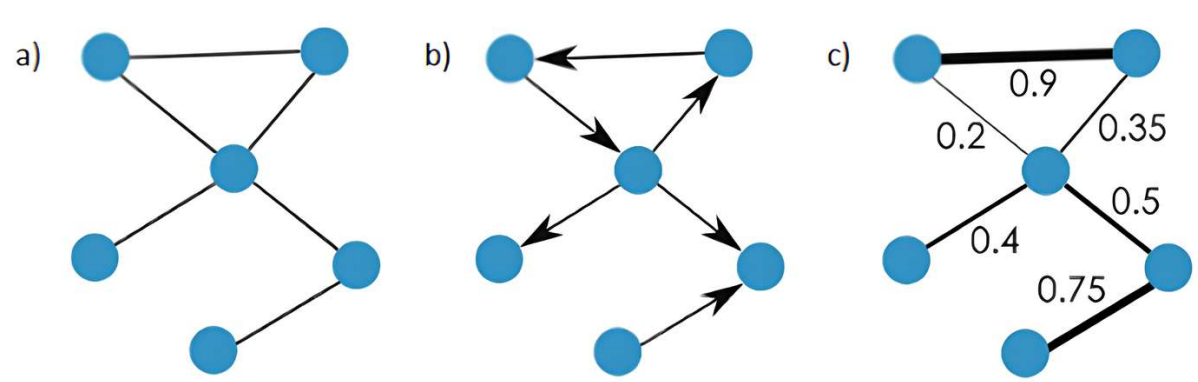

Abbildung 6: Verschiedene Kantenarten [13], abgeänderte Abbildung
6a) zeigt ein Netzwerk mit ungerichteten, ungewichteten Kanten. b) hat gerichtete Kanten (Pfeile) und c) hat gewichtete Kanten (erkennbar an der Stärke der Verbindungslinien).

2.3.2 Arten von Netzwerken

Grob lassen sich die genbasierten Netzwerke in drei Kategorien einteilen [17] (Abb. 7):

A. **Genkoexpressionsnetzwerke** (GCNs) haben ungerichtete, aber gewichtete Kanten und können infolgedessen auch die kausalen Zusammenhänge nicht wiedergeben. Sie kommen bei der Suche nach **funktional verwandten Genen** zum Einsatz.
B. **Genregulatorische Netzwerke** (GRNs) zeichnen sich dadurch aus, dass all ihre Kanten gerichtet sind. Sie geben Ausschluss darüber, welche Gene sich gegenseitig **regulieren**[5].
C. Eine Unterkategorie von GRNs sind die **transkriptionsregulatorischen Netzwerke** (TRNs). Deren gerichtete Kanten gehen von Transkriptionsfaktor-Genen aus.

[5] «Regulieren» bedeutet beispielsweise, dass Gen A ein Protein P_A herstellt, dass die Transkription von Gen B verhindert. Gleichzeitig kann P_A dafür verantwortlich sein, die Translation von Gen B zu fördern. Vergleiche Abb. 5.

Allerdings wird in der Forschung diese Einteilung nur lose verwendet. Die Bezeichnung «genregulatorisches Netzwerk» hat sich als Oberbegriff zu allen Netzwerken etabliert und wird teilweise auch für GCNs oder TRNs verwendet. Da in dieser Arbeit jedoch die Koexpression untersucht wird, wird der Ausdruck «Koexpressionsnetzwerk» (GCN) verwendet werden.

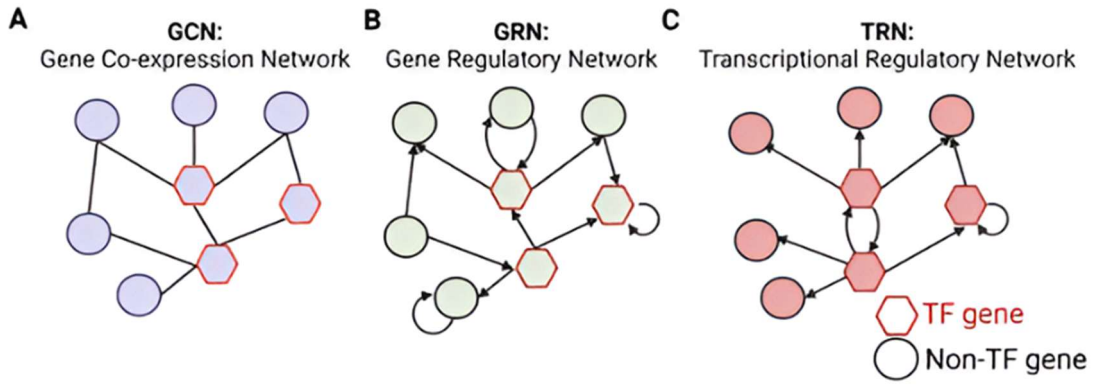

Abbildung 7: Arten von Gennetzwerken [17]
Das Koexpressionsnetzwerk (GCN) in A) stellt die Koexpression zwischen jeweils zwei Genen mittels einer Linie (Kante) dar. B): Genregulatorische Netzwerke (GRNs) hingegen zeigen, welches Gen i ein anderes Gen j beeinflusst (reguliert). Transkriptionsregulatorische Netzwerke (TRNs) wie das in C) zeigen, welche Gene durch die Herstellung von Transkriptionsfaktoren andere Gene beeinflussen.

2.3.3 Umwandlung in Matrizen

Häufig sind Netzwerke für weitere Berechnungen ungeeignet und werden daher als **Adjazenzmatrizen** dargestellt. Zeilen und Spalten entsprechen den Knoten, die numerischen Einträge den Kanten [18]; deren Werte variieren je nach Netzwerktyp (Abb. 8). Die Matrix ist symmetrisch, und die Diagonale wird auf null gesetzt.

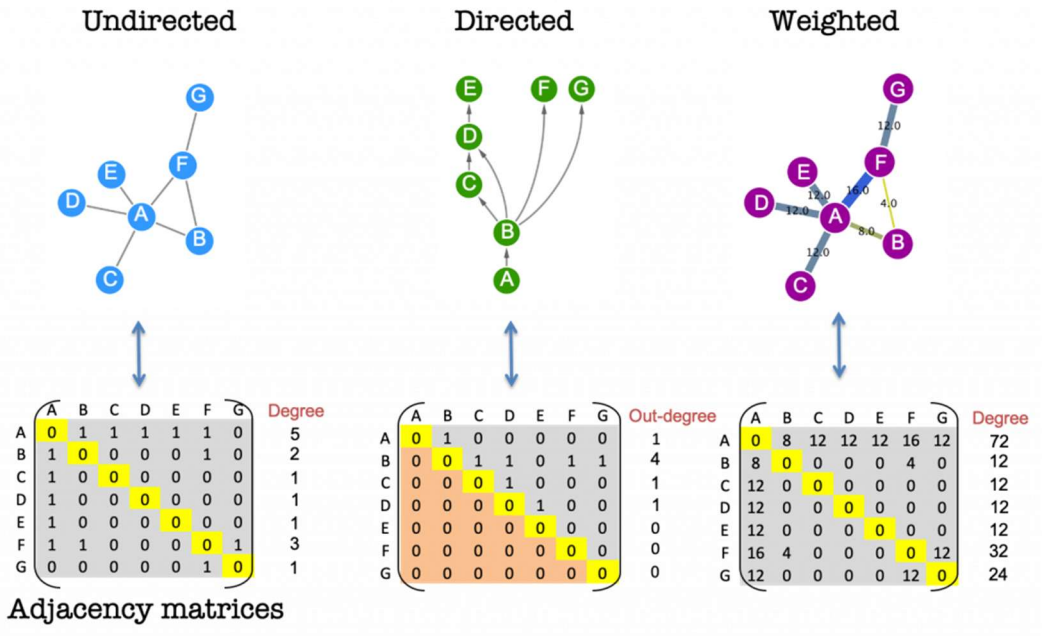

Abbildung 8: Graphen nach Kantenart und ihre Adjazenzmatrizen [18]
In ungerichteten, ungewichteten Netzwerken werden die Kanten mit 1 (vorhanden) oder 0 (nicht vorhanden) angegeben. Adjazenzmatrizen von gerichteten Netzwerken zeigen nur eingehende Kanten an. Für gewichtete Netzwerke liegen die Kantenwerte in einem bestimmten Bereich, bspw. $[0;\ 100]$ oder $[0;\ 1]$.

2.3.4 Die skalenfreie Topologie

Ein zentrales Kriterium zur **Validierung** von biologischen Netzwerken ist die Untersuchung ihrer globalen Struktur [19], [20], [21].

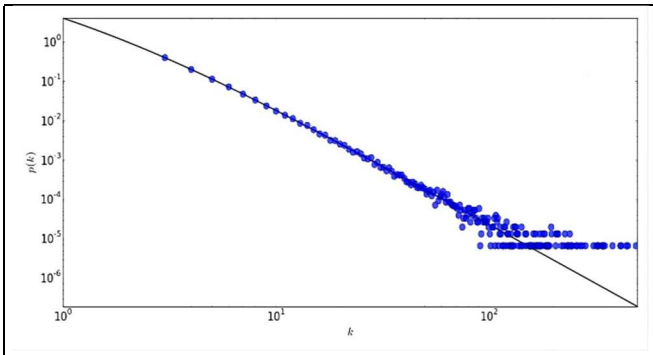

Abbildung 9: Optimale Gradverteilung
Jeder Punkt in diesem Graphen stellt die Anzahl Knoten (Gene) P(k) mit k Kanten (Verbindungen) dar. Für skalenfreie Netzwerke ergibt sich dadurch eine abfallende Gerade, an deren Ende sich eine Häufung von Punkten befinden kann.

(By Rsoaresp - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=34204143)

Eine skalenfreie Topologie (**scale-free topology**) ist ein fundamentales Merkmal vieler realer Netzwerke, einschliesslich biologischer Systeme. Solche Netzwerke zeichnen sich dadurch aus, dass die meisten Knoten (Gene) nur wenige Verbindungen aufweisen, während einige wenige Knoten, sogenannte **«Hubs»**, eine extrem hohe Anzahl von Verbindungen besitzen. Die Anzahl Knoten $P(k)$ eines Netzwerks, welche k Verbindungen (Kanten) zu anderen Knoten haben, ist gegeben durch das Potenzgesetz $P(k) \sim k^{-y}$ (Abb. 9). Der Parameter y liegt im Normalfall im Intervall $(2; 3)$.

Die biologische Relevanz dieser Topologie ist signifikant. Metabolische und regulatorische Netzwerke folgen oft diesem Prinzip [22], [23]. Es wird angenommen, dass diese Architektur ein grundlegendes Designprinzip zellulärer Organisation widerspiegelt, das den Systemen eine hohe **Robustheit** gegenüber zufälligen Störungen und Fehlern verleiht. Parameter, die zu einer

schlechten Anpassung an die skalenfreie Topologie führen, korrelieren oft mit einem **schwächeren «biologischen Signal»** (z.B. einer geringeren Korrelation zwischen Konnektivität und funktioneller Aussagekraft).

Aufgrund dieser starken biologischen Evidenz wird das Erreichen einer skalenfreien Topologie oft als **Qualitätskriterium** für die Netzwerk-Konstruktion verwendet. Methoden wie WGCNA nutzen dieses Kriterium (gemessen als $\boldsymbol{R^2}$ **-Wert**), um die Parameter für die Erstellung der Adjazenzmatrix zu optimieren.

2.3.5 Permutationstests

Eine weitere Art, die Qualität und statistische Aussagekraft eines Netzwerks zu überprüfen, ist ein sogenannter **Permutationstest**. Im Folgenden wird die von Tesson et al. (2010) beschriebene Anwendung vorgestellt [24].

Zunächst wird ein **Dispersionswert** berechnet: Dieser misst, wie sehr sich Korrelation zwischen zwei Genen aus den Modulen $c1$ und $c2$ der Test- und Kontrollbedingung unterscheidet. Falls $c1 = c2$, quantifiziert der Dispersionswert die Koexpressionsunterschiede innerhalb von $c1$.

Anschliessend werden alle **Proben** $\boldsymbol{1000}$ **-mal zufällig in zwei Gruppen** aufgeteilt, um gewissermassen eine neue «Test»- und «Kontrollgruppe» zu erstellen, und es wird für jede Zuteilung ein Dispersionswert berechnet. Die dadurch entstehende Verteilung zeigt an, wie gross die zufällig zu erwartenden Unterschiede wären. Dagegen vergleicht man den tatsächlich beobachteten Dispersionswert.

2.4 Mathematik und Statistik

Für **normale** Datenverteilungen kann zur Berechnung des Verwandtschaftsgrades zweier Gene mittels des Estimators «Spearman» die gegenseitige Information berechnet werden [25].

2.4.1 Der Spearman-Korrelationskoeffizient

Der Spearman-Korrelationskoeffizient untersucht, ob zwischen zwei Variablen ein **monotoner Zusammenhang** besteht. „Monoton" bedeutet, dass grössere Werte der einen Variable im Allgemeinen mit grösseren (oder kleineren) Werten der zweiten Variable einhergehen. Dabei muss der Zusammenhang **nicht linear** sein [26].

Im Gegensatz zur Pearson-Korrelation, die direkt mit den ursprünglichen Messwerten arbeitet, basiert die Spearman-Korrelation auf **Rängen**: Alle Werte werden der Grösse nach sortiert, und jeder Beobachtung wird eine **Rangnummer** zugeordnet[6]. Bei gleichen Werten erhalten diese entweder den gleichen Rang oder den Mittelwert der entsprechenden Rangpositionen [27].

Anschliessend wird auf diesen Rangwerten die Pearson-Korrelation berechnet. Die Formel für die Pearson-Korrelation lautet [28], [29]:

[6] Zum Beispiel erhält der kleinste Wert den Rang 1 und der zweitkleinste Wert den Rang 2. Der grösste von n Werten besitzt den Rang n.

Gleichung 1: Berechnung der Pearson-Korrelation

$$\rho_{x,y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2} \cdot \sqrt{\sum(y_i - \bar{y})^2}}$$

Dabei bezeichnen x_i und y_i die einzelnen Messwerte (hier: die Ränge) und $\bar{x}$ sowie $\bar{y}$ deren Mittelwerte. Die Terme $\sqrt{\sum(x_i - \bar{x})^2}$ bzw. $\sqrt{\sum(y_i - \bar{y})^2}$ entsprechen den jeweiligen Standardabweichungen der Rangwerte.

Da Spearman nicht mit den ursprünglichen Messwerten, sondern mit deren Rängen arbeitet, handelt es sich um ein **nicht-parametrisches Verfahren**. Das bedeutet, dass keine bestimmte Verteilung der Daten vorausgesetzt wird. Diese Eigenschaft macht den Spearman-Koeffizienten besonders geeignet für biologische Daten.

2.4.2 Mutual Information

Sind zwei Variablen nicht unabhängig voneinander, lässt sich von der einen Variable mit gewisser Wahrscheinlichkeit auf die andere schliessen [30]. Diese Wahrscheinlichkeit lässt sich mit der «gegenseitigen Information» (Mutual Information oder MI) quantifizieren.

Für normalverteilte Daten kann die gegenseitige Information I aus dem Spearman-Koeffizienten r berechnet werden [25], [31]:

Gleichung 2: MI-Berechnung auf Basis des Spearman-Koeffizienten

$$I(X_i, Y_i) = -\frac{1}{2}\log(1 - r^2)$$

Falls $I = 0$ ist, enthält eine Variable keine Information über eine andere. Ein höherer MI-Wert zwischen zwei Variablen bedeutet, dass die eine auf eine nicht zufällige Art mit der anderen verwandt ist. In diesem Sinne ist die Annahme sinnvoll, dass zwei Gene umso wahrscheinlicher eine biologische Verwandtschaft haben, desto höher ihr gemeinsamer MI-Wert ist [32].

2.4.3 Weitere Grundlagen der Statistik

Nullhypothese H_0

Eine Nullhypothese H_0 ist eine Annahme in der Statistik, die besagt, dass es keinen Effekt, keinen Zusammenhang oder **keinen Unterschied zwischen Variablen** gibt, und die mit einem Hypothesentest überprüft wird. Sie wird solange als wahr angesehen, bis statistische Beweise sie widerlegen, und bildet das Gegenstück zur Alternativhypothese H_1 [34], [35].

p-Wert

Der p-Wert ist definiert als die Wahrscheinlichkeit, die Teststatistik oder einen noch höheren Wert zu erhalten, unter der Annahme, dass die Nullhypothese wahr ist. Ein kleiner p-Wert signalisiert, dass die Beobachtung unter der Annahme des Zufalls sehr unwahrscheinlich ist, was zur Verwerfung der Nullhypothese führt.

Bernouilli-Experiment

Ein Bernoulli-Experiment ist ein Zufallsexperiment mit genau **zwei möglichen Ergebnissen**, die als «Erfolg» und «Misserfolg» bezeichnet werden. Die Erfolgswahrscheinlichkeit bleibt durch Prozesse wie beispielsweise «Zufälliges Ziehen mit Zurücklegen» konstant [36].

3 Methodik

3.1 Daten

Der verwendete Datensatz stammte aus der öffentlichen Datenbank **Gene Expression Omnibus** (GEO, Accession: GSE128816; $n\_test = 10, n\_control = 10$) [39]. Die Genexpressionsdaten basierten auf RNA-Seq-Analysen von humanen $CD4^+$-T-Zellen nach Kontakt mit VitD3-tolDC beziehungsweise mit mDCs. Damit sollten die von VitD3 ausgelösten Mechanismen der Toleranzinduktion untersucht werden.

3.1.1 Normalisierung

Die Daten wurden zunächst **log2(x+1)**[7] **verwandelt**, um aus einer exponentialähnlichen eine (möglichst) normale Verteilung zu erhalten [40].

Zur Vergleichbarkeit der Expressionswerte zwischen den Bedingungen wurde eine **Quantilnormalisierung** (qspline) durchgeführt [41]. Dieses Verfahren gleicht systematische Unterschiede in den Signalverteilungen der Arrays aus, um sicherzustellen, dass beobachtete Unterschiede biologischen und nicht technischen Ursprungs sind. Dabei werden die Quantile der ursprünglichen Verteilungen an eine gemeinsame Zielverteilung angepasst.

In einem QQ-Plot ist diese Anpassung graphisch erkennbar: Die x-Achse zeigt die «theoretical quantiles» (die Quantile, die man bei einer perfekten Normalverteilung erwarten würde) an. Nach der Normalisierung sollten ihnen die «sample quantiles» (die tatsächlich beobachteten Quantile) auf der y-Achse entsprechen – die Punkte sollten also genau auf der Diagonalen liegen.

3.2 Materialien und Software

Alle statistischen Berechnungen und bioinformatischen Analysen wurden in der Programmiersprache **R** (Version **4.4.3**) [42] in der Entwicklungsumgebung **RStudio** (Version **2025.09.0+387**) [43] durchgeführt. Für die jeweiligen Analyseschritte kamen die nachfolgenden R-Pakete aus dem Bioconductor- und dem CRAN-Repository zum Einsatz (Tab. 2):

Tabelle 2: verwendete Pakete

Paket	Version	Repository	Quelle
bc3net	1.0.5	CRAN	[44]
WGCNA	1.73	CRAN	[45]
RColorBrewer	1.1.3	CRAN	[46]
preprocessCore	1.68.0	Bioconductor	[47]
igraph	2.1.4	CRAN	[48]
moduleColor	1.8.4	CRAN	[49]
scales	1.4.0	CRAN	[50]
dplyr	1.1.4	CRAN	[51]
ggplot2	4.0.0	CRAN	[52]
flashClust	1.01-2	CRAN	[53]
clusterProfiler	4.14.6	Bioconductor	[54]
org.Hs.eg.db	3.20.0	Bioconductor	[55]
enrichplot	1.26.6	Bioconductor	[56]

[7] Der sogenannte „Pseudocount“ von 1 wird allen Werten addiert, um das nicht definierte log (0) zu vermeiden.

Die Analysen erfolgten auf einem Windows-Laptop (Intel i7-1165G7, 16 GB RAM).

Die folgende Github-Repository enthält ein PDF der Arbeit, den verwendeten Code und das Datenset sowie alle erhaltenen Ergebnisse:

https://github.com/py-sofia/Maturaarbeit_Sofia-Surace_2025

3.3 Netzwerkerstellung mit *bc3net*

Zur Erstellung des Genregulationsnetzwerks aus den vorliegenden Genexpressionsdaten wurde die Methode ***bc3net*** (*bagging c3net*) verwendet [31].

bc3net ist ein sogenanntes **Ensemble-Verfahren**, das speziell darauf ausgelegt ist, aus grossen, aber oft verrauschten biologischen Datensätzen robuste und zuverlässige Netzwerke zu extrahieren. Die Kernidee besteht darin, nicht nur ein einziges Netzwerk aus den Daten zu berechnen, sondern eine Vielzahl von Netzwerken zu erstellen und diese anschliessend zu einem finalen, statistisch validierten Netzwerk zu vereinigen. Dieser Ansatz erhöht die Zuverlässigkeit der Ergebnisse erheblich und reduziert die Anfälligkeit für zufällige Schwankungen oder Ausreisser in den Daten.

Der Algorithmus lässt sich in drei konzeptionelle Hauptschritte (Abb. 10) unterteilen, die im Folgenden erläutert werden.

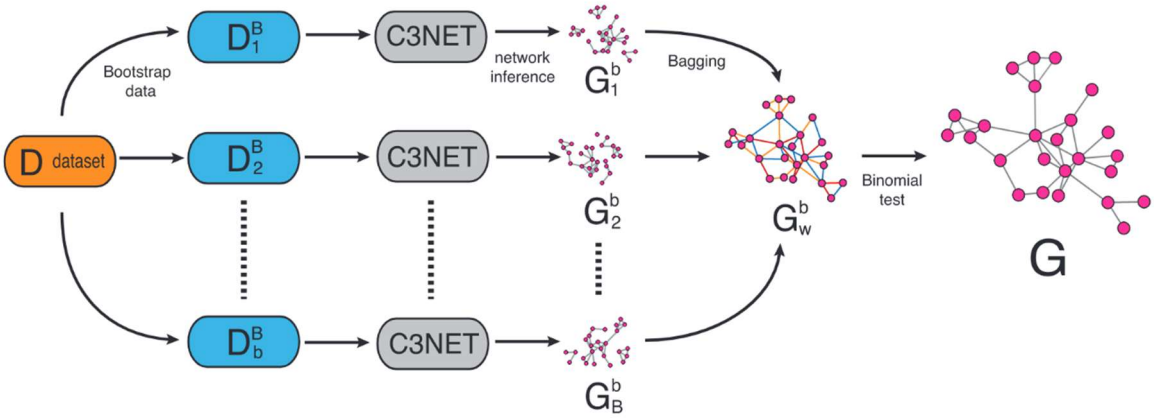

Abbildung 10: Schematische Abbildung des bc3net-Algorithmus [31]
Aus einem Datenset D werden B Bootstraps erstellt. Auf jeden dieser Bootstraps wird der C3NET-Algorithmus zur Korrelationsberechnung angewendet. Die entstehenden Netzwerke $\{G_k^b\}_{k=1}^B$ werden dann zu einem ungerichteten, gewichteten Finalnetzwerk G_w^b zusammengeführt. Anschliessend wird zur Validierung der Kanten ein Binomialtest durchgeführt.

3.3.1 Erstellung eines Bootstrap-Ensembles (Bagging)

Genexpressionsdaten sind durch komplexe Eigenschaften gekennzeichnet: Sie sind hochdimensional (Tausende von Genen), oft nicht linear und enthalten technisches sowie biologisches Rauschen. Zudem steht einer grossen Anzahl von Genen p (Variablen) typischerweise nur eine kleine Anzahl von Proben n gegenüber («Large p small n») [57].

Um diesen Herausforderungen zu begegnen, wendet *bc3net* das Prinzip der **Bootstrap Aggregation**, kurz **Bagging**, an. Aus dem ursprünglichen, einzelnen Genexpressionsdatensatz D wird nicht nur ein Netzwerk, sondern ein ganzes Ensemble von Netzwerken abgeleitet. Für diese

Arbeit wurden dafür zunächst aus dem Originaldatensatz wurden durch **zufälliges Ziehen mit Zurücklegen** hundert $(B = 100)$ neue, sogenannte **Bootstrap-Datensätze** $(D_1, D_2, \ldots, D_B)$ erstellt [58], [59]. Jeder dieser Bootstrap-Datensätze war dem Originaldatensatz sehr ähnlich, aber dennoch einzigartig in seiner Zusammensetzung. Dieses Vorgehen simulierte gewissermassen, wie die Ergebnisse aussehen würden, wenn das Experiment viele Male wiederholt werden würde.

3.3.2 Netzwerkinferenz mit c3net für jeden Bootstrap-Datensatz

Für jeden einzelnen der hundert erstellten Bootstrap-Datensätze wurde nun mithilfe des **c3net-Algorithmus** ein Koexpressionsnetzwerk berechnet. c3net ist ein recheneffizienter Algorithmus, der seinerseits in drei Schritten arbeitet:

1. **Berechnung der Mutual Information:** Zuerst wurden für alle möglichen Gen-Paare als Mass für ihre statistische Abhängigkeit die **Mutual Information** berechnet. *bc3net* berechnete zunächst den Spearman-Korrelationskoeffizienten und wandte diesen anschliessend in einen MI-Wert um[8].

2. **Eliminierung schwacher Verbindungen:** Für jedes Gen wurde nur die stärkste Verbindung zu einem anderen Gen beibehalten; **pro Gen** konnte es also **maximal eine Kante** geben. Alle anderen potenziellen Verbindungen dieses Gens verwarf *c3net*. Dieser **konservative** Schritt sorgte dafür, dass nur die signifikantesten Interaktionen im Netzwerk verblieben und verhinderte, dass das Netzwerk zu dicht und unübersichtlich wurde.

3. **Fehlerkontrolle:** Die Bonferroni-Korrektur für multiples Testen wurde angewendet, um die Anzahl der zufällig gefundenen Verbindungen zu kontrollieren (siehe Anhang).

Als Ergebnis dieses zweiten Schrittes lag ein Ensemble von B leicht unterschiedlichen Netzwerken $\{G_k^b\}_{k=1}^B$ vor, die jeweils die stärksten Verbindungen jedes Gens in dem entsprechenden Bootstrap-Datensatz repräsentierten.

3.3.3 Aggregation und statistische Validierung

Alle Netzwerke des Ensembles wurden zu einem **gewichteten Netzwerk** G_w^b zusammengeführt. Das Gewicht einer Kante n_{ij} entsprach dabei der absoluten Häufigkeit, mit der diese spezifische Kante in den B Einzelnetzwerken vorkam:

Gleichung 3: Kantengewicht in G_w^b

$$G_w^b(i,j) = \sum_{k=1}^{B} I\left(G_k^b(i,j)\right)$$

wobei die Indikatorfunktion $I()$ definiert ist als:

Gleichung 4: Die Indikatorfunktion I(x)

$$I(x) = \begin{cases} 1, x = 1, \\ 0, sonst. \end{cases}$$

[8] *bc3net* ist in der Lage, mit MI-Estimators anstatt von Korrelationskoeffizienten zu rechnen, um nichtlineare Daten zu berücksichtigen. Für normale Daten lässt sich jedoch auch aus den Korrelationskoeffizienten die Mutual Information berechnen.

Eine Kante, die beispielsweise in 95 von 100 Bootstrap-Netzwerken auftrat, wies ein hohes Gewicht auf und wurde entsprechend als robust betrachtet. Trat eine Kante dagegen nur in 20 Netzwerken auf, deutete dies auf eine geringere Zuverlässigkeit hin.

Statistischer Test für jede Kante

Anstatt einen willkürlichen Schwellenwert festzulegen (z.B. «Behalte alle Kanten, die in mehr als 30 der Netzwerke vorkommen»), führte *bc3net* für jede potentielle Kante einen **statistischen Hypothesentest** durch.

Die Nullhypothese H_0 lautete, dass die beobachtete Häufigkeit einer Kante nicht höher war als rein zufällig erwartet. Da die B Bootstrap-Datensätze voneinander unabhängig waren, konnte das wiederholte Schätzen eines Netzwerks als eine Serie von B **unabhängigen Bernoulli-Experimenten** betrachtet werden. In jedem dieser Experimente besass die Kante (i, j) genau zwei mögliche Ausgänge: «vorhanden» (Erfolg) oder «nicht vorhanden» (Misserfolg).

Unter H_0 folgte die Anzahl der Beobachtungen einer Kante, also die Teststatistik n_{ij}, somit einer **Binomialverteilung**[9] [37]

Gleichung 5: Verteilung der Teststatistik unter der Nullhypothese

$$n_{ij} \sim \mathrm{Bin}(B, p_c),$$

wobei p_c die Wahrscheinlichkeit bezeichnete, dass zwei Gene rein zufällig durch eine Kante verbunden waren. Dieser Wert wurde von *bc3net* u.a. basierend auf der Grösse des Datensets geschätzt.

Nachdem für jede Kante (i, j) die Anzahl der beobachteten Erfolge n_{ij} bestimmt wurde, erfolgte die statistische Bewertung. Entscheidend war, ob n_{ij} so hoch war, dass eine zufällige Fluktuation unwahrscheinlich wurde und auf eine echte biologische Interaktion hingedeutet werden konnte.

Der p-Wert für eine Kante (i, j) entsprach der Wahrscheinlichkeit, unter Annahme von H_0 mindestens so viele Erfolge zu erhalten wie beobachtet:

Gleichung 6: Berechnung des p-Werts für die Kante (i, j)

$$p(i,j) = \Pr(n \geq n_{ij}) = \sum_{n=n_{ij}}^{B} \binom{B}{n} p_c^n (1 - p_c)^{B-n}$$

Jeder Summand $\binom{B}{n} p_c^n (1 - p_c)^{B-n}$ war die Wahrscheinlichkeit, in genau n der B Bootstrap-Netzwerke das zufällige Auftreten der Kante zu beobachten. Die Binomialkoeffizienten $\binom{B}{n}$ gaben dabei die Anzahl möglicher Kombinationen dieser n Erfolge an.

[9] Die Binomialverteilung beschreibt die Anzahl Erfolge in einer Serie von **Bernoulli**-Experimenten. Ist p die Erfolgswahrscheinlichkeit bei einem Versuch, dann bezeichnet $Bin(k, p)$ die Wahrscheinlichkeit, genau k Erfolge zu erzielen.

Erstellung des finalen Netzwerks

Nur die Kanten, deren p-Wert unter einem strengen, für multiples Testen korrigiertes Signifikanzniveau lag (**Bonferroni-Korrektur**), wurden in das endgültige, ungerichtete Netzwerk G aufgenommen.

Bedeutung

Dieser statistisch fundierte Ansatz ist ein wesentlicher Vorteil von *bc3net*, da er eine objektive und reproduzierbare Methode zur Auswahl der finalen Kanten bietet und die Notwendigkeit **manueller Schwellenwerte eliminiert**. Das resultierende Netzwerk G repräsentierte somit ein konservatives, aber statistisch signifikantes Abbild der stärksten und robustesten Zusammenhänge in den Genexpressionsdaten.

3.3.4 Praktische Umsetzung

Für diese Arbeit wurde die Implementierung der Autoren im R-Paket «bc3net» genutzt (Code-Block 1). Als Ähnlichkeitsmass diente der **Spearman**-Korrelationskoeffizient. Die Anzahl der Bootstraps wurde auf 100 gesetzt, um eine stabile Netzwerkkonvergenz bei gleichzeitig akzeptabler Rechenzeit zu erreichen. Anschliessend wurde das erstellte igraph-Objekt (Netzwerk) in eine Adjazenzmatrix umgewandelt.

Code-Block 1: Berechnung der Netzwerke mit bc3net

```
library(igraph)
library(bc3net)
library(WGCNA)


netControl <- bc3net(datControl, verbose=TRUE, estimator="spearman",
boot=100)
netTest <- bc3net(datTreated, verbose=TRUE, estimator="spearman", boot=100)

save(netControl, file = "netControl.RData")
save(netTest, file = "netTest.RData")

adjMatControl <- as_adjacency_matrix(netControl, attr="weight", sparse=F)
adjMatTreated <- as_adjacency_matrix(netTest, attr="weight", sparse=F)

genesTreated <- V(netTest)$name # get the names of all the treated genes
adjMatControl <- adjMatControl[genesTreated, genesTreated] # only keep ge-
nes present in treated

collectGarbage()
```

3.4 Skalenfreie Topologie und der Parameter beta1

Der optimale Soft-Threshold beta1 wurde mithilfe der Funktion **pickSoftThreshold.fromSimilarity** bestimmt (Code-Block 2). Als skalenfrei wurde ein Netzwerk erachtet, wenn er einen R^2-Wert von über 0.85 erreicht.

Code-Block 2: Die Wahl von beta1

```
powers <- c(seq(1,10,1), seq(12,20,2))

sft = pickSoftThreshold.fromSimilarity(
  similarity = AdjDiff,
  powerVector = powers,
  RsquaredCut = 0.85,
  moreNetworkConcepts = TRUE,
  verbose = 5);

beta1 <- sft$powerEstimate
```

3.5 Grobe Netzwerkbeschreibung

Für eine grobe Beschreibung der mit *bc3net* erstellten Netzwerke erfolgte zunächst eine graphische Darstellung mit **igraph**. Das hierarchische Clustering wurde mit der **Louvain-Methode** [10] durchgeführt [60], [61]. Code-Block 3 visualisiert den grössten zusammenhängenden Teil, die sogenannte «**Giant Component**», des netTest-Netzwerks und färbt die Knoten und Kanten nach Clusterzugehörigkeit bzw. nach Kantengewicht ein. Bei der Layout-Erstellung kam **DrL**[11] zum Einsatz, da sich diese Methode gut für die Visualisierung grosser Netzwerke eignet [62], [63]. netControl wurde auf gleiche Weise visualisiert.

Code-Block 3: Visualisierung der Netzwerke

```
componentsInfoTest <- components(netTest)
giantTest <- induced_subgraph(netTest,
                              which(componentsInfoTest$membership ==
which.max(componentsInfoTest$csize)))

comTest <- cluster_louvain(giantTest)
V(giantTest)$color <- brewer.pal(n = length(comTest), name =
"Set3")[comTest$membership]

edgeColorFnTest <- col_numeric(palette = c("#fff", "#404040"), domain = NULL)
E(giantTest)$color <- edgeColorFnTest(E(giantTest)$weight)

pdf("netTest.pdf", width = 12, height = 12)
plot(giantTest,
     layout = layout_with_drl(giantTest),
     vertex.label = NA,
     vertex.size = 1,
     vertex.frame.color = NA,
     edge.arrow.mode = 0)
dev.off()
```

Anschliessend wurden für beide Netzwerke Dichte, Durchmesser und durchschnittlicher Knotengrad berechnet sowie die Anzahl Kanten bzw. Knoten gezählt.

[10] Die Louvain-Methode ist ein Greedy-Algorithmus und identifiziert sich nicht überschneidende Cluster.
[11] In DrL-Graphen sind nahverwandte Knoten nah aneinander, während entferntere Verbindungen nach aussen verschoben werden.

3.6 *DiffCoEx*

Um die Unterschiede in den Gen-Interaktionsmustern zwischen den beiden analysierten Zuständen zu identifizieren, kam die Methode ***DiffCoEx*** (Differentially CoExpressed gene modules) zum Einsatz [24]. *DiffCoEx* versucht, ganze Module von Genen zu identifizieren, deren gemeinsames Expressionsmuster (ihre Koexpression) sich signifikant zwischen den beiden Bedingungen ändert. Anders als bei der normalen Clustersuche enthalten die gefundenen Module also nicht koexprimierte Gene; stattdessen gruppiert *DiffCoEx* Gene, welche ihre **Verbindungsstärke** zu denselben anderen Genen auf eine ähnliche Art **ändern**[12].

DiffCoEx nutzt **WGCNA** (Weighted Gene Co-expression Network Analysis), ein R-Paket mit verschiedenen Funktionen zur Erstellung von gewichteten Koexpressionsnetzwerken, wendet es aber nicht auf die Expressionsdaten selbst an, sondern **auf die Veränderung der Korrelationen** zwischen den beiden Zuständen.

Der grosse Vorteil ist, dass die Methode **unvoreingenommen** ist. Sie kann also *de novo* Gen-Module auf Basis gemeinsamer Unterschiede in den Korrelationsmustern finden, ohne auf bereits bekannte Gen-Gruppen angewiesen zu sein. Dabei folgt *DiffCoEx* einem fünfteiligen Prozess (Abb. 11).

Abbildung 11: DiffCoEx-Workflow, eigene Abbildung
Aus den Adjazenzmatrizen der Test- und Kontrollgruppe wird eine Differenzmatrix berechnet. Diese dient als Grundlage für die TOM-basierte Dissimilaritätsmatrix, welche dann in Cluster aufgespalten wird. Als letzter Schritt wird die Signifikanz der entstandenen Cluster ermittelt (z.B. mit Permutationstests).

3.6.1 Adjazenzmatrizen als Ausgangslage

Die Grundlage für die Analyse bilden die für jeden Zustand erstellten Adjazenzmatrizen aus der *bc3net*-Analyse. In diesen Matrizen ist für jedes Genpaar ein Wert hinterlegt, der die Stärke seiner statistischen Verbindung (Koexpression) repräsentiert.

3.6.2 Berechnung der Adjazenzdifferenz-Matrix

Im zweiten Schritt werden die beiden Adjazenzmatrizen verrechnet, um eine einzige **Adjazenzdifferenz-Matrix** zu erstellen. Ein hoher Wert in dieser Matrix bedeutet, dass

[12] Bei der gewöhnlichen Clustersuche wird eine einzelne Adjazenzmatrix als Eingabe für das Topological Overlap Measure (TOM) verwendet. Diese beschreibt die Korrelationsstärke zwischen Genen in einem Zustand (Test- oder Kontrollgruppe). Die entstehende TOM-Matrix beschreibt daher Gene, welche in diesem Zustand dieselben Korrelationspartner (Nachbarn) haben. Das Ergebnis sind Module von Genen, die miteinander koexprimiert sind.

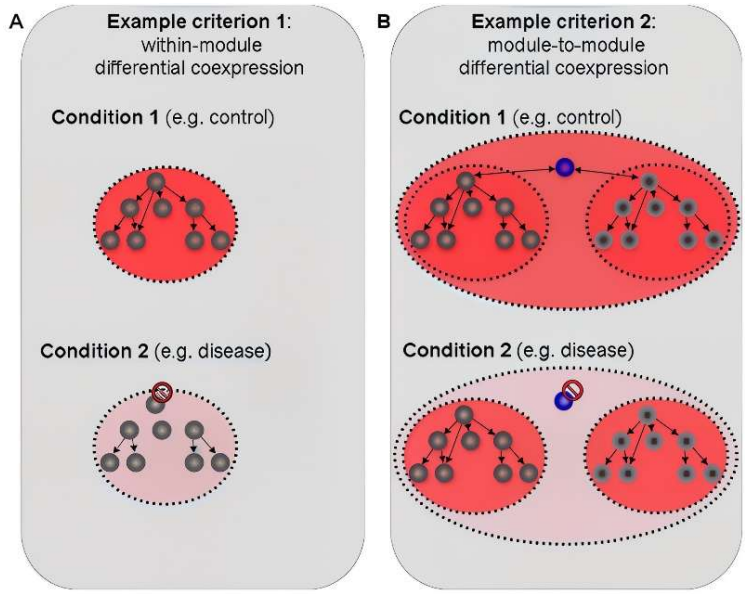

Abbildung 12: Differentialkoexpressions-Szenarios [24]
Fall A: Das Gennetzwerk ist in Bedingung 1 koexprimiert. In Bedingung 2 ist ein wichtiger Regulator jedoch inaktiv und das Modul ist nicht länger koexprimiert. Fall B: Zwei ansonsten unabhängige Signalwege werden in Bedingung 1 durch das blau markierte Gen koordiniert. Seine Inaktivität in Bedingung 2 entfernt die Korrelation zwischen den beiden Modulen – doch die Korrelation innerhalb der Module bleibt bestehen.

sich die Koexpression zwischen zwei Genen stark zwischen den beiden Zuständen unterscheidet (z.B. hohe Korrelation in der Testgruppe, aber tiefe Korrelation in der Kontrolle). Ein tiefer Wert bedeutet, dass ihre Verbindung stabil geblieben ist.

Hierbei wird ein sogenannter Soft-Threshold-Parameter (*beta1*) verwendet. Dieser dient dazu, grosse und damit bedeutsame Korrelationsunterschiede stärker zu gewichten als kleine, potenziell durch Rauschen verursachte Unterschiede. Konzeptionell funktioniert dieser Parameter wie ein Verstärker für die relevantesten Signale in den Daten. Für die Wahl von *beta1* wurde versucht, ein Netzwerk zu erstellen, welches das Skalenfreiheits-Kriterium erfüllte.

3.6.3 Topologischer Overlap zur Identifikation gemeinsamer Muster

In einem nächsten Schritt wird der **Topologische Overlap Measure (TOM)** berechnet. Zwei Gene werden hier als «ähnlich» betrachtet, wenn sie signifikante Korrelationsänderungen mit derselben Gruppe von Nachbargenen aufweisen.

Dieser Ansatz ermöglicht die Entdeckung von zwei wesentlichen Arten differentieller Koexpression (Abb. 12):

A. **Innerhalb-Modul-Änderung:** Eine Gruppe von Genen ist in Zustand 1 stark miteinander korreliert, in Zustand 2 jedoch kaum noch (oder umgekehrt). Das gesamte Modul verliert oder gewinnt also seinen inneren Zusammenhalt.

B. **Zwischen-Modul-Änderung:** Zwei Gen-Module A und B werden in beiden Zuständen intern stabil koexprimiert. Jedoch korrelieren die Gene aus Modul A in Zustand 1 stark mit den Genen aus Modul B, während diese Verbindung in Zustand 2 verloren geht. Die Methode erkennt dies, weil alle Gene in Modul A ein ähnliches Muster der «Verbindungsänderung» zu den Genen in Modul B aufweisen.

Aus dieser Berechnung resultiert eine Dissimilaritätsmatrix, welche die Unähnlichkeit zwischen allen Genpaaren auf Basis ihrer gemeinsamen «Veränderungs-Nachbarschaft» beschreibt.

In dieser Matrix stellt ein tiefer Wert eine hohe Ähnlichkeit (niedrige Unähnlichkeit) dar: Dies bedeutet, dass Gen i und j mit denselben Nachbarn eine signifikante Korrelationsänderung aufweisen. Diese Korrelationsänderung ist ihr «topologischer Overlap».

3.6.4 Clustering und Modul-Identifikation

Die im vorherigen Schritt erstellte Dissimilaritätsmatrix dient nun als Input für ein hierarchisches Clustering. Dabei werden die Gene in einem baumartigen Diagramm (Dendrogramm) angeordnet, wobei ähnliche Gene nahe beieinander liegen. Anschliessend wird dieser Baum mithilfe des Algorithmus *Dynamic Tree Cut* in einzelne Äste geschnitten, welche die finalen differentiell koexprimierten Module darstellen.

Die Schnitthöhe *cutHeight* wird dabei empirisch bestimmt: Für Werte zwischen 0.90 und 0.999 analysiert die Arbeit schrittweise, wie sich die Anzahl der Module und deren mediane Grösse verändern. Die Ergebnisse werden grafisch dargestellt, wodurch sich ein Bereich erkennen lässt, in dem sich beide Grössen stabilisieren.

3.6.5 Statistische Signifikanzprüfung

Permutationsbasierte Signifikanztests, wie sie von Tesson et al. (2010) beschrieben werden, wurden in dieser Arbeit nicht durchgeführt (siehe Diskussion).

3.7 Genset-Analyse

3.7.1 Over-Representation Analysis (ORA)

Die Over-Representation-Analyse (ORA) ist ein weit verbreitetes Verfahren, um zu bestimmen, ob bestimmte **bekannte biologische Funktionen** oder Prozesse in einer experimentell abgeleiteten Genliste überrepräsentiert (angereichert) sind.

Für diese Arbeit wurde untersucht, welche *a priori* definierten, ungeordneten Sammlungen von Genen (**Gensets**) [64], in einer Untermenge von „interessanten" Genen häufiger vorkamen, als es zufällig zu erwarten wäre [65], [66].

Der p-Wert für die **Überrepräsentation** (engl. enrichment) kann mit einem hypergeometrischen Verteilungstest berechnet werden[13]:

Gleichung 7: Hypergeometrischer Test für den p-Wert

$$p = P(X \geq x) = 1 - P(X \leq x - 1) = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

[13] Erklärung der Gleichung:

N: die Gesamtanzahl aller Gene im Experiment
n: die Anzahl „interessanter" Gene
M: die Anzahl Gene, die zu einem bestimmten Genset S gehören
x: die Anzahl der „interessante" Gene, welche tatsächlich in diesem Genset S vorkommen

Hier entspricht $\binom{M}{i}\binom{N-M}{n-i}$ der Anzahl Kombinationen, die es gibt, bei denen genau i der gezogenen n Gene im betrachteten Genset S liegen. $\binom{N}{n}$ gibt die Gesamtanzahl aller möglichen Stichproben der Grösse n an. Das Verhältnis Zähler zu Nenner beschreibt also die Wahrscheinlichkeit , dass bei einer zufälligen Auswahl von n Genen aus N genau i davon zum Genset S gehören.

Nach der Berechnung des p-Werts wurde die BH-Korrektur für multiples Testen angewandt werden (siehe Anhang).

3.7.2 Gene Ontology (GO)

Die Anreicherungsanalyse nutzte die Datenbank **Gene Ontology (GO)**. Diese hat drei Kategorien (Tab. 3) [64], [67]:

Tabelle 3: GO-Kategorien

Molecular Function (MF)	Cellular Component (CC)	Biological Process (BP)
Beschreibt die biochemischen Aktivitäten eines von einem einzelnen Genprodukt (z. B. Protein oder RNA). Sie gibt an, **was ein Genprodukt tun kann**, jedoch nicht, wo oder wann es dies tut.	Beschreibt den **Ort innerhalb der Zelle**, an dem eine molekulare Funktion stattfindet.	Beschreibt **übergeordnete biologische Abläufe**, die durch das Zusammenspiel vieler molekularer Funktionen entstehen.

3.7.3 Praktische Umsetzung

Die Berechnung erfolgte mithilfe der Funktion **enrichGO** des Moduls **clusterProfiler** (Code-Block 4).

Code-Block 4: Enrichment

```
# Convert gene symbols to Entrez IDs for each module
modulesMergedEntrez <- lapply(modulesMerged, function(gene_set) {
  result <- bitr(gene_set,
                 fromType="SYMBOL",
                 toType="ENTREZID",
                 OrgDb = org.Hs.eg.db)
  return(result$ENTREZID)
})


# Perform GO enrichment analysis for each module
enrichmentResults_ontALL <- lapply(modulesMergedEntrez, function(gene_set)
{
  enrichGO(
    gene = gene_set,
    OrgDb = org.Hs.eg.db,
    keyType = "ENTREZID",
    ont = "ALL",
    pAdjustMethod = "BH",
    readable = TRUE,
    pvalueCutoff = 0.05,
    qvalueCutoff = 1)
})
```

4 Ergebnisse

4.1 Normalisierung

Nach der **log2-Transformation** veränderte sich die Verteilung der Expressionswerte deutlich: Der ursprünglich exponentielle Verlauf mit einer ausgeprägten Spitze bei $x = 0$ (Abb. 13) flachte ab. Stattdessen zeigte sich eine breitere, gestreckte Hauptspitze bei etwa $x = 9$ (Abb. 14). Auffällig war zudem ein **lokales Maximum bei $x = 0$** sowie ein **Plateau zwischen $x = 0$ und $x = 7$**, bevor die Verteilung ab etwa $x = 7$ annähernd einer Normalverteilung entsprach.

Nach der **Quantilnormalisierung** zeigte sich im Q-Q-Plot (Abb. 14) eine deutlich verbesserte Übereinstimmung der empirischen mit den theoretischen Quantilen. Während die Daten vor der Normalisierung (Abb. 13) stark von der Diagonalen abwichen, verliefen die Punkte nach der Normalisierung weitgehend linear. Lediglich im unteren Bereich ($x < 0$) war eine Abflachung zu erkennen, was darauf hindeutet, dass sehr niedrige Intensitätswerte durch die Normalisierung auf einen engen Wertebereich zusammengezogen wurden.

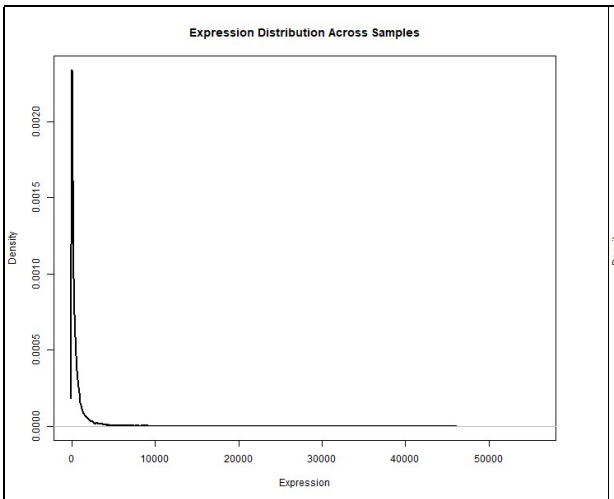

Abbildung 13: Rohdaten, eigene Abbildung
Auf der x-Achse ist der Wert der Expression angegeben. Die y-Achse beschreibt die Anzahl Gene, welche diesen Wert besitzen («density»). Man erkennt deutlich eine Spitze bei ca. x =0 und wie die Häufigkeit daraufhin exponentiell abnimmt.

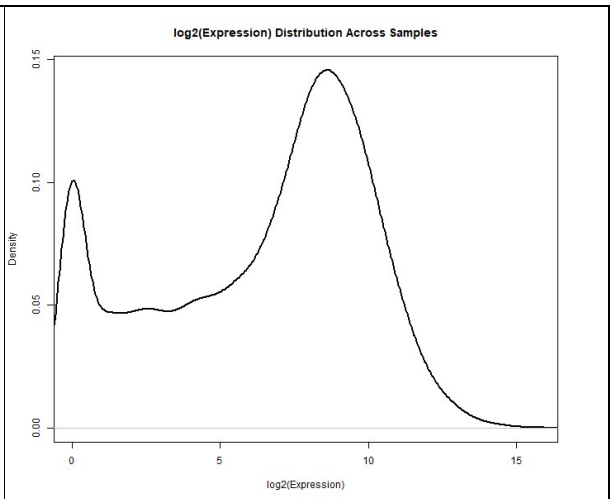

Abbildung 14: log2-verwandelte Daten, eigene Abbildung
Hier zeigt die x-Achse den log2-verwandelten Wert der Expression, während die y-Achse wieder die «density» beschreibt. Die Spitze wurde auf die Werte zwischen etwa 7 und 10 gestreckt. Ein lokales Maximum bei x = 0 wird in dieser Abbildung ebenfalls erkennbar.

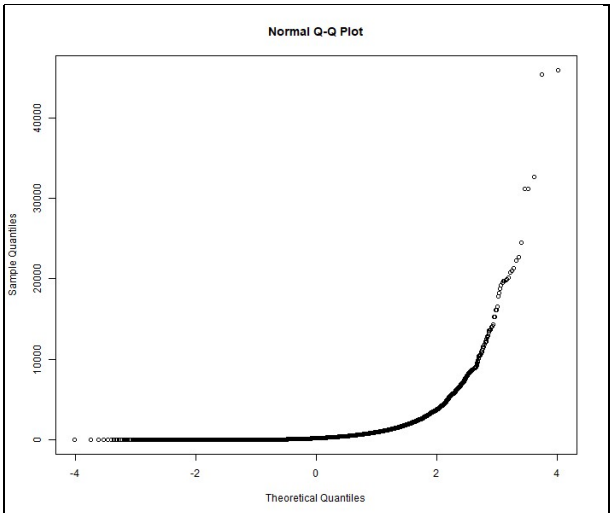

Abbildung 15: Q-Q-Plot der Rohdaten, eigene Abbildung
In der x-Achse sind die theoretisch erwarteten, in der y-Achse die tatsächlich beobachteten Quantile. Die Kurve steigt exponentiell an.

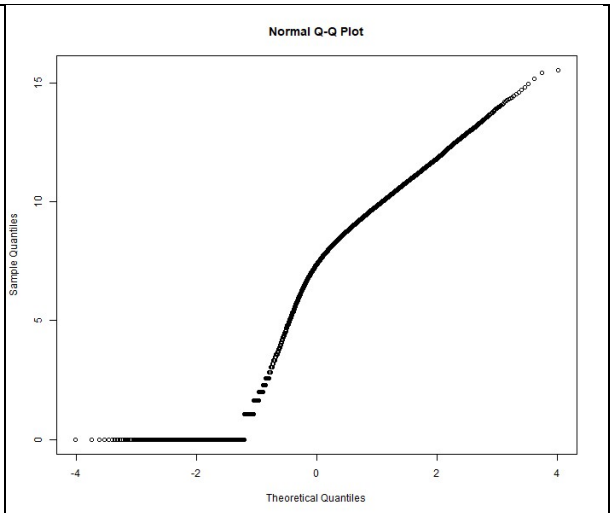

Abbildung 16: Q-Q-Plot der log2-verwandelten, quantilnormalisierten Daten, eigene Abbildung
Bis zum theoretischen Quantil -1 (x-Achse) sind die empirischen Quantile (y-Achse) gleich 0. Im Bereich von -1 bis 0 steigt Kurve stark an und verläuft ab ca. 0 nahezu linear entlang der Diagonalen.

4.2 *bc3net*-Netzwerke

Für beide Bedingungen (Test- und Kontrollgruppe) wurde die grösste zusammenhängende Komponente («Giant Component») der mit *bc3net* inferierten Koexpressionsnetzwerke bestimmt und mit **igraph** visualisiert (Abb. 17).

Beide Netzwerke umfassten 16333 Gene (Knoten), unterschieden sich jedoch leicht in ihrer Topologie (Tab. 4). Das Netzwerk der Testgruppe (netTest) enthielt 117706 Kanten und wies damit eine leicht höhere Dichte ($8.825 \cdot 10^{-4}$) als das Kontrollnetzwerk netControl mit 114537 Kanten (Dichte: $8.588 \cdot 10^{-4}$) auf. Der durchschnittliche Knotengrad war mit 14.41 gegenüber 14.03 geringfügig erhöht, während der Durchmesser mit 2.04 im Vergleich zu 1.73 leicht grösser war.

In Abbildung 20 ist zudem erkennbar, dass die Gradverteilungen der beiden Netzwerke nur ab einem x-Achsen-Wert von 2 annähernd dem **Potenzgesetz** (Kap. 2.3.4) entsprachen.

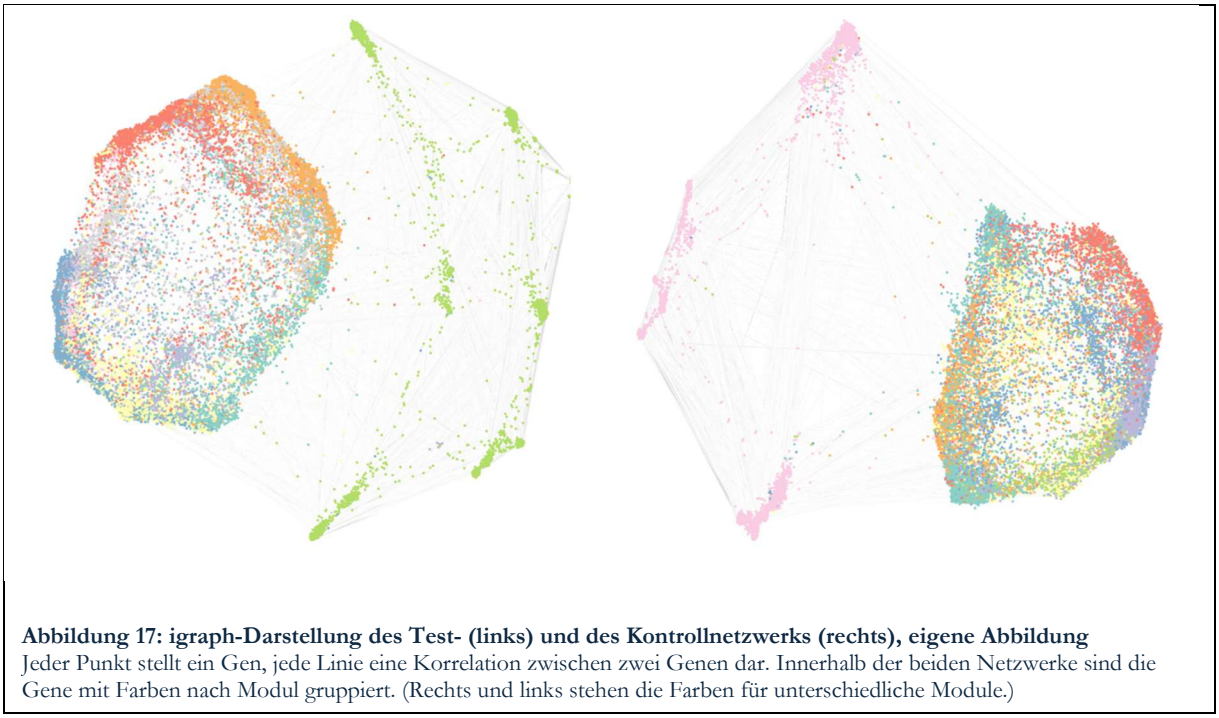

Abbildung 17: igraph-Darstellung des Test- (links) und des Kontrollnetzwerks (rechts), eigene Abbildung
Jeder Punkt stellt ein Gen, jede Linie eine Korrelation zwischen zwei Genen dar. Innerhalb der beiden Netzwerke sind die Gene mit Farben nach Modul gruppiert. (Rechts und links stehen die Farben für unterschiedliche Module.)

Tabelle 4: Eigenschaften der Netzwerke

	netTest	**netControl**
Anzahl Knoten	16333	16333
Anzahl Kanten	117706	114537
Dichte	$8.825 \cdot 10^{-4}$	$8.588 \cdot 10^{-4}$
Durchmesser	2.04	1.73
Ø Knotengrad	14.41	14.03

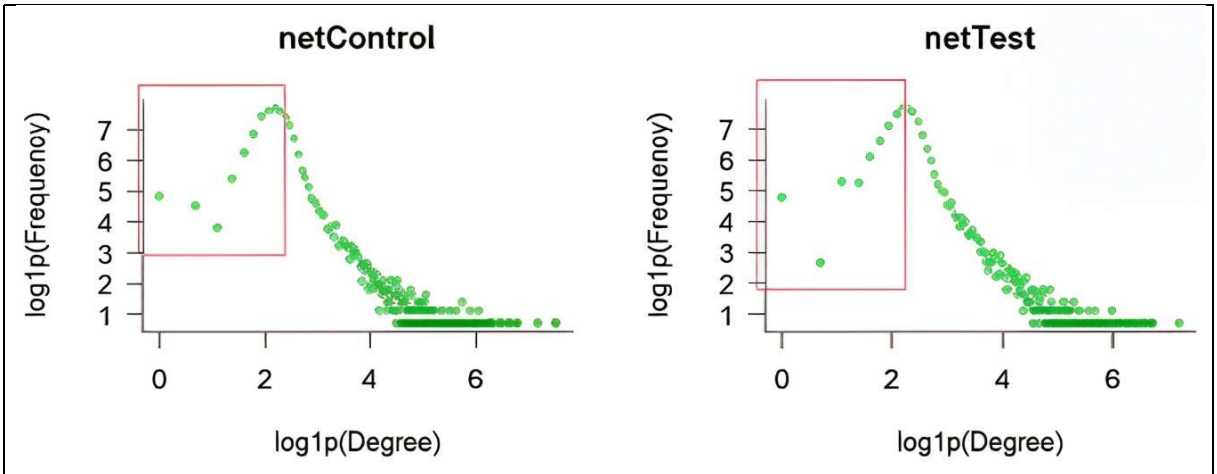

Abbildung 18: Beobachtete Gradverteilung, eigene Abbildung
Jeder Punkt in diesem Graphen stellt die Anzahl Knoten (Gene) auf der y-Achse mit einer bestimmten Anzahl Kanten (Verbindungen) auf der x-Achse dar. Die Gene im markierten Bereich weichen von der skalenfreien Optimalverteilung ab.

4.3 Skalenfreie Topologie

Keiner der vorgeschlagenen «powers» erreichte den R^2-Cut von 0.85; somit konnte mit keinem Parameter eine annähernd skalenfreie Topologie für die TOM-Dissimilaritätsmatrix erreicht werden. Der höchste Wert (0.54) trat bei beta1 $= 3$ auf (Abb. 19), sodass dieser Wert manuell als Soft-Threshold festgelegt wurde.

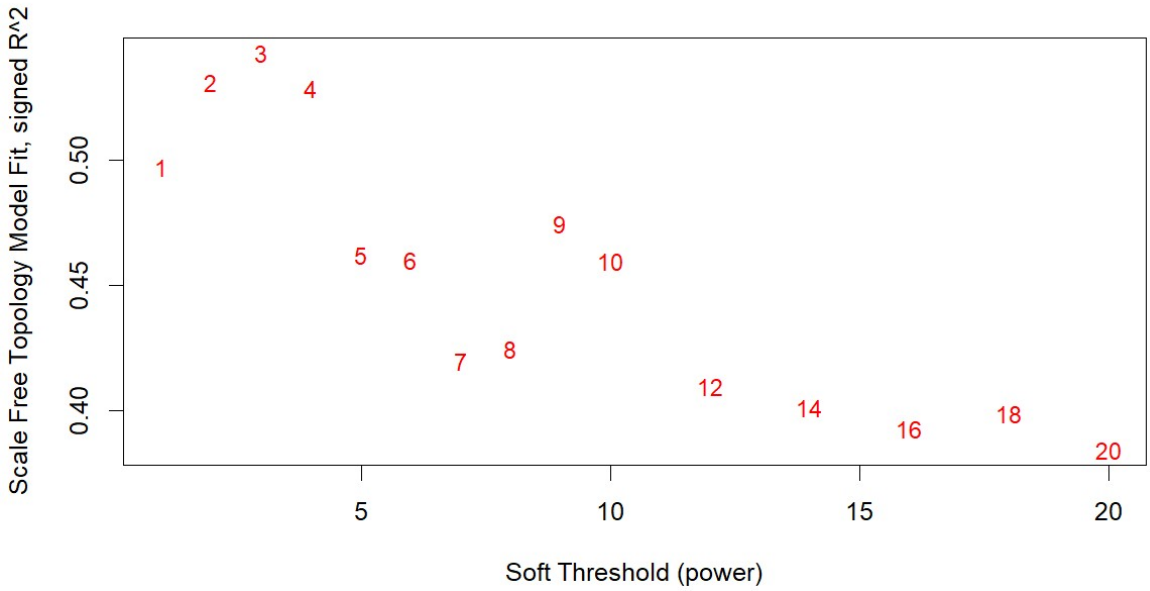

Abbildung 19: Erreichte R^2-Werte für verschiedene mögliche Thresholds, eigene Abbildung
Die Zahlen im Diagramm stehen für das getestete Threshold («power»). Je weiter oben sie sich befinden, desto mehr erfüllen die mit ihnen erstellten Dissimilaritätsmatrizen (dissTOM) das Kriterium der Skalenfreiheit. Unerwartet war, dass keiner der Thresholds einen Wert über 0.54 erreichte, obwohl die Erwartung bei über 0.85 lag.

4.4 Wahl des Parameters *cutHeight*

„Modules vs cutHeight“ und „Median size vs cutHeight“ (Abb. 20) zeigen, dass sich die Modulstruktur im Bereich von etwa $0.94 - 0.96$ stabilisierte. Oberhalb dieser Schwelle blieb die Modulanzahl weitgehend konstant, während der zuvor dominante Grosscluster zerfiel und die mediane Modulgrösse abrupt abnahm. Beide Grössen veränderten sich in diesem Bereich nur noch gering, was auf eine **robuste** und biologisch gut interpretierbare **Clustereinteilung** hinweist. Auf Grundlage dieser Beobachtungen wurde eine Schnitthöhe von $\mathbf{0.96}$ gewählt.

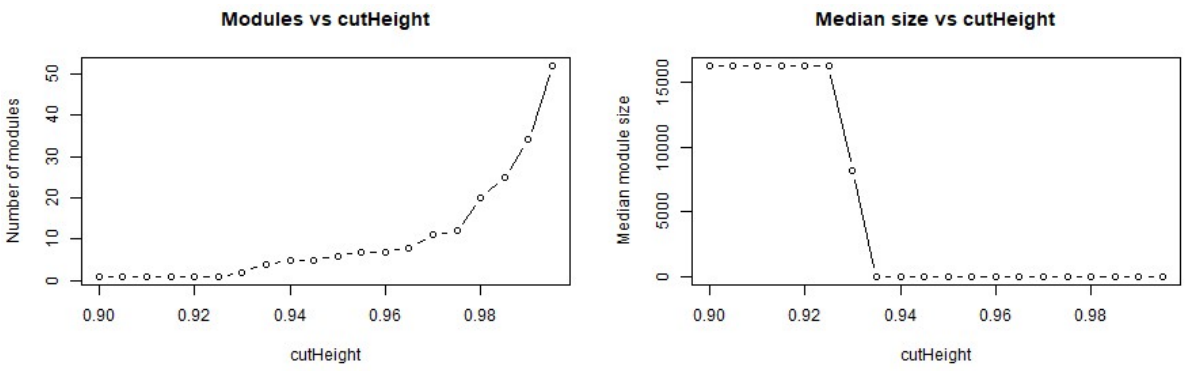

Abbildung 20: Analyse des Einflusses der Schnitthöhe, eigene Abbildung
Im Plot „Modules vs cutHeight" zeigte sich ab einer Schnitthöhe von etwa 0.94 bis 0.96 ein relatives Plateau, in dem die Anzahl der Module weitgehend stabil blieb, bei gleichzeitig noch akzeptabler Modulanzahl. Im Plot „Median size vs cutHeight" hingegen trat bei etwa 0.93 ein markanter Sprung auf: Unterhalb dieser Schwelle dominierte ein sehr grosses Modul, während oberhalb viele kleine, differenzierte Module gebildet wurden.

4.5 Identifizierte differentiell koexprimierte Module[14]

Mit der mit dem TOM berechneten Dissimilaritätsmatrix wurde eine **modulare Analyse** mittels hierarchischem Clustering durchgeführt. Das Verfahren ***Dynamic Tree Cut*** identifizierte dabei Gruppen von Genen mit ähnlichen differentiellen Koexpressionsmustern zwischen den Bedingungen.

Insgesamt konnten sechs Module mit signifikanter differentieller Expression bestimmt werden (Tab. 5). Jedes Modul umfasst Gene, die deren Korrelation sich auf eine ähnliche Weise ändert.

Die Module unterschieden sich sowohl in Grösse als auch in funktioneller Zusammensetzung. Beispiele sind das rote Modul, das u. a. *CABP7* und *CADM4* enthielt, das grüne Modul mit Genen wie *ERRFI1* und *FOXD1* sowie das türkisfarbene Modul, das zahlreiche Gene aus der RNA-Polymerase-I/III-Familie umfasste (*POLR1B*, *POLR3GL*). Weitere Module (gelb, blau, braun) beinhalteten Gene mit potenziellen Funktionen in Signaltransduktion, Transportprozessen und Ubiquitin-vermittelter Regulation.

Das resultierende **Dendrogramm** (Abb. 21) zeigt die hierarchische Gruppierung der Gene basierend auf ihren Differentialkoexpressionsprofilen. Die farbliche Markierung unterhalb der Zweige veranschaulicht die durch *Dynamic Tree Cut* identifizierten Module und erlaubt eine visuelle Zuordnung der Gencluster zu den jeweiligen Modulen. Gut erkennbar ist insbesondere, wie klein der Anteil von bedeutenden Genen (farbig dargestellt) im Vergleich zum unbedeutenden Anteil (grau) war.

Diese Module und deren Gencluster bildeten die Grundlage für die nachfolgende Anreicherungsanalyse zur Identifikation biologischer Prozesse, die durch die VitD3-Behandlung beeinflusst werden könnten.

[14] Achtung: Bei den hier besprochenen Modulen handelt es sich um diejenigen mit einer differentiellen Koexpression zwischen den Bedingungen (Test und Kontrolle), während es in 4.2 um Module von Koexpression innerhalb einer Bedingung ging. Folglich sind die Farben aus 4.2 und 4.4 gänzlich unverwandt.

Tabelle 5: gefundene Module

Modulname (Farbe)	Enthaltene Gene (Gensymbole)	Modulgrösse
red	C6orf136 C7orf60 C9orf64 CABP7 CADM4	5
green	EPS8L1 ERRFI1 FKBP14 FMR1NB FOXD1 NFXL1	6
turquoise	POLM POLR1B POLR1C POLR3GL POM121 POM121C POR POTEE POU2AF1 PP2D1 PPFIA2 PPP1R15B	12
yellow	RHOV RHPN1 RILPL1 RIMS2 RMDN2 RNASE10 RNF24 RORB RP9	9
blue	SKP2 SLC10A2 SLC10A7 SLC16A11 SLC16A12 SLC17A7 SLC18A2 SLC1A4 SLC24A3 SLC25A13	10
brown	USP44 UTP23 UTRN UTS2 UTS2R UXT VAC14 VANGL2 ZNF737	9

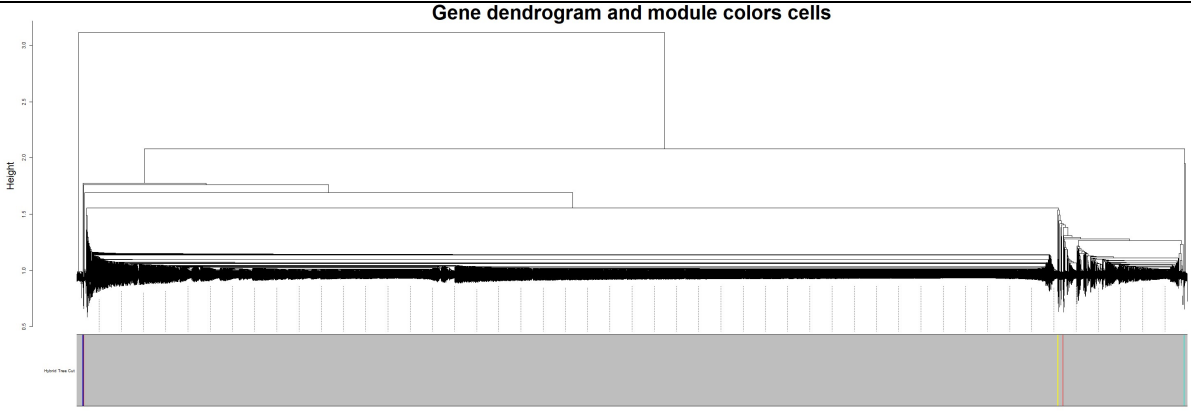

Abbildung 21: Dendrogramm mit Modulzuteilung, eigene Abbildung
Jede Linie stellt ein Modul dar. Auf der tiefsten Stufe haben diese Module die Grösse 1: Sie enthalten ein einziges Gen. Schneidet man das Dendrogramm hoch oben, entstehen grössere Module. Bei Height = 2.5 beispielsweise werden alle vorhandenen Gene in zwei Module aufgeteilt, bereits bei der Höhe 2 hat man drei verschiedene Module. Je tiefer man die Schnitthöhe festsetzt, desto mehr Module erhält man, diese sind jedoch fortschreitend kleiner.

4.6 Überrepräsentierte Signalwege

Um die biologischen Funktionen der identifizierten Module näher zu charakterisieren, wurde eine **funktionelle Anreicherungsanalyse** durchgeführt. Dabei identifizierte die Analyse für jedes Modul die überrepräsentierten Signalwege und funktionellen Kategorien aus der **Gene Ontology (GO)**. Aus den erhaltenen Ergebnissen wurden jeweils die drei Signalwege mit dem kleinsten adjustierten p-Wert ausgewählt[15] (Tab. 6).

Das **rote Modul** zeigte eine deutliche Anreicherung von Signalwegen, die mit der negativen Regulation der vaskulären endothelialen Wachstumsfaktor-(VEGF)-Signalübertragung und der Peptidyl-Threonin-Phosphorylierung assoziiert sind (alle $p = 0.006$). Diese Prozesse deuten auf eine mögliche Beteiligung des Moduls an der Modulation zellulärer Signalaktivität und Gefässneubildung (Angiogenese) hin.

[15] Die vollständigen Enrichment-Ergebnisse sind auf Github unter „results/GO_enrichment_tables“ auffindbar.

Im **grünen Modul** fanden sich unter anderem Gene mit Funktionen in der T-Zell-Rezeptorbindung, DNA-Biegung sowie GTPase-Regulatoraktivität ($p = 0.021$). Diese Befunde weisen auf eine potenzielle Rolle des Moduls in der Signaltransduktion und Transkriptionsregulation hin.

Das **türkisfarbene Modul** war stark mit Komponenten des RNA- und DNA-Polymerase-Komplexes sowie mit katalytischer Aktivität auf RNA assoziiert ($0.001 < p < 0.04$). Diese Anreicherungen deuten auf eine funktionelle Einbindung in die Transkription und RNA-Verarbeitung hin.

Für die **gelben** und **braunen** Module konnte keine signifikante funktionelle Anreicherung festgestellt werden, während das **blaue Modul** Gene mit aktiver transmembraner Transportaktivität enthielt ($p < 0.001$), darunter sekundäre aktive Transporter und Symporter.

Insgesamt zeigten die Ergebnisse, dass die durch *Dynamic Tree Cut* identifizierten Module klar unterschiedlichen biologischen Prozessen zugeordnet werden konnten.

Tabelle 6: überrepräsentierte Signalwege

Modul	Beschreibung	Ontologie	p.adjust
red	negative regulation of vascular endothelial growth factor receptor signaling pathway	BP	0.006
	negative regulation of peptidyl-threonine phosphorylation	BP	0.006
	negative regulation of vascular endothelial growth factor signaling pathway	BP	0.006
green	T cell receptor binding	MF	0.021
	DNA binding, bending	MF	0.021
	GTPase regulator activity	MF	0.021
turquoise	RNA polymerase complex	CC	< 0.001
	DNA polymerase activity	CC	< 0.001
	catalytic activity, acting on RNA	CC	0.043
yellow	*No significant enrichment*		
blue	secondary active transmembrane transporter activity	MF	<0.001
	active transmembrane transporter activity	MF	<0.001
	symporter activity	MF	<0.001
brown	*No significant enrichment*		

5 Diskussion

In der vorliegenden Diskussion werden die zentralen **biologischen Ergebnisse** dieser Arbeit eingeordnet und im Kontext der **bestehenden Forschung** zu VitD3, T-Zell-Regulation und Rheumatoider Arthritis interpretiert. Besonderes Augenmerk liegt auf der Auswahl und Aussagekraft des verwendeten Datensatzes, der Bedeutung der mit *DiffCoEx* identifizierten Module sowie der biologischen Relevanz der angereicherten Signalwege. Anschliessend werden **methodische Entscheidungen und Limitationen**, insbesondere im Hinblick auf die Netzwerkanalyse und statistische Verfahren, kritisch reflektiert.

5.1 Die Wahl des Datensets

Die Analyse nutzte ein Datenset (GSE128816), das die Genexpression in T-Zellen erfasste, welche mit VitD3-tolDCs ko-kultiviert wurden. Dies war **keine intuitive Wahl** für die Erforschung des Effektes von VitD3auf RA, basierte allerdings auf mehreren methodischen und biologischen Überlegungen.

T-Zellen spielen eine **zentrale Rolle** in der Pathogenese der Rheumatoiden Arthritis (RA) und anderer Autoimmunerkrankungen. Ein entscheidender Mechanismus zur Wiederherstellung der Immuntoleranz ist die Induktion von regulatorischen T-Zellen (T_{reg}-Zellen). Es ist bekannt, dass **tolDCs** die Fähigkeit besitzen, naive T-Zellen in T_{reg}-Zellen zu differenzieren [68], [69], [70]. Ein gängiger und effektiver Ansatz zur Generierung von tolDCs *in vitro* ist die Behandlung mit VitD3. Das gewählte Datenset (GSE128816) nutzt genau diesen Mechanismus.

TolDCs können auch durch **andere Stimuli** wie Retinsäure (Vitamin A) TGF-β induziert werden. Die resultierenden tolDC-Subtypen weisen allerdings **unterschiedliche molekulare Profile** und immunmodulatorische Eigenschaften auf. VitD3-induzierte tolDCs stellen dabei einen gut charakterisierten und häufig verwendeten tolDC-Typ dar, dessen Wirkmechanismen in der Regulation von T-Zellen und insbesondere in der Induktion von T_{reg}-Zellen **umfassend beschrieben** sind. Die Wahl eines Datensets, das explizit VitD3-induzierte tolDCs nutzt, ist daher sinnvoll, da es einen klar definierten und biologisch relevanten Signalweg repräsentiert.

Es muss angemerkt werden, dass das gewählte Datenset **nicht RA-spezifisch** war; die verwendeten Zellen stammen nicht von RA-Patienten. Die Relevanz von GSE128816 liegt jedoch darin, dass der untersuchte Mechanismus – die Modulation von T-Zellen durch tolDCs – einen fundamentalen Prozess abbildet, der bei der RA beeinträchtigt ist. Die Ergebnisse bieten daher Einblicke in grundlegende regulatorische Prozesse, die auf die RA **übertragbar** sind.

Ein weiteres wesentliches Auswahlkriterium war die **Grösse des Datensets**. Die Genexpressionsanalyse in dieser Arbeit erforderte eine ausreichende Anzahl an Samples pro Behandlungsgruppe, um statistisch robuste und zuverlässige Ergebnisse zu erzielen [71]. Altay et al. (2023) schlugen für verlässliche Gennetzwerke **zwischen 30 und 85 Samples** vor [72]. Obwohl GSE128816 dieses Kriterium mit jeweils zehn Samples pro Gruppe nicht erfüllte, verbesserte das Bagging mit *bc3net* die statistische Aussagekraft der Ergebnisse.

Schliesslich war zum Zeitpunkt der Recherche das analysierte Datenset das **einzige öffentlich verfügbare Set**, das die Kombination aus VitD3-Modulation, den RA-relevanten Immunzelltypen (T-Zellen und DCs) und einer vergleichsweise grossen Stichprobengrösse bot.

5.2 Die Bedeutung von *DiffCoEx*

Der Einsatz von *DiffCoEx* war notwendig, da eine gewöhnliche Clustersuche, die koexprimierte Gene in *einem* Zustand identifiziert, die hier relevante Dynamik nicht erfasst hätte.

Wie in Abschnitt 3.6 dargelegt, verschiebt *DiffCoEx* die Analyseebene: Es fragt nicht, welche Gene «verbunden» sind, sondern welche Gene ihr «Verbindungsmuster auf dieselbe Weise ändern».

Der entscheidende Vorteil dieses Vorgehens war die Fähigkeit, neben **Innerhalb-Modul-Änderungen** (Verlust des inneren Zusammenhalts) auch **Zwischen-Modul-Änderungen** (Verbindungsänderung eines Moduls zu anderen Modulen, ohne den eigenen inneren Zusammenhalt zu verlieren) zu erkennen. Letztere wären durch Standard-Koexpressionsanalysen unerkannt geblieben.

5.3 Biologische Interpretation der gefundenen Pathways

Unter den angereicherten Pathways waren insbesondere die folgenden drei Signalwege im Kontext der Rheumatoiden Arthritis interessant.

5.3.1 Der VEGF-Signalweg

Ein zentrales Ergebnis dieser Arbeit war die Anreicherung der Pathways aus dem Modul «**red**», die am **VEGF-Signalweg** beteiligt sind, insbesondere:

- «negative regulation of VEGF receptor signaling pathway»
- «negative regulation of VEGF signaling pathway»
- «negative regulation of VEGF cellular response to stimulus»

Der vaskuläre endotheliale Wachstumsfaktor (VEGF) ist ein Schlüsselfaktor in der Pathogenese der Rheumatoiden Arthritis, indem er die **synoviale Angiogenese** fördert. Dies bezeichnet die Bildung neuer Blutgefässe in der Gelenkinnenhaut (Synovia). Diese pathologische Gefässneubildung ist entscheidend für das Fortschreiten der Krankheit, da sie die Zufuhr von Sauerstoff und Nährstoffen zum Gelenk sicherstellt und so die Entzündung aufrechterhält. Die Anreicherung dieser Pathways deutet somit darauf hin, dass die in unseren Modulen gefundenen Gene an einer **aktiven Unterdrückung** dieses krankheitsfördernden Prozesses beteiligt sind.

Dieses Ergebnis stand in Einklang mit den Beobachtungen von Zhao et al. (2020), wonach VitD3 die Herstellung von VEGF in Mastzellen unterdrückt [73]. Eine klinische Studie von Irani et al. (2017) fand in Frauen mit PCOS eine bedeutende Herunterregulation von VEGF nach einer Behandlung mit VitD3 [74]. Allerdings beobachteten Cardus et al. (2009) nach Zugabe von VitD3 *in vitro* eine Hochregulierung von VEGF in vaskulären glatten Muskelzellen [75]. Möglicherweise spielt das Vitamin in verschiedenen **zellulären Kontexten** unterschiedliche Rollen.

Nichtsdestotrotz zeigte die Identifizierung dieser Pathways einen Mechanismus auf, durch den die **pathologische Angiogenese bei RA gehemmt** werden könnte. Dies stellt einen wichtigen Aspekt der immunmodulatorischen Wirkung der VitD3-tolDCs dar.

5.3.2 Die T-Zell-Rezeptor-Bindung

Im Modul «**green**» war der Pathway «T cell receptor binding» bedeutsam. Dies suggeriert eine direkte Modulation der **Signalübertragung bei Antigen-Rezeptor-Komponenten**. Dies

stimmte mit den Beobachtungen von Hafkamp et al. (2020) und van der Aar et al. (2011) überein, dass VitD3-tolDCs in der Lage sind, naive T-Zellen zur Entwicklung zu **T_{reg}-Zellen** zu beeinflussen [76], [77].

5.3.3 Transkriptionelle Prozesse

Des Weiteren zeigte sich eine signifikante Anreicherung von Signalwegen, die **transkriptionellen Mechanismen** zuzuordnen sind. Dazu gehörten fundamentale Mechanismen wie die Aktivität von RNA-Polymerase-Komplexen (I/III), die DNA-gerichtete RNA-Polymerase-Aktivität sowie generelle DNA-Bindungs- und Biegeprozesse. Im Modul «**turquoise**» wurden unter anderem die folgenden Pathways gefunden:

- «RNA polymerase complex»
- «DNA polymerase complex»
- «catalytic activity, acting on RNA»

Die Anreicherung dieser Kernprozesse der Genablesung ist ein starker Indikator für tiefgreifende Veränderungen in der Zellregulation. Dies lässt sich auf zwei Arten interpretieren:

Erstens könnte es eine **globale Herunterregulierung biosynthetischer Programme** widerspiegeln. Eine solche allgemeine Drosselung der Zellaktivität würde zu einem Zustand der **Hyporesponsivität** (verringerte Reaktionsfähigkeit) führen, was im Kontext der RA eine geringere Reaktion auf entzündliche Stimuli bedeuten würde. Navarro-Barriuso et al. (2021), die Autoren des in dieser Arbeit verwendeten Datensets, beobachteten diesen Effekt ebenfalls [78].

Zweitens könnte dies eine gezielte **Umgestaltung der transkriptionellen Kontrolle** anzeigen. Im Kontext der untersuchten T-Zellen deutet dieses Ergebnis darauf hin, dass nicht einfach alle Prozesse gehemmt, sondern stattdessen aktiv spezifische **regulatorische Programme** induziert werden. Diese Umprogrammierung der Genexpression ist ein plausibler Mechanismus, der der Entwicklung von naiven T-Zellen zu T_{reg}-Zellen durch Interaktion mit VitD3-tolDCs zugrunde liegen könnte.

5.3.4 Der Bezug zu Rheumatoider Arthritis

Zusammenfassend lässt sich festhalten, dass die Relevanz dieser Analyse für die Rheumatoide Arthritis in der Untersuchung eines potentiellen Toleranzinduktionsmechanismus lag. Das gewählte Datenset (GSE128816) modellierte exakt die Interaktion (VitD3-tolDCs und T-Zellen), die für die Wiederherstellung der Immuntoleranz, primär durch die Induktion von T_{reg}-Zellen, zentral sein könnte.

Die Anwendung von *DiffCoEx* ermöglichte es, über die reine Koexpression hinauszublicken und die **dynamische Neuausrichtung der Gen-Interaktionen** zu erfassen, die dieser immunologischen Modulation zugrunde liegt.

Die Ergebnisse lieferten spezifische, auf die RA-Pathogenese übertragbare Einsichten:

1. Die identifizierten Veränderungen im **VEGF-Signalweg** (Modul «red») deuten auf einen Mechanismus hin, der die krankheitsfördernde Angiogenese im Gelenk aktiv hemmt.

2. Die Modulation des **T-Zell-Rezeptor-Bindings** (Modul «green») und die tiefgreifende Umgestaltung **transkriptioneller Prozesse** (Modul «turquoise») spiegeln die molekulare Basis für die Umprogrammierung von T-Zellen wider.

Diese Mechanismen (die Hemmung der Angiogenese und die Induktion eines hyporesponsiven oder regulatorischen T-Zell-Status) stellen zentrale Ansatzpunkte dar, um den pathologischen Prozessen der RA entgegenzuwirken. Die Analyse lieferte somit, trotz der Verwendung gesunder Spenderzellen, wertvolle Einblicke in die **therapeutisch relevanten Grundlagen** der T-Zell-Modulation.

5.4 Methodische Überlegungen und Schwächen

5.4.1 Skalenfreie Topologie und der Parameter *beta1*

Das **Ausbleiben einer skalenfreien Topologie** im erstellten Netzwerk stellte allerdings dessen biologische Aussagekraft infrage. Ein möglicher Grund für diese Diskrepanz könnte in der Gradverteilung von netTest und netControl liegen. Die rot markierten Bereiche (Gene) in Abbildung 20 (Kap. 4.2) stimmen nicht mit dem Potenzgesetz überein. Eine strengere Vorfilterung niedrig exprimierter Gene könnte die Topologie verbessern.

Folglich ist es möglich, dass das hier generierte Netzwerk nicht die robuste, fehler-tolerante Struktur repräsentierte, die für biologische Systeme erwartet wird. Die identifizierten Verbindungen und Module könnten daher **weniger aussagekräftig** sein [16]. Zukünftige Analysen sollten insbesondere die erstellten Netzwerke vor der Anwendung von *DiffCoEx* systematisch neu bewerten, um den R^2-Wert für die skalenfreie Topologie zu maximieren und so die biologische Relevanz des Modells zu verbessern.

5.4.2 Transformation und Verteilungsproblematik

Die **log2-Transformation** spreizte den Wertebereich erfolgreich, sodass relevante Signalunterschiede deutlicher hervortraten (Abb. 14). Eine annähernde Normalverteilung zeigte sich jedoch erst ab einem Wert von ca. $x = 7$, während bei $x = 0$ ein lokales Maximum (Häufung niedriger Intensitäten) bestehen blieb.

Diese Abweichung ist problematisch, da der *bc3net*-Algorithmus eine **Normalverteilung voraussetzt**. Die Überrepräsentation niedriger Werte könnte die Schätzung der Zufallsverteilung verzerren: Es besteht das Risiko, dass schwach gestützte Kanten fälschlicherweise als signifikant eingestuft (Überschätzung des Rauschens) und echte Zusammenhänge in normalverteilten Bereichen unterschätzt werden. Dies mindert die Trennschärfe und könnte die **Netzwerkstruktur verfälschen**.

5.4.3 Normalisierung

Die anschliessende **Quantilnormalisierung** trug jedoch wesentlich zur Vereinheitlichung der Signalverteilungen bei. Der über weite Strecken lineare Verlauf des Q-Q-Plots belegt eine **erfolgreiche Angleichung** der empirischen an die theoretische Verteilung (Abb. 16). Damit

[16] Umso wichtiger ist der Vergleich mit der bereits vorhandenen Literatur (Kap. 5.3).

wurden technische Varianzen zwischen den Arrays reduziert und die Vergleichbarkeit der Proben sichergestellt.

Die Abflachung im linken Bereich des Plots (bei $x < 0$) deutet auf eine **starke Komprimierung** sehr niedriger Intensitätswerte hin. Dieser **Informationsverlust** ist methodisch akzeptabel, da er die Reproduzierbarkeit in den biologisch relevanteren mittleren und höheren Bereichen verbessert.

Zusätzlich zur Quantilnormalisierung sollte eine Korrektur möglicher **Batch-Effekte** in Betracht gezogen werden. Systematische Unterschiede zwischen den Versuchsgruppen könnten sonst die Expressionswerte verzerren und zu falschen biologischen Schlussfolgerungen führen [79].

5.4.4 Wahl der Schnitthöhe im hierarchischen Clustering

Die Auswahl der Schnitthöhe (*cutHeight*) im hierarchischen Clustering bestimmt massgeblich die resultierende Modulstruktur, da sie festlegt, bis zu welchem **Ähnlichkeitsgrad** Gene zu einem gemeinsamen Cluster zusammengefasst werden. Eine zu niedrige Schnitthöhe führt typischerweise zu einer Überfragmentierung des Netzwerks in zahlreiche kleine Cluster, während eine zu hohe Schwelle dazu neigt, funktionell heterogene Gene in wenigen grossen Modulen zusammenzufassen.

Die Schnitthöhe von $\mathbf{0.96}$ stellte einen sinnvollen Kompromiss zwischen **Kohärenz** und **Modulgranularität** dar. Diese Schwelle vermied sowohl die Bildung eines dominanten, biologisch schwer interpretierbaren Grossmoduls als auch eine übermässige Fragmentierung in zahlreiche Kleinstmodule. Obwohl diese Entscheidung empirisch ist, wies die Stabilität der Modulstruktur in benachbarten Bereichen ($0.95 - 0.97$) auf eine robuste und konsistente Clusterbildung hin.

5.4.5 Signifikanztest

Aus der empirischen Parameterwahl ergibt sich unmittelbar die Frage nach deren statistischer Absicherung: Da der Soft-Threshold (*beta1*) und die Schnitthöhe (*cutHeight*) experimentell gewählt werden, sieht der *DiffCoEx*-Algorithmus als fünften und letzten Schritt ein **Permutationsverfahren** vor, um die statistische Signifikanz der identifizierten differentiell koexprimierten Module zu bewerten. Dieses Verfahren wurde in der vorliegenden Arbeit nicht angewendet.

Der Grund dafür lag in der Abweichung der hier verwendeten Netzwerkmethodik von der Originalimplementierung: Die von den *DiffCoEx*-Autoren vorgeschlagene Permutationsroutine basiert auf einfachen **Korrelationsmatrizen**, während in dieser Arbeit Netzwerke mit *bc3net* erstellt wurden. Letztere Methode lieferte robustere Schätzungen der Beziehungen zwischen Genen, war jedoch **wesentlich rechenintensiver**[17].

Da für eine verlässliche Signifikanzabschätzung typischerweise **mehrere Hundert bis Tausend Permutationen** erforderlich wären, erschien eine vollständige Implementierung in Kombination mit *bc3net* **rechnerisch nicht praktikabel**. Die Bewertung der Ergebnisse stützte sich daher primär auf biologische Plausibilität und Konsistenz über verschiedene Analyseschritte hinweg. Möglicherweise wäre die Verwendung eines weniger rechenintensiven Verfahrens sinnvoll.

[17] Ein *bc3net*-Durchgang mit allen Daten dauerte etwa 24 Stunden.

5.5 Vergleichbarkeit und Reproduzierbarkeit

Ein zentrales Anliegen dieser Arbeit war, reproduzierbare Ergebnisse und vor allem **nachvollziehbare Methoden** bereitzustellen. Viele der wissenschaftlichen Publikationen, die im Rahmen der Recherche konsultiert wurden, erwiesen sich als nur eingeschränkt reproduzierbar: Teilweise war der Code fest an einen spezifischen Datensatz gebunden, teilweise wurde er überhaupt nicht offengelegt. Daher wurden für diese Arbeit sämtliche Analyseschritte vollständig auf **GitHub** (Kap. 3.2) dokumentiert, sodass sie transparent eingesehen und bei Bedarf erneut ausgeführt werden können.

Zum verwendeten Algorithmus *bc3net* ist eine Anmerkung erforderlich: Da *bc3net* auf Bootstrap-Aggregation basiert, ist das Verfahren **stochastisch** und kann bei wiederholten Durchläufen leicht unterschiedliche Netzwerke generieren. Diese Variabilität erschwert direkte Vergleiche zwischen einzelnen Ergebnissen, reflektiert jedoch gleichzeitig die inhärente Unsicherheit in den zugrunde liegenden Daten. Eine **grössere Anzahl an Bootstraps** könnte die Stabilität und Reproduzierbarkeit der Ergebnisse erhöhen.

Die Arbeit führte **mehrere unabhängige *bc3net*-Analysen** durch, um die Robustheit der Ergebnisse zu überprüfen. Dabei fiel eine unerwartete **Diskrepanz** zwischen zwei Berechnungsvarianten auf: In einem Fall wurde innerhalb der Funktion bc3net() ein *igraph*-Objekt erzeugt (igraph=TRUE) und anschliessend in eine Adjazenzmatrix konvertiert; im anderen Fall wurde direkt eine Adjazenzmatrix erstellt (igraph=FALSE). Trotz identischer Eingabedaten führten diese beiden Ansätze zu unterschiedlichen Netzwerkstrukturen.

Zudem zeigten sich deutliche Unterschiede in der Rechenleistung zwischen **Betriebssystemen**. Während Windows-Systeme bc3net(igraph=TRUE) innerhalb von etwa 24 Stunden erfolgreich ausführten, konnte auf macOS-Systemen die Berechnung aufgrund von **Speicherlimitierungen** nicht abgeschlossen werden. Umgekehrt verlief die Berechnung mit igraph=FALSE auf Macs deutlich schneller (ca. 2 Stunden) als auf Windows-Rechnern.

Für die Analysen wurden daher die Netzwerke verwendet, die auf dem Windows-System mit igraph=TRUE erzeugt wurden, da diese Variante konsistent mit der **Originalimplementierung** des *bc3net*-Workflows ist. Künftige Arbeiten könnten systematisch untersuchen, ob die beobachteten Unterschiede durch betriebssystemspezifische Speicherverwaltung, Bibliotheksversionen oder Unterschiede in der parallelen Verarbeitung bedingt sind. Eine Wiederholung der Analyse unter **Linux** wäre ein geeigneter Ansatz, um diese technische Unsicherheit weiter zu minimieren.

5.6 Ausblick

Zusammenfassend sind die **identifizierten Module biologisch plausibel** und stimmen mit der Literatur überein. Sie sind jedoch aufgrund der genannten methodischen Einschränkungen **explorativ** zu verstehen.

Zukünftige Arbeiten könnten auf dieser Analyse aufbauen, indem sie grössere, **RA-spezifische Datensätze** verwenden und die Effekte von VitD3 auf dendritische Zellen vertieft untersuchen. Die **experimentelle Validierung** der identifizierten Signalwege, idealerweise direkt an Proben von RA-Patienten, könnte die Relevanz der Befunde untermauern. Methodisch sollte darauf geachtet werden, eine skalenfreie Netzwerktopologie zu erhalten, die Datenvorverarbeitung zu optimieren

und die gewählten Parameter mittels Permutationstests zu validieren. Angesichts der begrenzten Robustheit der aktuellen R-Implementierung von *bc3net* könnte zudem der Einsatz **alternativer Inferenzmethoden** die Stabilität und Zuverlässigkeit der Ergebnisse verbessern.

6 Fazit

Diese Arbeit untersuchte, wie VitD3-behandelte tolerogene dendritische Zellen (tolDCs) die Gen-Koexpressionsnetzwerke von T-Zellen beeinflussen. Mittels der Algorithmen *bc3net* und *DiffCoEx* wurden differenziell aktive Genmodule identifiziert.

Die biologische Auswertung deutet auf eine Beeinflussung wichtiger Signalwege hin, darunter die **negative Regulation des VEGF-Signalwegs**, die **T-Zell-Rezeptor-Bindung** sowie die **Polymerase-Aktivität**. Diese Ergebnisse stützen die immunmodulatorische Rolle von VitD3 und bieten potenzielle Ansatzpunkte für Therapien.

Die Aussagekraft ist durch wesentliche methodische Limitationen eingeschränkt. Insbesondere erreichte das konstruierte Netzwerk keine skalenfreie Topologie, was dessen biologische Relevanz infrage stellt.

Zukünftige Forschung muss daher die Methodik verbessern:

1. Die Daten sollten besser vorverarbeitet werden und die Normalisierung sollte durch eine Batch-Effekt-Korrektur ergänzt werden.
2. Die Netzwerk-Konstruktion muss optimiert werden, um eine skalenfreie Topologie zu erreichen.
3. Möglicherweise sollte eine alternative Methode zu *bc3net* gefunden werden.
4. Zur Validierung der Parameter sollte ein Permutationstest durchgeführt werden.

Der nächste Schritt ist die **experimentelle Validierung** der bioinformatischen Hypothesen. Insbesondere die funktionelle Relevanz der identifizierten Signalwege (z.B. die Hemmung des VEGF-Pathways) könnte im Labor überprüft werden.

7 Verzeichnisse

7.1 Quellenverzeichnis

[1] 'Induktion tolerogener dendritischer Zellen'. 2015. [Online]. Available: https://epub.uni-regensburg.de/31977/1/Promotion_epub_MP_21.06.2015.pdf

[2] D. Eckebrecht, 'Immunbiologie', in *Natura 9-12: Grundlagen der Biologie für Schweizer Maturitätsschulen*, 5. Auflage (unveränderter Nachdruck der 1. Auflage 2018)., Klett und Balmer Verlag.

[3] M. Bscheider and E. C. Butcher, 'Vitamin D immunoregulation through dendritic cells', *Immunology*, vol. 148, no. 3, pp. 227–236, 2016, doi: 10.1111/imm.12610.

[4] D. Eckebrecht, 'Anatomie und Physiologie der Lebewesen', in *Natura 9-12: Grundlagen der Biologie für Schweizer Maturitätsschulen*, 5. Auflage (unveränderter Nachdruck der 1. Auflage 2018)., Klett und Balmer Verlag.

[5] S. T. Galatage, 'Rheumatoid Arthritis: Severity Classification, Factors Responsible, Pathophysiology, Current and Herbal Treatment', in *Rheumatoid Arthritis*, Sant Gajanan Maharaj College of Pharmacy, India, 2022.

[6] D. A. Walsh, 'Angiogenesis and arthritis.', *Rheumatology*, vol. 38, no. 2, pp. 103–112, Feb. 1999, doi: 10.1093/rheumatology/38.2.103.

[7] P. C. Taylor, 'VEGF and imaging of vessels in rheumatoid arthritis', *Arthritis Res.*, vol. 4, no. Suppl 3, pp. S99–S107, 2002, doi: 10.1186/ar582.

[8] 'Fibroblast', *Wikipedia*. Dec. 16, 2023. Accessed: Oct. 13, 2025. [Online]. Available: https://de.wikipedia.org/w/index.php?title=Fibroblast&oldid=240247939

[9] 'Grundlagen der Genetik'. 2014. Accessed: Oct. 07, 2025. [Online]. Available: https://www.ernaehrungs-umschau.de/fileadmin/Ernaehrungs-Umschau/pdfs/pdf_2014/05_14/EU05_2014_M258_M265.pdf

[10] D. Eckebrecht, 'Genetik', in *Natura 9-12: Grundlagen der Biologie für Schweizer Maturitätsschulen*, 5. Auflage (unveränderter Nachdruck der 1. Auflage 2018)., Klett und Balmer Verlag.

[11] Y. Zhao, C. V. Forst, C. E. Sayegh, I.-M. Wang, X. Yang, and B. Zhang, 'Molecular and Genetic Inflammation Networks in Major Human Diseases', *Mol. Biosyst.*, vol. 12, no. 8, pp. 2318–2341, Jul. 2016, doi: 10.1039/c6mb00240d.

[12] K. Kuchta *et al.*, 'Predicting proteome dynamics using gene expression data', *Sci. Rep.*, vol. 8, no. 1, p. 13866, Sep. 2018, doi: 10.1038/s41598-018-31752-4.

[13] G. Sanguinetti and V. A. Huynh-Thu, *Gene Regulatory Networks: Methods and Protocols*. in Springer Protocols. 2019.

[14] K. P. Singh, C. Miaskowski, A. A. Dhruva, E. Flowers, and K. M. Kober, 'Mechanisms and Measurement of Changes in Gene Expression', *Biol. Res. Nurs.*, vol. 20, no. 4, pp. 369–382, Jul. 2018, doi: 10.1177/1099800418772161.

[15] 'Gensonde', *Wikipedia*. Mar. 19, 2025. Accessed: Oct. 22, 2025. [Online]. Available: https://de.wikipedia.org/w/index.php?title=Gensonde&oldid=254345203

[16] 'Network Analysis and Visualization with R and igraph'. Accessed: Oct. 08, 2025. [Online]. Available: https://kateto.net/netscix2016.html

[17] Y. Uzun, *Approaches for benchmarking single-cell gene regulatory network inference methods*. 2023. doi: 10.48550/arXiv.2307.08463.

[18] EMBL-EBI, 'Graph theory: adjacency matrices | Network analysis of protein interaction data'. Accessed: Nov. 15, 2025. [Online]. Available: https://www.ebi.ac.uk/training/online/courses/network-analysis-of-protein-interaction-data-an-introduction/introduction-to-graph-theory/graph-theory-adjacency-matrices/

[19] B. Zhang and S. Horvath, 'A General Framework for Weighted Gene Co-Expression Network Analysis', *Stat. Appl. Genet. Mol. Biol.*, vol. 4, no. 1, Jan. 2005, doi: 10.2202/1544-6115.1128.

[20] 'Scale-free network', *Wikipedia*. Oct. 05, 2025. Accessed: Oct. 25, 2025. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Scale-free_network&oldid=1315229129
[21] E. Almaas, A. Vazquez, and A.-L. Barabasi, 'Scale-free networks in biology', *Biol. Netw.*, vol. 3, Jan. 2013, doi: 10.1142/9789812772367_0001.
[22] S. Bergmann, J. Ihmels, and N. Barkai, 'Similarities and differences in genome-wide expression data of six organisms', *PLoS Biol.*, vol. 2, no. 1, p. E9, Jan. 2004, doi: 10.1371/journal.pbio.0020009.
[23] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabási, 'The large-scale organization of metabolic networks', *Nature*, vol. 407, no. 6804, pp. 651–654, Oct. 2000, doi: 10.1038/35036627.
[24] B. M. Tesson, R. Breitling, and R. C. Jansen, 'DiffCoEx: a simple and sensitive method to find differentially coexpressed gene modules', *BMC Bioinformatics*, vol. 11, no. 1, p. 497, Oct. 2010, doi: 10.1186/1471-2105-11-497.
[25] C. Olsen, P. E. Meyer, and G. Bontempi, 'On the Impact of Entropy Estimation on Transcriptional Regulatory Network Inference Based on Mutual Information', *EURASIP J. Bioinforma. Syst. Biol.*, vol. 2009, no. 1, p. 308959, Nov. 2008, doi: 10.1155/2009/308959.
[26] 'Spearman's rank correlation coefficient', *Wikipedia*. Oct. 01, 2025. Accessed: Oct. 31, 2025. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Spearman%27s_rank_correlation_coefficient&oldid=1314372934
[27] 'Ranking (statistics)', *Wikipedia*. Jun. 10, 2025. Accessed: Oct. 31, 2025. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Ranking_(statistics)&oldid=1294851829
[28] 'Pearson correlation coefficient', *Wikipedia*. Oct. 30, 2025. Accessed: Oct. 31, 2025. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Pearson_correlation_coefficient&oldid=1319520508
[29] 'Pearson Correlation - an overview | ScienceDirect Topics'. Accessed: Oct. 31, 2025. [Online]. Available: https://www.sciencedirect.com/topics/computer-science/pearson-correlation
[30] V. Demberg, 'Mathematische Grundlagen III - Informationstheorie', 2012.
[31] R. de M. Simoes and F. Emmert-Streib, 'Bagging Statistical Network Inference from Large-Scale Gene Expression Data', *PLOS ONE*, vol. 7, no. 3, p. e33624, Mar. 2012, doi: 10.1371/journal.pone.0033624.
[32] A. J. Butte, 'Mutual Information Relevance Networks: Functional Genomic Networks Built From Pair-wise Entropy Measurements', 2002.
[33] 'Quantile', *Wikipedia*. Aug. 11, 2025. Accessed: Nov. 19, 2025. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Quantile&oldid=1305273839
[34] M. bei DocCheck, 'Nullhypothese', DocCheck Flexikon. Accessed: Nov. 19, 2025. [Online]. Available: https://flexikon.doccheck.com/de/Nullhypothese
[35] S. Turney, 'Nullhypothese: Definition, Alternativhypothese, Beispiele', Scribbr. Accessed: Nov. 19, 2025. [Online]. Available: https://www.scribbr.ch/statistik-ch/nullhypothese/
[36] 'Bernoulli-Prozess', *Wikipedia*. Dec. 13, 2024. Accessed: Nov. 19, 2025. [Online]. Available: https://de.wikipedia.org/w/index.php?title=Bernoulli-Prozess&oldid=251205110
[37] 'Binomialverteilung', *Wikipedia*. Oct. 17, 2025. Accessed: Nov. 19, 2025. [Online]. Available: https://de.wikipedia.org/w/index.php?title=Binomialverteilung&oldid=260675866
[38] 'Hypergeometrische Verteilung', *Wikipedia*. Sep. 08, 2025. Accessed: Nov. 19, 2025. [Online]. Available: https://de.wikipedia.org/w/index.php?title=Hypergeometrische_Verteilung&oldid=259576182
[39] 'GEO Accession viewer'. Accessed: Oct. 14, 2025. [Online]. Available: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE128816

[40] 'Microarray normalization'. Accessed: Oct. 31, 2025. [Online]. Available: https://www.cs.cmu.edu/~epxing/Class/10810/lecture/recitation7.pdf
[41] C. Workman *et al.*, 'A new non-linear normalization method for reducing variability in DNA microarray experiments', *Genome Biol.*, vol. 3, no. 9, p. research0048.1, Aug. 2002, doi: 10.1186/gb-2002-3-9-research0048.
[42] 'R: The R Project for Statistical Computing'. Accessed: Oct. 26, 2025. [Online]. Available: https://www.r-project.org/
[43] 'Posit | The Open-Source Data Science Company', Posit. Accessed: Oct. 26, 2025. [Online]. Available: https://posit.co/
[44] R. de M. Simoes and F. Emmert-Streib, *bc3net: Gene Regulatory Network Inference with Bc3net*. (May 06, 2025). Accessed: Oct. 26, 2025. [Online]. Available: https://cran.r-project.org/web/packages/bc3net/index.html
[45] P. Langfelder and S. Horvath, 'WGCNA: an R package for weighted correlation network analysis', *BMC Bioinformatics*, vol. 9, no. 1, p. 559, Dec. 2008, doi: 10.1186/1471-2105-9-559.
[46] E. Neuwirth, *RColorBrewer: ColorBrewer Palettes*. (Apr. 03, 2022). Accessed: Oct. 26, 2025. [Online]. Available: https://cran.r-project.org/web/packages/RColorBrewer/index.html
[47] 'preprocessCore', Bioconductor. Accessed: Oct. 26, 2025. [Online]. Available: http://bioconductor.org/packages/preprocessCore/
[48] G. Csárdi *et al.*, *igraph: Network Analysis and Visualization*. (Oct. 13, 2025). Accessed: Oct. 26, 2025. [Online]. Available: https://cran.r-project.org/web/packages/igraph/index.html
[49] P. L. and S. Horvath, *moduleColor: Basic Module Functions*. (Apr. 09, 2022). Accessed: Oct. 26, 2025. [Online]. Available: https://cran.r-project.org/web/packages/moduleColor/index.html
[50] H. Wickham, T. L. Pedersen, D. Seidel, P. Software, PBC [cph, and fnd, *scales: Scale Functions for Visualization*. (Apr. 24, 2025). Accessed: Oct. 26, 2025. [Online]. Available: https://cran.r-project.org/web/packages/scales/index.html
[51] H. Wickham *et al.*, *dplyr: A Grammar of Data Manipulation*. (Nov. 17, 2023). Accessed: Oct. 26, 2025. [Online]. Available: https://cran.r-project.org/web/packages/dplyr/index.html
[52] 'Create Elegant Data Visualisations Using the Grammar of Graphics'. Accessed: Oct. 26, 2025. [Online]. Available: https://ggplot2.tidyverse.org/
[53] P. Langfelder and S. Horvath, 'Fast R Functions for Robust Correlations and Hierarchical Clustering', *J. Stat. Softw.*, vol. 46, pp. 1–17, Mar. 2012, doi: 10.18637/jss.v046.i11.
[54] G. Yu, L.-G. Wang, Y. Han, and Q.-Y. He, 'clusterProfiler: an R package for comparing biological themes among gene clusters', *Omics J. Integr. Biol.*, vol. 16, no. 5, pp. 284–287, May 2012, doi: 10.1089/omi.2011.0118.
[55] 'org.Hs.eg.db', Bioconductor. Accessed: Oct. 26, 2025. [Online]. Available: http://bioconductor.org/packages/org.Hs.eg.db/
[56] 'enrichplot', Bioconductor. Accessed: Oct. 26, 2025. [Online]. Available: http://bioconductor.org/packages/enrichplot/
[57] M. West, 'Bayesian Factor Regression Models in the "Large p, Small n" Paradigm', in *Bayesian Statistics 7*, V. Lindley, J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. Smith, and M. West, Eds., Oxford University PressOxford, 2003, pp. 733–742. doi: 10.1093/oso/9780198526155.003.0053.
[58] J. L. Horowitz, 'Bootstrap Methods in Econometrics', *Annu. Rev. Econ.*, vol. 11, no. Volume 11, 2019, pp. 193–224, Aug. 2019, doi: 10.1146/annurev-economics-080218-025651.
[59] P. Bůžková, T. Lumley, and K. Rice, 'Permutation and parametric bootstrap tests for gene—gene and gene—environment interactions', *Ann. Hum. Genet.*, vol. 75, no. 1, pp. 36–45, Jan. 2011, doi: 10.1111/j.1469-1809.2010.00572.x.
[60] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, 'Fast unfolding of communities in large networks', *J. Stat. Mech. Theory Exp.*, vol. 2008, no. 10, p. P10008, Oct. 2008, doi: 10.1088/1742-5468/2008/10/P10008.

[61] 'Louvain method', *Wikipedia*. Sep. 29, 2025. Accessed: Nov. 17, 2025. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Louvain_method&oldid=1314054933
[62] 'The DrL graph layout generator — layout_with_drl'. Accessed: Nov. 17, 2025. [Online]. Available: https://r.igraph.org/reference/layout_with_drl.html
[63] G. Stulp, *Chapter 5 layouts | Network Data Visualisation in R – The Patio*. Accessed: Nov. 20, 2025. [Online]. Available: https://stulp.gmw.rug.nl/patio/layoutt.html
[64] G. Yu, *Chapter 5 Overview of enrichment analysis | Biomedical Knowledge Mining using GOSemSim and clusterProfiler*. 2022. Accessed: Oct. 15, 2025. [Online]. Available: https://yulab-smu.top/biomedical-knowledge-mining-book/enrichment-overview.html
[65] E. I. Boyle *et al.*, 'GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes', *Bioinforma. Oxf. Engl.*, vol. 20, no. 18, pp. 3710–3715, Dec. 2004, doi: 10.1093/bioinformatics/bth456.
[66] D. W. Huang, B. T. Sherman, and R. A. Lempicki, 'Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists', *Nucleic Acids Res.*, vol. 37, no. 1, pp. 1–13, Jan. 2009, doi: 10.1093/nar/gkn923.
[67] 'Gene Ontology overview', Gene Ontology Resource. Accessed: Oct. 22, 2025. [Online]. Available: http://geneontology.org/docs/ontology-documentation/
[68] M. A. A. Jansen, R. Spiering, I. S. Ludwig, W. van Eden, C. M. U. Hilkens, and F. Broere, 'Matured Tolerogenic Dendritic Cells Effectively Inhibit Autoantigen Specific CD4+ T Cells in a Murine Arthritis Model', *Front. Immunol.*, vol. 10, Aug. 2019, doi: 10.3389/fimmu.2019.02068.
[69] S. Wang *et al.*, 'TolDC Restores the Balance of Th17/Treg via Aryl Hydrocarbon Receptor to Attenuate Colitis', *Inflamm. Bowel Dis.*, vol. 30, no. 9, pp. 1546–1555, Sep. 2024, doi: 10.1093/ibd/izae022.
[70] R. Volchenkov, M. Karlsen, R. Jonsson, and S. Appel, 'Type 1 Regulatory T Cells and Regulatory B Cells Induced by Tolerogenic Dendritic Cells', *Scand. J. Immunol.*, vol. 77, no. 4, pp. 246–254, 2013, doi: 10.1111/sji.12039.
[71] F. Maleki, K. Ovens, I. McQuillan, and A. J. Kusalik, 'Size matters: how sample size affects the reproducibility and specificity of gene set analysis', *Hum. Genomics*, vol. 13, no. 1, p. 42, Oct. 2019, doi: 10.1186/s40246-019-0226-2.
[72] G. Altay, J. Zapardiel-Gonzalo, and B. Peters, 'RNA-seq preprocessing and sample size considerations for gene network inference', *BioRxiv Prepr. Serv. Biol.*, p. 2023.01.02.522518, Jan. 2023, doi: 10.1101/2023.01.02.522518.
[73] J.-W. Zhao *et al.*, 'Vitamin D suppress the production of vascular endothelial growth factor in mast cell by inhibiting PI3K/Akt/p38 MAPK/HIF-1α pathway in chronic spontaneous urticaria', *Clin. Immunol.*, vol. 215, p. 108444, Jun. 2020, doi: 10.1016/j.clim.2020.108444.
[74] M. Irani, D. B. Seifer, R. V. Grazi, S. Irani, Z. Rosenwaks, and R. Tal, 'Vitamin D Decreases Serum VEGF Correlating with Clinical Improvement in Vitamin D-Deficient Women with PCOS: A Randomized Placebo-Controlled Trial', *Nutrients*, vol. 9, no. 4, p. 334, Apr. 2017, doi: 10.3390/nu9040334.
[75] A. Cardus *et al.*, '1,25-Dihydroxyvitamin D3 regulates VEGF production through a vitamin D response element in the VEGF promoter', *Atherosclerosis*, vol. 204, no. 1, pp. 85–89, May 2009, doi: 10.1016/j.atherosclerosis.2008.08.020.
[76] A. M. G. van der Aar *et al.*, 'Vitamin D3 targets epidermal and dermal dendritic cells for induction of distinct regulatory T cells', *J. Allergy Clin. Immunol.*, vol. 127, no. 6, pp. 1532-1540.e7, Jun. 2011, doi: 10.1016/j.jaci.2011.01.068.
[77] F. M. J. Hafkamp, E. W. M. Taanman-Kueter, T. M. M. van Capel, T. G. Kormelink, and E. C. de Jong, 'Vitamin D3 Priming of Dendritic Cells Shifts Human Neutrophil-Dependent Th17 Cell Development to Regulatory T Cells', *Front. Immunol.*, vol. 13, Jul. 2022, doi: 10.3389/fimmu.2022.872665.

[78] J. Navarro-Barriuso, M. J. Mansilla, B. Quirant-Sánchez, A. Teniente-Serra, C. Ramo-Tello, and E. M. Martínez-Cáceres, 'Vitamin D3-Induced Tolerogenic Dendritic Cells Modulate the Transcriptomic Profile of T CD4+ Cells Towards a Functional Hyporesponsiveness', *Front. Immunol.*, vol. 11, p. 599623, 2020, doi: 10.3389/fimmu.2020.599623.
[79] J. Luo *et al.*, 'A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data', *Pharmacogenomics J.*, vol. 10, no. 4, pp. 278–291, Aug. 2010, doi: 10.1038/tpj.2010.57.
[80] 'R (Programmiersprache)', *Wikipedia*. Sep. 17, 2025. Accessed: Oct. 29, 2025. [Online]. Available: https://de.wikipedia.org/w/index.php?title=R_(Programmiersprache)&oldid=259813255
[81] T. Sagendorf, *8.2 Over-Representation Analysis | Proteomics Data Analysis in R/Bioconductor*. Accessed: Oct. 22, 2025. [Online]. Available: https://pnnl-comp-mass-spec.github.io/proteomics-data-analysis-tutorial/ora.html#ora-drawbacks
[82] J. Frost, 'What is the Bonferroni Correction and How to Use It', Statistics By Jim. Accessed: Oct. 16, 2025. [Online]. Available: https://statisticsbyjim.com/hypothesis-testing/bonferroni-correction/
[83] S.-Y. Chen, Z. Feng, and X. Yi, 'A general introduction to adjustment for multiple comparisons', *J. Thorac. Dis.*, vol. 9, no. 6, pp. 1725–1729, Jun. 2017, doi: 10.21037/jtd.2017.05.34.
[84] 'Falscherkennungsrate', *Wikipedia*. Feb. 11, 2025. Accessed: Oct. 29, 2025. [Online]. Available: https://de.wikipedia.org/w/index.php?title=Falscherkennungsrate&oldid=253237508
[85] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. L. Barabási, 'Hierarchical organization of modularity in metabolic networks', *Science*, vol. 297, no. 5586, pp. 1551–1555, Aug. 2002, doi: 10.1126/science.1073374.

7.2 Abbildungsverzeichnis

Die Bildqualität einiger Abbildungen wurde mit IMGUpscaler (https://imgupscaler.ai/) verbessert.

7.3 Gleichungsverzeichnis

7.4 Tabellenverzeichnis

7.5 Code-Verzeichnis

7.6 Begriffs- und Abkürzungsverzeichnis

Begriff / Abkürzung	**Definition**
A priori (Adj.)	Im Voraus bekannt oder theoretisch begründet.
Adjazenzmatrix	Eine Matrix, die die Verbindungsstärke (Koexpression) zwischen allen Genpaaren in einem Netzwerk darstellt.
Ähnlichkeitsmass	Quantifiziert die Ähnlichkeit zwischen Variablen, z. B. mittels Korrelation.
Angiogenese	Bildung neuer Blutgefässe aus bestehenden.
Anti-Citrullinated Peptide Antibodies (ACPA)	Hochspezifische Autoantikörper für Rheumatoide Arthritis.
Antigenpräsentierende Zelle (APC)	Zelle (z.B. dendritische Zelle), die Antigene aufnimmt, verarbeitet und den T-Zellen präsentiert.
Autolog (Adj.)	Vom selben Individuum stammend.
Bagging / Bootstrap Aggregation	Verfahren zur Stabilisierung von Modellen durch Aggregation vieler Bootstrap-Stichproben.
bc3net	Ein Algorithmus (Bagging c3net) zur Erstellung robuster Genregulationsnetzwerke aus Expressionsdaten.
Benjamini-Hochberg-Korrektur (BH-Korrektur)	Verfahren zur Korrektur multipler Tests zur Kontrolle der FDR.
Binomialtest	Ein statistischer Test, der bei einer Serie von Experimenten mit zwei möglichen Ausgängen (z.B. Kante vorhanden/nicht vorhanden) verwendet wird.
Biological Process (BP)	Biologischer Vorgang, an dem mehrere Gene oder Proteine beteiligt sind.
Bonferroni-Korrektur	Verfahren zur Korrektur multipler Tests zur Kontrolle der FWER.
Bootstrap Aggregation (Bagging)	Ein Ensemble-Verfahren, bei dem durch wiederholtes Ziehen mit Zurücklegen aus einem Datensatz viele neue

	Datensätze erstellt werden, um die Stabilität von Modellen zu erhöhen.
c3net	Ein recheneffizienter Algorithmus zur Netzwerkinferenz, der die Grundlage für *bc3net* bildet.
$CD4^+$-T-Zelle	Naive T-Helferzelle; eine zentrale Schaltstelle der adaptiven Immunantwort, die andere Immunzellen steuert.
Cellular Component (CC)	Ort innerhalb der Zelle, an dem ein Genprodukt wirkt.
Chondrozyten	Knorpelzellen, die Gelenkknorpel bilden.
Citrullinierung	Posttranslationale Modifikation, bei der Arginin in Citrullin umgewandelt wird.
Clusterbildung	Gruppierung ähnlicher Datenpunkte oder Gene.
clusterProfiler	R-Paket zur funktionellen Anreicherung und Visualisierung von Genlisten.
Container-Umgebung (z.B. Docker)	Isolierte Umgebung zur reproduzierbaren Ausführung von Software.
Dendritische Zelle (DC)	Eine antigenpräsentierende Zelle, die als Brücke zwischen angeborenem und adaptivem Immunsystem fungiert.
Dendrogramm	Baumdiagramm zur Darstellung hierarchischer Clusterstrukturen.
DiffCoEx	Methode zur Identifizierung von Genmodulen, deren Koexpressionsmuster sich signifikant zwischen zwei Zuständen ändern.
Dispersionswert	Mass für die Veränderung der Korrelation zwischen zwei Bedingungen.
Dissimilaritätsmass	Quantifiziert Unterschiede zwischen Objekten; je höher, desto unähnlicher.
Dynamic Tree Cut	Algorithmus zur automatischen Erkennung von Clustern in hierarchischen Bäumen.
Enrichment	Überrepräsentation, Anreicherung; Überdurchschnittliche Häufigkeit bestimmter Gene in einer Funktionskategorie.
Ensemble-Verfahren	Kombination mehrerer Modelle zur Erhöhung der Stabilität und Genauigkeit.
False Discovery Rate (FDR)	Anteil der fälschlich als signifikant identifizierten Tests unter allen signifikanten Tests.
Family-Wise Error Rate (FWER)	Wahrscheinlichkeit, mindestens einen Fehler 1. Art in einer Testfamilie zu begehen.
Fehler 1./2. Art	Falsch-positiver bzw. falsch-negativer Entscheid in einem statistischen Test.
Gelenkschmiere	Flüssigkeit in der Gelenkhöhle zur Schmierung und Ernährung des Knorpels.
Gene Expression Omnibus (GEO)	Eine öffentliche Datenbank für Genexpressionsdaten.
Gene Ontology (GO)	Eine Datenbank, die Genfunktionen und deren Beziehungen standardisiert beschreibt.
Gene Set Enrichment Analysis (GSEA)	Methode zur Identifikation überrepräsentierter Gensets entlang eines geordneten Rankings.
Genexpression	Der Prozess, bei dem die genetische Information eines Gens abgelesen (transkribiert) und zur Herstellung eines

	funktionellen Genprodukts (meist Protein) verwendet wird.
Genkoexpressionsnetzwerk (GCN)	Netzwerk, in dem Knoten Gene sind und ungerichtete Kanten eine statistische Ähnlichkeit im Expressionsmuster (Koexpression) darstellen.
Genregulatorisches Netzwerk (GRN)	Netzwerk, das die regulatorischen Beziehungen zwischen Genen mit gerichteten Kanten darstellt.
Genset	Gruppe von Genen, die eine gemeinsame biologische Funktion oder Regulation teilen.
Giant Component	Die grösste zusammenhängende Komponente eines Netzwerks.
Hämatopoetisch (Adj.)	Blutbildend; von Stammzellen im Knochenmark ausgehend.
Hierarchisches Clustering	Ein Algorithmus, der Datenpunkte (hier: Gene) basierend auf ihrer Ähnlichkeit in einer baumartigen Struktur (Dendrogramm) anordnet.
HLA-Klasse-II-Moleküle	Proteine auf der Oberfläche von antigenpräsentierenden Zellen. Bestimmte Varianten (z.B. HLA-DR4) sind ein Risikofaktor für RA.
Hub-Gen	Ein Gen in einem Netzwerk, das eine besonders hohe Anzahl an Verbindungen (Kanten) zu anderen Genen aufweist und oft eine zentrale regulatorische Rolle spielt.
Hypergeometrischer Test	Ein statistischer Test, der verwendet wird, um die Überrepräsentation (Anreicherung) einer Gengruppe in einer vordefinierten Liste zu berechnen (z.B. bei ORA).
Hyporesponsivität	Ein Zustand verminderter Reaktivität des Immunsystems auf einen Stimulus.
igraph	R-Bibliothek zur Erstellung und Analyse von Netzwerken.
Immune-Mediated Inflammatory Disease (IMID)	Gruppe chronisch entzündlicher, autoimmuner Erkrankungen.
Immunogene dendritische Zelle (mDC)	Dendritische Zelle, die Immunreaktionen aktiviert.
Interleukin (IL)	Eine Gruppe von Zytokinen, die als Botenstoffe zwischen den Zellen des Immunsystems dienen (z.B. IL-10, IL-12).
Knoten	Element (z. B. Gen) in einem Netzwerk.
Koexpression	Ein Mass zur Beschreibung der Verbindungsstärke zwischen zwei Genen. Diese Verbindungsstärke entspricht der statistischen Ähnlichkeit im Expressionsmuster.
Large p small n	Datensituation mit mehr Variablen als Stichproben.
log2-Transformation	Eine mathematische Transformation von Daten, die oft bei Genexpressionsdaten angewendet wird, um die Verteilung zu normalisieren und die Visualisierung zu verbessern.
Louvain-Methode	Algorithmus zur Erkennung von Modulen (Communities) in Netzwerken.
Molecular Function (MF)	Beschreibt die molekulare Aktivität eines Proteins.
Multiples Testen	Die Durchführung vieler statistischer Tests gleichzeitig zur Verbesserung der statistischen Signifikanz, was eine

	Korrektur (z.B. Bonferroni) erfordert, um die Rate falsch-positiver Ergebnisse zu kontrollieren.
Mutual Information (MI)	Ein Mass aus der Informationstheorie, das die statistische Abhängigkeit zwischen zwei Variablen quantifiziert. Ein hoher MI-Wert zwischen zwei Genen deutet auf eine biologische Verwandtschaft hin.
Myeloisch/lymphatisch (Adj.)	Zwei Hauptlinien der hämatopoetischen Differenzierung.
Nachbarn	Direkt verbundene Knoten eines Knotens.
Naiv (Adj.)	Eine naive Immunzelle wurde noch nicht durch Antigenbindung aktiviert und ist somit inaktiv.
Normalverteilung	Symmetrische, glockenförmige Wahrscheinlichkeitsverteilung mit definiertem Mittelwert und Varianz.
Nullhypothese (H_0)	Die Annahme in einem statistischen Test, dass es keinen Effekt oder Zusammenhang gibt (z.B. dass eine Kante nur zufällig beobachtet wird).
Nullverteilung	Erwartete Verteilung einer Teststatistik unter der Nullhypothese.
Over-Representation Analysis (ORA)	Ein statistisches Verfahren, um zu prüfen, ob bekannte biologische Funktionen oder Pathways in einer Liste von Genen überrepräsentiert (angereichert) sind.
Pannus	Aggressives, wucherndes Gewebe, das sich bei Rheumatoider Arthritis in der Gelenkinnenhaut bildet und Knorpel sowie Knochen zerstört.
Pathway / Signalweg	Eine Kette von molekularen Interaktionen in einer Zelle, die eine bestimmte Funktion ausführt.
Permutationstest	Nichtparametrischer Test, der Signifikanz durch zufälliges Vertauschen von Gruppenlabels bestimmt.
Potenzgesetz	Mathematische Beziehung, bei der eine Grösse als Potenz einer anderen skaliert.
Pseudocount	Eine kleine Konstante, die zu Daten addiert wird (z.B. vor einer log-Transformation), um mathematische Probleme wie log(0) zu vermeiden.
p-Wert	Wahrscheinlichkeit, ein gleich starkes oder stärkeres Ergebnis unter der Nullhypothese zu erhalten.
Quantilnormalisierung	Eine Methode zur Normalisierung von hochdimensionalen Daten (z.B. Microarrays), um technische Variationen zwischen Proben zu entfernen.
R	R ist eine freie Programmiersprache für statistische Berechnungen und Grafiken [80].
R^2	Mass für die Güte der Anpassung an das skalenfreie Modell, wobei Werte über 0,85 anzeigen, dass die Netzwerkstruktur gut durch ein skalenfreies Topologiemodell beschrieben wird.
Rangbasierte Methode	Verfahren, das auf der Reihenfolge statt den absoluten Werten basiert.
Regulatorische T-Zelle (T_{reg})	Eine Subpopulation von T-Zellen, die für die Unterdrückung von Immunreaktionen und die Verhinderung von Autoimmunität entscheidend ist.

Rheumatoide Arthritis (RA)	Eine chronische, systemische Autoimmunerkrankung, die hauptsächlich durch Entzündungen der Gelenke gekennzeichnet ist.
RNA-/DNA-Polymerase	Enzym, das RNA bzw. DNA synthetisiert.
RNA-Sequencing (RNA-Seq)	Eine Methode zur Messung der Genexpression durch Zählen und Sequenzieren von RNA-Molekülen in einer Probe.
Schwellenwert	Grenzwert zur Entscheidung, ob eine Verbindung oder ein Ergebnis berücksichtigt wird.
Selbsttoleranz	Fähigkeit des Immunsystems, körpereigene Strukturen nicht anzugreifen.
Serumkonzentration	Konzentration einer Substanz im Blutserum.
Signaltransduktion	Übertragung eines extrazellulären Signals in eine zelluläre Antwort.
Soft-Threshold	Schwellenwert, der Kanten im Netzwerk graduell gewichtet statt strikt abschneidet.
Spearman-Korrelationskoeffizient	Ein nicht-parametrisches Mass für die statistische Abhängigkeit zwischen zwei Variablen, das auf den Rängen der Datenwerte basiert.
Stammzellen	Undifferenzierte Zellen mit Selbsterneuerungs- und Differenzierungspotential.
Synoviale Fibroblasten	Bindegewebszellen der Gelenkinnenhaut, die Entzündungen und Gewebezerstörung fördern.
Synovium	Die Gelenkinnenhaut, die das Gelenk auskleidet und bei RA Ziel der Entzündung ist.
Systemische Autoimmunerkrankung	Erkrankung, bei der das Immunsystem körpereigene Strukturen in mehreren Organen angreift.
Testbedingung	Experimentelle Bedingung, in der eine Variable (z.B. Behandlung) verändert wird.
T-Helferzelle	Eine Art von T-Zelle, die eine pro-inflammatorische Immunantwort koordiniert.
Tolerogene dendritische Zelle (tolDC)	Ein Typ dendritischer Zelle, der Immunreaktionen aktiv unterdrückt und zur Aufrechterhaltung der Selbsttoleranz beiträgt.
Topological Overlap Measure (TOM)	Mass für die Ähnlichkeit zweier Knoten unter Berücksichtigung gemeinsamer Nachbarn.
Transkriptionsfaktor	Protein, das die Transkription spezifischer Gene reguliert.
Transkriptionsregulatorisches Netzwerk (TRN)	Netzwerk von Transkriptionsfaktoren und deren Zielgenen.
Transkriptom	Die Gesamtheit aller RNA-Moleküle (Transkripte), die zu einem bestimmten Zeitpunkt in einer Zelle vorhanden sind.
Tumornekrosefaktor-α (TNF-α)	Ein zentrales pro-inflammatorisches Zytokin, das bei RA eine wichtige Rolle bei der Gelenkzerstörung spielt.
Vascular Endothelial Growth Factor (VEGF)	Wachstumsfaktor, der Angiogenese fördert.
Vitamin D3 (VitD3)	Eine Substanz, die unter anderem die Entwicklung von tolerogenen dendritischen Zellen fördern kann.
Weighted Gene Co-expression Network Analysis (WGCNA)	Ein weit verbreitetes R-Paket und Framework zur Analyse von Gen-Koexpressionsnetzwerken.

Zytokin	Ein Protein, das als Botenstoff im Immunsystem dient und Entzündungsreaktionen reguliert.
α-Fehler-Kumulierung	Zunahme der Gesamtfehlerwahrscheinlichkeit bei mehrfachen Tests.

8 Beiträge

Name	E-Mail	Beitrag
Dr. Izaskun Mallona	izaskun.mallona@mls.uzh.ch	Mentoring und Anleitung zur Erstellung von Gennetzwerken, Korrekturlesen
Prof. Dr. Dr. Caroline Ospelt	caroline.ospelt@usz.ch	Hilfe bei der Definition der Forschungsfrage
Prof. Dr. Amedeo Caflisch	caflisch@bioc.uzh.ch	Hilfe bei der Eingrenzung des Themas

9 Verwendung generativer KI

Gemäss den Richtlinien für den Nationalen Wettbewerb von Schweizer Jugend Forscht werden hier die verschiedenen Anwendungen von generativen KI-Tools in diesem Projekt transparent dargelegt.

Grundsätzlich wurde KI ausschliesslich als unterstützendes Werkzeug eingesetzt. Alle generierten Inhalte wurden sorgfältig auf ihre Richtigkeit überprüft, mit Quellen abgeglichen und grundlegend überarbeitet.

- **Ideenfindung**: ChatGPT 5 wurde zur Evaluation von Forschungsansätzen, zur Ermittlung von Suchbegriffen für Fachdatenbanken und zur Identifikation von Wissenslücken genutzt.
- **Verständnis von Fachliteratur**: ChatGPT half ebenfalls dabei, komplexe wissenschaftliche Publikationen auf ihre Kernaussagen zu reduzieren, um gezielte Recherchen zu erleichtern.
- **Textbearbeitung**: Gemini 2.5 Pro diente der sprachlichen Überarbeitung von Textabschnitten. Zudem wurden aus eigenen Zusammenfassungen Textentwürfe generiert, die anschliessend manuell weiterverarbeitet wurden.
- **Code-Entwicklung**: Claude (Sonnet 4.5) wurde für die Erstellung, Fehlersuche und Optimierung von Code konsultiert. Ergänzend kam GitHub Copilot (Education) in RStudio zur Code-Annotation und beschleunigten Entwicklung zum Einsatz.
- **Abbildungsqualität**: IMGUpscaler wurde verwendet, um die Qualität von Abbildungen zu verbessern, ohne den Inhalt zu verändern.

Anhang

1 Over-Representation Analysis (ORA)

Trotz ihrer weiten Verbreitung ist ORA mit mehreren konzeptionellen und statistischen Schwächen behaftet und wird daher in der Literatur zunehmend kritisch betrachtet [66], [81]. Ein zentrales Problem liegt in der **Abhängigkeit von der gewählten Signifikanzschwelle**: Bereits kleine Änderungen im Schwellenwert (p-Wert) oder in der Methode zur Mehrfachtestkorrektur können zu stark unterschiedlichen Ergebnissen führen.

Zudem berücksichtigt ORA nicht die **Richtung der Expressionsänderung**; es wird nicht unterschieden, ob Gene innerhalb eines Sets überwiegend hoch- oder herunterreguliert sind. Eine Aufteilung der Gene nach Expressionsrichtung und anschliessende getrennte Analyse wäre zwar rechnerisch möglich, ist aber methodisch problematisch und kann zu Fehlinterpretationen führen.

Ein weiterer Nachteil besteht darin, dass ORA **bei kleinen Genmodulen instabil** ist. Wenn beispielsweise nur sehr wenige Gene als „interessant" klassifiziert werden, können minimale Änderungen in der Anzahl dieser Gene die Signifikanz einzelner Anreicherungen drastisch verändern. Das führt dazu, dass scheinbar signifikante Ergebnisse häufig nicht robust sind.

Schliesslich setzt ORA voraus, dass jedes Gen nur einmal im Datensatz vorkommt. Doppelte Einträge oder redundante Zuordnungen – etwa wenn mehrere Proteine demselben Gen zugeordnet werden – können zu einer **künstlichen Überrepräsentation** führen.

Aufgrund dieser Einschränkungen wird ORA in aktuellen Studien nicht als optimale Methode nach einer differentiellen Expressionsanalyse empfohlen. Methoden wie die **Gene Set Enrichment Analysis** (GSEA), die auf kontinuierlichen Rangstatistiken beruhen und sowohl Effektgrösse als auch Richtung der Regulation berücksichtigen, bieten hier eine methodisch überlegene Alternative. .

2 Begründung der Estimatorwahl «Spearman» für *bc3net*

Die Wahl der **Spearman-Rangkorrelation** für die Analyse der Gen-Koexpression in dieser Arbeit stützt sich auf die vergleichende Studie von Kumari et al. (2012). Die Autoren untersuchten acht statistische Methoden und zeigten, dass Spearman eine der leistungsstärksten ist, um biologisch sinnvolle Zusammenhänge in Genexpressionsdaten aufzudecken.

Die Studie belegt dies anhand von zwei zentralen Kriterien:

1. **Empirische Leistung:** Spearman identifizierte äusserst effektiv Gene, die zu denselben biologischen Signalwegen gehören. Ebenso war die Methode sehr erfolgreich bei der Konstruktion von GRNs. In allen Tests schnitt sie durchweg besser ab als die weitverbreitete Pearson-Korrelation.
2. **Theoretische Eignung:** Der entscheidende Vorteil von Spearman ist seine Eigenschaft als nicht-parametrische, rangbasierte Methode. Das bedeutet, sie verwendet die Rangfolge der Genexpressionswerte anstelle der Absolutwerte. Dies macht sie unempfindlich gegenüber Ausreissern (*outliers*) und setzt keine Normalverteilung der Daten voraus.

Neben der Leistungsstärke wurde auch die **Berechnungsdauer** berücksichtigt: Im Vergleich zu vielen anderen Methoden, einschliesslich der in der Studie vorgeschlagenen Alternativen, ist Spearman relativ rechengünstig.

3 Korrektur für multiples Testen

Wenn im Rahmen einer Studie **mehrere statistische Hypothesentests** durchgeführt werden, steigt das Risiko für Zufallsbefunde (α-Fehler-Kumulierung) [82]. Um dies zu kontrollieren, werden zwei Masse unterschieden:

Die **Family-Wise Error Rate (FWER)** ist die Wahrscheinlichkeit, **mindestens einen** falsch-positiven Befund (Fehler 1. Art) in der gesamten Test-Familie zu erhalten. Sie ist ein strenges Mass, um jeglichen Fehlalarm zu vermeiden.

Die **False Discovery Rate (FDR)** bezeichnet den erwarteten **Anteil** der falsch-positiven Ergebnisse an allen tatsächlich als signifikant («positiv») erklärten Ergebnissen. Sie erlaubt einzelne Fehler, solange der Grossteil der Entdeckungen korrekt ist [83].

3.1 Die Bonferroni-Korrektur

Die Bonferroni-Korrektur ist eine Methode zur FWER-Kontrolle [82].

Berechnung und Anwendung

Die Korrektur wird durchgeführt, indem das ursprüngliche Signifikanzniveau durch die Anzahl der durchgeführten Tests dividiert wird.

Gleichung 8: Bonferroni-Korrektur

$$\alpha_{corr} = \frac{\alpha}{n}$$

Ein Ergebnis wird nur dann als statistisch signifikant betrachtet, wenn sein p-Wert kleiner oder gleich diesem neuen, strengeren Wert von α_{corr} ist.

Kritik

Der Hauptnachteil der Bonferroni-Korrektur ist, dass sie **sehr konservativ** ist, insbesondere bei einer grossen Anzahl von Tests. Indem sie das Kriterium für Signifikanz so stark verschärft, reduziert sie die statistische Power. Das bedeutet, die Wahrscheinlichkeit, einen tatsächlich existierenden Effekt zu übersehen (ein Fehler 2. Art oder falsch-negatives Ergebnis), steigt an.

3.2 Die Benjamini-Hochberg-Korrektur

Die Benjamini-Hochberg-Korrektur (BH-Korrektur) ist ein Verfahren zur FDR-Kontrolle [83], [84].

Berechnung

Der Prozess der BH-Korrektur erfolgt in mehreren Schritten:

1. Sortierung der p-Werte: Zunächst werden die p-Werte aller getesteten Hypothesen in aufsteigender Reihenfolge geordnet.

2. Berechnung des Schwellenwerts: Für jeden p-Wert wird ein kritischer Schwellenwert q berechnet, der in der Form $q = \frac{k}{m} \cdot Q$ definiert wird, wobei k die Position des jeweiligen p-Werts in der sortierten Liste, m die Gesamtzahl der Tests und Q die gewünschte maximale FDR ist.
3. Abweisung der Nullhypothesen: Alle Nullhypothesen mit einem p-Wert, der kleiner oder gleich dem berechneten Schwellenwert ist, werden abgelehnt.

Anwendung

Die BH-Korrektur ermöglicht eine **flexiblere Kontrolle** der Fehlerquote, was insbesondere bei einer grossen Anzahl von Tests von Vorteil ist. Durch diese Methode wird die Wahrscheinlichkeit von falsch positiven Ergebnissen reduziert, ohne die Teststärke drastisch zu beeinträchtigen. Daher ist die BH-Korrektur besonders in Studien mit vielen simultan getesteten Hypothesen von Bedeutung.

4 Der Topological Overlap Measure (TOM)

4.1 Mathematische Grundlagen

Der Topological Overlap Measure (TOM) beschreibt, wie ähnlich oder miteinander verbunden zwei Knoten eines Netzwerks sind; nicht nur auf direktem Wege durch eine Kante, sondern auch **indirekt** durch gemeinsame Nachbarn [19]. Es kann mit folgender Formel berechnet werden:

Gleichung 9: Berechnung des TOMs

$$\omega_{ij} = \frac{l_{ij} + a_{ij}}{\min\{k_i, k_j\} + 1 - a_{ij}}$$

$l_{ij} = \sum_u a_{iu} a_{ju}$ entspricht der Anzahl Knoten u, die sowohl mit i als auch mit j verbunden sind.

$k_i = \sum_u a_{iu}$ beschreibt die Anzahl Nachbarn (Grad) von i.

a_{ij} stellt den Koexpressionswert da. Es entsteht ein nicht-gewichtetes TOM ($w_{ij} = \{0, 1\}$), wenn a_{ij} nur die Werte 0 und 1 annehmen kann [85]. Falls allerdings $0 \leq a_{ij} \leq 1$, liegt auch w_{ij} im Intervall $[0;\ 1]$.

Ist $\omega_{ij} = 1$, dann besitzen die zwei Knoten i und j eine perfekte topologische Überschneidung: Alle Nachbarn des Knotens i, das die wenigsten Kanten besitzt, sind also mit dem anderen Knoten j verbunden. Ausserdem sind i und j sind untereinander direkt verbunden. Im Gegensatz bedeutet $\omega_{ij} = 0$, dass i und j nicht direkt verbunden sind und keine gemeinsamen Nachbarn besitzen.

In der Praxis wird häufig mit Dissimilarität gehandelt. Um das Ähnlichkeitsmass TOM umzuwandeln, kann es von 1 subtrahiert werden:

Gleichung 10: Dissimilaritätsmass

$$d_{ij}^{\omega} = 1 - \omega_{ij}$$

4.2 Praktische Umsetzung

Code-Block 5: Praktische Umsetzung des TOMs

```
dissTOM <- TOMdist((AdjDiff)^(beta1/2))
```

Die verwendete Funktion TOMdist() berechnete zunächst das TOM und wandelte dieses anschliessend in ein Dissimilaritätsmass um (Code-Block 5).

Um eine skalenfreie Topologie (Kap. 2.3.4) zu erreichen, wurde der Parameter *beta1* empirisch gesetzt. Diese Entscheidung wurde in Kapitel 5.4.1 ausführlich diskutiert.

5 Vergrösserte Darstellung des Dendrogramms