

Über die Differentielle Analyse von Protein-Protein-Interaktionsnetzwerken

Inauguraldissertation

zur

Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

der

Mathematisch-Naturwissenschaftlichen Fakultät

der

Ernst-Moritz-Arndt-Universität Greifswald

vorgelegt von

Gregor Warsow

geboren am 30.03.1982

in Karlsburg

Greifswald, 03.09.2013

Dekan: Professor Dr. Klaus Fesser

1. Gutachter: Professor Dr. Georg Füllen

2. Gutachter: Professor Dr. Tim Beißbarth

Tag der Promotion: 14.04.2014

„Es geht doch nichts über die Freude, die uns das Studium der Natur gewährt. Ihre Geheimnisse sind von einer unergründlichen Tiefe; aber es ist uns Menschen erlaubt und gegeben, immer weitere Blicke hineinzutun. Und gerade, dass sie am Ende doch unergründlich bleibt, hat für uns einen ewigen Reiz, immer wieder zu ihr heranzugehen und immer wieder neue Einblicke und Entdeckungen zu versuchen.“

Johann Wolfgang von Goethe
im Gespräch mit Frédéric Jacob Soret

Inhaltsverzeichnis

Abbildungsverzeichnis	III
Abkürzungsverzeichnis	V
I Zusammenfassung	1
1 Einführung	3
1.1 Netzwerke in der Bioinformatik	3
1.2 Die Erstellung von Netzwerken	4
1.3 Netzwerke in der Erforschung von Krankheiten	6
1.4 Einschränkungen von Netzwerken und mögliche Lösungsansätze	7
1.5 Transkriptomdaten	9
2 Ergebnisse	11
2.1 ExprEssence	11
2.2 Der ExprEssence LinkScore	13
2.3 ExprEssence-Erweiterungen	16
2.3.1 MetaExprEssence und ExprEsSector	16
2.3.2 MovieMaker	18
2.4 ExprEssence im Kontext anderer Methoden	19
2.5 Podozyten und das PodNet	21
3 Ausblick	23
3.1 Informationsaustausch in der Wissenschaft im Zeitalter der Bioinformatik	23
3.2 Ausblicke bezüglich der netzwerkbasierten Forschung und ExprEssence .	24

II	Publikationen	27
4	Publikation #1	31
4.1	ExprEssence – Revealing the essence of differential experimental data in the context of an interaction/regulation network	31
5	Publikation #2	49
5.1	PodNet, a protein-protein interaction network of the podocyte	49
6	Publication #3, in Überarbeitung	61
6.1	Differential Network Analysis Applied to Preoperative Breast Cancer Chemotherapy Response	61
III	Appendix	87
7	Hintergrundinformationen zu den Berechnungen	89
7.1	Details zur LinkScore-Berechnung durch ExprEssence	89
7.2	Details zur Berechnung der Varianz der Interaktionsstärke	93
7.3	Details zur Schnittmengenbildung mit ExprEsSector	95
Literaturverzeichnis		97

Abbildungsverzeichnis

2.1	Prinzip der Netzwerkkondensierung mit <i>ExprEssence</i>	11
2.2	Screenshot von <i>ExprEssence</i>	12
2.3	Prinzip des Schneidens kondensierter Netzwerke mit <i>ExprEsSector</i>	17

Abkürzungsverzeichnis

BioGRID	Biological General Repository for Interaction Datasets [5]
ChIP	Chromatin-Immunoprecipitation – Chromatin-Immunpräzipitation
ChIP-Chip	Chromatin-Immunoprecipitation Chip [2]
ChIP-Seq	Chromatin-Immunoprecipitation DNA-Sequencing [1]
cDNA	complementary DNA – komplementäre DNA
DNA	deoxyribonucleic acid – Desoxyribonukleinsäure
etc.	et cetera – und so weiter
GRN	genregulatorisches Netzwerk
HPRD	Human Protein Reference Database [4]
KEGG	Kyoto Encyclopedia of Genes and Genomes [7]
mRNA	messenger RNA – Boten-RNA
NGS	Next Generation Sequencing – Sequenzierung der nächsten Generation [31]
RNA	ribonucleic acid – Ribonukleinsäure
STITCH	Search Tool for Interactions of Chemicals [78]
STRING	Search Tool for the Retrieval of Interacting Genes/Proteins [6]
SVM	Support Vector Maschine

Teil I

Zusammenfassung

1 Einführung

1.1 Netzwerke in der Bioinformatik

Netzwerke werden häufig verwendet, um Zusammenhänge, Wechselwirkungen, Ähnlichkeiten, Abstände und vieles mehr abzubilden. Sie ermöglichen die Formalisierung und Untersuchung komplexer Systeme verschiedenster Arten, beispielsweise im Transportwesen, in der Telekommunikation, im Finanzwesen aber auch in der Biologie und Medizin. Hier werden sie als vereinfachende Modelle biologischer Prozesse genutzt. Formal lassen sie sich als Graphen beschreiben, die aus einer Knotenmenge V (vertex set) und einer Kantenmenge E (edge set) bestehen. Die Knoten repräsentieren in biologischen Netzwerken häufig Gene, Proteine, Metabolite oder andere biologisch relevante Moleküle. Die Kanten verbinden die Knoten miteinander und drücken so die Zusammenhänge, Abhängigkeiten und Wechselwirkungen zwischen den Knoten aus.

Graphen können gerichtet oder ungerichtet sein. Gerichtete Graphen werden genutzt, wenn die Richtung der Kanten von Bedeutung ist. Protein-Protein-Interaktionsnetzwerke beispielsweise sind ungerichtet. Sie beschreiben lediglich, welche Proteine miteinander interagieren, treffen aber keine Aussage über die Qualität der Interaktion, etwa, welches Protein die Expression welchen Gens stimuliert oder inhibiert. Hierfür sind nur gerichtete Netzwerke geeignet.

Neben den durch das Netzwerk selbst repräsentierten Informationen können zu jedem Knoten und jeder Kante Daten aus verschiedenen Quellen in das Netzwerk integriert werden. Zu den gängigsten in Interaktions- oder Regulationsnetzwerke integrierten Daten zählen Hochdurchsatzdaten der Expression von Genen, der Proteinmengen, des Epigenoms (DNA-oder Histon-Modifikation), von Genvariationen sowie der Evidenz der im Netzwerk enthaltenen Interaktionen/Kanten (Chromatin-Immunpräzipitation, also ChIP-Seq [1] beziehungsweise ChIP-Chip [2] für Protein-DNA-Interaktionen oder Yeast-Two-Hybrid für Protein-Protein-Interaktionen).

Trotz der simplifizierenden Darstellung biologischer Prozesse/Zustände durch Netzwerke erlaubt dieser integrative Ansatz eine möglichst ganzheitliche Sicht auf die dem untersuchten Zustand zugrundeliegende Biologie [3]. Netzwerke haben daher - unter anderem bei der Suche nach molekularen Mechanismen, die mit Krankheiten assoziiert werden können - eine breite Anwendung gefunden. Durch Gegenüberstellung eines Netzwerks, das einen krankhaften Zustand beschreibt mit einem gesunden Kontrollzustand repräsentierenden Netzwerk können Unterschiede identifiziert werden, die mit der Krankheit korreliert und möglicherweise sogar ursächlich für diese sind. In der Regel sind solche

Netzwerke jedoch aufgrund der Vielzahl der bekannten Interakteure und ihrer komplexen Wechselwirkungen nicht intuitiv erfassbar. Hier stellt sich der Bioinformatik die Aufgabe der Reduktion der Netzwerke auf die für die jeweilige Fragestellung wesentlichen Teile, um hieraus Hypothesen, etwa über die Pathogenese, ableiten zu können. Dieser Prozess wird derzeit insbesondere durch die Identifizierung sogenannter aktiver Subnetzwerke realisiert. Hierfür wurden zahlreiche Methoden entwickelt, zu denen auch die hier vorgestellte Methode *ExprEssence* zählt, auf die später noch genauer eingegangen wird.

1.2 Die Erstellung von Netzwerken

Netzwerke werden unter anderem durch die Aggregation von in der Literatur oder in Datenbanken (zum Beispiel HPRD [4], BioGRID [5], STRING [6], oder KEGG [7]) vorhandenem Wissen gewonnen. Das manuelle Sichten der Literatur ermöglicht eine sehr hohe Qualität des Netzwerks, da uneindeutige Bezeichnungen präzisiert und fehlerhafte Aussagen direkt korrigiert werden können. Allerdings ist dieser Prozess äußerst aufwendig.

Deutlich ressourcensparender, jedoch weniger treffsicher, lassen sich Netzwerke auch durch automatisches Auswerten der wissenschaftlichen Literatur (Text-Mining) generieren [8]. Da der Text-Mining Ansatz sämtliche für den Text-Miner verfügbare Fachliteratur berücksichtigt, ist das so zusammengetragene Wissen potentiell vollständiger als ein manuell erstelltes Netzwerk: Es können Nebenbefunde in einem unerwarteten Zusammenhang und somit aus bei manueller Erstellung eventuell nicht untersuchten Fachartikeln Berücksichtigung finden. Sämtliche in den Wissenschaftsliteratur-Datenbanken PubMed (23 Millionen Kurzzusammenfassungen, Stand August 2013) und PubMed Central (2,8 Millionen Volltextartikel, Stand August 2013) verfügbare Literatur kann in kurzer Zeit zur Erstellung von reichhaltig annotierten Netzwerken prozessiert werden.

Text-Miner müssen vielfältige Schwierigkeiten meistern, etwa den Umgang mit Homonymen und Synonymen. Diese erschweren die eindeutige Identifizierung der genannten Gene oder Proteine, welche häufig erst bei Berücksichtigung des Kontextes, in dem die Bezeichnungen genannt werden, möglich wird. Methoden wie GENO stellen sich dieser Herausforderung und nehmen in 85-90 % der Fälle korrekte Zuordnungen vor [9].

Eine weitere Herausforderung für Text-Miner ist das Erkennen der Qualität von Interaktionen (Aktivierung, Repression, Modifikation etc.). Wenn solche den Interaktionstyp ausdrückenden Beschreibungen in einer vom Text-Miner nicht erwarteten, komplizierten Satzstruktur auftreten, führt dies häufig zu fehlerhaften Interpretationen des Textes.

Folglich ist beim Arbeiten mit einem durch Text-Mining generierten Netzwerk stets die

KAPITEL 1. EINFÜHRUNG

Korrektheit der einzelnen Knoten und Kanten zu prüfen. Ist das gesamte Netzwerk hierfür zu umfangreich, sollte darauf dennoch nicht gänzlich verzichtet werden. Eine mögliche Strategie ist, das unkorrigierte Netzwerk für die durchzuführende Analyse zu nutzen und anschließend eine Prüfung der für die Ergebnisse relevanten Proteine und Interaktionen auf Korrektheit vorzunehmen. Vom Text-Miner fehlerhaft interpretierte Knoten oder Interaktionen sind zu berichtigen beziehungsweise zu entfernen. Fehlerhaft bedeutet hierbei nicht, dass eine Aussage im Netzwerk nicht zu den in der Netzwerkanalyse genutzten Daten passt, sondern, dass die Aussage *per se* nach fachlichen Aspekten unkorrekt ist. Die Analyse wird dann mit dem berichtigten Netzwerk erneut ausgeführt. Diese Schritte werden wiederholt, bis sich keine Informationen mehr unter den für die Hypothesenbildung genutzten Netzwerkkomponenten befinden, die als fehlerhaft bekannt sind. Die so gewonnene Hypothese wird einer experimentellen Prüfung unterzogen, deren Ergebnisse wiederum in das Netzwerk einfließen können, um es weiter zu verbessern. Bei dieser Vorgehensweise ist zu berücksichtigen, dass sich zwar keine als unkorrekt bekannten Knoten und Kanten mehr im Ergebnisnetzwerk befinden, jedoch fälschlicherweise nicht als für das Ergebnisnetzwerk relevant erkannte - und somit nicht korrigierte - fehlerhafte Knoten/Kanten das Ergebnisnetzwerk indirekt beeinflussen können.

Ist ein bereits bestehendes - beispielsweise auf einen bestimmten Zelltyp fokussiertes - Netzwerk um zusätzliche, ebenso spezifisch zum Netzwerk passende Proteine zu erweitern, kann dies mit Hilfe von PILGRM (platform for interactive learning by genomics results mining [10]) erfolgen. Die Methode erhält als Eingabedaten eine Liste der im Netzwerk repräsentierten Gene (sogenannte Goldstandard-Gene) sowie eine Menge von uninteressanten Genen. Mit Hilfe von Support Vector Maschinen (SVM) werden diejenigen Gene identifiziert, die sich in ihrem Expressionsprofil ähnlich zu den Goldstandard-Genen verhalten. Die Expressionsdaten liegen PILGRM bereits als Kompendium aus verschiedensten Studien vor. Das Ergebnis ist eine Liste ähnlich exprimierter Gene, die jedoch noch in das Netzwerk integriert werden müssen. Die Suche nach Interaktionen zwischen Genen aus dem Netzwerk und den Genen der Ergebnisliste kann durch die Nutzung von Text-Minern vereinfacht werden.

Eine alternative Erstellung von Netzwerken ist als (*de novo*)-Synthese auch ohne biologisches Vorwissen unter Nutzung von Hochdurchsatzdaten möglich. Ein einfaches Beispiel hierfür sind gewichtete Koregulationsnetzwerke, bei denen die Knoten durch gewichtete Kanten verbunden werden. Das Kantengewicht beschreibt die Ähnlichkeit der Genexpression (siehe Abschnitt 1.5) der verbundenen Gene über verschiedene Zustände hinweg (Koexpression).

Eine Familie weiterentwickelter Methoden zur Netzwerkerstellung, die unter dem Begriff *reverse engineering* zusammengefasst werden, verwendet ebenso Hochdurchsatzdaten [11–13]. Aus Genexpressionsdaten werden so beispielsweise genregulatorische Netz-

1.3. NETZWERKE IN DER ERFORSCHUNG VON KRANKHEITEN

werke (GRN) entwickelt, die erklären sollen, wie regulierende Elemente - etwa Transkriptionsfaktoren - das beobachtete Expressionsverhalten generieren [14]. Die *de novo*-Synthese eines vollständigen regulatorischen Netzwerks auf Basis von ChIP-Seq und Genexpressions-Daten wurde von Galagan et al. vorgestellt [15]. Mit dem erstellten Netzwerk konnten Expressionsänderungen in Abhängigkeit von der Sauerstoffverfügbarkeit im *Mycobacterium tuberculosis* korrekt vorhergesagt werden. Trotz erfolgreicher Anwendung von *reverse engineering* Methoden in Einzelfällen ist deren Präzision bei der Prädiktion der Netzwerkmotive zum Großteil jedoch noch gering (50%) [16].

Die im Rahmen dieser Arbeit genutzten Interaktionsnetzwerke wurden manuell erstellt (PodNet oder KEGG [7], PluriNet), stammen aus einer Datenbank (GlobalNet basierend auf der STRING Datenbank [6]) oder sind manuell erstellt und um direkte Interaktionspartner aus STRING ergänzt worden (XPodNet) [17].

1.3 Netzwerke in der Erforschung von Krankheiten

Die erfolgreiche Behandlung von Krankheiten auf molekularer Basis wird durch ein umfangreiches und detailliertes Wissen über den Ablauf biologischer Prozesse gefördert. Krankheiten können als irreguläre, den Organismus negativ beeinflussende Zustände betrachtet werden. Folglich müssen der Normalzustand und der krankhafte Zustand des Organismus, seiner Zellen und der in diesen ablaufenden molekularen Prozesse bekannt sein, um gesund und krank auf molekularer Ebene miteinander in Relation setzen zu können.

In vielen Fällen von Krankheitsentstehung ist das im gesunden Zustand vorhandene Interaktionsgeflecht zwischen Proteinen, DNA, Metaboliten, verschiedenen Arten von RNA, Co-Faktoren und weiteren Molekülen aus dem Gleichgewicht geraten. Bildlich ausgedrückt liegt Krankheiten eine Erkrankung der Regulations- und Interaktionsnetzwerke zugrunde. Der Einsatz von Netzwerken in der Untersuchung der Krankheiten zielt darauf ab, diese deregulierten Bereiche zu identifizieren und zu charakterisieren.

Mit diesem Ansatz wird es möglich, Proteine mit der Krankheit in Verbindung zu bringen, die allein durch die Analyse der in das Netzwerk integrierten Daten nicht aufgefallen wären [18, 19]. Hierbei wird die Beobachtung ausgenutzt, dass Proteine, die mit einem bestimmten Phänotyp, so auch mit der Entstehung einer Krankheit, assoziiert sind, nicht zufällig im Netzwerk verteilt sind. Vielmehr liegen die diese Proteine repräsentierenden Knoten im Netzwerk topologisch nahe beieinander, was bedeutet, dass sie über relativ wenig Kanten (oder Knoten) miteinander verbunden sind [20, 21]. Neben dem Einsatz klassischer sogenannter Biomarker - Gene, Proteine, Hormone oder andere Moleküle, die

für eine eine (frühzeitige) Diagnose oder eine Verlaufsprägnose der Krankheit relevant sind - können auch die bei einer Krankheit als verändert erkannten Subnetzwerke als Biomarker genutzt werden. Sie repräsentieren nicht das Verhalten einzelner Gene oder Proteine, sondern vielmehr die gemeinsame Veränderung einer funktional zusammengehörigen Gruppe. Am Beispiel von Brustkrebs konnte gezeigt werden, dass sie gegenüber klassischen Biomarkern eine erhöhte Vorhersagegenauigkeit (Korrektklassifikationsrate) haben [22, 23].

Sollte bei einer Person mit Hilfe von Biomarkern auf den möglicherweise bevorstehenden Ausbruch einer Krankheit geschlossen werden, könnten ihr gezielte, relativ engmaschige Untersuchungen angeboten werden, um das tatsächliche Ausbrechen frühzeitig erkennen und gegebenenfalls dagegen intervenieren zu können. Dieser Ansatz der individuellen Gesundheitsvorsorge bildet einen Teil der Personalisierten/Individualisierten Medizin, die sich zur Zeit noch in der Entwicklung befindet. Sie hat - neben der möglichst frühzeitigen Diagnose einer Krankheit - zum Ziel, den Therapieplan spezifisch auf den Patienten zuzuschneiden, um diesen möglichst effizient und nebenwirkungsarm zu gestalten. Um das hierfür notwendige Wissen zusammenzutragen, wird eine große Anzahl verschiedener Datenquellen integrativ miteinander verknüpft, wobei Netzwerke eine tragende Rolle innehaben [24, 25].

In der vorliegenden Arbeit wird eine Methode vorgestellt und angewandt, die unter anderem zur Identifizierung krankheitsrelevanter Interaktionen beitragen kann. Diese Methode kombiniert vorhandenes Wissen in Form eines Interaktionsnetzwerks mit verschiedenen Zustände (gesund, krank) beschreibenden Hochdurchsatzdaten. Diese Zustände werden einander gegenübergestellt, um ein Subnetzwerk der am stärksten differentiell regulierten Interaktionen zu erhalten. Die Methode ist vielseitig anwendbar und somit nicht auf die Ursachenforschung bei Krankheiten beschränkt. Implementiert wurde die Methode als Computerprogramm unter dem Namen *ExprEssence* (siehe Publikation #1).

1.4 Einschränkungen von Netzwerken und mögliche Lösungsansätze

Ein offensichtlicher Nachteil von nicht-*do novo*-Netzwerken ist, dass sie nicht direkt zur Aufdeckung bisher vollständig unbekannter, das heißt, nicht im Netzwerk enthaltener, zellbiologischer Mechanismen beitragen können. Interaktionen oder Interakteure, die im untersuchten Kontext von Bedeutung, im Netzwerk jedoch nicht repräsentiert sind, können in den Ergebnissen von Netzwerkanalysen - etwa der Suche nach aktiven Subnetzwerken, siehe Abschnitt 2.4 - nicht auftreten (falsch-negative Resultate). Daher

1.4. EINSCHRÄNKUNGEN VON NETZWERKEN UND MÖGLICHE LÖSUNGSANSÄTZE

ist eine regelmäßige Aktualisierung der Netzwerke auf den aktuellen Wissensstand, was auch die Neuberechnung von *reverse engineering*-Netzwerken mit neu verfügbaren Hochdurchsatzdaten beinhaltet, vorzunehmen. Dies kann dazu beitragen, die Unvollständigkeit sukzessive abzubauen.

Neben falsch-negativen können allerdings auch falsch-positive Resultate auftreten. Biologische Systeme sind dynamisch und kontextabhängig, während Netzwerke überlagerte Schnappschüsse sind - eine Sammlung von Wechselwirkungen, die unter verschiedensten Bedingungen nachgewiesen wurden, nicht notwendigerweise jedoch gleichzeitig auftreten müssen. Selbst wenn die Gene von als interagierend bekannten Proteinen in einer Zelle zeitgleich exprimiert werden, können beispielsweise verschiedene Lokalisierungen der Proteine in der Zelle, eine fehlende Proteinmodifikation (etwa einer Phosphorylierung) oder das Vorhandensein eines die Wechselwirkung störenden kompetitiven Moleküls die Interaktion unterbinden. Dies kann bei der Interpretation folglich zu falsch-positiven Ergebnissen führen. Ist bekannt, dass eine Interaktion im untersuchten Kontext nicht auftritt, sollte sie aus dem verwendeten Netzwerk entfernt werden. Insbesondere bei der Untersuchung gewebs- oder zelltypspezifischer biologischer Prozesse ist, aufgrund von in den untersuchten Zellen *nicht* auftretenden beziehungsweise *exklusiv* in diesen Zellen auftretenden Proteinen und Interaktionen, die Anwendung eines gewebs-/zelltypspezifischen Netzwerks angeraten [26].

Statische Netzwerke sind zudem wenig geeignet für die Modellierung dynamischer Prozesse wie der verzögerten Expression von Genen, die durch einen Transkriptionsfaktor kontrolliert werden, der wiederum erst nach Durchlaufen einer Kaskade diverser vorgesetzter Genregulations-Kontrollmechanismen exprimiert wird. Sind derartige Prozesse abzubilden, sollten dynamische Netzwerke eingesetzt werden, deren Topologie veränderlich ist [27, 28]. Im Idealfall erfolgt die Anwendung von für die experimentelle Fragestellung (zelltyp)spezifisch entwickelten Interaktions- und Regulationsnetzwerken mit dynamischer Topologie [29, 30].

Nicht zuletzt ist zu erwähnen, dass Netzwerke ein verzerrtes Bild bezüglich der repräsentierten Proteine und Interaktionen aufweisen können. Dies ist der sehr intensiven Untersuchung einiger als besonders wichtig erachteter Proteine oder Gene geschuldet, während bisher weniger auffällige kaum untersucht, jedoch für die mit dem Netzwerk untersuchte Fragestellung von mindestens ebenso großer Bedeutung sein können. Das führt zum erwähnten Problem, dass einerseits die weniger untersuchten Proteine im Netzwerk fehlen, die intensiv untersuchten Proteine andererseits stark vernetzt mit vielen Interaktionspartnern im Netzwerk repräsentiert sind (sogenannte Hubs).

Folglich kann den Hubs eine möglicherweise ungerechtfertigt zentrale Bedeutung zugemessen werden. Häufig hilft bei der Bewertung der Bedeutung eines Hub-Knotens in einem Netzwerk nur die biologische Expertise weiter.

1.5 Transkriptomdaten

Bei den in der vorliegenden Arbeit genutzten Hochdurchsatzdaten handelt es sich überwiegend um Transkriptomdaten. Daher soll hier eine sehr kurze Einführung in die Thematik gegeben werden.

Die unterschiedlichen Fähigkeiten der verschiedenen Zelltypen werden zu wesentlichen Teilen durch Proteine realisiert, die selektiv hergestellt und abgebaut werden. Die Synthese eines Proteins erfolgt, sehr stark vereinfacht dargestellt, indem die DNA des Ziel-Gens, die den Code für die Proteinsequenz enthält, abgelesen wird, um hiervon Kopien in Form von messenger RNA (mRNA) herzustellen (Transkription). Diese wird dann für die eigentliche Proteinbiosynthese genutzt (Translation). Soll bestimmt werden, welche Gene in welchem Umfang abgelesen werden (Genexpression), können die mRNA-Moleküle quantitativ bestimmt werden. Hierfür gibt es eine Vielzahl etablierter Verfahren, von denen das Next Generation Sequencing (NGS) zu der fortschrittlichsten Familie solcher Verfahren zählt [31]. NGS Daten standen in der vorliegenden Arbeit jedoch nicht zur Verfügung. Es wurden Daten aus Microarray-Experimenten genutzt, welche ebenfalls verlässliche und wertvolle Informationen liefern [32].

Es wurden größtenteils Daten von GeneChip Microarrays der Firma Affymetrix (Affymetrix, Santa Clara, CA, USA) genutzt. Ihre Funktionsweise wird im Folgenden stark vereinfacht dargestellt. Als Einstieg in weiterführende Literatur sei auf [33] verwiesen. Affymetrix GeneChips bestehen aus einer Glasplatte, die in mehrere hunderttausend Felder unterteilt ist. Auf jedem Feld sind viele kurze einzelsträngige DNA-Moleküle am Glasboden fixiert, deren Sequenz jeweils Abschnitten einzelner Gene entspricht und pro Feld identisch ist. Somit repräsentiert jedes Feld einen Abschnitt eines Gens.

Wurde ein Gen in der Zelle exprimiert, befindet sich in dieser die zugehörige mRNA. Alle mRNA-Moleküle werden aus der Zelle entnommen und in einzelsträngige cDNA übersetzt. Jedes cDNA-Molekül wird bei diesem Prozess mit einem fluoreszierenden Molekül ligiert. Die cDNA-Moleküle werden anschließend mit dem Microarray inkubiert und binden spezifisch durch komplementäre Basenpaarung an einen der fixierten DNA-Stränge auf jenem Feld, das einen Abschnitt ihres jeweiligen Gens repräsentiert. Nach diesem Hybridisierungsvorgang und einem anschließenden Waschgang zum Entfernen von nicht spezifisch gebundener cDNA erfolgt die Ermittlung der Genexpression. Je stärker ein Gen exprimiert wurde, um so mehr cDNA hat an die fixierte DNA der entsprechenden Felder gebunden und um so stärker ist dort die Fluoreszenz. Die Intensität der Fluoreszenz jedes Feldes wird mit einer Kamera ermittelt und in einen numerischen Wert übersetzt.

1.5. TRANSKRIPTOMDATEN

Das Ergebnis der Ermittlung des Transkriptoms sind Intensitätswerte pro Feld, welche noch vorbereitenden Datenverarbeitungs- und Qualitätssicherungsschritten unterzogen werden, um schließlich für jedes auf dem Microarray vertretene Gen einen Expressionswert zu erhalten. Neben verschiedenen statistischen Normalisierungs- und Aufbereitungsschritten werden die Daten auch zur Basis 2 logarithmiert. Hierdurch wird eine annähernde Normalverteilung der (logarithmierten) Genexpressionswerte erreicht [34]. Diese so aufbereiteten Daten stehen dann für weitere Analysen zur Verfügung.

2 Ergebnisse

2.1 ExprEssence

Ein Protein-Protein-Interaktionsnetzwerk mit hunderten bis tausenden von Proteinen und zehn- bis hunderttausenden Interaktionen ist in seiner Komplexität nicht intuitiv erfassbar: Zusammenhänge können hier nicht direkt erkannt werden. Erst die Reduktion eines Netzwerks auf die für einen zellbiologischen Zustand wesentlichen Komponenten ermöglicht es, die diesem Zustand zugrundeliegenden molekularen Mechanismen nachzu vollziehen - insoweit diese im Netzwerk repräsentiert sind. Das Entfernen unwesentlicher und Herausstellen zustandsrelevanter Interaktionen ist die Aufgabe von Methoden, die nach sogenannten *aktiven* Subnetzwerken suchen. Das vom Verfasser dieser Arbeit entwickelte Programm *ExprEssence* gehört zu dieser Gruppe von Methoden.

ExprEssence wurde als Plugin für die frei verfügbare Netzwerk-Visualisierungs- und -Analyse-Plattform Cytoscape [35] entwickelt. Es dient der Identifizierung von Interaktionen in biologischen Netzwerken, deren Interaktionsstärke (Auftretenshäufigkeit, Aktivität) sich zwischen zwei Zuständen (beispielsweise krank versus gesund) am stärksten unterscheidet. Diese Interaktionen werden als zwischen den beiden Zuständen *differentiell reguliert* bezeichnet. *ExprEssence* reduziert die Netzwerkgröße, indem es nur die am stärksten differentiell regulierten Interaktionen im Netzwerk belässt.

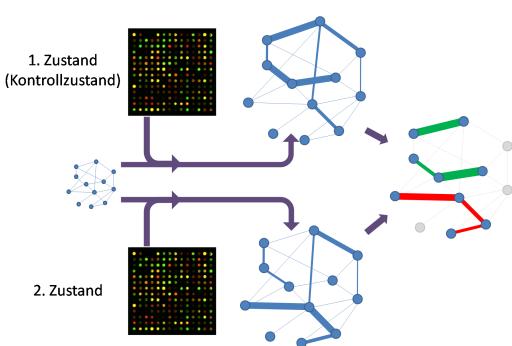


Abbildung 2.1: Prinzip der Netzwerkkonkondensierung mit *ExprEssence*. In ein Protein-Protein-Interaktions- oder Regulationsnetzwerk werden Proteom-/Transkriptomdaten von zwei zu vergleichenden Zuständen integriert. *ExprEssence* ermittelt die Interaktionsstärke/Aktivität jeder Interaktion im Netzwerk für jeweils beide Zustände (dargestellt durch die Dicke der blauen Kanten). Anschließend wird für jede Interaktion der Aktivitätsunterschied zwischen beiden Zuständen ermittelt. Nur Interaktionen mit am stärksten veränderter Aktivität bilden das kondensierte Ergebnisnetzwerk. Die Stärke und Richtung der Änderung wird durch die Kantendicke und -farbe dargestellt (grün: abnehmende, rot: zunehmende Aktivität).

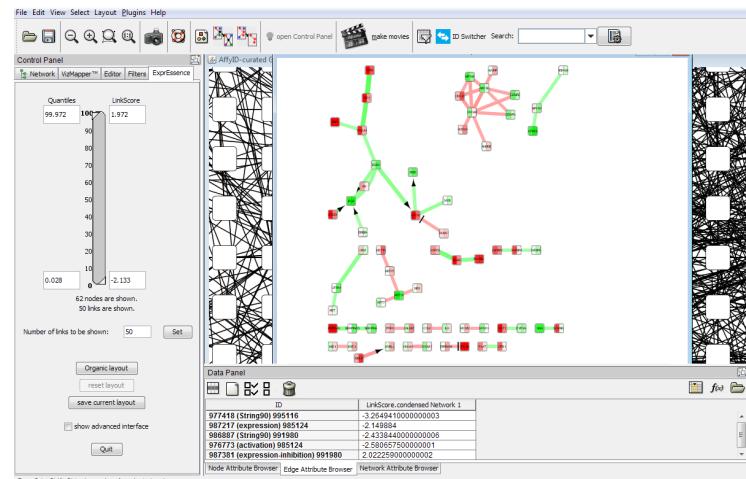
Um die am stärksten differentiell regulierten Interaktionen zu ermitteln, werden neben einem Interaktionsnetzwerk Hochdurchsatzdaten benötigt, die sich als Proteinmenge der durch die Knoten des Netzwerks repräsentierten Proteine interpretieren lassen. Da Proteomdaten für die untersuchten Fragestellungen nicht zur Verfügung standen, wurden Genexpressionsdaten von Microarrays genutzt. Ein theoretischer Nachteil der Nutzung von Genexpressions- gegenüber Proteomdaten bei der Analyse von Protein-Protein-

2.1. EXPRESSENCE

Interaktionen ist eine insgesamt nur mäßige Korrelation ($r : 0,4$ bis $0,75$) zwischen mRNA und Protein-Menge [36]. Jedoch berichteten Greenbaum et al., dass die mRNA-Mengen differentiell regulierter Gene eine höhere Korrelation mit den zugehörigen Proteinmengen aufweisen ($r = 0,89$) [37]. Da gerade die differentielle Regulation von Genen bei *ExprEssence* zu differentiell regulierten Interaktionen und somit zum Verbleiben der Interaktionen im verkleinerten Netzwerk führt, ist die vereinfachende Gleichsetzung von mRNA- mit Protein-Menge gerechtfertigt.

Auf Grundlage der die Proteinmengen beschreibenden Daten wird für jede Interaktion des Netzwerks ein numerischer Wert, der LinkScore (siehe Abschnitt 2.2), berechnet. Er beschreibt die Änderung der Interaktionsstärke/Aktivität¹ vom ersten zum zweiten Zustand. Nur eine vom Nutzer bestimmte Anzahl der am stärksten differentiell regu-

Abbildung 2.2: Screenshot von *ExprEssence*. Mittig ist ein kondensiertes Netzwerk abgebildet. Es enthält die am stärksten differentiell regulierten Kanten aus dem Netzwerk, das im Hintergrund abgebildet ist. Die Richtung der differentiellen Regulation einer Kante wird durch deren Farbe beschrieben (grün: abnehmende, rot: zunehmende Aktivität). Die LinkScores können dem Datenbereich (unten) entnommen werden - alternativ werden sie auch angezeigt, wenn der Mauszeiger über einer Kante zum Stehen kommt. Links befindet sich die Bedienoberfläche, über die unter anderem festgelegt wird, aus wie vielen Kanten das kondensierte Netzwerk bestehen soll.



lierten Interaktionen verbleibt im Netzwerk und bildet somit das *kondensierte* Netzwerk (siehe Abbildung 2.1). Dabei können die Anzahl der aktiver werdenden und der in ihrer Interaktionsstärke abnehmenden Interaktionen jeweils separat festgelegt werden. Die Interaktionen werden im Netzwerk gefärbt, um die Zunahme (rot) beziehungsweise Abnahme (grün) der Aktivität zu visualisieren. Das Ausmaß der differentiellen Regulation (der Betrag des LinkScores) einer Interaktion wird im Netzwerk durch die Kantendicke dargestellt, sodass Kanten mit einem großen Absolutbetrag des LinkScores dicker sind als Kanten mit kleineren Beträgen.

Damit die als differentiell reguliert identifizierten Interaktionen bei Bedarf auch in einem erweiterten Kontext interpretiert werden können, ermöglicht *ExprEssence* das fallweise

¹Es ist zu beachten, dass der Terminus der Aktivität einer Interaktion nicht gleichzusetzen ist mit einem aktiven Subnetzwerk. Die Aktivität einer Interaktion beschreibt die Häufigkeit ihres Auftretens. Ein Subnetzwerk hingegen wird als aktiv bezeichnet, wenn es differentiell regulierte Interaktionen enthält – Interaktionen deren Aktivität sich zwischen beiden Zuständen hinreichend stark ändert.

Hinzufügen von Proteinen, deren Gene unter beiden verglichenen Zuständen stark exprimiert wurden und somit in beiden Zuständen potentiell als Interaktionspartner zur Verfügung stehen. Auch wenn diese Proteine keinen direkten Beitrag zu den Unterschieden der Phänotypen leisten, können sie als eventuelle Haushaltsproteine (housekeeping proteins) dennoch von wesentlicher funktionaler Bedeutung für die Zelle sein. Weiterhin lassen sich bei Bedarf Kanten anzeigen, die zwischen im kondensierten Netzwerk vorhandenen Knoten liegen, jedoch nicht ausreichend differentiell reguliert sind, um selbst im Subnetzwerk vertreten zu sein (einige der grauen Kanten im kondensierten Netzwerk in Abbildung 2.1). Durch Nutzung beider Optionen kann die Interpretierbarkeit des Subnetzwerks verbessert werden.

Eine detaillierte Beschreibung der Methode ist in Warsow et al. [38] zu finden; jedoch ist zu beachten, dass die Formeln zur Berechnung des LinkScores mittlerweile, wie im nächsten Abschnitt beschrieben, verändert wurden.

2.2 Der ExprEssence LinkScore

Der von *ExprEssence* berechnete LinkScore ist ein Maß für die Änderung der Interaktionsstärke (im Sinne der Häufigkeit des Auftretens der Interaktion) zweier Proteine A und B beim Übergang der Zelle von einer in eine andere Bedingung. Die Formeln zur Berechnung des LinkScores wurden aus der Stoßtheorie/dem Massenwirkungsgesetz abgeleitet. Für physische und genregulatorisch positiv wirkende Interaktionen lautet sie:

$$\begin{aligned} LS_{1 \rightarrow 2} &= I_2 - I_1 \\ &= (E_2^A + E_2^B) - (E_1^A + E_1^B), \end{aligned} \tag{2.1}$$

wobei $I_{1/2}$ die Interaktionsstärke unter Bedingung 1 beziehungsweise 2 und $E_{1/2}^{A/B}$ der log2-transformierte Genexpressionswert von A beziehungsweise B unter Bedingung 1 beziehungsweise 2 seien.

Für genregulatorisch inhibierend wirkende Interaktionen gilt für den LinkScore:

$$\begin{aligned} LS_{1 \rightarrow 2}^i &= I_2^i - I_1^i \\ &= (E_2^R - E_2^Z) - (E_1^R - E_1^Z), \end{aligned} \tag{2.2}$$

wobei I_1^i bzw. i_2 die Stärke der inhibierend wirkenden Interaktion unter Bedingung 1 beziehungsweise 2 und E_1^R bzw. Z der log2-transformierte Genexpressionswert des Repressor-Proteins R beziehungsweise des Ziel-Gens Z unter der Bedingung 1 beziehungsweise 2 seien. Einzelheiten zur Herleitung der Formeln befinden sich im Anhang, Abschnitt 7.1.

2.2. DER EXPRESSENCE LINKSCORE

In der Publikation zu *ExprEssence* wurde der LinkScore über die Differenz $D_i, i \in \{A, B\}$ der Expression eines Gens $i \in \{A, B\}$ beim Übergang von der ersten zur zweiten Bedingung hergeleitet [38]. Diese Differenz wurde durch Anwendung der Welch-Formel [39, 40] modifiziert, um zu erreichen, dass der Betrag des LinkScores bei großen Varianzen kleiner wird:

$$LS = D_A + D_B \text{ mit } D_i = \frac{E_2^i - E_1^i}{\sqrt{\frac{Var(E_1^i)}{n_1} + \frac{Var(E_2^i)}{n_2}}}, i \in \{A, B\} \quad (2.3)$$

Jedoch führen hier kleine Varianzen zu einer Inflation des LinkScores. Dies ist insbesondere bei kleinen Gruppengrößen (n_1, n_2) zu berücksichtigen, da hier die geschätzte Varianz zufällig sehr gering ausfallen kann [41]. Kleine Genexpressions-Änderungen können so bei kleinen Varianzen zu einem extremeren LinkScore führen als im Mittel ausgeprägtere Änderungen bei großer Varianz.

Tusher et al. haben vorgeschlagen, diesem Umstand durch die Einführung eines Korrekturfaktors (s_0) entgegenzuwirken, der den Einfluss kleiner Varianzen dämpft und dessen Anwendung zu folgender Berechnung führt [42]:

$$D_i = \frac{E_2^i - E_1^i}{\sqrt{\frac{Var(E_1^i)}{n_1} + \frac{Var(E_2^i)}{n_2}} + s_0}, i \in \{A, B\} \quad (2.4)$$

Hierbei wird s_0 so gewählt, dass der Variationskoeffizient von D_i minimiert wird. Der Wert des Faktors liegt oft im 5-10% Quantil der gemeinsamen Standardabweichung (Nenner in Formel 2.3) [43].

Das primäre Interesse bei der Anwendung von *ExprEssence* ist jedoch, jene Interaktionen mit den absolut stärksten Änderungen zu erkennen. Eine Umsortierung der Wertigkeit der Änderung der Interaktionsstärken durch die Einbeziehung der Varianzen ist hier kontraintuitiv. Daher wurde die LinkScore-Berechnung optimiert, indem der LinkScore ohne die Korrektur nach Formel 2.3, jedoch zusätzlich mit einem Signifikanzmaß berechnet wird.

Die unmodifizierten LinkScores werden genutzt, um die am stärksten differentiell regulierten Interaktionen zu identifizieren und diese, bis zur eingestellten Anzahl an Kanten, im kondensierten Interaktionsnetzwerk zu belassen. Anschließend wird für diese Interaktionen ein Signifikanzmaß/die Irrtumswahrscheinlichkeit des Einstufens als differentiell regulierte Interaktion berechnet. Interaktionen mit einer Irrtumswahrscheinlichkeit $> 5\%$ (kann von Nutzer geändert werden) werden nicht sofort verworfen, sondern mit steigender Irrtumswahrscheinlichkeit zunehmend ausgeblendet, um deren Änderung als weniger verlässlich kenntlich zu machen. Ziel der beschriebenen Trennung in Selektion von maximal differentiell regulierten Interaktionen und deren anschließender Reliabi-

litätsbestimmung ist, die Resultate von *ExprEssence* noch intuitiver interpretierbar zu machen.

Im Folgenden wird die Berechnung des Signifikanzmaßes beschrieben. Ihr liegt ein statistischer Test zugrunde, dessen Nullhypothese, dass die Differenz der Interaktionsstärken zwischen der ersten und zweiten Bedingung den Wert Null habe, der Alternativhypothese, die Differenz sei ungleich 0, gegenübersteht:

$$H_0 : I_1 = I_2 \text{ sowie } H_A : I_1 \neq I_2 \quad (2.5)$$

Es handelt sich hierbei um eine zweiseitige Fragestellung, da *a priori* nicht bekannt ist, ob der LinkScore positiv oder negativ sein wird. Ein t-Test, der gleiche Varianzen in beiden Gruppen voraussetzt, ist für diesen Test im Allgemeinen ungeeignet. Zwar zeigten Simulationsstudien (Monte Carlo), dass sich der t-Test robust gegenüber Verletzungen der Varianzhomogenität verhält, wenn die Gruppengrößen annähernd gleich sind [44], dennoch ist ein Test anzuwenden, der mit unbekannten und nicht als gleich vorausgesetzten Varianzen genutzt werden kann, da die Anzahl der zur Verfügung stehenden Replikate in beiden Zustandsgruppen wesentlich voneinander abweichen kann.

Das führt zum Behrens-Fischer-Problem, für dessen Lösung eine Approximation nach Welch existiert - bekannt unter der Bezeichnung Welch's t-Test. Dieser Test kann mit ungleichen Stichprobengrößen und ungleichen Varianzen genutzt werden. Die Wahl dieses Tests ist unabhängig von der oben beschriebenen Verwendung in Formel 2.3. Trotz des Fallenlassens der Annahme der Varianzhomogenität und gleicher Gruppengrößen setzt dieser Test jedoch noch die Normalverteilung der zu vergleichenden Zufallsvariablen - der Interaktionsstärken - voraus. Da in der einschlägigen Literatur eine log-Normalverteilung der Genexpression angenommen wird [34] und die Interaktionsstärke als Summe der jeweils normalverteilten logarithmierten Expressionswerte der beteiligten Gene wieder normalverteilt ist, ist die Annahme einer Normalverteilung der Interaktionsstärken gerechtfertigt. Zudem ist diese Voraussetzung bei Stichprobenumfängen ≥ 30 vernachlässigbar [44].

Somit kann ein zweiseitiger t-Test nach Welch angewendet werden. Die Teststatistik für den Test nach Welch auf Gleichheit der Interaktionsstärken - oder einen LinkScore von 0,0 - lautet

$$t_{LS_{1 \rightarrow 2}} = \frac{|I_2 - I_1|}{\sqrt{\frac{Var(I_1)}{n_1} + \frac{Var(I_2)}{n_2}}}, \quad (2.6)$$

wobei n_1, n_2 weiterhin die Anzahl der Messungen in Bedingung 1 beziehungsweise 2 beschreiben. Um die Formel anwenden zu können, ist die Varianz der Interaktionsstärke zu ermitteln. Hierauf wird im Anhang, Abschnitt 7.2, näher eingegangen. $t_{LS_{1 \rightarrow 2}}$ ist t verteilt mit der Anzahl an Freiheitsgraden ν , die nach der Formel von Welch-Satterthwaite

2.3. EXPRESSENCE-ERWEITERUNGEN

berechnet wird:

$$\nu_{t_{LS_{1 \rightarrow 2}}} = \frac{\left(\frac{Var(I_1)}{n_1} + \frac{Var(I_2)}{n_2}\right)^2}{\frac{Var(I_1)^2}{n_1^2(n_1-1)} + \frac{Var(I_2)^2}{n_2^2(n_2-1)}} \quad (2.7)$$

Für jeden LinkScore kann nun sein p-Wert berechnet werden. Er gibt an, mit welcher Wahrscheinlichkeit die Ablehnung der Nullhypothese bei einer Interaktion falsch wäre, da diese Interaktion tatsächlich nicht differentiell reguliert ist. Der p-Wert ist gleich dem doppelten Anteil (aufgrund der zweiseitigen Fragestellung) der Fläche der t-Verteilung (mit der in Formel 2.7 berechneten Anzahl an Freiheitsgraden), die rechts von dem Wert $t_{LS_{1 \rightarrow 2}}$ liegt. Die p-Werte werden anschließend mit der Methode nach Benjamini-Hochberg (False Discovery Rate, FDR) für multiples Testen korrigiert [45]. Als signifikant differentiell reguliert gelten Interaktionen, deren korrigierter p-Wert $\leq 0,05$ ist. Sie werden im kondensierten Netzwerk vollwertig dargestellt, während Interaktionen bis zu einem korrigierten p-Wert von 0,1 kontinuierlich ausgeblichen werden. Beide Schwellenwerte können vom Nutzer angepasst werden.

2.3 ExprEssence-Erweiterungen

2.3.1 MetaExprEssence und ExprEsSector

Seit der Publikation von *ExprEssence* sind Erweiterungen, ebenso Plugins für Cytoscape, entwickelt worden. Zum einen kann *ExprEssence* nun im sogenannten Batch-Modus mehrere Aufträge gleichzeitig entgegennehmen, was durch das Tool *MetaExprEssence* realisiert wird. Zum anderen kann mit *ExprEsSector* (*ExprEssence Intersector*) nach Interaktionen gesucht werden, die unter verschiedenen Zustands-Übergängen *gemeinsam* differentiell reguliert werden. Diese Fragestellung ist beispielsweise von Belang, wenn verschiedene Krankheiten auf das Vorhandensein mechanistischer Ähnlichkeiten im Sinne gemeinsam differentiell regulierter Interaktionen überprüft werden sollen.

Vorbereitend für die Nutzung von *ExprEsSector* werden mittels *ExprEssence* die krankheitsspezifisch differentiell regulierten Interaktionen durch Gegenüberstellung des Krankheitszustands mit einem (gesunden) Kontrollzustand identifiziert. Die so erhaltenen Subnetzwerke werden anschließend mit *ExprEsSector* miteinander geschnitten (siehe Abbildung 2.3). Es entsteht ein Netzwerk, das nur solche Knoten und Kanten enthält, die in allen kondensierten Netzwerken vertreten sind. Das heißt, diese Interaktionen sind in allen verglichenen Krankheiten bezogen auf den Kontrollzustand differentiell reguliert. Je mehr Netzwerke miteinander geschnitten werden, umso größer ist die Wahrscheinlichkeit, dass es sich um eine gemeinsame Mechanismen handelt.

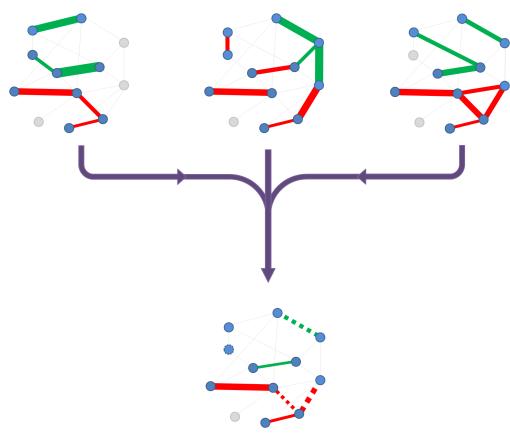


Abbildung 2.3: Prinzip des Schneidens kondensierter Netzwerke mit *ExprEsSector*. Dargestellt ist die Anwendung von *ExprEsSector* auf drei kondensierte Netzwerke (oben) unter der relaxierenden Bedingung, dass die Knoten und Kanten in nur zwei anstatt allen drei Netzwerken präsent sein müssen, um im Schnittmengen-Netzwerk enthalten zu sein.

Das Schnittmengen-Netzwerk (unten) besteht somit genau aus den Knoten und Kanten, die in allen drei oberen Netzwerken (Linien durchgehend gezeichnet) oder in nur zwei der drei oberen Netzwerke enthalten sind (gestrichelte Linien).

Der LinkScore der Kanten im unteren Netzwerk berechnet sich als Mittelwert der zugehörigen LinkScores aus den oberen Netzwerken. Er wird durch die Kantenfarbe (grün: negativ, rot: positiv) und die Kantendicke (Absolutbetrag des LinkScores) visualisiert.

lichkeit, dass einzelne Kanten nicht in sämtlichen kondensierten Netzwerken vorhanden sind. Bereits das Fehlen einer Kante in nur einem Netzwerk führt zu deren Verlust im Schnittmengennetzwerk. Soll diesem Umstand entgegengewirkt werden und die Schnittmengenbildung robuster gegenüber dem nur vereinzelten Fehlen von Kanten werden, kann die Anforderung des Vorhandenseins aller Knoten und Kanten in *jedem* ausgewählten *ExprEssence*-Netzwerk auch auf eine beliebige kleinere Anzahl von Netzwerken relaxiert werden: Fehlen Knoten oder Kanten in nicht mehr Netzwerken als erlaubt, bleiben sie im Schnittmengennetzwerk erhalten. Dabei ist es unbedeutend, in genau welchen *ExprEssence*-Netzwerken sie fehlen. Um solche Kanten von Kanten unterscheiden zu können, die in jedem Netzwerk vertreten sind, werden sie nicht durchgehend sondern gestrichelt dargestellt.

Weiterhin ermöglicht *ExprEsSector*, vor dem Schneiden ein Vereinigungsnetzwerk ausgewählter Netzwerke zu bilden. Dies ist von Interesse, wenn für die gleiche Krankheit Datensätze verschiedener Qualität (beispielsweise Transkriptom- und Proteomdaten) vorliegen. Lassen sich die Daten nicht sinnvoll kombinieren, um mit ihnen ein Netzwerk mit *ExprEssence* zu kondensieren, wird für die Transkriptom- und Proteomdaten jeweils ein kondensiertes Netzwerk erstellt. Beide kondensierten Netzwerke können anschließend mit der Vereinigungsoperation von *ExprEsSector* genutzt werden, um die jeweils als differentiell reguliert identifizierten Kanten in einem Netzwerk zu aggregieren, welches anschließend mit anderen Netzwerken geschnitten werden kann. Ebenso kann jedoch auch die Schnittmengenbildung beider Netzwerke von Interesse, wenn die Übereinstimmung der aus den Transkriptom- und Proteomdaten gewonnenen Signale geprüft werden soll. Eine detaillierte Beschreibung der Schnittmengenberechnungen wird im Anhang, Abschnitt 7.3, gegeben.

2.3. EXPRESSION-ERWEITERUNGEN

Der LinkScore der Interaktionen in den geschnittenen Netzwerken entspricht dem Mittelwert der LinkScores aus den zu schneidenden Netzwerken. Handelt es sich bei einem der zu schneidenden Netzwerke um ein Vereinigungsnetzwerk, werden für dieses die LinkScores zuvor aus den Mittelwerten der LinkScores der zu vereinigenden Netzwerke gebildet.

Die Anzahl der in den ausgewählten *ExprEssence*-kondensierten Netzwerken vorhandenen Kanten kann vom Nutzer mit den von *ExprEssence* bekannten Bedienelementen jederzeit manipuliert werden. Sämtliche zu vereinigende und zu schneidende *ExprEssence*-Netzwerke werden bei einer solchen Änderung sofort angepasst, sodass alle die eingestellte Anzahl an Interaktionen besitzen. Anschließend werden automatisch erneut die Vereinigungs- und Schnittmengenberechnungen durchgeführt, sodass die Änderungen im Schnittmengennetzwerk unmittelbar umgesetzt werden. Durch die Anwendung von *ExprEsSector* wird so eine gemeinsame Betrachtung verschiedener Krankheiten auf einem gemeinsamen molekularen Niveau möglich: Molekulare Mechanismen, die von verschiedenen Krankheiten geteilt werden, werden direkt erkennbar.

2.3.2 MovieMaker

Eine zusätzliche Erweiterung von *ExprEssence* ist das Plugin *MovieMaker*. Dieses Programm vereinfacht durch eine Animation das Auffinden von sonst erst nach längerer Betrachtung erkennbaren Gemeinsamkeiten und Unterschieden zwischen verschiedenen *ExprEssence*-kondensierten Netzwerken. Im Allgemeinen identifiziert der Mensch Unterschiede in zwei Bildern durch häufiges Wechseln zwischen beiden Bildern. Das Erkennen von Unterschieden wird mit *MovieMaker* einfacher, da hinzukommende und verlorengehende Knoten und Kanten langsam ein- beziehungsweise ausgeblendet werden und dadurch leichter identifizierbar sind. Knoten, die beim Wechsel zwischen zwei kondensierten Netzwerken erhalten bleiben, sind durch eine Markierung leicht identifizierbar. Für alle visualisierten Netzwerke wird ein gemeinsames Layout genutzt, sodass der Überblick nicht durch Knoten beeinträchtigt wird, die die Position wechseln. Die Reihenfolge, in der die Animation durch die ausgewählten *ExprEssence*-Netzwerke führen soll, wird zu Beginn vom Nutzer festgelegt; jedoch kann er jederzeit direkt zu einem bestimmten Netzwerk wechseln.

Diese Art der Darstellung eignet sich aufgrund der linearen Struktur des von *MovieMaker* erstellten Films besonders gut für die Visualisierung von *ExprEssence*-Netzwerken, deren Zeitreihendaten zugrundeliegen.

2.4 ExprEssence im Kontext anderer Methoden

Wie bereits dargestellt, wird die Kombination von Netzwerken mit Expressionsdaten häufig bei der Suche nach den molekularen Grundlagen unterschiedlicher Phänotypen angewandt. Einige der Methoden, die diesem integrativen Ansatz folgen, bewerten die Aktivität bereits vordefinierter Stoffwechsel- oder Regulationswege (Pathways) unter den betrachteten Zuständen. Diese Strategie wird beispielsweise bei der *Ingenuity Pathway Analysis* (IPA, Ingenuity®Systems, www.ingenuity.com) oder für Pathways von KEGG angewandt, um jene Pathways zu identifizieren, die unter einer betrachteten Bedingung mutmaßlich aktiv sind [46]. Für die vorhandenen vordefinierten Pathways ist dieser Ansatz durchaus nützlich. Jedoch werden die als aktiv identifizierten Pathways hier isoliert betrachtet und nicht miteinander in Verbindung gebracht.

Das ist jedoch bei einer zweiten Gruppe von Methoden möglich, deren Vertreter aktive Subnetzwerke aus einem Protein-Protein-Interaktionsnetzwerk oder einem regulatorischen Netzwerk extrahieren. Sind die Gene/Proteine der erwähnten vordefinierten Pathways in dem Netzwerk enthalten, können hier ihre gegenseitigen Wechselwirkungen und Abhängigkeiten Berücksichtigung finden.

Einige der Methoden, die in solchen Netzwerken nach aktiven Komponenten suchen, setzen Vorwissen über den untersuchten Phänotyp voraus. So müssen Gruppen von Genen oder Funktionsgruppen von Proteinen (GeneOntology) definiert werden, die bereits als zum Phänotyp zugehörig gelten. Entsprechend des in der jeweiligen Methode definierten Distanzmaßes wird schließlich die Ähnlichkeit/Zugehörigkeit der Kandidaten zum vordefinierten Kontext ermittelt. Die hierzu am besten passenden Proteine und deren Interaktionen werden als Subnetzwerk ausgegeben [47, 48]. Das Wissen über die biologische Zusammengehörigkeit topologisch nahe beieinander liegender Knoten wird neben der Identifizierung aktiver Subnetzwerke auch genutzt, um Proteine unbekannter Funktionalität in einen funktionalen Kontext zu setzen [49, 50].

Bei der Suche nach aktiven Subnetzwerken ohne Vorwissen über den untersuchten Phänotyp werden Methoden eingesetzt, die ausschließlich das zugrundeliegende Netzwerk und die darin integrierten differentiellen Hochdurchsatzdaten nutzen. Zu dieser Gruppe gehört *ExprEssence*. Die Aktivität der Subnetzwerke wird häufig knotenbasiert aus den Genexpressionsunterschieden, den zugehörigen Signifikanzwerten (p-Werte) oder anderen die differentielle Genexpression repräsentierenden Maßen abgeleitet [51–53]. Auch kantenbezogene Informationen - wie die Anzahl an publizierten Experimenten, die eine Interaktion nachweisen [54], oder die Korrelation der Genexpression [55] - werden für die Suche nach aktiven Subnetzwerken genutzt. Letzteres Maß berücksichtigt jedoch nicht die Qualität der Interaktionen (stimulierend oder inhibierend). Ein wichtiger Vorteil von *ExprEssence*

2.4. EXPRESSENCE IM KONTEXT ANDERER METHODEN

gegenüber solchen Verfahren ist daher die Berücksichtigung des Interaktionstyps (physisch, Stimulation, Inhibition).

Aufgrund der Interaktionstyp-spezifischen Berechnung des LinkScores sind neben Protein-Protein-Interaktionsnetzwerken auch genregulatorische Netzwerke für *ExprEssence* nutzbar. Verfahren wie jActiveModules [51], OptDis [52] oder KeyPathwayMiner [53] berücksichtigen ausschließlich die Unterschiede in der Genexpression. So können dort auch inhibitorische Interaktionen im Subnetzwerk auftreten, obwohl die Änderung der Genexpression der zugehörigen Gene gleichgerichtet ist.

Im Gegensatz zu den erwähnten Methoden ist die Anwendung von *ExprEssence*, abgesehen von der festzulegenden Anzahl an Kanten, die das kondensierte Netzwerk bilden sollen, frei von Parametern, deren genutzte Werte vom Nutzer zu legitimieren wären.

Ein weiterer Unterschied von *ExprEssence* zu anderen etablierten Verfahren, welche das Auffinden möglichst umfangreicher zusammenhängender Subnetzwerke zum Ziel haben, ist die ausschließliche Nutzung des LinkScores ohne weitere Nebenbedingungen bezüglich des Zusammenhangs² des identifizierten Subnetzwerks. Die Suche des aktivsten zusammenhängenden Subnetzwerks ist ein NP-schweres Problem³ [56]. Um das Problem in akzeptabler Zeit näherungsweise zu lösen, werden daher Heuristiken angewandt [51, 57]. Hierdurch wird jedoch die Optimalität der Subnetzwerke nicht von jeder Methode sichergestellt, wodurch die identifizierten Subnetzwerke bei jeder erneuten Anwendung verschieden sein können. Dies schränkt die Nachvollziehbarkeit für den Anwender erheblich ein.

Der Verzicht auf die Zusammenhangsforderung vereinfacht den Rechenaufwand enorm, da keine optimalen Subnetzwerke gefunden werden müssen. Bemerkenswerterweise tritt bei Anwendung von *ExprEssence* dennoch häufig die Bildung von Zusammenhangskomponenten auf. Das kann darauf zurückzuführen sein, dass die die Zusammenhangskomponente bildenden differentiell regulierten Interaktionen tatsächlich mechanistisch verknüpft und somit gemeinsam differentiell reguliert sind. *ExprEssence* ist am besten in der Lage diese Zusammengehörigkeit aufzudecken, wenn die genutzten Netzwerke auf die der jeweiligen Untersuchung zugrundeliegenden Biologie fokussiert sind (zelltyp-/gewebsspezifische Netzwerke) (siehe kondensierte *PodNet*- (spezifisch) und *GlobalNet*-Netzwerke (unspezifisch) in Publikation #3, Abbildungen 5 und 6, Seite 57).

Doch auch bei der Nutzung von nichtfokussierten Netzwerken werden mit *ExprEssence* plausible Subnetzwerke gewonnen: Im Anwendungsfall der Unterscheidung von Erfolg

²Ein Netzwerk ist zusammenhängend, wenn es von jedem Knoten eine Verbindung zu jedem beliebigen anderen Knoten gibt.

³Als NP-schwer wird eine Klasse von Fragestellungen bezeichnet, für die keine Lösung in polynomieller Zeit bekannt ist. Auf die Suche nach optimalen aktiven Subnetzwerken bezogen bedeutet das, dass der Zeitbedarf bei der Suche nach ihnen um so schneller ansteigt, je größer das Ausgangsnetzwerk ist. Eine vorgegebene Lösung lässt sich jedoch in polynomieller Zeit prüfen.

und Misserfolg einer Chemotherapie bei Brustkrebs (siehe Publikation #3) wurde ein unfokussiertes, auf der Protein-Protein-Interaktionsdatenbank STRING [6] basierendes Netzwerk genutzt, da kein Brustkrebs-fokussiertes Netzwerk zur Verfügung stand, das nicht ebenso auf Interaktionen aus Datenbanken zurückgeht [23,58]. Eine Gegenüberstellung der Subnetzwerke von *ExprEssence* und OptDis hat gezeigt, dass das *ExprEssence*-Subnetzwerk direkter mit dem Wirkungsprinzip der Therapie korrespondiert, als es bei OptDis der Fall ist. Aus dem erhaltenen *ExprEssence*-Subnetzwerk konnte eine Hypothese über einen am Therapie-Erfolg beteiligten Mechanismus abgeleitet werden, der bisher nicht in der Literatur erwähnt wurde. Die Hypothese wurde bereits in Teilen experimentell untermauert (siehe Publikation #3).

2.5 Podozyten und das PodNet

Die vorgestellten selbstentwickelten Methoden wurden unter anderem bei der Untersuchung des Podozyten, eines für die Filterfunktion der Niere essentiellen Zelltyps, angewandt. Diese Zellen befinden sich im Glomerulum (Bowman-Kapsel) und bilden dort zusammen mit der Basalmembran und dem Kapillarendothel die Blut-Harn-Schranke. Der Name des Podozyten leitet sich von seiner äußeren Form ab, welche an einen Fuß (griechisch: Podos) erinnert. Diese Form wird durch das Aktin-Zytoskelett der Zellen ermöglicht [59]. Die Fußfortsätze (Zehen) zweier Podozyten interdigitieren miteinander wie die Finger zusammengefalteter Hände und bilden so den Filtrationsschlitz, welcher vom extrazellulären Schlitzdiaphragma bedeckt ist [60]. Dieses hat die Funktion eines größenselektiven Filters, der hochmolekulare Substanzen wie Albumine im Blut zurückhält, während niedermolekulare Substanzen diese Barriere durchdringen und zu Urin angereichert werden können. Durch die geschlungene Form der Filtrations-schlüsse wird die Filtrationsfläche stark vergrößert. Wenn Podozyten, beispielsweise durch überhöhten Blutdruck, einer erhöhten Dehnung ausgesetzt sind, versuchen sie durch Reorganisation ihres Aktin-Zytoskeletts, dem Druck standzuhalten und ihre Funktion weiterhin auszuüben [61, 62]. Sollten die Podozyten jedoch Schaden nehmen und sich ihre Fußfortsätze zurückbilden [63], verkleinert dies die Filtrationsfläche und somit die Filtrationsrate. Schließlich können sich die Podozyten aufgrund des gestiegenen Drucks nicht mehr an der Basalmembran halten. Sie werden abgesprengt und die Filtrationskompetenz der Niere ist folglich eingeschränkt bis nicht mehr vorhanden, was die Dialysepflicht und gegebenenfalls eine Nierentransplantation zur Folge hat.

Ein Ziel der Forschung an Podozyten ist aufgrund dieser medizinischen Relevanz die Erlangung von Wissen über die molekularen Prozesse bei der Rückbildung der Fußfortsätze.

2.5. PODOZYTEN UND DAS PODNET

Hierbei spielen das Aktin-Zytoskelett und mit diesem assoziierte Proteine wie Cd2ap, Palladin oder Synaptopodin eine zentrale Rolle [64, 65].

Um die Auswirkungen experimenteller Stimuli sowie verschiedener Entwicklungsstadien des Podozyten auf diese Proteine sowie deren Interaktionsumfeld abschätzen zu können, wurde ein funktionell annotiertes Protein-Protein-Interaktionsnetzwerk des Podozyten (*PodNet*) erstellt (Publikation #2, Abbildungen 1, Seite 51). Mit diesem Netzwerk wurde eine wichtige Grundlage für die systembiologische Untersuchung von Podozyten geschaffen. Mit Hilfe des *PodNets* wurden zahlreiche Erkenntnisse, etwa bezüglich einer möglichen Relevanz zweier bisher nicht näher betrachteten Proteine (Cldn5, Pak1) sowie der Ähnlichkeiten und Unterschiede zwischen in der Entwicklung befindlichen und zellkultivierten Podozyten, gewonnen [17].

Allerdings ist es unserer sowie anderen Forschungsgruppen bisher nicht gelungen, die Rückbildung der Fußfortsätze zu unterbinden oder sogar ein therapeutisch anwendbares Konzept zur Podozyten-Regenerierung zu entwickeln. Das Dogma, dass eine Regenerierung verlorengegangener Podozyten nicht möglich sei, wurde bisher nicht wissenschaftlich zweifelsfrei widerlegt. Zwar sind Publikationen erschienen, die mögliche Ansätze zur Regenerierung von Podozyten aufzuzeigen versuchen [66–69], jedoch sind diese noch unzureichend experimentell belegt und werden im Fach kontrovers diskutiert.

3 Ausblick

3.1 Informationsaustausch in der Wissenschaft im Zeitalter der Bioinformatik

Das *PodNet* wurde im Cytoscape-Format (zu Cytoscape: siehe oben) erstellt und steht unter www.podnet.de zum Download zur Verfügung. Zudem ist es auf WikiPathways [70] vorhanden und kann somit von anderen Personen ergänzt oder gegebenenfalls korrigiert werden. Bis jetzt ist dies jedoch noch nicht geschehen, was als Indiz dafür aufgefasst werden kann, dass diese Form der Kooperation zwischen den Forschungsgruppen entweder unzureichend bekannt ist, als zu aufwendig empfunden wird oder neu gewonnene Erkenntnisse noch nicht geteilt werden sollen. Folglich werden vorhandene Synergien nicht genutzt. Erschwerend kommt hinzu, dass die publizierten Ergebnisse häufig nicht maschinenlesbar verfügbar gemacht werden. Dieser Zustand entspricht einem Vergraben (im Fließtext einer Publikation) und späterem Wiederausgraben (manuell oder mit Text-Mining) von Informationen, um sie dann mit Hochdurchsatzdaten kombiniert analysieren zu können.

Hier versuchen innovative Plattformen wie Nano-Publications Abhilfe zu schaffen [71]. Nano-Publications stellen die kleinste Einheit einer publizierbaren Information dar – ein Triplet aus Subjekt, Prädikat und Objekt (A aktiviert B; C ist ein D; E verursacht F und Entsprechendes mehr). Diese Triplets werden in einer Datenbank gesammelt und stehen für Suchanfragen zur Verfügung. Derartige Dienste werden jedoch bisher in der Breite kaum angenommen.

Langfristig ist die Verpflichtung anzustreben - ähnlich wie bereits bei Genexpressions- oder Sequenzierungsdaten -, die in einer eingereichten Publikation vorgestellten wesentlichen Erkenntnisse an eine Nano-Publications Plattform zu übermitteln, um sie direkt maschinell verwertbar vorzuhalten. Dies würde einen essentiellen Fortschritt bei der Verwendung des verfügbaren Wissens in bioinformatischen Methoden bedeuten, ohne deren Hilfe neu erhobene Daten, wenn überhaupt, nur sehr viel schwerer in ein vom Menschen interpretierbares Konzept übersetzt werden können.

Ein Bereich, in dem dieser Ansatz gerade Fuß zu fassen beginnt, ist die pharmazeutische Forschung. In einem kollaborativen Projekt mit über zwei Dutzend Beteiligten aus dem akademischen, pharmazeutischen und biotechnologischen Umfeld wurde die Plattform OpenPHACTS gegründet [72], mit deren Hilfe neue Medikamente effizienter entwickelt werden sollen. Hierzu werden Informationen aus diversen Quellen in einem seman-

3.2. AUSBLICKE BEZÜGLICH DER NETZWERKBASIERTEN FORSCHUNG UND EXPRESSENCE

tisch einheitlichen Konzept zusammengeführt und somit algorithmisch direkt nutzbar gemacht.

3.2 Ausblicke bezüglich der netzwerkbasierten Forschung und ExprEssence

Biologische Netzwerke, unter ihnen Protein-Protein sowie genregulatorische Netzwerke, stellen in der Biologie und Medizin zunehmend einen integralen Bestandteil der ganzheitlichen Modellierung der jeweils untersuchten Prozesse und Zustände dar. Sie ermöglichen es, biologisch aktive Komponenten wie Gene, Proteine, Metabolite, Signallstoffe etc. nicht isoliert, sondern im funktionalen Kontext zu interpretieren und sehr komplexe Verflechtungen und Abhängigkeiten zwischen ihnen abzubilden. Zudem sind Netzwerke besonders gut als Gerüst für die Integration von Daten diverser Quellen und Qualitäten (Genom, Transkriptom, Proteom, Metabolom, Epigenom) geeignet. Hierauf aufbauend sind Methoden entwickelt worden, die in Netzwerken nach Bereichen suchen, die unter bestimmten Bedingungen - beispielsweise dem Auftreten einer Krankheit - vom Normalzustand abweichen. Die Nutzung dieser aktiven Subnetzwerke resultiert in einer höheren Korrektklassifikationsrate gegenüber der Klassifikation mit üblichen Biomarkern. Die so gewonnenen Erkenntnisse können bei der Diagnose, für prognostische und somit therapierelevante Zwecke und nicht zuletzt für die Ursachenforschung der Krankheit genutzt werden.

Im Rahmen des Ausbaus und der großflächigen Umsetzung der Individualisierten Medizin wird die Arbeit mit biologischen Netzwerken angesichts der bisherigen erfolgreichen Anwendungen weiteren Auftrieb erhalten. Damit aus den Netzwerken Resultate abgeleitet werden können, die der Realität möglichst nahekommen, sollten die Netzwerke die für die jeweilige Fragestellung wesentlichen molekularbiologischen Aspekte berücksichtigen. Derzeit ist die Anzahl solch fokussierter Netzwerke jedoch überschaubar. Sollten die Spezifität und Sensitivität von für die Erstellung kontextabhängiger Netzwerke genutzten Text-Miner hinreichend groß werden (~95-98 %), würde dies für die Arbeit mit biologischen Netzwerken einen großen Fortschritt bedeuten: Die manuelle Prüfung der Netzwerkinhalte auf Korrektheit könnte entfallen und die Netzwerke selbst würden für Hochdurchsatzverfahren zur Verfügung stehen. Von mindestens ebenso großer Bedeutung ist die experimentelle Bestimmung von Protein-Protein-Interaktionen im Hochdurchsatzverfahren, etwa durch Chromatin-Immunpräzipitation.

KAPITEL 3. AUSBLICK

In Folge könnten sogenannte Graphlet-basierte⁴ vergleichende Analysen von biologischen Netzwerken neue Einblicke in die Architektur zellbiologischer Prozesse liefern [73–75]. Eine für die Pharmazeutische Industrie lukrative Anwendung von biologischen Netzwerken ist das sogenannte *drug repositioning*. Hierbei wird für bereits bekannte biologisch aktive Wirkstoffe nach bisher unbekannten Einsatzmöglichkeiten gesucht [76]. Dieses Vorgehen reduziert den Aufwand der Entwicklung und Zulassung der für den neuen Anwendungsfällen vorgesehenen Medikamente erheblich. Neben der bereits erwähnten Plattform OpenPHACTS stellen die Datenbanken PROMISCUOUS und STITCH Erkenntnisse über die Wirkung von über 25.000 beziehungsweise 300.000 Wirkstoffen auf Proteine in einer Datenbank zur Verfügung (Stand August 2013) [77, 78]. Diese Wechselwirkungen lassen sich in Protein-Protein-Interaktionsnetzwerke integrieren, um den Einfluss eines Wirkstoffs auf im Netzwerk weiter entfernt liegende Bereiche abschätzen zu können.

Eine mögliche Anschlussverwendung von mit *ExprEssence* ermittelten aktiven krankheitsassoziierten Subnetzwerken in Kombination mit derartigen Wirkstoff-Protein-Interaktionsnetzwerken ist die Suche nach Wirkstoffen, die in ihrer Wirkung den vom Normalzustand abweichenden krankhaften Änderungen entgegengerichtet sind. Lautet die Vorhersage, dass die im Krankheitsfall in ihrer Aktivität gesteigerten Interaktionen durch Applikation eines Wirkstoffs oder einer Wirkstoffkombination wieder gedämpft werden können, stellen diese Wirkstoffe Kandidaten für eine mögliche Therapie der Krankheit dar und können experimentell geprüft werden.

⁴Graphlets sind Subnetzwerke mit etwa bis zu 5 Knoten. Über die Verteilung der verschiedenen Graphlets im Netzwerk sollen Aussagen beispielsweise bezüglich der Ähnlichkeit von Netzwerken getroffen werden. Hierbei werden ausschließlich die Netzwertopologie betreffende Daten genutzt.

Teil II

Publikationen

Ausweisung der Eigenanteile an den Publikationen

Publikation #1

ExprEssence - revealing the essence of differential experimental data in the context of an interaction/regulation net-work.

Warsow G, Greber B, Falk SS, Harder C, Siatkowski M, Schordan S, Som A, Endlich N, Schöler H, Repsilber D, Endlich K, Fuellen G.

BMC Syst Biol. 2010 Nov 30;4:164. doi: 10.1186/1752-0509-4-164.

Anteil: Entwicklung, Implementierung und Anwendung von ExprEssence; Prozessierung der Genexpressionsdaten; Vergleich mit jActiveModules; Schreiben eines mehrheitlichen Anteils des Manuskripts.

Publikation #2

PodNet, a protein-protein interaction network of the podocyte.

Warsow G, Endlich N, Schordan E, Schordan S, Chilukoti RK, Homuth G, Moeller MJ, Fuellen G, Endlich K.

Kidney Int. 2013 Jul;84(1):104-15. doi: 10.1038/ki.2013.64. Epub 2013 Apr 3.

Anteil: Aufbau von PodNet aus vorgegebenen Interaktionslisten; Anbindung von Cytoscape an die STRING Datenbank zur Erweiterung von PodNet zu XPodNet sowie zur Generierung von GlobalNet; Prozessierung der Genexpressionsdaten; Erstellung und Analyse der kondensierten Netzwerke; Schreiben der methodischen, nichtinterpretativen Teile des Manuskripts.

Publikation #3

Differential Network Analysis Applied to Preoperative Breast Cancer Chemotherapy Response.

Warsow G, Struckmann S, Kerkhoff C, Reimer T, Engel N, Fuellen G.

eingereicht bei PLOS ONE; in Überarbeitung

Anteil: Erstellung des Ausgangsnetzwerks; Aufbereitung der Transkriptomdaten; Anwendung von ExprEssence; Interpretation des kondensierten Netzwerks; Vergleich mit OptDis und KeyPathwayMiner; Schreiben eines mehrheitlichen Anteils des Manuskripts.

SOFTWARE**Open Access**

ExprEssence - Revealing the essence of differential experimental data in the context of an interaction/regulation net-work

Gregor Warsow^{1,2,3}, Boris Greber⁴, Steffi SI Falk¹, Clemens Harder¹, Marcin Siatkowski^{5,1}, Sandra Schordan², Anup Som¹, Nicole Endlich², Hans Schöler^{4,6}, Dirk Repsilber⁷, Karlhans Endlich², Georg Fuellen^{1*}

Abstract

Background: Experimentalists are overwhelmed by high-throughput data and there is an urgent need to condense information into simple hypotheses. For example, large amounts of microarray and deep sequencing data are becoming available, describing a variety of experimental conditions such as gene knockout and knockdown, the effect of interventions, and the differences between tissues and cell lines.

Results: To address this challenge, we developed a method, implemented as a Cytoscape plugin called *ExprEssence*. As input we take a network of interaction, stimulation and/or inhibition links between genes/proteins, and differential data, such as gene expression data, tracking an intervention or development in time. We condense the network, highlighting those links across which the largest changes can be observed. Highlighting is based on a simple formula inspired by the law of mass action. We can interactively modify the threshold for highlighting and instantaneously visualize results. We applied *ExprEssence* to three scenarios describing kidney podocyte biology, pluripotency and ageing: 1) We identify putative processes involved in podocyte (de-)differentiation and validate one prediction experimentally. 2) We predict and validate the expression level of a transcription factor involved in pluripotency. 3) Finally, we generate plausible hypotheses on the role of apoptosis, cell cycle deregulation and DNA repair in ageing data obtained from the hippocampus.

Conclusion: Reducing the size of gene/protein networks to the few links affected by large changes allows to screen for putative mechanistic relationships among the genes/proteins that are involved in adaptation to different experimental conditions, yielding important hypotheses, insights and suggestions for new experiments. We note that we do not focus on the identification of 'active subnetworks'. Instead we focus on the identification of single links (which may or may not form subnetworks), and these single links are much easier to validate experimentally than submodules. *ExprEssence* is available at <http://sourceforge.net/projects/expronsense/>.

Background

The pace of data generation in the life sciences is steadily increasing. Primary data sets grow in depth and accuracy, covering more and more aspects of life. In molecular biology and biomedicine, these include large-scale measurements of DNA/Histone acetylation, transcriptional activity, gene expression and protein abundance (e.g. [1]). Measuring epigenetic patterns (DNA methylation, DNA/Histone acetylation) on a large

scale has become possible only recently [1,2]. Measuring transcription is entering a new era with the introduction of deep (or next-generation, RNA-seq) sequencing [3,4]. Proteomics is becoming possible at unprecedented depth, covering ever-larger parts of the proteome on a routine basis [5]. For these primary data, repositories such as the Gene Expression Omnibus database (GEO [6]) or ArrayExpress [7] are constantly expanding.

Often, measurements are differential: they are made for two or more conditions (such as gene knockdown or knockout [8]), for two or more time points (such as time series tracking the consequences of some experimental intervention, [9]), or for two or more species

* Correspondence: fuellen@uni-rostock.de

¹Institute for Biostatistics and Informatics in Medicine and Ageing Research, University of Rostock, Ernst-Heydemann-Str. 8, 18057 Rostock, Germany
Full list of author information is available at the end of the article

(such as mouse and human, [10]). *Exploiting differential measurements is one key to cope with the flood of data, by focusing on the most pronounced differences.*

Life scientists also have to handle a deluge of secondary data, in the form of papers, reviews and curated databases. These may be integrated by automated systems such as STRING [11], or by manual efforts [12-14]. *Exploiting secondary data provides another key to cope with the flood of primary data, by putting them into context and focusing on the most pronounced confirmations and contradictions to what is known already.*

In this paper, we propose to interpret differential data in the context of knowledge, yielding the ‘essence’ of an experiment. Differential data may be provided by two microarrays, and knowledge may be provided by a network describing gene/protein interaction and regulation. In this case, data tracking gene expression in the course of an experiment can be used to identify the most pronounced putative mechanisms. They are identified as those known links between genes/proteins along which expression changes indicate that there may have been some regulatory change, such as the startup or shutdown of an interaction, a stimulation or an inhibition. *ExprEssence* highlights these links, and it enables the user to filter out all links with no or negligible change. The higher the filter threshold on the amount of change to be displayed, the fewer links are shown, making it straightforward to examine the ‘essence’ of the experiment. Network condensations are illustrated by pairs of figures (original network - condensed network) in the section on Case Studies. The condensed network contains good candidates for interpreting the experiment in mechanistic terms, giving rise to the design of new experiments. However, all inferences are hypotheses derived from correlations in the experimental data in the context of the *a priori* knowledge encoded in the network, and it must be kept in mind that correlative data do *not necessarily* entail mechanistic causality. Moreover, the validity of the hypotheses generated by our method will depend on the coverage and correctness of the network, and on the accuracy of the experimental data.

Related Work

Starting with the pioneering work of Ideker et al. [15], there is a plethora of methods that combine network data with high-throughput data (such as microarrays), in order to highlight pathways or subnetworks, see the excellent recent reviews of Minguez & Dopazo [16], Wu et al. [17] and Yu & Li [18]. Notably, few of these methods are readily available as publicly accessible software packages, plugins or web services (see Table and in [17]). Also, there does not seem to be a gold standard that can be used for validation purposes (see, e.g., Tarca et al. [19] for a recent

discussion). Some methods lack validation except for the example for which they were developed for, while others are studied for an array of specific examples. In these cases, strong enrichment in plausible Gene Ontology categories or detection of known pathways or annotations is often used to demonstrate utility, as in [19-25]. We found two articles including a comparison of different subnetwork identification methods. The first one by Parkkinen and Kaski [26] introduces variants of the Interaction Component Model (ICM) method, comparing them to the original ICM method, to a method based on hidden modular random fields (HMof) [27] and to Matisse [28], using identification of Gene Ontology classes and coverage of protein complexes for two selected data sets (osmotic shock response and DNA damage data) to judge one method over the other. An evaluation of ClustEx [29], jActiveModules [15], GXNA [21] and a simple approach based on fold change can be found in [29], taking identification of gene sets, pathways and microarray targets known from the literature and from the Gene Ontology for comparison.

In general, it is exceedingly difficult to validate the detection of (sub-) networks or (sub-) pathways: these are complex entities, and ultimate experimental validation is impossible because of this complexity: experimentalists are usually limited to investigating only few components in isolation at any given time. Nevertheless, we will compare results of our method with results obtained by *jActiveModules*, in a separate section following the case studies. In contrast, by just highlighting single links in networks, we tackle a more primitive task, but in this case results can be validated directly by experiment, or by identifying corroborative statements in the literature. In particular, as can be seen from our case studies, the single links that we highlight give rise to predictions about single genes and about single one-step mechanisms that can be investigated in isolation. Therefore, we would like to emphasize the direct utility of our focus on single links and genes, complementing the (sub-)network centric view that is usually employed; to the best of our knowledge, the ‘single link and gene’ focus is not employed by other methods combining network and high-throughput (‘omics’) data. In fact, we propose a ‘winning combination’ of ‘network’/‘omics’ and ‘classical’ biology, using networks and high-throughput data to highlight single genes and links that may then be validated directly by classical molecular biology, as will be demonstrated in our case studies.

As future work, our formula for link highlighting can, however, be *integrated* into current methods for pathway/subnetwork detection, possibly improving these considerably. In particular, no such method treats inhibitions and stimulations in a distinct way, as we do. In particular, we envision that the edge score formula of

Guo et al. [20], which is based on measuring co-variance, may be replaced by our formula (see below), emphasizing a different aspect of differential gene expression: While Guo et al. identify *coordinated* changes using their formula, integration of our formula into their framework would identify subnetworks with changes that are *consistent* with an input network of interactions, stimulations and inhibitions. In any case, we wish to stress that for the identification of coordinated changes, correlation coefficients are most suitable. Our approach, however, identifies a different biological message, namely startups/shutdowns of interactions, stimulations and inhibitions, using an input network that is informative about biological relationships such as stimulations and inhibitions.

Implementation

ExprEssence is implemented in Java Standard Edition 6. It is a plugin for Cytoscape [30], an easy-to-install tool for biological network analysis and visualization. Cytoscape is an open source software project and provides basic features such as network layout and modification. Cytoscape can be enhanced for analysis purposes by straightforward installation of plugins.

Input data

ExprEssence analyses are based on a network of genes and/or proteins, in a format readable by Cytoscape, such as cys, sif, xgmml or gpml. It may be imported from databases using web services such as the Pathway Commons Web Service Client or the WikiPathways Web Service Client [31,32] as a ‘simplified binary model’ (see Fig. Five in [33]) or it may be downloaded directly from the web. Usually, it reflects expert-curated interaction/regulation data concerning a particular signaling pathway or molecular phenomenon.

The network data must follow a simple specification defined by two constraints:

- Each link (edge) must be typed to represent either an interaction, stimulation or inhibition. It is possible that all links represent physical interactions, as is the case in a pure protein-protein interaction network. Stimulations and inhibitions are directional, whereas interactions can be interpreted to be unidirectional as well as bi-directional.
- For each gene (node) at least two numerical values must be given on which a meaningful comparison can be based. For example, these may be expression values, derived from measurements in two experiments E_1 and E_2 .

By default, for better data interchangeability, *ExprEssence* recognizes Systems Biology Ontology terms [34],

also included in the activity flow language of the Systems Biology Graphical Notation (SBGN, [35]), for the specification of interaction types. Thus, each link (edge) must include an attribute called *Interactiontype*, whose values can be either *stimulation* (corresponding to SBO:0000170), *inhibition* (SBO:0000169) or *interaction* (SBO:0000231). In the networks discussed in this article, a single node is used for a gene and its protein product, and the exact nature of the links (edges) denoting stimulations, inhibitions and interactions depends on the evidence underlying the link. For example, a stimulation may be due to the modification of one protein by another, but it may also be the transcriptional stimulation of a target gene by a transcription factor.

The differential measurement data used for comparison may be integrated into the network as described in the Cytoscape manual [36]. Usually, integration is accomplished by mapping unique gene/protein identifiers in the data to unique gene/protein identifiers in the network. The measurements may be gene expression values, but they may also denote protein abundance, methylation levels, etc.

If the numerical data result from multiple measurements (replicates), the number of replicates has to be declared for each experiment, and for each experiment and for each node (gene/protein), the mean value and its corresponding variance have to be given. More specifically, for two experiments E_1 and E_2 to be compared, node A has either two or four numerical values: If the data consist of a single measurement, for node A these are the two values $M_A^{E_1}$, $M_A^{E_2}$. If replicates are analyzed, the two values $M_A^{E_1}$, $M_A^{E_2}$ are the mean values and the two variances $Var_A^{E_1}$, $Var_A^{E_2}$ are also provided. The number of replicates are n_1 and n_2 . *ExprEssence* analyses based on replicated measurements, where mean values and variances are used as input, are more reliable than analyses based on single measurements. Specifically, as the variances are used for calculations, feature variation within and between groups is considered and evaluated appropriately. However, also comparisons based on single measurements can be used to suggest underlying mechanisms.

Identifying change in a network, motivation

For each link in the network we want to measure the amount of change between experiments E_1 and E_2 , where ‘change’ is a modification in the intensity with which one gene/protein may be influencing another gene/protein; depending on the input data, such influence may be direct physical interaction (in the case of proteins), transcriptional stimulation or inhibition. Therefore, for all links connecting two genes/proteins A

and B in the network under consideration, *ExprEssence* uses the measurements $M_A^{E_1}$, $M_A^{E_2}$ and $M_B^{E_1}$, $M_B^{E_2}$ for the two experiments E_1 and E_2 to calculate a link score proportional to the amount of change from E_1 to E_2 . The formulae are given in the next section. The sign of the score corresponds to the direction of change giving a positive score for startups and a negative score for shutdowns. The magnitude of this signed change corresponds to the absolute value of the score. Links with a link score whose absolute value does not exceed a user-defined threshold are deleted from the network. Hence, only those links are kept, where changes (startups or shutdowns) are pronounced.

Following the heatmap metaphor, large measurement values for genes are indicated by red color and small values are indicated by green color. Similarly, links with a positive value of the link score are colored in red and indicate startups. Links with a negative value are colored in green and indicate shutdowns.

More specifically, in case of a stimulation of gene/protein T (target) by gene/protein S (stimulator), abbreviated $S \rightarrow T$, we suppose that the stimulation starts up (from E_1 to E_2), if the values of both genes increase (see Figure 1(a) and Figure 2(a); values for both genes S and T go up, green to red). If the values of both genes decrease, we suppose that the stimulation shuts down (Figure 1(b) and Figure 2(b)). In short, we reward correlated change. In case of an inhibition of gene/protein T (target) by gene/protein I (inhibitor), abbreviated $I \rightarrow T$, we suppose that the

inhibition starts up (from E_1 to E_2), if the value of the inhibitor increases from E_1 to E_2 , and the value of the target goes down (Figure 2(c)). If the value of the inhibitor decreases from E_1 to E_2 , and the value of the target goes up, we suppose that the inhibition is shut down (Figure 2(d)). In short, we reward anti-correlated change. Other cases, such as no change of values or an inconsistent change, that is an anticorrelated change in case of a stimulation or a correlated change in case of an inhibition, give rise to a link score with a reduced absolute value, see Figure 1(c) and 1(d), Figure 2(e)-(m), and below.

Note that stimulations are treated in a symmetrical way: $S \rightarrow T$ is treated the same way as $T \rightarrow S$. Indeed, we do not and cannot distinguish $S \rightarrow T$ and $T \rightarrow S$, because in both cases we expect increments in S to be correlated with increments in T : Higher amounts of the stimulator go hand in hand with higher amounts of the target. A similar argument holds for decrements. Motivated by this argument, interaction links ($S \leftrightarrow T$) are treated in the same way as stimulation links. This makes sense in general, because the amount of A and B interacting with each other increases in proportion to the amount of both interactors. More generally, if the interaction represents a biochemical reaction, a straightforward interpretation of our reasoning is given by the law of mass action, see the next section ‘Calculation of the amount of change’.

Calculation of the amount of change

Recall that for measurements of two experiments E_1 and E_2 , and two genes/proteins A and B , we denote the mean of the measured values for A , or, if only data of one measurement exists, the single value for A in experiment E_1 by $M_A^{E_1}$, and in experiment E_2 by $M_A^{E_2}$, respectively. The values for B are $M_B^{E_1}$ and $M_B^{E_2}$. We can then calculate the amount of change as described in the following. For gene/protein A , we determine the differential of A , D_A , that is the difference of the measured values between experiments E_1 and E_2 :

$$D_A = M_A^{E_2} - M_A^{E_1}. \quad (1)$$

In case of replicates, D_A is corrected for the variance within the replicates for both experimental conditions, employing Welch's formula [37]:

$$D_A = \frac{M_A^{E_2} - M_A^{E_1}}{\sqrt{\frac{Var_A^{E_1}}{n_1} + \frac{Var_A^{E_2}}{n_2}}}, \quad (2)$$

where $M_A^{E_1}$, $M_A^{E_2}$: Mean value of gene/protein A under experimental condition E_1 , E_2 ;

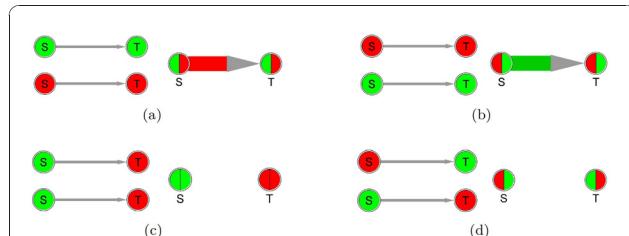


Figure 1 Network condensation - exemplified for stimulations.
For each of the panels (a) to (d), in the graphs on the left side a stimulator (S) and a target (T) are connected by a stimulation link. Values in the first (E_1 , upper graph) and second experiment (E_2 , lower graph) are indicated, where low values are marked green and high values are red (following the heat map metaphor). The coloring scheme is red, pale red, white, pale green and green, in order of decreasing values (see text). The graphs on the right side of each panel show the resulting link after applying our method. For each gene, its values for E_1 and E_2 are now inlineed simultaneously in the circle (value of E_1 on the left, of E_2 on the right side). The link connecting both nodes describes the direction and the amount of change between E_1 and E_2 . Links with startup of stimulation are colored in red (a), shutdown links in green (b). If values do not change for both the source and the target, or if they change in a completely inconsistent way, the link is removed, see (c) and (d).

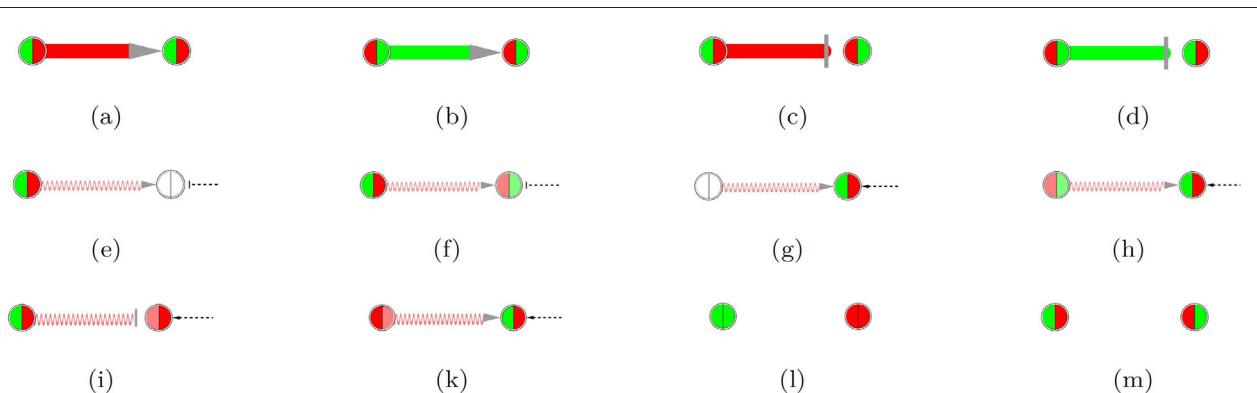


Figure 2 Network condensation - a gallery of various scenarios. For each gene, its (expression) values are represented by color. For each link, its score is represented by color and thickness. The coloring scheme is red, pale red, white, pale green and green, in order of decreasing values (see text). Links connecting genes with measurement values changing in an inconsistent way are marked by wavy lines. As in Fig. 1, if the interacting genes are linked by a stimulation $S \rightarrow T$, the stimulation is assumed to start up, if for both genes, the values go up from E_1 to E_2 ((a), E_1 value: left side of circle, E_2 : right side of circle), and it is assumed to be shut down if both values go down (b). An inhibition $I \rightarrow T$ is assumed to start up, if the inhibitor value goes up, but the target value goes down (c); it is assumed to shut down if the inhibitor value goes down and the target value goes up (d). In cases (e) and (f) the startup of the stimulation as presented is still a justified hypothesis, even though the target does not go up. For example, in (e) and (f), the stimulation by the source (the stimulator) goes up but it may be counteracted by other inhibiting effects (dashed T-Bar arrow) on the target, as the target does not change (e) or even goes down slightly (f) (*source principle*, see text). In cases where the amount of the stimulator is constant (g) or goes down slightly (h), the startup of a stimulation is still a justified hypothesis based on the target value. Strictly speaking, we hypothesize the startup of the stimulatory effect on the target gene. For example, in (g), the startup is not concluded from the change in the value of the stimulator, but it may be due to stimulator accumulation in time, and/or due to cooperation of the stimulator with other stimulations of the target which go up at the same time; startup of the stimulating effect is concluded from the behaviour of the target (*target principle*, see text). Scenario (i) is another example of the *source principle*: it is a justified hypothesis that the inhibition starts up because the amount of inhibitor increases, even though counteracting stimulations drive the amount of the target. Scenario (k) is another example of the *target principle*: it is a justified hypothesis that the stimulatory effect goes up, observing the target and assuming other cooperating effects on it. Lastly, if values do not change at all, or if they change in a completely inconsistent way, the amount of change is zero or near to zero, as in (l) and (m) (also see Fig. 1 (c) and (d)). Note that cases (e)-(m) all result in reduced link scores. Hence, inconsistent links tend to be removed from the network, as the link score threshold is made more stringent.

$Var_A^{E_1}$, $Var_A^{E_2}$: Variance of values of gene A under experimental condition E_1 , E_2 ;

n_1 , n_2 : Number of replicates done in experiment E_1 , E_2 .

D_B is determined analogously. This equation corresponds to the Welch t-test for comparison of mean values of two samples with unequal variances. As we do not want to make strong preconditions about the statistical distribution of the samples, we do not calculate p-values. The weaker preconditions for Welch's t-test are fulfilled if, for both experiments, independent samples are measured, and if their values are approximately normally distributed. Given D_A and D_B , the amount of change for an interaction link is the sum of the two differentials:

$$\text{LinkScore}_{Int} = D_A + D_B. \quad (3)$$

Taking the difference $M_A^{E_2} - M_A^{E_1}$ and not $M_A^{E_1} - M_A^{E_2}$ reflects the motivation to denote startups of interactions by a positive score and shutdowns by a negative score.

The formula gives scores with high absolute value for correlated changes of the values of A and B from E_1 to E_2 . Depending on the direction of the correlated change, the score becomes positive or negative which denotes a startup or shutdown of the interaction/stimulation. Anti-correlated changes are given a reduced absolute value of the score (see below and Figure 2(f) and 2(h)). The formula is simple, yet powerful:

1. In the specific case of a physical interaction between two proteins, and *log*-transformed data, the formula above corresponds to the law of mass action, as follows. The 'activity' of a physical interaction of protein A with protein B can be expressed by the product of the abundances of both, assuming that the expression values correspond to the 'amount' of protein. The 'amount' of the complex AB in experiment 1 can then be compared to the 'amount' of the complex AB in experiment 2, by taking the ratio. Large changes in this ratio indicate that there will be much more or much less of the protein complex, comparing experiment 1 with experiment 2. (Note that we do neither calculate

equilibrium constants nor reaction kinetics.) As we have two experimental conditions and are interested in the change from E_1 to E_2 , startup of 'activity' is thus proportional to the ratio of the products of the abundances of A and B , taking experiment E_2 over experiment E_1 : $([A]^{E_2} \cdot [B]^{E_2}) / ([A]^{E_1} \cdot [B]^{E_1})$. In case of log-transformed values, this is the difference of the sums of the measurement values under both conditions: $(M_A^{E_2} + M_B^{E_2}) - (M_A^{E_1} + M_B^{E_1})$. This can be written as $(M_A^{E_2} - M_A^{E_1}) + (M_B^{E_2} - M_B^{E_1})$ and corresponds to $D_A + D_B$ from formula (3). Hence, our formula for the link score of interaction links can be connected directly to the law of mass action.

2. As explained above, we can treat the stimulation of a gene/protein A by a gene/protein B in the same way as an interaction of the two proteins with each other and therefore use the same formula to determine the link score:

$$\text{LinkScore}_{\text{Stim}} = \text{LinkScore}_{\text{Int}} = D_A + D_B. \quad (4)$$

3. Formula (4) can be modified to capture inhibitions $A \rightarrow B$ (A inhibits B), where A and B are expected to be anticorrelated in their expression/amount:

$$\text{LinkScore}_{\text{Inh}} = D_A - D_B. \quad (5)$$

This equation honors the case where higher amounts of the inhibitor A go hand in hand with lower amounts of the target B and vice versa, whereas correlated changes are penalized (see Figure 2(c) and 2(d)).

4. Our formulae deliver justified hypotheses also in the cases that are not as straightforward as the cases in Figure 1(a)-(b)/Figure 2(a)-(d), given two additional assumptions, that we call the *source principle* and the *target principle*. It is important to note that these complicated cases are characterized by relatively low link scores and additionally they will be marked by wavy lines. Furthermore, they can be identified by inspecting the color-coded measurement values (Figure 2(e)-(k)), which can be made explicit by addition of gene/node labels as in the condensed networks of case studies 2 and 3.

The *source principle* maintains that changes in the source, if they are large enough, are sufficient for a

hypothesis regarding startup/shutdown of a stimulation/inhibition. Even if the value of the target is inconsistent, putting trust into the network data (that is, the stimulation/inhibition link is not questioned), the link then describes a startup/shutdown which is assumed to act on the target, even though it is counteracted by other effectors (Figure 2(e),(f), (i)). The other effectors may or may not be included in the network: we assume that the network is correct, but not necessarily complete. In case of transcriptional stimulations/inhibitions, a simple example for counteracting effectors are transcription factors that act in an opposite way at a different position of the regulatory region of the target gene. Here, we view gene regulation as a 'transcription factor battlefield' [38,39]. In fact, the target gene may not be observable (expressed) at all without the stimulation that is highlighted. There is alternative interpretation for an inconsistent target value: The stimulation may not be in the scope of what is being measured. For example, if the values refer to expression levels, a stimulation of the target by phosphorylation goes undetected.

The *target principle* holds that large changes in the target are sufficient for a hypothesis regarding startup/shutdown of a stimulation/inhibition, even if the value of the source (the stimulator/inhibitor) is inconsistent. Again trusting the network data, the link then describes a startup/shutdown that is becoming relevant because other effectors are now cooperating on the target (Figure 2(g),(h),(k)). Then, strictly speaking, in all these three cases we hypothesize that it is not the stimulation itself that goes up, but its effect on the target gene. Again, we view gene regulation as a 'transcription factor battlefield'. Also, the other effectors may or may not be part of the network. Of course, the inconsistent change in the source has to be lower than the tale-telling change in the target. Also, the startup of the stimulating effect is assumed to require only a low amount of the stimulator, which is however still exceeded. There is an alternative interpretation for an inconsistent source value: The stimulating effect may simply be delayed in case of a time series, where the stimulator (protein) needs time to accumulate, which may also happen during a period of constant or down-regulated gene expression of the stimulator.

Naturally, inconsistencies can also give rise to revision of the network. However, our formula is not designed to reveal severe inconsistencies (since such links receive scores close to zero and are removed from the network as in Figure 2(m)).

5. To distinguish straightforward from inconsistent cases by inspection, and to aid the interpretation of links, our plugin offers multi-colored nodes, inlineing directly the measurement values of a gene for a pair of experiments within a single node as a pie-chart as explained in Figure 1 and inlineed in Figure 2. To calculate the color for visualization of the values in the pie-chart, we take the 10%, 50% and 90% quantiles of the ordered list of all attribute values. The value associated with the 10% quantile defines the lower threshold. All values below this threshold are visualized by green color of same intensity. Values above this threshold and up to the value corresponding to the 50% quantile get a color defined by linear interpolation between the 10% quantile (green color) and the 50% quantile (white). Analogously, values are visualized by a color between white (50% quantile) and red (90% quantile). Values above the 90% quantile are represented by red color of same intensity. The thresholds and the coloring scheme can be redefined by the user. Furthermore, our plugin provides labeling of selected genes/nodes with the measurement data used for node coloring as shown in the condensed networks of case studies 2 and 3.

Finally, depending on the value of the *Interactiontype* attribute for a link, the respective formula for the link score is as follows:

$$\text{LinkScore} = \begin{cases} \text{LinkScore}_{\text{Int}} & \text{if } \text{Interactiontype} = \text{Interaction}, \\ \text{LinkScore}_{\text{Stim}} & \text{if } \text{Interactiontype} = \text{stimulation}, \\ \text{LinkScore}_{\text{Inh}} & \text{if } \text{Interactiontype} = \text{inhibition}. \end{cases} \quad (6)$$

We will use this link score to identify those links along which there is a large change between E_1 and E_2 . Links with a link score exceeding a user-defined threshold are colored in red or green; the other links are deleted from the network.

Condensation of networks

After importing the network and measurement data into Cytoscape, the *ExprEssence* dialog window is used to define which data shall be taken for calculation of the link score and hence for network condensation. As discussed above, the network must include at least two numerical attributes for each gene/protein, so that the formulae can be employed. These two attributes are explicitly selected by the user, indicating their order (E_1 versus E_2 , or E_2 versus E_1). After selecting two attributes, the user may then indicate that there is variance data available and specify the number of replicates. In this case, the measured values are implicitly assumed to be the mean values for which the variances are

provided. Finally, calculations are started and results are inlineed in a new network window in Cytoscape. Links with a positive change (startups) are rendered in red, and negative change (shutdown) is rendered in green. Color saturation and link thickness are directly linked to the link score calculated.

In the user control interface of *ExprEssence*, a slider (Figure 3) is provided to define the threshold to keep all links with link score exceeding the threshold, on both the positive (startup) and negative (shutdown) side of the spectrum of link score values. Using this slider, the user can cut the number of links in the network. The more stringent the threshold, the more links are removed and only links with high absolute value of the link score will remain. Genes which have no link left after removal of links are also removed from the network. Using the condensed network, the user can investigate components of the network where interactions, stimulations or inhibitions start up or shut down, comparing experiment E_1 with E_2 .

Results

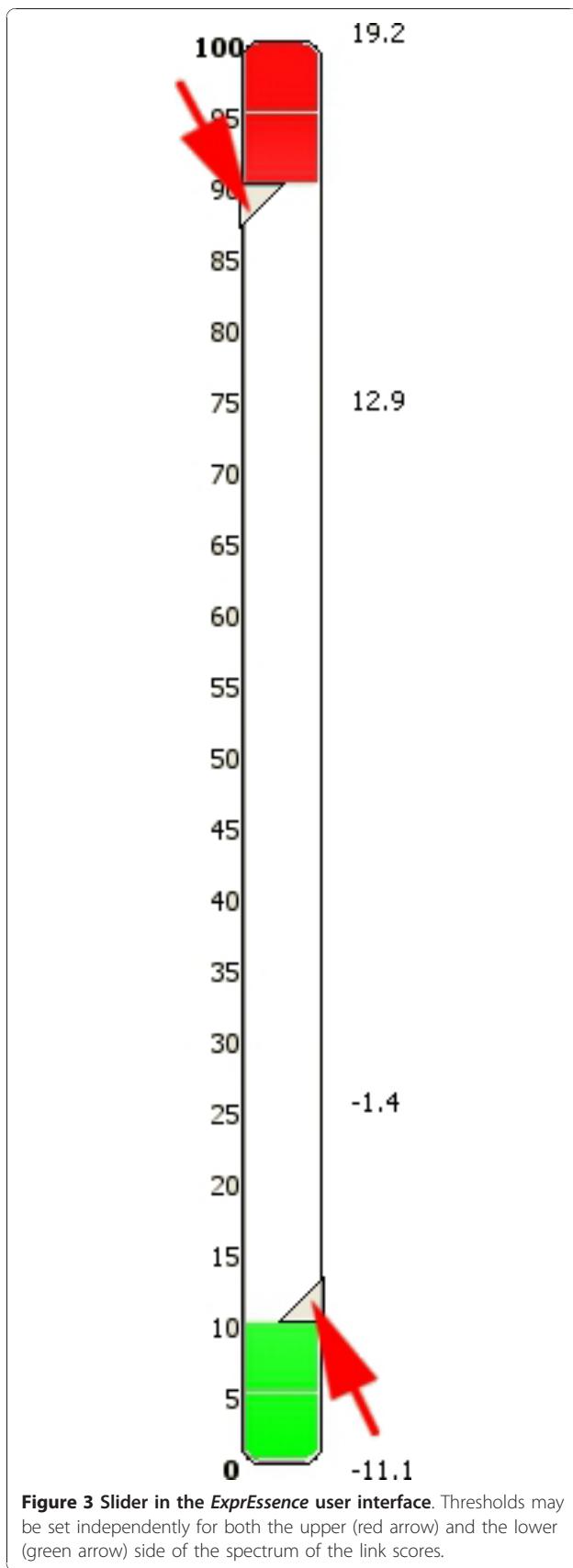
We present results of the application of *ExprEssence* in three case studies.

Case Studies

We will describe three application scenarios, condensing networks and describing the insights gained from these. As a first example, we condense a network based on literature-curated interaction data of proteins involved in structure and function of the podocyte, which is the cell forming the kidney filtration barrier. The second example will describe how a hand-curated network of interaction and regulation of genes maintaining the pluripotent state of stem cells can be condensed using microarray data tracking an early transition process of embryonic stem cells, yielding a mechanistic hypothesis that was then confirmed experimentally. In a third application, we will take a biological network describing ageing-related processes from the WikiPathways database, integrate publicly available microarray data, and confirm some basic insights into ageing. Cytoscape session files PodocyteCellMatrix.cys, Epiblast.cys and DNA_Damage.cys are provided as Additional Files 1, 2 and 3, and they enable reproduction of figures following the instructions given there.

Case Study 1 - Interaction network of podocyte cell-matrix proteins

Podocytes cover the outer aspect of the capillaries in the kidney glomerulus, where the ultra filtration of blood takes place. The filtration barrier is composed of endothelial cells, the glomerular basement membrane (GBM) and podocytes. The proper function of podocytes is essential for the ultrafiltration process.



Podocytes synthesize the majority of extracellular matrix molecules that are present in the GBM. The podocyte-GBM interface is crucial for mechanical anchorage and inside-out as well as outside-in signaling. Damage or loss of podocytes is estimated to be responsible for about 90% of kidney diseases in humans [40]. To date several hereditary kidney diseases are known that are caused by mutations in genes involved in the podocyte-GBM interface, e.g. Alport syndrome. Thus, the podocyte-GBM interface is of central importance in kidney biology and pathology.

We constructed a protein interaction network of the podocyte-GBM interface based on expert knowledge.

We collected proteins and experimentally well-described protein-protein interactions of the podocyte-GBM interface by a comprehensive survey of the podocyte literature. The expert network consists of 42 nodes (proteins) and 33 edges (protein-protein interactions). The proteins of the expert network were screened for further interaction partners utilizing the STRING database [11], to extend the expert network by further experimentally verified interactions involving at least one node (protein) of the network. If not yet existent in the network, the respective interaction partners were also added. The extended network consists of 124 nodes and 206 edges (Figure 4).

Podocyte cell lines are a frequently employed tool to study podocyte biology. However, it is well known that podocyte cell lines are partially dedifferentiated as compared to *in vivo* podocytes. To extract the main differences between the podocyte-GBM interface of *in vivo* vs. cultured podocytes, we mapped microarray gene expression data of *in vivo* and cultured mouse podocytes onto the extended network shown in Figure 4. We used publicly available microarray data (GSE10017, [41]) generated from a podocyte cell line and from *in vivo* podocytes, which were isolated as podocalyxin-positive cells in a cell suspension of enzymatically digested mouse glomeruli. By condensing a protein interaction network using gene expression data, we implicitly assume that protein abundance is correlated to gene expression. We *log*-transformed and quantile-normalized these data.

By interactive use of *ExprEssence* we removed 94% of the edges keeping the 3% quantiles of the most strongly differentially altered interactions between *in vivo* and cultured podocytes (Figure 5). *ExprEssence* revealed that the interactions of semaphorin 3 d (Sema3d), fibroblast growth factor receptor 1 (Fgfr1) and Gipc1 PDZ domain-containing protein (Gipc1) with neuropilin 1 (Nrp1) as well as the interaction between pinch 2 (Lims2) and α -parvin (Parva) are most strongly diminished (green links) in cultured podocytes as compared to the *in vivo* situation. On the other hand, the interactions of integrin β_3 (Itgb3) and myelin-associated

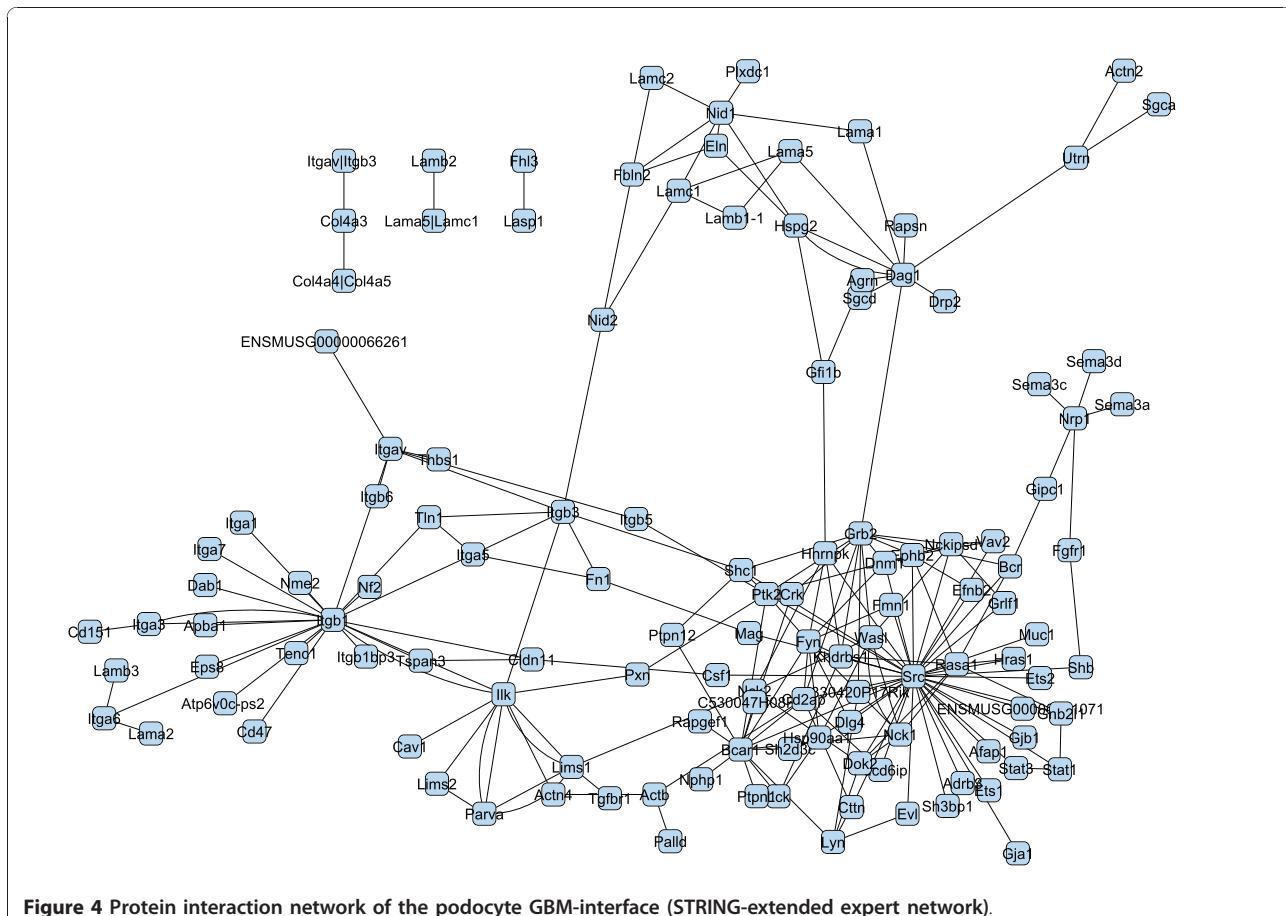


Figure 4 Protein interaction network of the podocyte GBM-interface (STRING-extended expert network).

glycoprotein (Mag) with fibronectin 1 (Fn1) are most strongly up-regulated in cultured podocytes. As Mag had so far not been reported as a podocyte protein, we analyzed Mag expression by RT-PCR in a podocyte cell line. Indeed, Mag expression was easily detected in cultured podocytes (Figure 6), revealing a novel candidate for the podocyte-GBM interface.

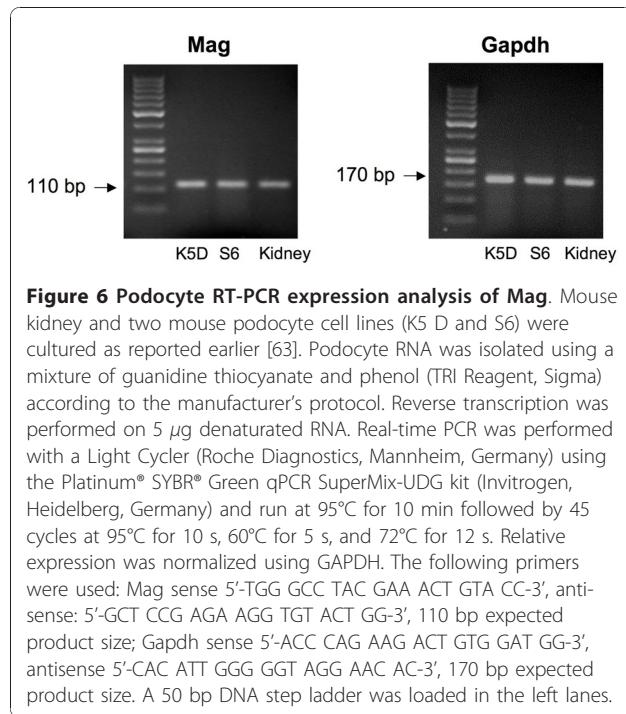
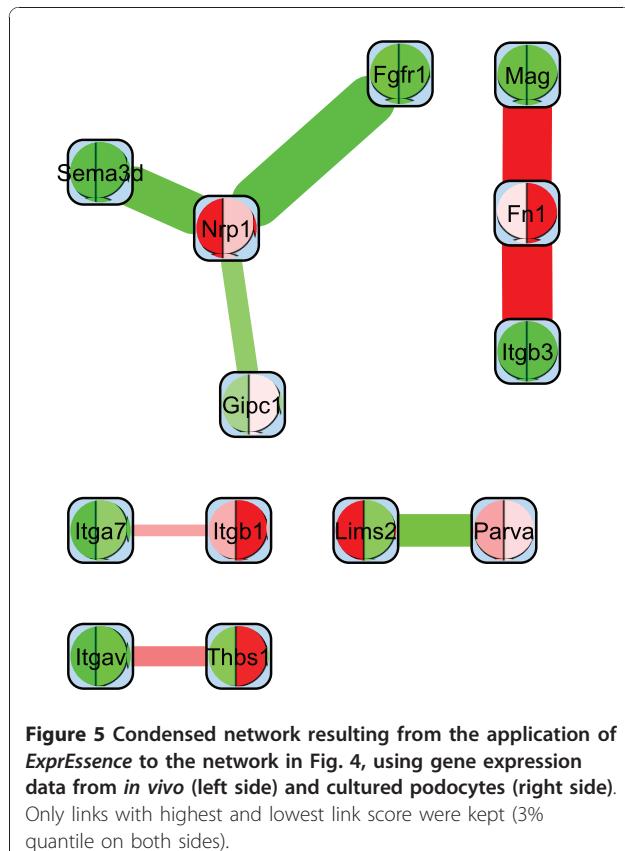
Podocytes dedifferentiate under cell culture conditions. Dedifferentiation of podocytes in culture may recapitulate dedifferentiation of podocytes *in vivo* during kidney disease. Thus, comparing gene expression between cultured and *in vivo* podocytes may give important clues about essential proteins and protein interactions needed for proper podocyte function. *ExprEssence* segregates the most strongly differentially altered interactions between cultured and *in vivo* podocytes, corroborating previous findings and discovering novel protein interactions that might be involved in the podocyte-GBM interface:

1. Pinch and parvin participate in integrin signaling via integrin-linked kinase. This pathway is essential for podocyte function, since mice with podocyte-

specific knockout of integrin-linked kinase die from renal failure at the age of 16 weeks [42]. The pinch/parvin interaction is shut down in cultured podocytes (see Figure 5), making it a candidate key interaction reflecting podocyte dedifferentiation in cell culture. In the healthy kidney, pinch and parvin may have an important role in transmitting signals from the extracellular matrix through integrin-linked kinase, to maintain podocytes in a differentiated state [43].

2. Neuropilin and its interaction with the guidance molecule semaphorin have been implicated in podocyte differentiation [44,45]. The interaction of neuropilin with several proteins, including semaphorin, is greatly diminished in cultured podocytes (see Figure 5). *ExprEssence* uncovers that loss of neuropilin interaction with extracellular molecules also participates in the dedifferentiation of podocytes in culture as suggested by the *in vivo* findings [46].

3. Massive up-regulation in cultured (= dedifferentiating) podocytes of the interaction between fibronectin 1 and the membrane protein Mag, suggest an important and hitherto unknown function of Mag in



the regulation of podocyte differentiation through the podocyte-GBM interface. Indeed, we could confirm podocyte expression of myelin-associated glycoprotein (Mag), which has so far not been implicated in podocyte biology. Since myelin proteins are known to be expressed only in glial cells of the nervous system, it is also notable that knockout of myelin protein zero, another myelin protein preferentially expressed in podocytes within the glomerulus, has been shown to result in proteinuria [47].

Case Study 2 - Analysis of a pluripotency-related experiment

Stem cell research is currently one of the most active areas in molecular biology and biomedicine, based in part on recent breakthroughs in generating 'induced pluripotent stem cells' (iPS cells) from somatic cells like fibroblasts (reviewed in [48,49]). Such a 'reprogramming' of differentiated cells into 'pluripotent' ones is possible by directly manipulating gene regulation in the cell, confronting the differentiated cell with artificial amounts of key transcription factors such as Oct4 (also known as POU5F1), Sox2 and Nanog. These 'ectopic' factors then re-direct the overall network of interaction and regulation into a direction that is so close to the 'embryonic state' that mice can be obtained, in which some (or even

all) of their cells derive from the manipulated somatic cells [50]. A mouse suffering from sickle-cell anemia was healed by reprogramming fibroblasts from its tail, correcting the genetic defect, and re-differentiating the iPS cells into blood-building cells that were then injected [51]. In human, iPS technology already allows to study a patient-specific disease in the 'petridish', and to regenerate tissues by re-differentiating iPS cells. Safety concerns currently hinder the engraftment of 'healed' tissue, and triggering the re-direction of the regulatory network by chemical compounds is one avenue to improve safety. Consequently, molecular analyses of the induction of pluripotency and of (re-)differentiation triggered by small chemical compounds is of high interest in the human as well as in the mouse system. Over the past year, we have assembled a network of molecular interactions, stimulations and inhibitions from 135 publications until March 2010, involving 262 genes/proteins of mouse. The network includes the core circuit of Oct4, Sox2 and Nanog, its periphery (such as Klf4, Esrrb, and c-myc), connections to upstream signaling pathways (such as Activin, Wnt, Fgf, Bmp, Insulin, Notch and LIF), and epigenetic regulators (Figure 7). An updated (June 2010) version of this 'PluriNetWork' is described in [14].

Applying ExprEssence to our expert network, we analyzed recently published data (GSE17136 [52]) on the effect of a pharmacological inhibitor (JAKi, Janus kinase Inhibitor I, Merck) on embryonic stem cells, which triggers a transition process from the embryonic stem cell

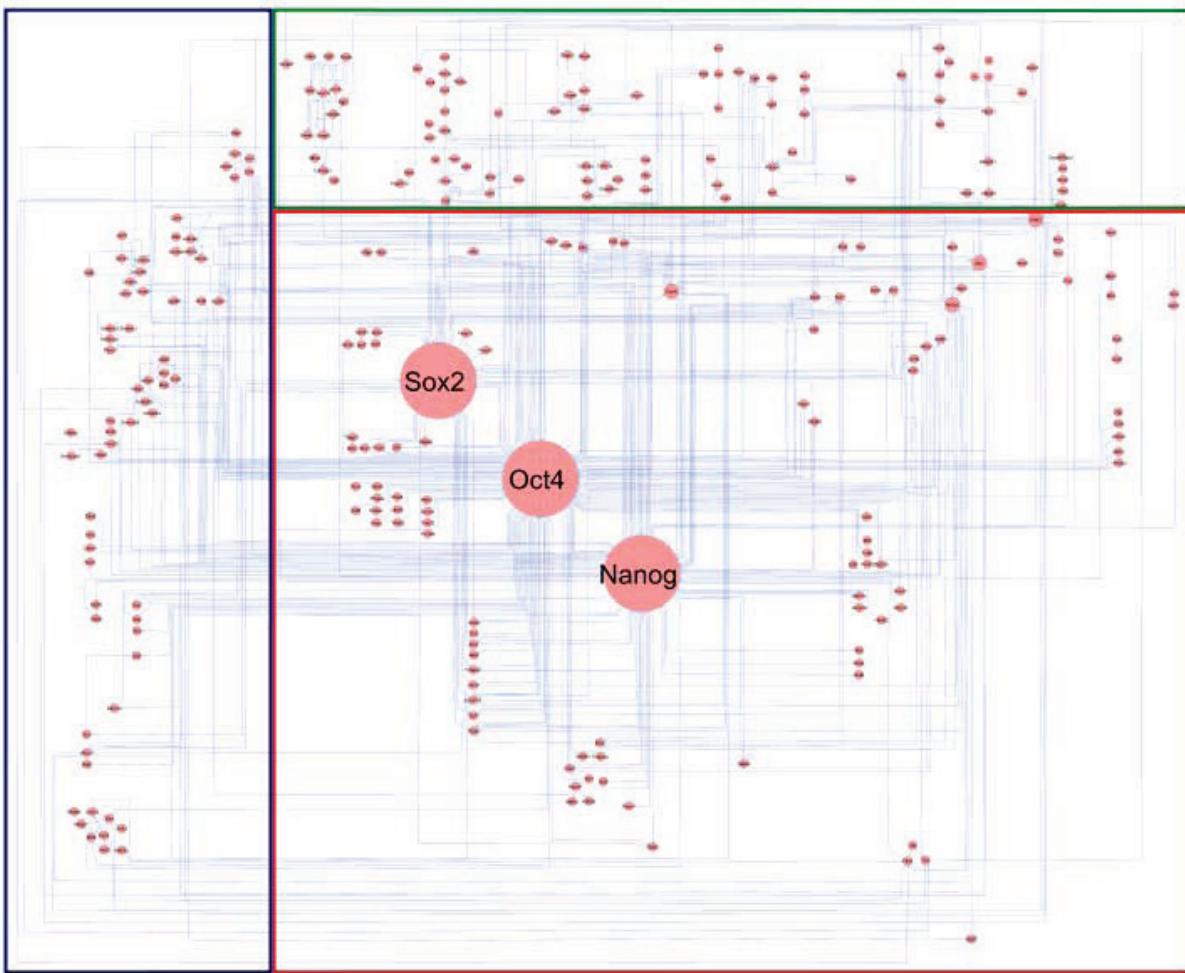


Figure 7 A network describing pluripotency-related interaction and regulation data assembled from the literature. The core part describes gene regulation, the upper part signaling pathways, and the part on the left epigenetic phenomena.

to another pluripotent cell state, the epiblast stem cell state. The effect is described by microarrays taken before, and 12 hours after the intervention. We kept the 5% quantiles of links with the largest amount of change. We observed that shutdown of stimulations is centered around the protein Esrrb, the expression of which is just slightly diminished (see Figure 8). Cooperative Esrrb regulation by a variety of transcription factors such as Klf4, Klf2 and Klf5 has already been observed by Jiang *et al.* [53]. Thus, we predict Esrrb down-regulation at a later time point. More specifically, Figure 8 inlines the condensed expert network, describing the effects of inhibition of the LIF/Jak/Stat3 signaling pathway [52] by the JAK inhibitor I. Notably, the stimulations of Esrrb by Nanog [54], Klf2, Klf4, Klf5 [53] and by itself [55] are shut down. These

shutdowns are the result of down-regulation of these stimulators within the first 12 hours. Klf2, Klf4, Klf5 and Nanog are known to be upstream of the ES cell-specific transcription factor Esrrb [53,55,56]. However, a strong effect on Esrrb was not yet seen at the 12 hour time-point, but according to Figure 8 our model suggested a down-regulation of Esrrb as a consequence of JAK inhibitor-mediated down-regulation of its upstream factors. To test this hypothesis, we carried out real-time PCR analysis of JAKi-treated ES cells at a later time-point, 48 hours. As can be inferred from Figure 9 Klf4 was already down-regulated at 12 hours but its downstream target gene Esrrb was not. At 48 hours, however, we did observe significant down-regulation of Esrrb, confirming the idea of its shutdown via other members of the ES cell self-renewal network.

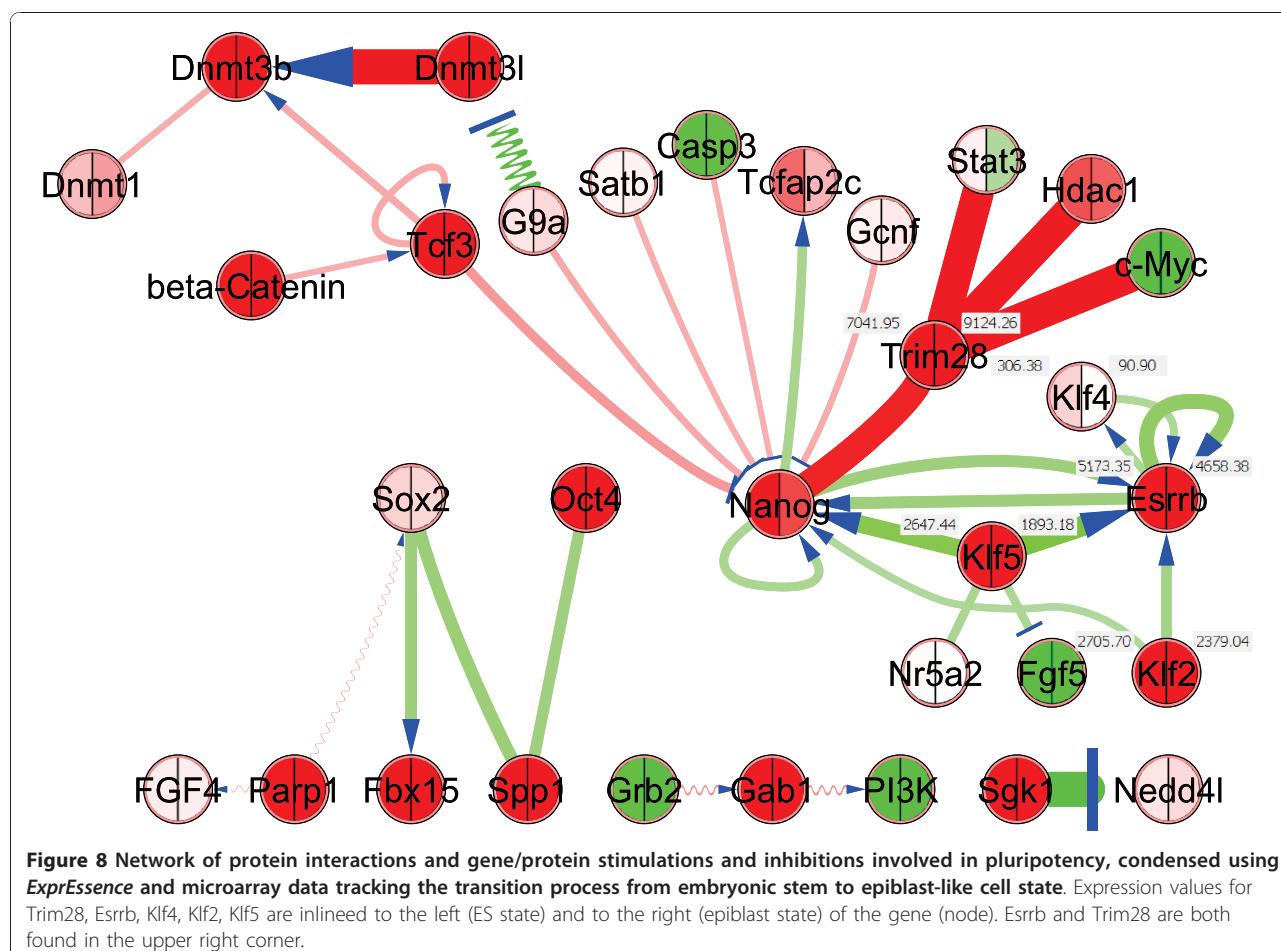


Figure 8 Network of protein interactions and gene/protein stimulations and inhibitions involved in pluripotency, condensed using ExprEssence and microarray data tracking the transition process from embryonic stem to epiblast-like cell state. Expression values for Trim28, Esrrb, Klf4, Klf2, Klf5 are inline to the left (ES state) and to the right (epiblast state) of the gene (node). Esrrb and Trim28 are both found in the upper right corner.

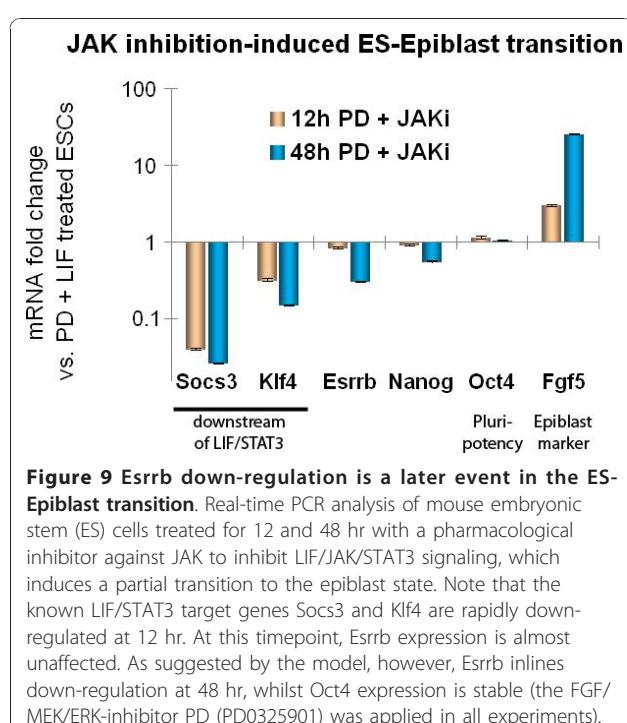
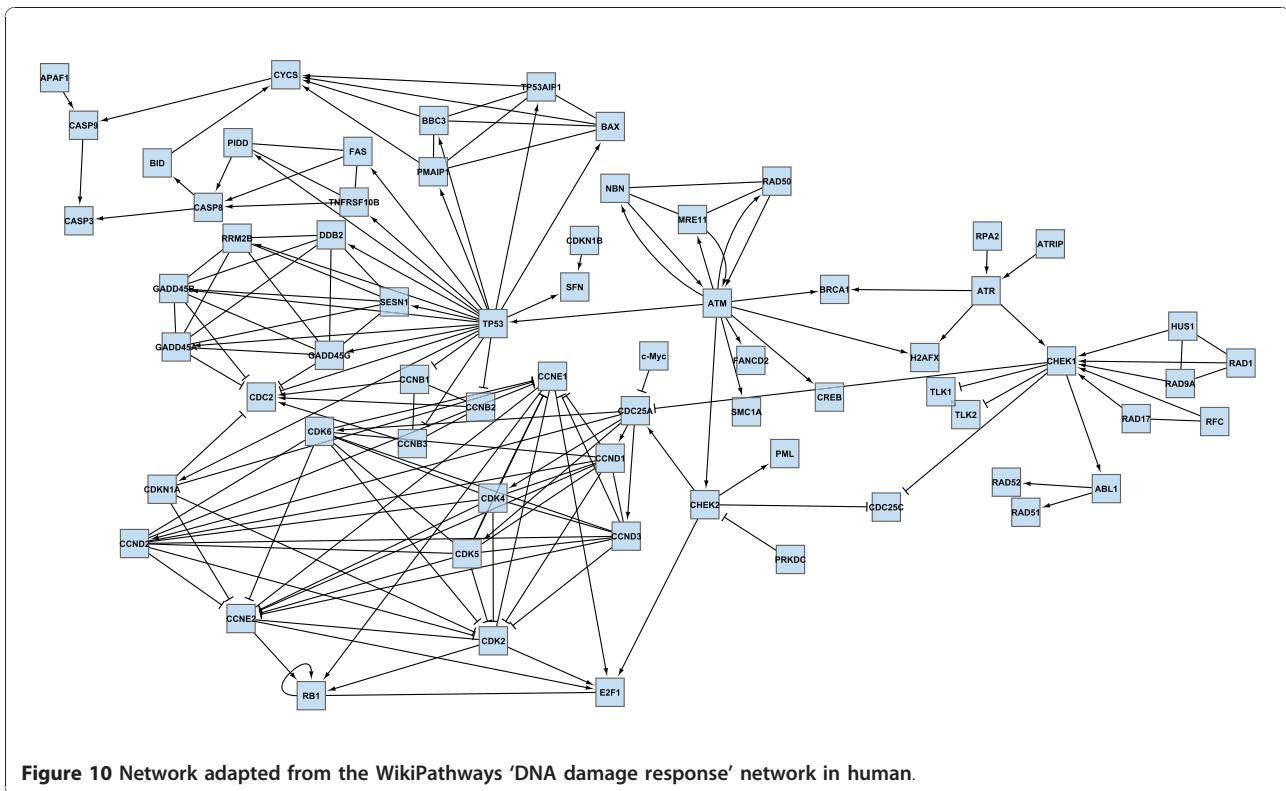


Figure 9 Esrrb down-regulation is a later event in the ES-Epiblast transition. Real-time PCR analysis of mouse embryonic stem (ES) cells treated for 12 and 48 hr with a pharmacological inhibitor against JAK to inhibit LIF/JAK/STAT3 signaling, which induces a partial transition to the epiblast state. Note that the known LIF/STAT3 target genes Socs3 and Klf4 are rapidly down-regulated at 12 hr. At this timepoint, Esrrb expression is almost unaffected. As suggested by the model, however, Esrrb inlines down-regulation at 48 hr, whilst Oct4 expression is stable (the FGF/MEK/ERK-inhibitor PD (PD0325901) was applied in all experiments).

As Klf4 and Nanog are known to be stimulated by Esrrb [55,56], these stimulations are also shut down (Figure 8*target principle*). Finally, interactions between the transcription factors Stat3, Hdac1, c-Myc & Nanog and Trim28 (also known as TIF1 β , a transcription co-regulator (co-repressor) and chromatin modifier [57,58]) are started. These startups are highlighted because the Trim28 expression value goes up strongly, from 7041 to 9124. The role of these startups is unknown, though they may reflect the general repression of components of the ES cell-specific self-renewal network by Trim28.

Case Study 3 - Analysis of ageing-related experiments

To study the effects of ageing on DNA damage response, we retrieved a network from WikiPathways [31], 'DNA damage response' in human, as of May 22, 2010. After importing it to Cytoscape, we expanded all complexes yielding the network in Figure 10. For example, for a complex in the original network such as CDK2, CCNE1 and CCNE2, all genes were connected pairwise to each other. We then integrated log-transformed and quantile normalized microarray data from GSE11882 [59]. From this dataset we used only the data



obtained from the hippocampus. We considered the same four age categories (20-39, 40-59, 60-79, and 80-99 years) as in [59]. Using *ExprEssence*, we analyzed the changes between the first (20-39 years) and the last age category (80-99 years) and kept the 3% quantiles of the most strongly differentially altered links. The startup of the stimulation of CASP8 by FAS (Figure 11 top red link) and the shutdown of the inhibition of CCNE1 by CCND3 are the largest changes. The up-regulation of apoptosis, highlighted by the red link between FAS and CASP8 just mentioned, is the result of stimulation by p53 (TP53), and is a known phenomenon in ageing processes [60]. Note that the expression value of CASP8 is going up slightly (from 4.56 to 5.38), whereas the up-regulation of FAS is more pronounced (from 5.37 to 7.15). The down-regulation of the inhibition of CCNE1 by CCND3 [61] and CCND1 as well as by their corresponding kinase CDK6 may trigger the higher expression of CCNE1, indicating a deregulation of the cell cycle. Finally, we found ageing-related up-regulation of a DNA repair pathway, that is, stimulation of DDB2 by p53 [62].

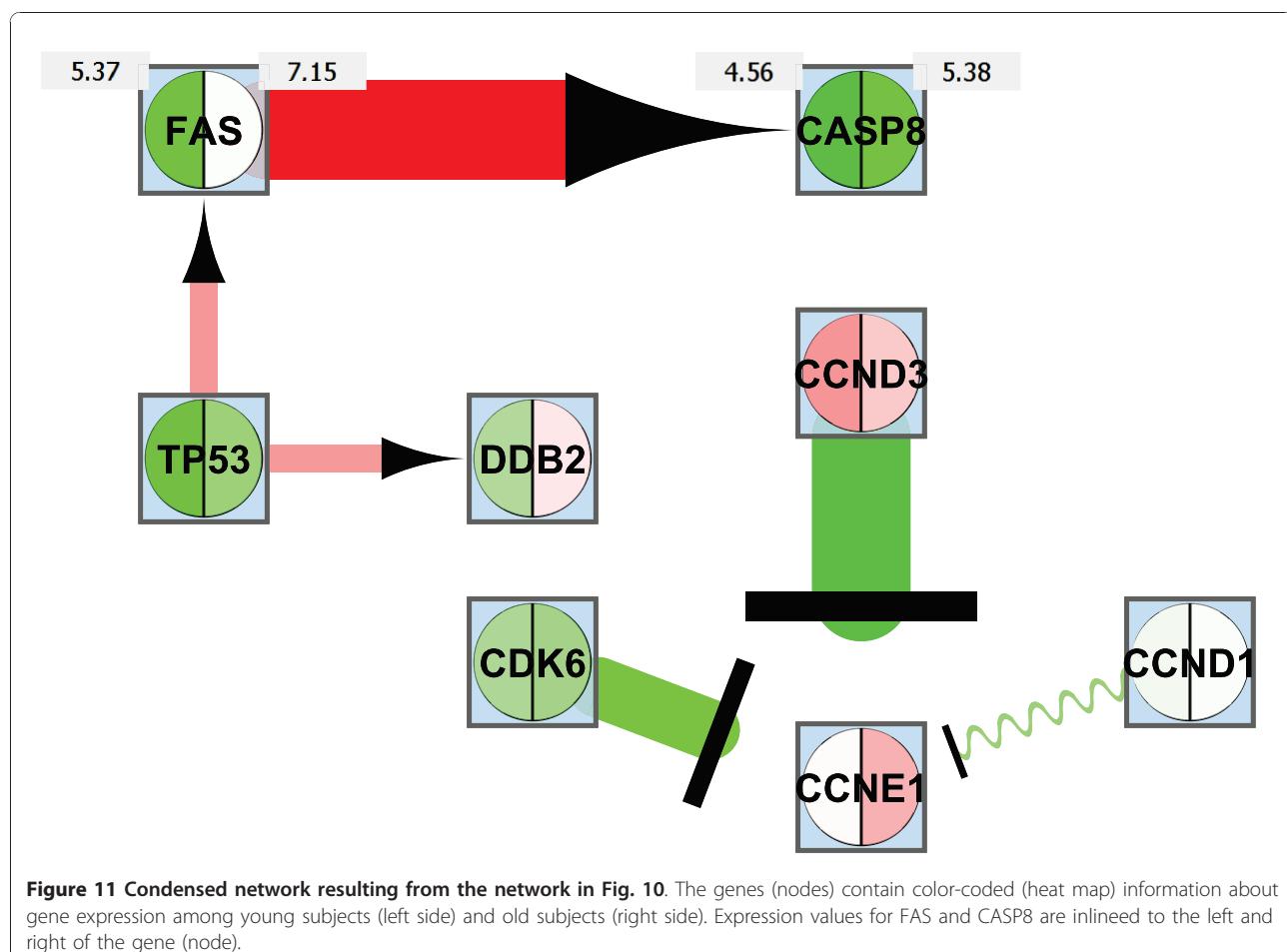
Subnetwork identification by *jActiveModules* for Case Studies 1-3

To put the results obtained in case studies 1-3 into the context of related work, we used *jActiveModules* [15] to

analyze the same data, identifying 'active modules', that are subnetworks where the constituent genes show significant changes in expression over the two conditions we investigate. As discussed in the section on 'Related Work', the aim of *ExprEssence* is quite different, namely the identification of single links (interactions, stimulations, inhibitions) and genes affected in the course of an experiment, where the links do not necessarily have to build up a connected subnetwork. Furthermore, *ExprEssence* exploits the knowledge about stimulations and inhibitions that may be encoded in the network.

We used *jActiveModules* with default parameters. In contrast to *ExprEssence*, which takes two expression values per gene (one for each experimental condition), *jActiveModules* requires one p-value per gene (describing the statistical significance of the expression change between the two experimental conditions; p-values were used as calculated while processing the raw expression data for the case studies).

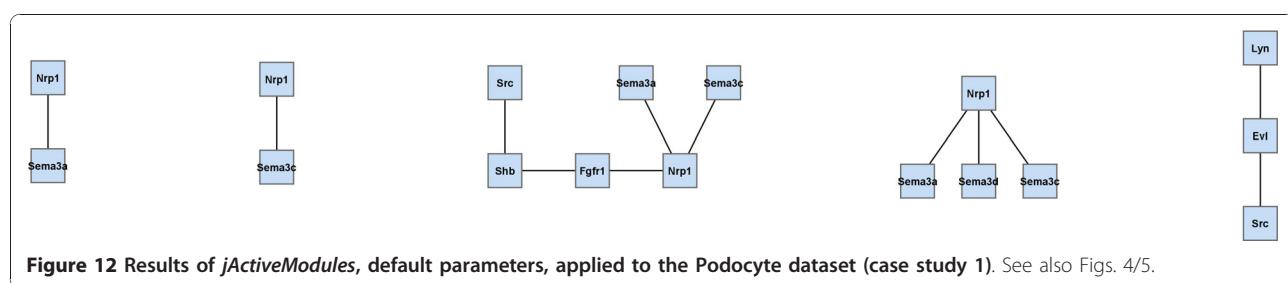
Figure 12 inlines the results of *jActiveModules* applied to the data of case study 1 (podocyte cell-matrix proteins). Module scores are (from left to right) 4.048, 3.384, 2.927, 2.861, 2.761. The first four subnetworks are overlapping. The Nrp1 gene/protein that is found in these four subnetworks is also found in our condensed network (Figure 5). In Figure 12 Nrp1 is linked to Sema3a & Sema3c, as well as to Fgfr1 & Sema3 d, and



expression of these four genes indeed changes significantly. Links to the latter two genes are highlighted by *ExprEssence*, because change of expression is correlated as in Figure 1(b), even though Fgfr1 & Sema3 d change only slightly. Links to the first two genes are not highlighted by *ExprEssence*, because change of expression is anti-correlated as in Figure 1(d), and the link threshold is not exceeded (Figure 1(d) describes a case of perfect anti-correlation yielding a link score of 0). Similarly, the subnetwork Lyn-Evl-Src is not highlighted, because the link scores are below threshold. In turn, the links between Fn1 and Mag/Itgb3 are not picked up by

jActiveModules, because Mag/Itgb3 do not change with sufficient significance (*p*-value); the same holds true for Itga7, Parva and Itgav.

The results of *jActiveModules* for case study 2 (transition from the embryonic stem cell to the epiblast stem cell state) are shown in Figure 13. Interestingly, we discover one small and two very large modules, scoring 3.612, 3.386, 2.768, respectively. The small network is composed of Klf4 (which is also a focus of highlighting by *ExprEssence*, due to its strong down-regulation) and Arid3a, which is the protein linked to Klf4 that changes most significantly. The two large modules have



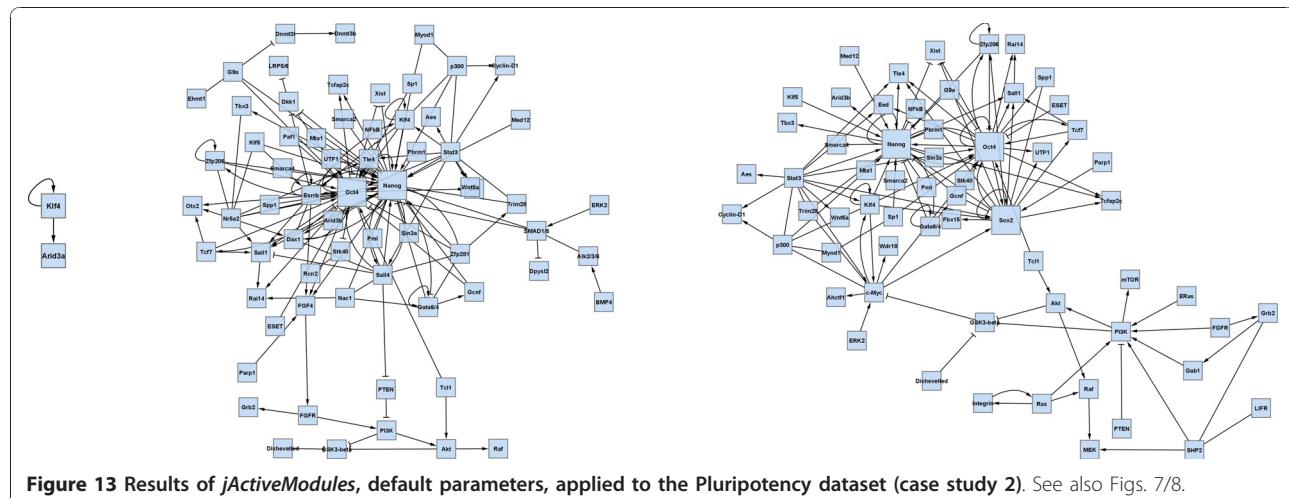


Figure 13 Results of *jActiveModules*, default parameters, applied to the Pluripotency dataset (case study 2). See also Figs. 7/8.

significant overlap with each other, and also with the *ExprEssence*-condensed network (Figure 8), but it can immediately be seen that the latter is more informative than the results of *jActiveModules*, due to link thickness and coloring, allowing easier identification and interpretation of mechanisms behind the observed expression change.

Finally, we put together the active modules found for the Ageing example of case study 3 (Figure 14). As in case study 2, large overlapping networks are obtained. Module scores are (from left to right) 1.899, 1.868, 1.387, 0.786, 0.547. The majority of the modules identified by *jActiveModules* include the link between TP53 and FAS, which is also highlighted by *ExprEssence*. The link between FAS and CASP8 is only considered marginally active (it is found in one module), because CASP8 does not feature a change with a high p-value. The link between CCND3 and CCNE1 is not considered by *jActiveModules*, because change of CCNE1 is not sufficiently significant.

Overall, we observe an overlap of results between our tool and *jActiveModules*. In all case studies, *jActiveModules* did not identify many of the links/effects on genes that we discovered and validated. However, it identified

interesting subnetworks (around Nrp1; Klf4-Arid3a; around TP53) that are plausible and worth investigating. Most importantly, however, *ExprEssence* can distinguish stimulations and inhibitions, and by marking links in thick green or red color, we enable a more informed focus on single links and genes, directly yielding suggestions for experiments that may test the hypotheses we generate.

Conclusions

The most important limitation of our approach is that highlighting is neither necessary nor sufficient for detecting mechanistic change. More specifically, it is quite possible that no change (no startup or shutdown of an interaction, stimulation or inhibition) happens across a highlighted link, or that change happens across a link that is not highlighted. The main reason for this problem is missing accuracy (in terms of sensitivity, i.e. false negatives, and specificity, i.e. false positives) of both network and measured data. In particular, many networks are seriously incomplete, so that we cannot highlight the 'essential' mechanisms simply because there are no links in the network that represent them. For example, the main mechanism may be mediated by

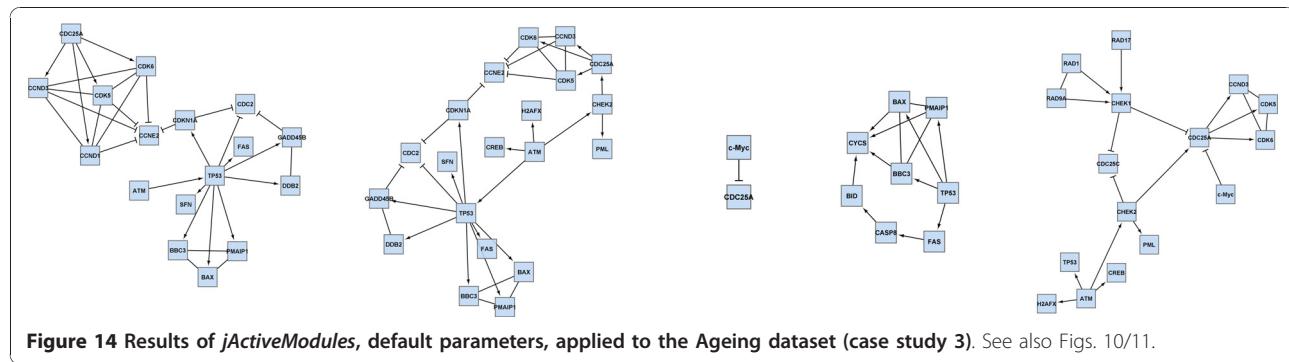


Figure 14 Results of *iActiveModules*, default parameters, applied to the Ageing dataset (case study 3). See also Figs. 10/11.

a regulatory RNA, which may be neither represented in the network, nor in the expression data gained by microarray experiments. Then, we simply cannot discover it, and the mechanisms that are highlighted will be either minor, or simply false positive. To give another example, imagine that the network data do not cover a gene *C* that acts on both *A* and *B*, but it includes the link *A* → *B*. Then, the link may be highlighted even though *C* is acting on both *A* and *B*, and nothing more. Since it to say, hypotheses generated with the help of *ExprEssence* have to be validated experimentally. On the other hand, in a signaling cascade, the mode of change (information flow) may be via phosphorylation events that cannot be measured by expression data. Then, *A* may stimulate *B* via the link *A* → *B*, but no change is detectable in the differential expression data, and no highlighting occurs.

With our approach towards identification of the critical parts of a gene/protein network using differential data, we offer a means to easily become aware of changes in gene/protein relationships that can be observed by contrasting two experimental conditions. We do not only consider physical interactions between proteins but are able to take into account stimulations or inhibitions and treat them accordingly in order to get specific insights into regulatory aspects. *ExprEssence* identifies startup/shutdown along all three different link types (interaction, stimulation, inhibition) in a coherent manner. Our method does not depend on a specific type of network or experimental data as long as edges in the network connect entities influencing each other and the experimental data can be interpreted as measurements proportional to the abundance of the entities.

The statistical basis for comparison of link scores of different edges depends on the input data: if no replicates are available, the plugin works without any measurement of variability, and allows exploration of the dataset. If replicates are given, the plugin uses Welch's formula to improve comparability of link scores by considering the variability of the measurements.

Despite its limitations, we developed a simple, straightforward and easy-to-use tool for hypothesis building, towards a mechanistic interpretation of experiments, seeing the forest for the trees in a large amount of data.

Additional Files

PodocyteCellMatrix.cys, Epiblast.cys, DNA_Damage.cys. Cytoscape Session files containing the original network, expression data and condensed network from case studies 1-3.

Additional material

Additional file 1: PodocyteCellMatrix.cys. To reproduce Figure 5 open the file 'PodocyteCellMatrix.cys' in Cytoscape, select the uncondensed network, start *condense!*, select 'Shaw_Ensembl' (left side) and 'Mundel_Ensembl' (right side) and *No variance data*. After submitting, click on *Organic Layout*.

Additional file 2: Epiblast.cys. To reproduce Figure 8 open the file 'Epiblast.cys' in Cytoscape, select the uncondensed network, start *condense!*, select '12 h PD+LIF Signal' (left side) and '12 h PD+JAK1 Signal' (right side) and *No variance data*. Modify the position of the two sliders to select the 5% and 95% quantiles, and choose *Organic layout*.

Additional file 3: DNA_Damage.cys. To reproduce Figure 11 open the file 'DNA_Damage.cys' in Cytoscape, select the uncondensed network, start *condense!*, select 'HC 1 mean' (left side) and 'HC_4_mean' (right side) and, as variance data, select 'HC_1_var' (left side) and 'HC_4_var' (right side). The number of replicates is 10 (left side) and 16 (right side). After submitting, click on *Organic Layout*.

Acknowledgements

Funding by the DFG SPP 1356, *Pluripotency and Cellular Reprogramming* (FU583/2-1), by the BMBF (01GN0901 & 01GN0805 *Generation of pluri- and multipotent stem cells*) and by European Foundation for the Study of Diabetes (EFSD)/Novo Nordisk is gratefully acknowledged.

Author details

¹Institute for Biostatistics and Informatics in Medicine and Ageing Research, University of Rostock, Ernst-Heydemann-Str. 8, 18057 Rostock, Germany.

²Institute for Anatomy and Cell Biology, Ernst Moritz Arndt University Greifswald, Friedrich-Loeber-Str. 23c, 17487 Greifswald, Germany. ³Department of Mathematics and Informatics, Ernst Moritz Arndt University Greifswald, Jahnstr. 15a, 17487 Greifswald, Germany. ⁴Department of Cell and Developmental Biology, Max Planck Institute for Molecular Biomedicine, Röntgenstrasse 20, 48149 Münster, Germany. ⁵DZNE, German Center for Neurodegenerative Disorders, Gehlsheimer Str. 20, 18147 Rostock, Germany. ⁶Medical Faculty, University of Münster, Domagkstr. 3, 48149 Münster, Germany. ⁷Leibniz Institute for Farm Animal Biology, Research Unit Biomathematics and Bioinformatics, 18196 Dummerstorf, Germany.

Authors' contributions

GW wrote the software, CH, MS, AS, GF contributed to software development and testing, GW, DR, MS, KE, GF contributed to method development, GW, BG, SF, NE, KE, GF contributed to data analysis, BG, SS conducted experiments, GF, GW, KE, SF wrote the paper, GF, KE, HS designed and supervised research. All authors read and approved the final version of the manuscript.

Competing interests

The authors declare that they have no competing interests.

Received: 5 July 2010 Accepted: 30 November 2010

Published: 30 November 2010

References

1. Lu R, Markowitz F, Unwin RD, Leek JT, Airoldi EM, MacArthur BD, Lachmann A, Rozov R, Ma'ayan A, Boyer LA, Troyanskaya OG, Whetton AD, Lemischka IR: *Systems-level dynamic analyses of fate change in murine embryonic stem cells*. *Nature* 2009, **462**(7271):358-362.
2. Laurent L, Wong E, Li G, Huynh T, Tsirigos A, Ong CT, Low HM, Sung KWK, Rigoutsos I, Loring J, Wei CL: *Dynamic changes in the human methylome during differentiation*. *Genome Res* 2010, **20**(3):320-331.
3. Mortazavi A, Williams BA, McCue K, Schaeer L, Wold B: *Mapping and quantifying mammalian transcriptomes by RNA-Seq*. *Nat Methods* 2008, **5**(7):621-628.

4. Sultan M, Schulz MH, Richard H, Magen A, Klingenhöfer A, Scherf M, Seifert M, Borodina T, Soldatov A, Parkhomchuk D, Schmidt D, O'Keeffe S, Haas S, Vingron M, Lehrach H, Yaspo ML: A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* 2008, **321**(5891):956-960.
5. Graumann J, Hubner NC, Kim JB, Ko K, Moser M, Kumar C, Cox J, Schöler H, Mann M: Stable isotope labeling by amino acids in cell culture (SILAC) and proteome quantitation of mouse embryonic stem cells to a depth of 5,111 proteins. *Mol Cell Proteomics* 2008, **7**(4):672-683.
6. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Marshall KA, Phillippe KH, Sherman PM, Muertter RN, Edgar R: NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res* 2009, , **37** Database: D885-D890.
7. Parkinson H, Kapushesky M, Kolesnikov N, Rustici G, Shojatalab M, Abeygunawardena N, Berube H, Dylag M, Emam I, Farne A, Holloway E, Lukk M, Malone J, Mani R, Pilicheva E, Rayner TF, Rezwan F, Sharma A, Williams E, Bradley XZ, Adamusiak T, Brandizi M, Burdett T, Coulson R, Krestyaninova M, Kurnosov P, Maguire E, Neogi SG, Rocca-Serra P, Sansone SA, Sklyar N, Zhao M, Sarkans U, Brazma A: ArrayExpress update—from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res* 2009, , **37** Database: D868-D872.
8. Sridharan R, Tchieu J, Mason MJ, Yachechko R, Kuoy E, Horvath S, Zhou Q, Plath K: Role of the murine reprogramming factors in the induction of pluripotency. *Cell* 2009, **136**(2):364-377.
9. Schulz H, Kolde R, Adler P, Aksoy I, Anastassiadis K, Bader M, Billon N, Boeuf H, Bourillot PY, Buchholz F, Dani C, Doss MX, Forrester L, Gitton M, Henrique D, Hescheler J, Himmelbauer H, Hübner N, Karantzali E, Kretsovali A, Lubitz S, Pradier L, Rai M, Reimand J, Rolletschek A, Sachinidis A, Savatier P, Stewart F, Storm MP, Trouillas M, Vilo J, Welham MJ, Winkler J, Wobus AM, Hatzopoulos AK, in Embryonic Stem Cells Consortium FG: The FunGenES database: a genomics resource for mouse embryonic stem cell differentiation. *PLoS One* 2009, **4**(9):e6804.
10. Cai J, Xie D, Fan Z, Chipperfield H, Marden J, Wong WH, Zhong S: Modeling co-expression across species for complex traits: insights to the difference of human and mouse embryonic stem cells. *PLoS Comput Biol* 2010, **6**(3):e1000707.
11. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, Doers T, Julien P, Roth A, Simonovic M, Bork P, von Mering C: STRING 8-a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res* 2009, , **37** Database: D412-D416.
12. Xu H, Schaniel C, Lemischka IR, Ma'ayan A: Toward a complete *in silico*, multi-layered embryonic stem cell regulatory network. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* 2010.
13. Macarthur BD, Ma'ayan A, Lemischka IR: Systems biology of stem cell fate and cellular reprogramming. *Nat Rev Mol Cell Biol* 2009, **10**(10):672-681.
14. Som A, Harder C, Greber B, Siatkowski M, Paudel Y, Warsow G, Cap C, Scholer H, Fuellen G: The PluriNetWork: An electronic representation of the network underlying pluripotency in mouse, and its applications. *PLoS One* 2010.
15. Ideker T, Ozier O, Schwikowski B, Siegel AF: Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* 2002, **18**(Suppl 1):S233-S240.
16. Minguez P, Dopazo J: Functional genomics and networks: new approaches in the extraction of complex gene modules. *Expert Rev Proteomics* 2010, **7**:55-63.
17. Wu Z, Zhao X, Chen L: Identifying responsive functional modules from protein-protein interaction network. *Mol Cells* 2009, **27**(3):271-277.
18. Yu H, Li YY: Recovering context-specific gene network modules from expression data: A brief review. *Frontiers of Biology in China* 2009, **4**(4):414-418.
19. Tarca AL, Draghici S, Khatri P, Hassan SS, Mittal P, Kim JS, Kim CJ, Kusanovic JP, Romero R: A novel signaling pathway impact analysis. *Bioinformatics* 2009, **25**:75-82.
20. Guo Z, Wang L, Li Y, Gong X, Yao C, Ma W, Wang D, Li Y, Zhu J, Zhang M, Yang D, Rao S, Wang J: Edge-based scoring and searching method for identifying condition-responsive protein-protein interaction sub-network. *Bioinformatics* 2007, **23**(16):2121-2128.
21. Nacu S, Critchley-Thorne R, Lee P, Holmes S: Gene expression network analysis and applications to immunology. *Bioinformatics* 2007, **23**(7):850-858.
22. Thomas R, Gohlke JM, Stopper GF, Parham FM, Portier CJ: Choosing the right path: enhancement of biologically relevant sets of genes or proteins using pathway structure. *Genome Biol* 2009, **10**(4):R44.
23. Ulitsky I, Shamir R: Identifying functional modules using expression profiles and confidence-scored protein interactions. *Bioinformatics* 2009, **25**(9):1158-1164.
24. Qiu YQ, Zhang S, Zhang XS, Chen L: Detecting disease associated modules and prioritizing active genes based on high throughput data. *BMC Bioinformatics* 2010, **11**:26.
25. James K, Wipat A, Hallinan J: Integration of Full-Coverage Probabilistic Functional Networks with Relevance to Specific Biological Processes. *DILS '09: Proceedings of the 6th International Workshop on Data Integration in the Life Sciences, Volume 5647 of Lecture Notes in Computer Science Berlin*, Heidelberg: Springer-Verlag; 2009, 31-46.
26. Parkkinen JA, Kaski S: Searching for functional gene modules with interaction component models. *BMC Syst Biol* 2010, **4**:4.
27. Shiga M, Takigawa I, Mamitsuka H: Annotating gene function by combining expression data with a modular gene network. *Bioinformatics* 2007, **23**(13):i468-1478.
28. Ulitsky I, Shamir R: Identification of functional modules using network topology and high-throughput data. *BMC Syst Biol* 2007, **1**:8.
29. Gu J, Chen Y, Li S, Li Y: Identification of responsive gene modules by network-based gene clustering and extending: application to inflammation and angiogenesis. *BMC Syst Biol* 2010, **4**:47.
30. Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, Christmas R, Avila-Campilo I, Creech M, Gross B, Hanspers K, Isserlin R, Kelley R, Killcoyne S, Lotia S, Maere S, Morris J, Ono K, Pavlovic V, Pico AR, Vailaya A, Wang PL, Adler A, Conklin BR, Hood L, Kuiper M, Sander C, Schmulevich I, Schwikowski B, Warner GJ, Ideker T, Bader GD: Integration of biological networks and gene expression data using Cytoscape. *Nat Protoc* 2007, **2**(10):2366-2382.
31. Kelder T, Pico AR, Hanspers K, van Iersel MP, Evelo C, Conklin BR: Mining biological pathways using WikiPathways web services. *PLoS One* 2009, **4**(7):e6447.
32. Cytoscape Web Service Clients Workflow. [<http://cytoscape.wodaklab.org/wiki/WebServiceWorkflow>].
33. Cerami E, Demir E, Schultz N, Taylor BS, Sander C: Automated network analysis identifies core pathways in glioblastoma. *PLoS One* 2010, **5**(2):e8918.
34. Novére NL: Model storage, exchange and integration. *BMC Neurosci* 2006, **7**(Suppl 1):S11.
35. Novére NL, Hucka M, Mi H, Moodie S, Schreiber F, Sorokin A, Demir E, Wegener K, Aladjem MI, Wimalaratne SM, Bergman FT, Gauges R, Ghazal P, Kawaiji H, Li L, Matsuo Y, Villéger A, Boyd SE, Calzone L, Courtot M, Dogrusoz U, Freeman TC, Funahashi A, Ghosh S, Jouraku A, Kim S, Kolpakov F, Luna A, Sahle S, Schmidt E, Watterson S, Wu G, Goryanin I, Kell DB, Sander C, Sauro H, Snoep JL, Kohn K, Kitano H: The Systems Biology Graphical Notation. *Nat Biotechnol* 2009, **27**(8):735-741.
36. Cytoscape 2.7 Manual. [http://www.cytoscape.org/manual/Cytoscape2_7Manual.pdf].
37. Welch BL: The generalisation of student's problems when several different population variances are involved. *Biometrika* 1947, **34**(1-2):28-35.
38. Silva J, Smith A: Capturing pluripotency. *Cell* 2008, **132**(4):532-536.
39. Fuellen G: Evolution of gene regulation-on the road towards computational inferences. *Brief Bioinform* 2010.
40. Wiggins RC: The spectrum of podocytopathies: a unifying view of glomerular diseases. *Kidney Int* 2007, **71**(12):1205-1214.
41. Akilesh S, Huber TB, Wu H, Wang G, Hartleben B, Kopp JB, Miner JH, Roopenian DC, Unanue ER, Shaw AS: Podocytes use FcRN to clear IgG from the glomerular basement membrane. *Proc Natl Acad Sci USA* 2008, **105**(3):967-972.
42. El-Aouni C, Herbach N, Blattner SM, Henger A, Rastaldi MP, Jarad G, Miner JH, Moeller MJ, St-Arnaud R, Dedhar S, Holzman LB, Wanke R, Kretzler M: Podocyte-specific deletion of integrin-linked kinase results in severe glomerular basement membrane alterations and progressive glomerulosclerosis. *J Am Soc Nephrol* 2006, **17**(5):1334-1344.
43. Yang Y, Guo L, Blattner SM, Mundel P, Kretzler M, Wu C: Formation and phosphorylation of the PINCH-1-integrin linked kinase-alpha-parvin complex are important for regulation of renal glomerular podocyte

- adhesion, architecture, and survival. *J Am Soc Nephrol* 2005, 16(7):1966-1976.
44. Bondeva T, Rüster C, Franke S, Hammerschmid E, Klagsbrun M, Cohen CD, Wolf G: Advanced glycation end-products suppress neuropilin-1 expression in podocytes. *Kidney Int* 2009, 75(6):605-616.
 45. Reidy KJ, Villegas G, Teichman J, Veron D, Shen W, Jimenez J, Thomas D, Tufro A: Semaphorin3a regulates endothelial cell number and podocyte differentiation during glomerular development. *Development* 2009, 136(23):3979-3989.
 46. Guan F, Villegas G, Teichman J, Mundel P, Tufro A: Autocrine class 3 semaphorin system regulates slit diaphragm proteins and podocyte survival. *Kidney Int* 2006, 69(9):1564-1569.
 47. Plaisier E, Mougenot B, Verpont MC, Jouanneau C, Archelos JJ, Martini R, Kerjaschki D, Ronco P: Glomerular permeability is altered by loss of P0, a myelin protein expressed in glomerular epithelial cells. *J Am Soc Nephrol* 2005, 16(11):3350-3356.
 48. Lau F, Ahfeldt T, Osaifune K, Akutsu H, Cowan CA: Induced pluripotent stem (iPS) cells: an up-to-the-minute review. *F1000 Biology Reports* 2009, 2009, 1:84.
 49. Do JT, Schöler HR: Regulatory circuits underlying pluripotency and reprogramming. *Trends Pharmacol Sci* 2009, 30(6):296-302.
 50. Zhao XY, Li W, Lv Z, Liu L, Tong M, Hai T, Hao J, long Guo C, wen Ma Q, Wang L, Zeng F, Zhou Q: iPS cells produce viable mice through tetraploid complementation. *Nature* 2009, 461(7260):86-90.
 51. Hanna J, Wernig M, Markoulaki S, Sun CW, Meissner A, Cassady JP, Beard C, Brambrink T, Wu LC, Townes TM, Jaenisch R: Treatment of sickle cell anemia mouse model with iPS cells generated from autologous skin. *Science* 2007, 318(5858):1920-1923.
 52. Greber B, Wu G, Bernemann C, Joo JY, Han DW, Ko K, Tapia N, Sabour D, Sternbeckert J, Tesar P, Schöler HR: Conserved and divergent roles of FGF signaling in mouse epiblast stem cells and human embryonic stem cells. *Cell Stem Cell* 2010, 6(3):215-226.
 53. Jiang J, Chan YS, Loh YH, Cai J, Tong GQ, Lim CA, Robson P, Zhong S, Ng HH: A core Klf circuitry regulates self-renewal of embryonic stem cells. *Nat Cell Biol* 2008, 10(3):353-360.
 54. Zhou Q, Chipperfield H, Melton DA, Wong WH: A gene regulatory network in mouse embryonic stem cells. *Proc Natl Acad Sci USA* 2007, 104(42):16438-16443.
 55. Feng B, Jiang J, Kraus P, Ng JH, Heng JCD, Chan YS, Yaw LP, Zhang W, Loh YH, Han J, Vega VB, Cacheux-Rataboul V, Lim B, Lufkin T, Ng HH: Reprogramming of fibroblasts into induced pluripotent stem cells with orphan nuclear receptor Esrrb. *Nat Cell Biol* 2009, 11(2):197-203.
 56. van den Berg DLC, Zhang W, Yates A, Engelen E, Takacs K, Bezstarost K, Demmers J, Chambers I, Poot RA: Estrogen-related receptor beta interacts with Oct4 to positively regulate Nanog gene expression. *Mol Cell Biol* 2008, 28(19):5986-5995.
 57. Kidder BL, Yang J, Palmer S: Stat3 and c-Myc genome-wide promoter occupancy in embryonic stem cells. *PLoS One* 2008, 3(12):e3932.
 58. Satou A, Taira T, Iguchi-Ariga SM, Ariga H: A novel transrepression pathway of c-Myc. Recruitment of a transcriptional corepressor complex to c-Myc by MM-1, a c-Myc-binding protein. *J Biol Chem* 2001, 276(49):46562-46567.
 59. Berchtold NC, Cribbs DH, Coleman PD, Rogers J, Head E, Kim R, Beach T, Miller C, Troncoso J, Trojanowski JQ, Zielke HR, Cotman CW: Gene expression changes in the course of normal brain aging are sexually dimorphic. *Proc Natl Acad Sci USA* 2008, 105(40):15605-15610.
 60. Matheu A, Maraver A, Serrano M: The Arf/p53 pathway in cancer and aging. *Cancer Res* 2008, 68(15):6031-6034.
 61. Geng Y, Whoriskey W, Park MY, Bronson RT, Medema RH, Li T, Weinberg RA, Sicinski P: Rescue of cyclin D1 deficiency by knockin cyclin E. *Cell* 1999, 97(6):767-777.
 62. Fitch ME, Cross IV, Ford JM: p53 responsive nucleotide excision repair gene products p48 and XPC, but not p53, localize to sites of UV-irradiation-induced DNA damage, *in vivo*. *Carcinogenesis* 2003, 24(5):843-850.
 63. Schiwek D, Endlich N, Holzman L, Holthöfer H, Kriz W, Endlich K: Stable expression of nephrin and localization to cell-cell contacts in novel murine podocyte cell lines. *Kidney Int* 2004, 66:91-101.

doi:10.1186/1752-0509-4-164

Cite this article as: Warsow et al.: ExprEssence - Revealing the essence of differential experimental data in the context of an interaction/regulation net-work. *BMC Systems Biology* 2010 4:164.

Submit your next manuscript to BioMed Central and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit



PodNet, a protein–protein interaction network of the podocyte

Gregor Warsow^{1,2,3}, Nicole Endlich¹, Eric Schordan¹, Sandra Schordan¹, Ravi K. Chilukoti⁴, Georg Homuth⁴, Marcus J. Moeller⁵, Georg Fuellen² and Karlhans Endlich¹

¹Department of Anatomy and Cell Biology, University Medicine Greifswald, Greifswald, Germany; ²Institute for Biostatistics and Informatics in Medicine and Ageing Research, University of Rostock, Rostock, Germany; ³Department of Mathematics and Informatics, Ernst Moritz Arndt University Greifswald, Greifswald, Germany; ⁴Interfaculty Institute for Genetics and Functional Genomics, Department of Functional Genomics, Ernst Moritz Arndt University Greifswald, Greifswald, Germany and ⁵Division of Nephrology and Immunology, University Hospital of the Aachen University of Technology (RWTH), Aachen, Germany

Interactions between proteins crucially determine cellular structure and function. Differential analysis of the interactome may help elucidate molecular mechanisms during disease development; however, this analysis necessitates mapping of expression data on protein–protein interaction networks. These networks do not exist for the podocyte; therefore, we built PodNet, a literature-based mouse podocyte network in Cytoscape format. Using database protein–protein interactions, we expanded PodNet to XPodNet with enhanced connectivity. In order to test the performance of XPodNet in differential interactome analysis, we examined podocyte developmental differentiation and the effect of cell culture. Transcriptomes of podocytes in 10 different states were mapped on XPodNet and analyzed with the Cytoscape plugin ExprEssence, based on the law of mass action. Interactions between slit diaphragm proteins are most significantly upregulated during podocyte development and most significantly downregulated in culture. On the other hand, our analysis revealed that interactions lost during podocyte differentiation are not regained in culture, suggesting a loss rather than a reversal of differentiation for podocytes in culture. Thus, we have developed PodNet as a valuable tool for differential interactome analysis in podocytes, and we have identified established and unexplored regulated interactions in developing and cultured podocytes.

Kidney International advance online publication, 3 April 2013;
doi:10.1038/ki.2013.64

KEYWORDS: microarray analysis; podocyte; protein interaction; transcriptional profiling

Podocytes have a crucial role for the glomerular filtration barrier¹ as well as in the development of chronic kidney disease.² One current aim of nephrology research is therefore to define the molecular mechanisms that underlie podocyte function and to understand the molecular mechanisms that are deranged in glomerular disease. The constant improvement of ‘omics’ technologies and developments in systems biology importantly contribute to this endeavor.^{3–5} Among the many available bioinformatics tools for the analysis of expression data, methods that allow for the analysis of differential changes of the interactome are especially well suited to advance our biological understanding. This is because cell function largely depends on protein–protein interactions (PPIs). To perform differential interactome analyses on expression data, algorithms have been developed by others⁶ and by us.⁷ These algorithms have been successfully applied to PPI networks, for example, to identify molecular mechanisms of pluripotency.⁸

In 2008, GlomNet was introduced as the first approach to analyze glomerular ‘omics’ data in a PPI network.^{9,10} However, GlomNet does not discriminate between glomerular cell types, that is, mesangial cells, endothelial cells, and podocytes. Thanks to the recently developed isolation techniques for podocytes from wild-type^{11,12} or transgenic mice,^{13–15} and thanks to existing mouse podocyte cell lines,^{16,17} an increasing number of expression data for mouse podocytes under various conditions have become available. The goal of our study was therefore to build a podocyte-specific PPI network allowing for the podocyte-specific analysis of expression data. To date, only a small fraction of all existing PPIs has been determined and curated for public databases,¹⁸ and especially PPIs relevant for highly specialized cell types, such as the podocyte, cannot be assembled easily. Therefore, we decided to build PodNet as an expert curated podocyte PPI network based on the findings reported in the literature. We show that PodNet, besides providing functional insight into the podocyte PPI network, identifies relevant transitions in podocyte development and culture by differential interactome analysis using podocyte expression data.

Correspondence: Karlhans Endlich, Institut für Anatomie und Zellbiologie, Universitätsmedizin Greifswald, Friedrich-Loeffler-Strasse 23c, Greifswald, D-17487, Germany. E-mail karlhans.endlich@uni-greifswald.de

Received 15 August 2012; revised 10 December 2012; accepted 13 December 2012

RESULTS

Generation and characterization of PPI networks

PodNet, our expert curated PPI network of the podocyte, consists of 315 genes/proteins (nodes) and 223 interactions (edges) extracted exclusively from the podocyte literature (see Supplementary Methods and Supplementary Data Files S1 and S2 online). Approximately 40% of the nodes and 80% of the edges form the largest connected component with a mean number of 2.7 interactions per node (Table 1). Of the nodes, 46% are not connected to any other node. In order to incorporate them into the network, we added further interactions from the STRING PPI database¹⁹ in an unbiased manner to build XPodNet. Unlike for PodNet, the added interactions are not necessarily known to be relevant for podocytes. By this approach, besides incorporating isolated nodes into the interconnected parts of the network, we further wanted to identify podocyte-relevant PPIs that have not yet been described for the podocyte, leading to novel hypotheses.

XPodNet (Figure 1) is made up of 839 nodes and 1048 edges with considerably improved connectivity: ~80% of the nodes and 95% of the edges are part of the largest connected component (Table 1). Of the 146 isolated nodes in PodNet, 44 could be integrated into one of the three largest connected components of XPodNet, whereas 84 nodes remained isolated. Hence, STRING interactions integrated approximately one-third of the single PodNet nodes into a PPI network. Of the 223 interactions in PodNet, we found only 23 in STRING, emphasizing the need for setting up expert-curated, cell type-specific networks as a solid basis for interaction-focused network investigations.

To perform comparative analyses, we generated GlobalNet as an unbiased network not focused on the podocyte. GlobalNet was based on all mouse interactions in STRING above a certain confidence score (see Supplementary Methods online). It contains 3607 nodes and 5657 edges, and its network characteristics, such as the relative size of the largest connected component, number of hubs, and the mean number of edges per node, are comparable to those of XPodNet (Table 1). It has to be kept in mind that GlobalNet contains only 10% of the PPIs in PodNet, as GlobalNet was built exclusively with database PPIs. All three PPI networks were analyzed for Gene Ontology term enrichments (Table 2; Supplementary Table S1 online). As can be seen in Table 2, Gene Ontology terms such as cytoskeletal protein and actin binding are highly enriched in PodNet and XPodNet, most likely reflecting a focus of podocyte research. All PPI networks are provided in Cytoscape²⁰ format and can be downloaded from www.PodNet.de.

Podocyte core network

During the past decade, many genes were identified that are essential for podocyte function. These genes were discovered by analyzing human disease, by gene knockout in mice, or by gene knockdown in zebrafish. We wondered whether these genes might be interconnected in a podocyte core network,

Table 1 | Characteristic numbers of PodNet, expanded PodNet (XPodNet), and GlobalNet

	PodNet	XPodNet	GlobalNet
Nodes (genes)	315	839	3607
Edges (protein-protein interactions)	223	1048	5657
Mean number of edges per node	1.4	2.5	3.1
Number of hubs, that is, nodes with ≥ 10 edges (fraction of network nodes)	5 (1.6%)	42 (5.0%)	196 (5.4%)
<i>Largest connected component</i>			
Nodes (fraction of network nodes)	130 (41%)	682 (82%)	2872 (80%)
Edges (fraction of network edges)	183 (82%)	994 (95%)	5085 (90%)

forming a functional module that is critical for podocyte function. To this end, we extracted the component of XPodNet that contained the largest number of essential podocyte genes held together by the lowest possible number of neighboring genes. The extracted podocyte core network consists of 34 nodes, of which 28 are essential nodes (Figure 2). Thus, almost half of the currently known essential podocyte genes are interconnected in a small network. The podocyte core network can be subdivided into three modules: an actin module, an adhesion module, and a slit diaphragm/signaling module. CD2AP and α -actinin-4 (Actn4) connect two and three modules, respectively. All nodes and edges in the podocyte core network are contained in PodNet, underlining the relevance of our expert-curated approach.

Most abundant interactions in podocytes

We mapped two sets of gene expression data of differentiated adult *in vivo* podocytes (Adult S-Pod and Adult G-Pod, see Table 3) on both PodNet and GlobalNet. By using the gene expression values, the abundance of the individual PPIs in both networks was estimated (see Supplementary Methods online). The most abundant interactions, that is, protein complexes with the highest concentrations, were extracted for both sets of gene expression data and intersected for PodNet and GlobalNet, respectively. Such an intersection yields the interactions being most abundant for both data sets, and thus increases the reliability of reflecting the *in vivo* situation.

As can be seen in Figure 3, the most abundant interactions in XPodNet include the interaction between the two main slit diaphragm proteins Nphs1 (nephrin) and Nphs2 (podocin). Nephrin is connected via α -II spectrin (Spna2) and Actn4 to the actin cytoskeleton. Given the fact that podocytes possess a huge number of the actin-based foot processes, it is comprehensible that structural PPIs of the actin cytoskeleton form the largest group among the most abundant interactions in PodNet. The most abundant interactions in GlobalNet are related to more general cellular functions like ribosomal, mitochondrial, and proteasomal functions. As claudin 5 (Cldn5) has only recently been reported to be expressed in podocytes,²¹ it is of interest that the PPIs between Actn4, Zona Occludens 1 (ZO-1; Tjp1), and Cldn5 were detected among the most abundant interactions in both XPodNet and GlobalNet.

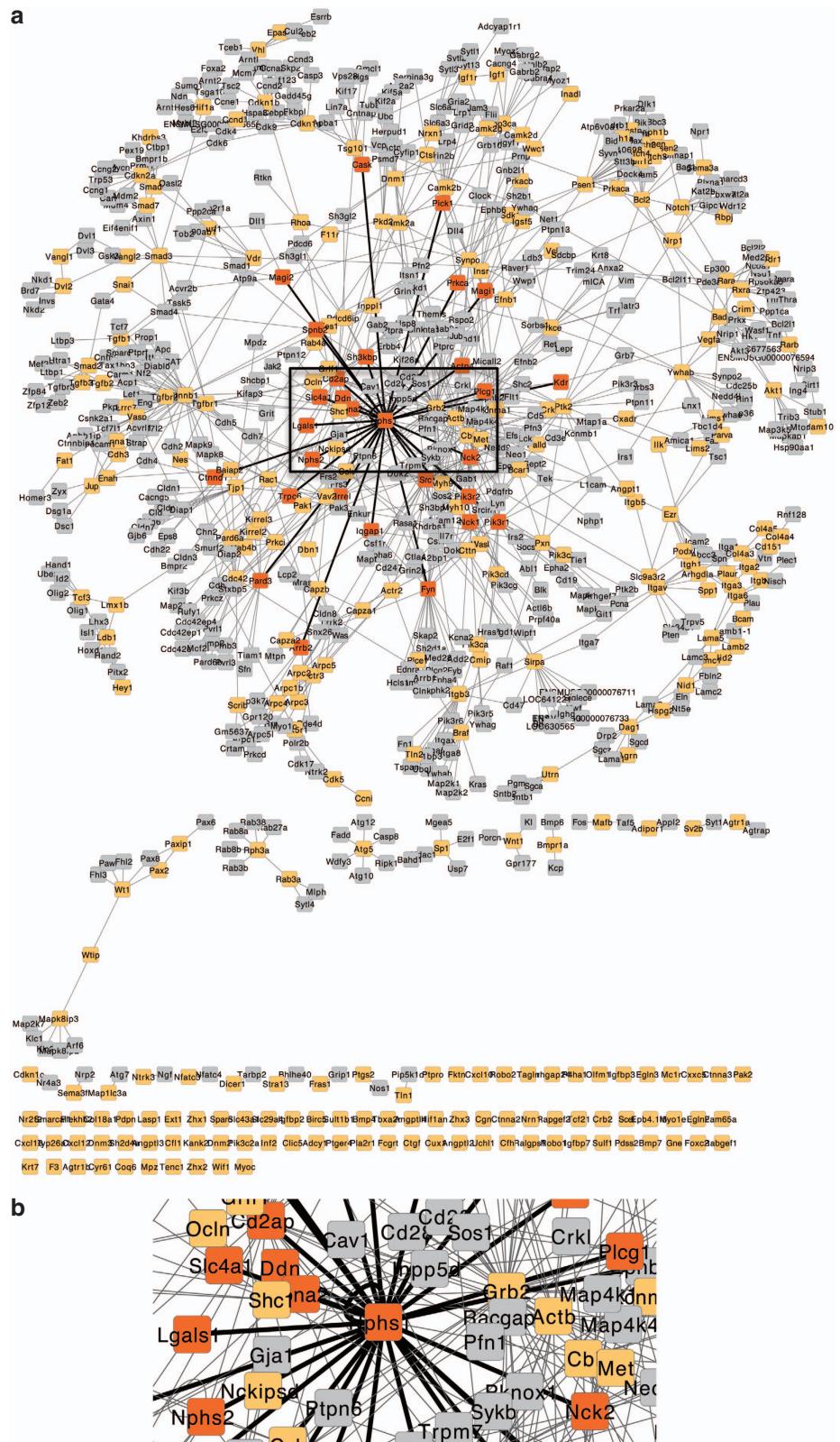


Figure 1 | XPodNet. Genes are represented by nodes and interactions of the corresponding proteins by edges. Nphs1 and its interaction partners are colored in dark orange. Nodes of PodNet are colored in pale orange and nodes from STRING are gray. (a) Complete network and (b) zoom-in centered to Nphs1.

Table 2 | Overrepresented Gene Ontology (GO) terms (molecular function) in PodNet, expanded PodNet (XPodNet), and GlobalNet

Molecular function GO term	PodNet			XPoNet		GlobalNet	
	Nodes (%)	P-value	Enrichment (fold)	Nodes (%)	Enrichment (fold)	Nodes (%)	Enrichment (fold)
Protein binding	77.9	0.0E + 00	4.2	81.4	4.4	68.3	3.7
Binding	90.6	3.8E - 89	2.5	92.8	2.6	87.4	2.5
Protein complex binding	12.7	8.1E - 33	15.8	9.2	11.4	3.9	4.9
Protein domain-specific binding	12.1	3.8E - 26	11.6	12.5	11.9	5.6	5.3
Cytoskeletal protein binding	13.4	7.4E - 26	9.5	10.2	7.3	5.4	3.8
Enzyme binding	13.0	1.1E - 25	9.8	12.3	9.3	6.0	4.4
Actin binding	11.1	3.0E - 24	11.7	7.6	8.0	3.3	3.5
Kinase activity	16.6	8.0E - 24	6.2	15.8	5.9	8.1	3.0
Kinase binding	8.5	1.2E - 21	15.4	7.0	12.7	2.7	4.9
Transferase activity transferring phosphorus-containing groups	16.6	4.4E - 21	5.4	16.1	5.2	9.7	3.1
Protein kinase binding	7.8	2.7E - 20	15.8	6.5	13.1	2.6	5.1
Phosphotransferase activity; alcohol group as acceptor	13.4	2.2E - 18	5.9	13.1	5.8	6.9	3.1
Receptor binding	14.0	3.7E - 18	5.5	12.7	5.0	7.7	3.0
Protein kinase activity	11.4	7.4E - 16	6.0	11.8	6.2	6.5	3.4
Integrin binding	3.9	1.1E - 13	29.8	2.4	18.0	0.6	4.8
Transferase activity	18.2	1.5E - 13	3.2	18.4	3.3	12.2	2.2
Protein heterodimerization activity	6.5	4.5E - 13	10.0	5.2	8.0	3.2	4.8
Protein dimerization activity	8.8	3.8E - 12	6.0	8.6	5.8	6.1	4.0
SH3 domain binding	4.6	1.9E - 11	14.7	3.9	12.4	1.5	4.7
Ribonucleotide binding	16.9	1.2E - 10	2.9	18.1	3.1	15.1	2.6
Purine ribonucleotide binding	16.9	1.2E - 10	2.9	18.1	3.1	15.1	2.6
Cell adhesion molecule binding	2.9	1.3E - 10	31.4	1.7	18.7	0.6	5.9
Purine nucleotide binding	17.3	1.6E - 10	2.8	18.3	3.0	15.4	2.5
Catalytic activity	32.2	3.9E - 10	1.9	32.1	1.9	30.0	1.8
Cadherin binding	2.3	4.5E - 10	50.8	1.0	22.2	0.2	5.4
Extracellular matrix binding	2.9	4.7E - 10	27.4	1.6	15.1	0.4	3.6

The P-value for overrepresentation of the GO term is false discovery rate (FDR) corrected (see Supplementary Table S1 online for the complete list of P-values). The fold values describe the factor of enrichment of the GO term in the corresponding network with respect to GO annotation of all genes.

Analysis of differentially regulated interactions

A total of 33 transcriptomes, including replicates, of 12 different entities were used for analysis (Table 3). Of the 33 transcriptomes, 30 are publicly available and 3 were generated specifically for this study. The Affymetrix 430 2.0 and 1.0 ST mouse arrays were used in 16 and 17 cases, respectively. First, transcriptomes were subjected to principal component analysis (PCA; Figure 4). For both array platforms, the first principal component separates *in vivo* podocytes from podocytes in culture. Regarding the *in vivo* podocytes, the second principal component separates developing podocytes from mature podocytes.

Subsequent to PCA, we performed 11 pairwise differential analyses of PPIs with ExprEssence, the formula of which is based on the law of mass action⁷ (see Supplementary Methods online). Six of the differential analyses describe podocyte development, three describe cultured versus *in vivo* podocytes, one describes differentiated versus undifferentiated cultured podocytes, and one describes whole glomeruli versus podocytes (Table 3; Supplementary Table S2 and Supplementary Figures S2 and S3 online). For the first two groups, the ExprEssence link scores have been ranked and the ranks have been averaged. Table 4 shows 10 most differentially upregulated and downregulated interactions for the development and cell culture comparisons, respectively. This

set of interactions has been curated manually from a larger set of differentially regulated interactions (Supplementary Table S3 online) to eliminate erroneous entries of STRING and to eliminate interactions of genes that are probably not expressed in podocytes.

Most differentially regulated interactions in podocyte development

The largest group (60%) of most upregulated PPIs during development consists of the interactions between slit diaphragm proteins like Nphs1 (nephrin) and Nphs2 (podocin) (Table 4). The strong upregulation of slit diaphragm PPIs during development is observed in both XPoNet and GlobalNet. The differential analysis between adult and embryonic day 13.5 (E13.5) podocytes further illustrates this finding (Figure 5). The apical podocalyxin (Podxl)/Nherf2 (Slc9a3r2)/ezrin (Ezr) complex, which is essential for podocyte function,²² is also strongly upregulated during podocyte development (Table 4). Finally, the unbiased GlobalNet reveals that the differentiation of podocytes is associated with a strong upregulation of Nbr1/Sqstm1/Cyld interactions (Table 4). Nbr1 has been implicated in targeting ubiquitinated proteins to the autophagosome,²³ an organelle that is vital for long-term podocyte function.²⁴

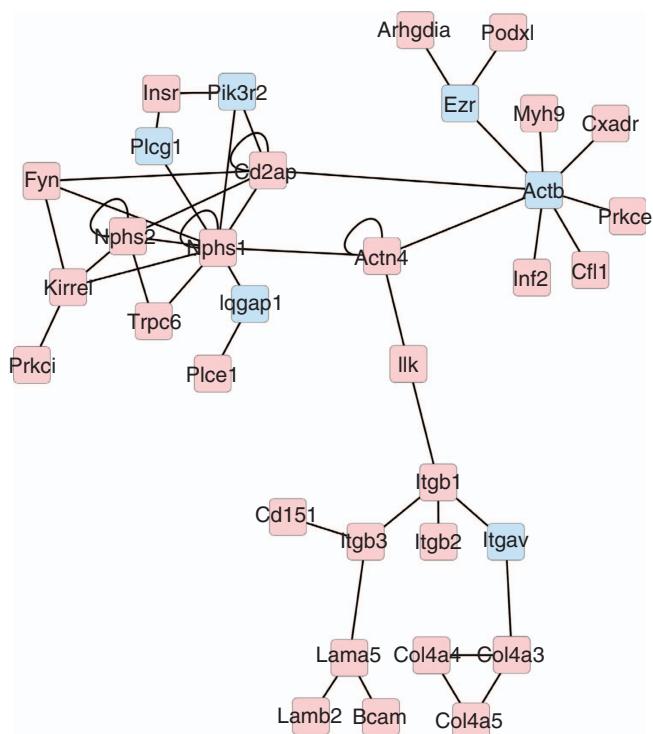


Figure 2 | Podocyte core network. Largest connected component of XPodNet after removal of all nodes not known to be essential (essential genes were determined by association of human gene mutation, gene knockout in mice, or gene knockdown in zebrafish with a podocyte phenotype such as foot process effacement), which can be omitted without breaking the component apart. Red nodes represent essential genes, and other nodes are colored in blue.

Among the interactions that are downregulated during podocyte development (Table 4), one finds the interaction between Grip1 and Fras1, which is an extracellular matrix protein required for kidney as well as glomerular development.²⁵ Several interactions of proteins involved in cell-cell contacts are downregulated during podocyte development: the junctional adhesion molecules 3 (Jam3) and 4 (Jam4 or Igsf5) and the gap junction-forming proteins connexin 31 (Gjb3) and 43 (Gja1). Interestingly, if podocytes are injured, these proteins are upregulated and become involved in junction formation, replacing slit diaphragms.^{26,27} Similarly, Notch1, of which several interactions are downregulated during podocyte development (Table 4), becomes reactivated in injured podocytes of diabetic patients.²⁸ As a novel finding of potential relevance, our analysis identifies interactions of Efs (embryonal Fyn-associated substrate) to be downregulated during podocyte development. The docking protein Efs may orchestrate signaling of cell-matrix adhesions in immature podocytes by binding the tyrosine kinases Src and Ptk2 (Fak (focal adhesion kinase)).²⁹

Most differentially regulated interactions in podocyte culture

In addition to the six differential analyses of podocyte development, we performed three differential analyses to assess the effect of putting podocytes into cell culture. The

most differentially regulated PPIs are compiled in Table 4, again averaging over all differential analyses. The differential analysis for Prim M-Pod versus Adult G-Pod is shown in Figure 6 as an illustrating example. The largest group (60%) of most downregulated PPIs in culture consists of the interactions between the slit diaphragm proteins. The loss of slit diaphragm PPIs in culture is detected in both XPodNet and GlobalNet (Table 4 and Figure 6). The interactions between ZO-1 (Tjp1) and Cldn5 and between Pak1 and the Rho-GTPases Rac1 and Cdc42, which belong to the most abundant PPIs in fully differentiated podocytes (Figure 3), are consistently downregulated in cultured podocytes.

Among the interactions that are most upregulated in cultured podocytes, interactions between the cell cycle proteins (45%) are prevailing, a finding that is again observed in both XPodNet and GlobalNet (Table 4 and Figure 6). The cell cycle-promoting interactions between cyclin-dependent kinase 1 (Cdk1) and cyclin A2 (Ccna2) or B2 (Ccnb2) are upregulated. However, the cell cycle in differentiated cultured podocytes is arrested by the even stronger upregulation of the interactions between Cdkks 2, 4, and 6 and the Cdk inhibitors 1a (Cdkn1a or p21^{Cip1}) or 2a (Cdkn2a or p16^{Ink4a}) (data not shown). Cdkn1a (p21^{Cip1}) is known to be upregulated in podocytes under many conditions of glomerular disease.³⁰ Besides the upregulation of PPIs among cell cycle proteins in cultured podocytes, interactions of the extracellular matrix proteins fibronectin 1 (Fn1) and the Pax-interacting protein 1 (Paxip1) are upregulated in cultured podocytes. Of note, expression of Pax8 has been observed in podocytes in a TGF-β1 transgenic mouse model of global glomerulosclerosis.³¹

Relationship between podocyte development and culture

As highlighted in Table 4, podocytes gain during development the highly podocyte-specific PPIs of the slit diaphragm, which are lost in cultured podocytes. However, interactions that are lost during podocyte development from renal vesicle to fully differentiated adult podocytes seem not to be regained when podocytes are put into culture. Hence, our differential PPI analysis does not support the notion that podocytes in cell culture dedifferentiate into an earlier developmental stage.

For a more comprehensive picture, we compared the mean changes in PPI abundance during podocyte development with those during dedifferentiation in culture using all interactions contained in XPodNet and GlobalNet. The scatter plot in Figure 7 further supports the results obtained by the most differentially regulated PPIs: several interactions that are upregulated during podocyte development are downregulated in culture. However, on the whole, changes in PPI abundance are merely weakly correlated, especially for the unbiased PPIs contained in GlobalNet ($r^2 = 0.13$). Thus, the interactome changes of cultured podocytes match the interactome changes of podocytes in early developmental stages to a minor extent only.

Table 3 | GEO data sets used for analysis of differential gene expression

	GEO accession	Array	Comparisons						
Freshly isolated Pod (Adult S-Pod) ¹¹	GSM253173	430 2.0	a						
Differentiated cultured Pod (Diff S-Pod) ¹¹	GSM253174	430 2.0	a	b					
Undifferentiated cultured Pod (Undiff S-Pod) ¹¹	GSM588087	430 2.0		b					
Differentiated cultured Pod (Diff E-Pod) ¹⁷	GSM1018531-33	1.0 ST			d				
Primary cultured Pod (Prim M-Pod) ¹⁴	GSM833242,44	1.0 ST		c					
Freshly isolated Pod (Adult G-Pod) ¹³	GSM429038-40	1.0 ST		c	d	e		g	h
Freshly isolated E13.5 Pod (E13.5 G-Pod) ¹³	GSM429041-43	1.0 ST				e	f		
Freshly isolated E15.5 Pod (E15.5 G-Pod) ¹³	GSM429046-48	1.0 ST					f	g	
E12.5 renal vesicles (Ren Ves) ⁴²	GSM144585-89	430 2.0	i		k				
E15.5 renal S-shaped bodies (S-Shaped) ⁴²	GSM144602-05	430 2.0	i	j					
E15.5 capillary loop/maturing renal corpuscle (Cap Loop) ⁴²	GSM144590-93	430 2.0		j	k				
Adult glomeruli (Glom) ⁴²	GSM429035-37	1.0 ST							h

Abbreviations: E13.5, embryonic day 13.5; GEO, Gene Expression Omnibus; Pod, podocyte.

A total of 33 transcriptomes of 12 different entities were used for analysis of differential gene expression. Transcriptomes were determined on two platforms: Affymetrix Mouse Genome 430 2.0 Arrays (430 2.0) or Affymetrix Mouse Gene 1.0 ST Arrays (1.0 ST). Eleven pairwise comparisons (a-k) were performed as indicated in the table. Pairwise comparisons colored in red (a, c, and d) were combined to study cultured versus *in vivo* podocytes. Pairwise comparisons colored in green (e, f, g, i, j, and k) were combined to study podocytes during development.

Differentiation of podocytes in culture

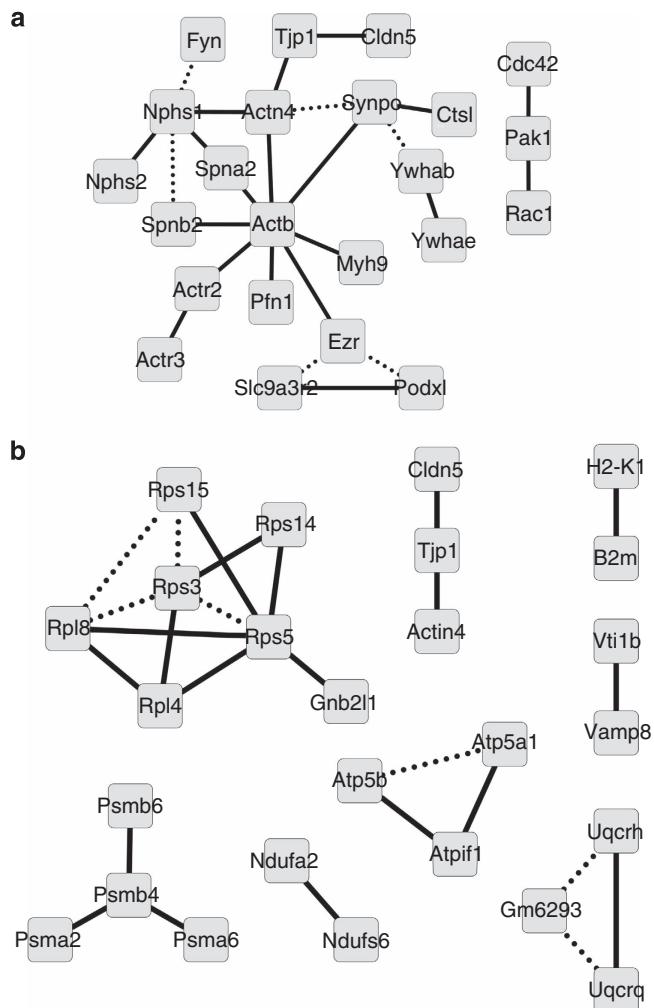
In 1997, Mundel *et al.*¹⁶ introduced the concept of conditional immortalization to podocyte cell culture. Thereafter, this concept has also been used by others for the generation of podocyte cell lines.^{11,17,32} Conditionally immortalized podocytes proliferate under the thermo-sensitive SV40 large T antigen that is active at 33 °C and whose expression can be stimulated by interferon-γ. If cells are shifted to 37 or 38 °C in the absence of interferon-γ, they stop proliferating. This process has frequently been termed ‘differentiation.’ Similarly, podocytes have been called ‘undifferentiated’ and ‘differentiated,’ respectively, under the two culture conditions. However, differentiation of cultured podocytes has never been examined in relation to differentiation of podocytes during development.

When comparing Diff S-Pod with Undiff S-Pod (Supplementary Figures S2 and S3 online), among the most differentially regulated interactions in XPodNet, we found two interactions between the proteins Oasl2-Smurf1 (140-fold downregulation) and Prmt1-Smad6 (100-fold downregulation) that are thought to mediate resistance to virus infection and to repress Stat1 transcriptional activity in the

late phase of interferon-γ signaling, respectively. Among the most differentially regulated interactions in GlobalNet, we found five interactions between the proteins H2-DMa, H2-Eb1, H2-Aa, Rmcs2, and Cd74 of the major histocompatibility complex II that were downregulated 730- to 47 000-fold (Supplementary Figure S3B online). Interactions that were most differentially regulated during podocyte development, for example, slit diaphragm interactions, were only affected to a negligible degree. Thus, ‘differentiation’ of conditionally immortalized podocytes is a means to achieve quiescence after induction of proliferation in culture, but it does not reflect the differentiation process of podocytes during development.

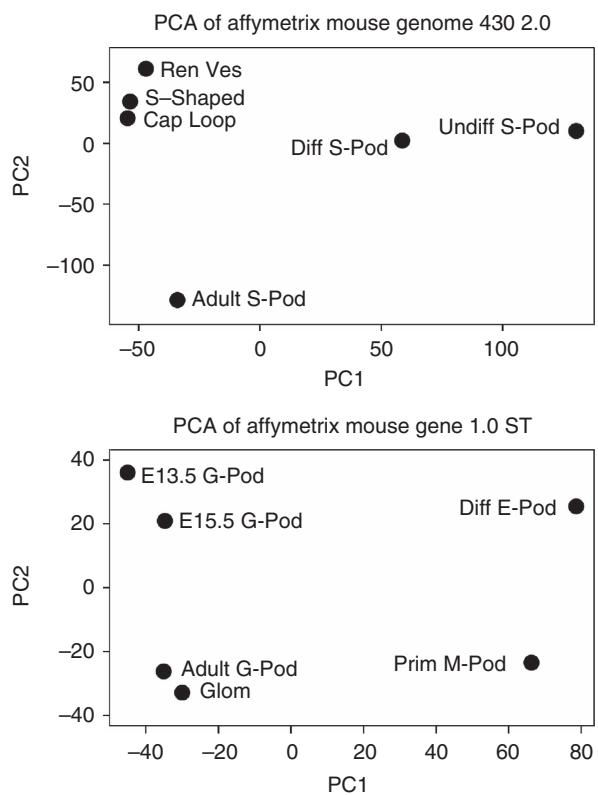
Analysis of differential interactions in whole glomeruli versus podocytes

As the glomerulus is formed by three different cell types, all three cell types contribute to the glomerular transcriptome. If we map the glomerular transcriptome on XPodNet or GlobalNet, many spurious interactions will be detected; for example, the ‘interaction’ of a protein expressed in the glomerular endothelium with one expressed in podocytes.

**Figure 3 | Networks of most abundant interactions.**

(a) Intersection of the 50 most abundant interactions for Adult S-Pod and Adult G-Pod in XPodNet (Supplementary Figure S1 online). Pod, podocyte. The resulting network contains of 19 most abundant interactions (solid lines). Dotted lines represent interactions that are not most abundant in at least one of both intersected networks (gray line in Supplementary Figure S1 online). (b) The same procedure has been applied to the 75 most abundant interactions in GlobalNet, to also give 19 common most abundant interactions for both conditions.

Nevertheless, if we perform a differential analysis of the interactions in whole glomeruli versus podocytes (Glom vs. Adult G-Pod), relevant information can be extracted (Supplementary Figures S2 and S3 online). Because of cell size and number, let us assume that podocyte transcripts contribute about half to the glomerular transcriptome. Under this assumption, we would expect that podocyte-specific proteins are ‘downregulated’ by a factor of two in the glomerular versus the podocyte transcriptome. Indeed, the estimated abundance of interaction between the podocyte-specific proteins Nphs1 (nephrin) and Nphs2 (podocin) is ‘downregulated’ in XPodNet by factor 2.8, close to the expected value of 4. For a protein that is not expressed in podocytes, we will encounter a strong upregulation of the interaction. As an example, the interaction between Pecam1 and β-catenin 1 (Ctnnb1) in XPodNet is ‘upregulated’ 56-fold

**Figure 4 | Principal component analysis.** Principal component analysis of several biological conditions (for abbreviations see Table 3) using gene expression data from the Affymetrix Mouse Genome 430 2.0 and Affymetrix Mouse Gene 1.0 ST microarrays.

in glomeruli versus podocytes (Supplementary Figure S2 online, link scores are not shown), as Pecam1 is absent in podocytes, but highly expressed in the glomerular endothelium. Thus, the analysis of differential interactions in glomeruli versus podocytes may help to identify genes that are specifically expressed in mesangial or glomerular endothelial cells.

DISCUSSION

In this study, we present PodNet, an expert-curated PPI network for the analysis of expression data of mouse podocytes. Nodes and edges in PodNet were extracted from the literature, as in the beginning of the project there were good reasons to assume that databases contain a rather low number of podocyte-specific PPIs. Indeed, only 10% of the PPIs in PodNet could be found in STRING, corroborating our approach. In addition, we expanded PodNet to XPodNet by adding database PPIs to enhance connectivity, and we generated GlobalNet as an unbiased network for the purpose of comparison by taking all mouse PPIs in STRING above a certain confidence score. XPodNet and GlobalNet yielded complementary, as well as overlapping, insight into the functionality of the podocyte PPI network and its regulation.

We used ExprEssence to perform differential interactome analysis on the PPI networks.⁷ The calculation of the link score in ExprEssence is based on the law of mass action,

Table 4 | Manually curated subset of the most differentially regulated podocyte protein–protein interactions in development and culture

Adult podocytes versus earlier stages				Cultured podocytes versus adult podocytes			
XPodNet	GlobalNet	XPodNet	GlobalNet				
Upregulated							
Interaction	Rank	Interaction	Rank	Interaction	Rank	Interaction	Rank
Nphs2-Cd2ap	0.994	Nphs1-Nphs2	0.997	Itgb3-Fn1	0.988	Itga5-Fn1	0.990
Nphs1-Nphs2	0.990	Nphs2-Sh3kbp1	0.981	Cdkn1a-Cdk4	0.987	Cdk1-Ccna2	0.984
Nphs2-Kirrel	0.979	Cd2ap-Cbl	0.959	Paxip1-Pax8	0.969	Itgb3-Fn1	0.984
Nphs1-Cd2ap	0.956	Nphs1-lqgap1	0.957	Cdkn2a-Cdk4	0.967	Cdkn1a-Cdk2	0.981
Nphs2-Kirrel3	0.951	Anxa2-Mras	0.938	Enah-Zyx	0.954	Fn1-Mag	0.981
Nphs2-Trpc6	0.951	Nbr1-Sqstm1	0.937	Cdkn1a-Casp3	0.949	Cdkn2a-Cdk4	0.976
Podxl-Ezr	0.944	Sqstm1-Cyld	0.934	Igf1r-Grb10	0.938	Cdkn1a-Cdk4	0.974
Podxl-Slc9a3r2	0.938	Tjp1-Cldn5	0.931	Cdkn1a-Cdk6	0.935	Paxip1-Pax8	0.972
Cd2ap-Ddn	0.938	Ilk-Parva	0.911	Actb-Pfn2	0.935	Cdk1-Ccnb2	0.970
Nphs2-Sh3kbp1	0.937	Cflar-Itch	0.910	Smad3-Smad4	0.932	Glis2-Hdac3	0.969
Downregulated							
Interaction	Rank	Interaction	Rank	Interaction	Rank	Interaction	Rank
Src-Efs	0.092	Cdt1-Gmnn	0.086	Nphs1-Nphs2	0.003	Nphs1-Nphs2	0.001
Notch1-Wdr12	0.110	Gmnn-Tbpl1	0.089	Nphs2-Cd2ap	0.006	Tjp1-Cldn5	0.003
Cdkn2a-Mycn	0.110	Gja1-Sgsm3	0.097	Nphs1-Ddn	0.007	Nphs2-Sh3kbp1	0.009
Fras1-Grip1	0.111	Npy-Npy1r	0.104	Nphs2-Trpc6	0.008	Magi2-Dll4	0.010
Igsf5-Magi1	0.115	Cdk9-Myc	0.104	Nphs2-Kirrel3	0.010	Nphs1-lqgap1	0.017
Notch1-Kat2a	0.117	Tle1-Hes6	0.111	Nphs1-Magi2	0.010	Pak1-Cdc42	0.018
Nid2-Fbln2	0.140	Gjb3-Gja1	0.112	Tjp1-Cldn5	0.017	Pak1-Rac1	0.019
Ptk2-Efs	0.150	Cdc7-Orc6l	0.113	Cd2ap-Ddn	0.019	Dynlt3-Dync1i1	0.038
Pick1-Jam3	0.157	Orc1l-Cdc45l	0.113	Nphs1-Slc4a1	0.027	Cflar-Itch	0.044
Notch1-Smarcd3	0.161	Mcm3-Cdc45l	0.114	Nphs2-Kirrel	0.029	Taf5-Mafb	0.046

The rank of each interaction is the mean of the normalized ranks of the link scores originating from our ExprEssence analyses (see Table 3). If an interaction appears in both XPodNet and GlobalNet in the development or culture comparison, it is written in bold.

Gray cells mark inversely regulated interactions, which are upregulated during development and downregulated in culture. There are no inversely regulated interactions that are lost during development and are regained in culture.

which relies on protein concentrations and the affinity constant. Thus, ExprEssence detects the coordinated upregulation or downregulation of interacting proteins, resulting, for example, from transcriptional coregulation. The formula for the calculation of the link score in ExprEssence does not contain the affinity constant, as in the case of differential analysis, when relative changes are analyzed, the law of mass action becomes independent of the affinity constant. However, if we analyze a single condition, as for the determination of the most abundant interactions, the reaction constants matter. Affinity constants of PPIs may vary over a certain range,³³ and hence may affect the outcome. Concerning the protein concentrations, we use transcriptome data in our present study, assuming a linear correlation between mRNA and protein concentration. It has been shown that mRNA and protein levels are correlated to a fair degree,^{34,35} although several relevant exceptions for some mRNA–protein pairs exist. Ideally, direct measurements of protein abundance should be used. Given the recent accomplishment by several groups to isolate podocytes,^{11–15} podocyte proteomes will become available in the future. A further limiting factor is that mRNA transcripts not represented on the microarray are dismissed. Finally, our approach does not take into consideration the fact that proteins are modified (for example, by phosphorylation) or compartmentalized in the cell. Our podocyte PPI networks are mouse networks, as

expression data of freshly isolated podocytes are available for the mouse only. The generation of human podocyte PPI networks will be worthwhile, as soon as expression data for human *in vivo* podocytes become available.

Other possible sources of errors arise from wrong entries in STRING and from contaminations of isolated podocytes with other cell types. Although we expanded PodNet and built GlobalNet only with experimentally described PPIs, we still noted wrong entries in STRING. Concerning the publicly available transcriptomes of freshly isolated podocytes, we noted contaminations by other cell types such as endothelial cells or distal tubular cells. Transcripts of highly expressed genes in contaminating cells will be present in relevant amounts in the transcriptome even if the purity of the podocyte preparation is above 95%. For example, transcripts of Pecam1 originating from endothelial cells (see Results) and of uromodulin originating from distal tubular cells are present in the podocyte transcriptomes in relevant amounts. In order to minimize potential errors, we used network intersection (Figure 3), averaging (Figure 7), and manual curation (Table 4).

Despite the limitations mentioned above, we demonstrate that podocyte PPI networks considerably extend previous approaches, such as the pioneering three cell-mixture PPI network of GlomNet⁹ and the recent gene expression-based definition of the molecular equipment of podocytes.¹³ The

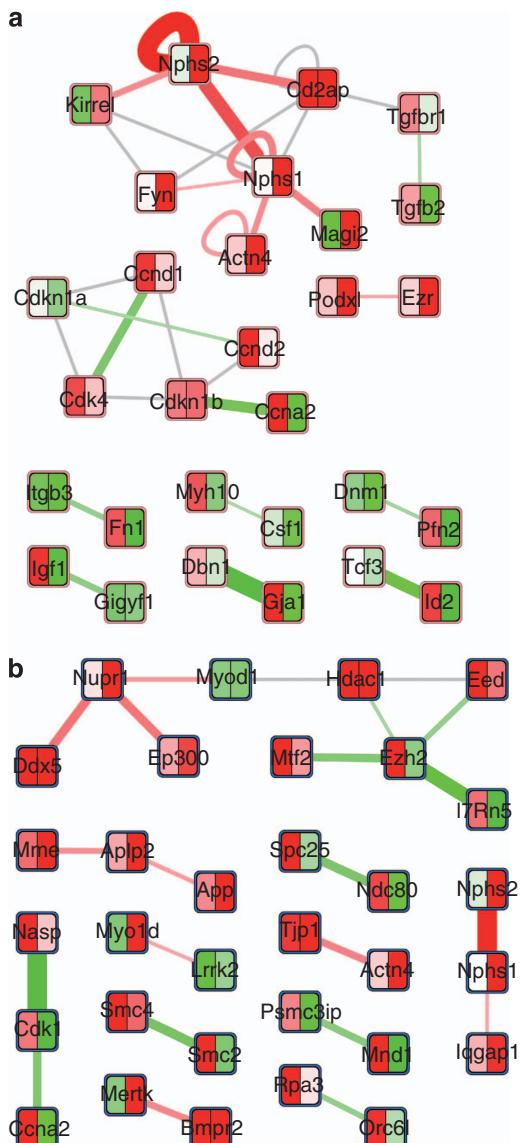


Figure 5 | ExprEssence analyses for Adult G-Pod versus E13.5 G-Pod. The 10 most upregulated (red) and downregulated (green) interactions, including connecting interactions (gray), are shown for (a) XPoDNet and (b) GlobalNet. The node color describes expression levels (left side, E13.5, G-Pod; right side, Adult G-Pod). The thicker an edge, the more it is upregulated/downregulated. E13.5, embryonic day 13.5; Pod, podocyte.

analysis of the podocyte PPI networks presented in this study provides a holistic view of the current knowledge about podocyte biology and a differential interactome analysis, delivering testable hypotheses. This will be illustrated by two examples. The podocyte core network (Figure 2) serves as the first example. Intriguingly, approximately half of the proteins known to be essential for podocyte function are held together by six additional proteins in an interconnected cluster. It is tempting to speculate that some of the additional proteins might also prove to be essential for podocyte function, such as ezrin (Ezr) or Iqgap1. Moreover, as more podocyte PPIs will be identified, the podocyte core network may grow and further essential nodes will be incorporated. Thus, a rather

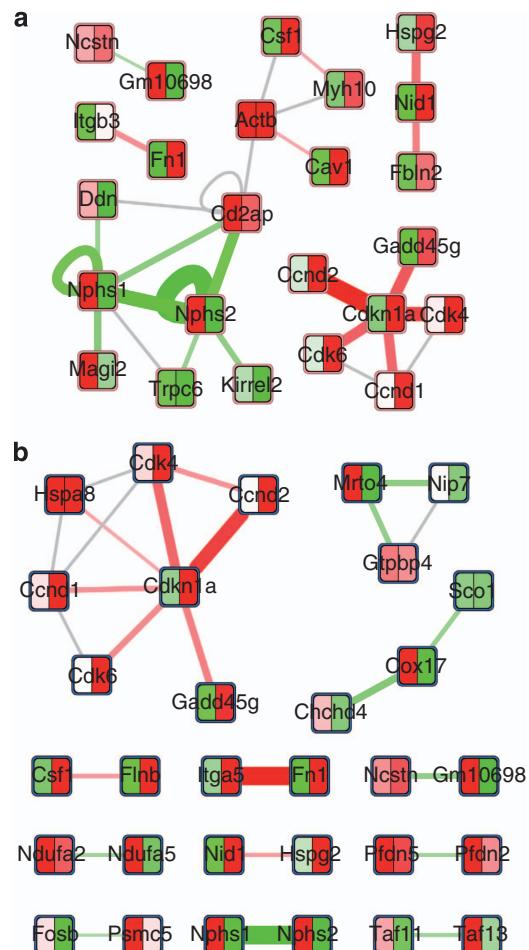


Figure 6 | ExprEssence analyses for Prim M-Pod versus Adult G-Pod. The 10 most upregulated (red) and downregulated (green) interactions, including connecting interactions (gray), are shown for (a) XPoDNet and (b) GlobalNet. The node color describes expression levels (left side, Adult G-Pod; right side, Prim M-Pod). The thicker an edge, the more it is upregulated/downregulated. Pod, podocyte; Prim, primary.

large core network may underlie podocyte function in health and disease. As a second example, we will consider the most abundant interactions in podocytes (Figure 3). In XPoDNet, several PPIs were identified among the most abundant interactions that have up to now received only very limited attention: Cldn5,²¹ interacting with ZO-1, p21-activated kinase (Pak1),³⁶ interacting with the GTPases Rac1 and Cdc42, and Spna2, connecting nephrin (Nphs1) to the actin cytoskeleton. As the Cldn5 and Pak1 interactions also belong to the most differentially regulated interactions in development and culture (Table 4), and as the most abundant interactions are of major relevance for podocyte structure, these proteins and their interactions are suggested to have an important role in podocyte biology.

As further applications, we analyzed existing podocyte transcriptomes mainly with regard to podocyte differentiation during development and to podocyte dedifferentiation in cell culture. In order to analyze podocyte differentiation during development, transcriptomes of podocytes isolated

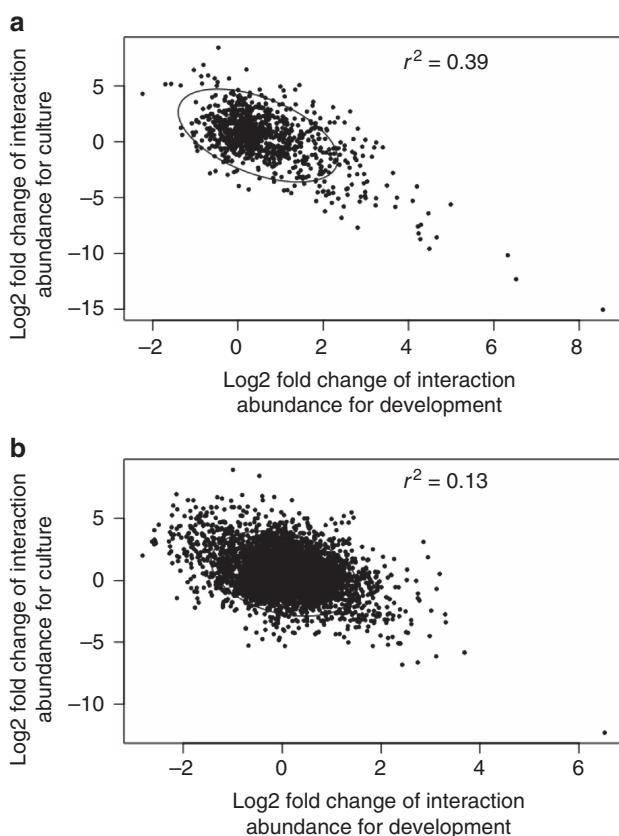


Figure 7 | Scatter plots of mean log fold changes in abundance of interaction (mean ExprEssence link score). The x axis describes changes in abundance of interactions during podocyte development and the y axis describes them for *in vivo* podocytes that are put into culture for interactions in (a) XPodNet and (b) GlobalNet. The ellipse contains 95% of the data points.

by fluorescence-activated cell sorting and transcriptomes of microdissected tissue were used. Both sources possess specific advantages: a pure cell population is provided by podocytes isolated by fluorescence-activated cell sorting and a defined developmental stage is provided by microdissected tissue. The advantages of both preparations were combined by extracting the most differentially regulated interactions (Table 4) in the complete data set. The identification of the slit diaphragm protein complex among the most differentially regulated interactions during development demonstrates the power of the differential interactome analysis with XPodNet and ExprEssence.^{13,37,38} Interactions between slit diaphragm proteins are massively downregulated in three independently derived podocyte cultures, in agreement with previous measurements of transcript and protein levels.³⁹ Thus, podocytes in culture are dedifferentiated. As interactions that are shut down during development do not come up in cultured podocytes (Table 4 and Figure 7), dedifferentiation of podocytes in culture represents rather a loss of differentiation than a reversal of development. This finding is further supported by PCA (Figure 4), by regression analysis of gene expression (Supplementary Figures S4 and S5 online), and by analysis of the most differentially expressed

genes (Supplementary Table S4 online). It is important to note that differential interactome analysis provides immediate insight into biological mechanisms by identifying differentially regulated pairs and clusters of PPIs (see Figures 5 and 6). In contrast, if the most differentially regulated genes are mapped on XPodNet and GlobalNet (Supplementary Figure S6 online), none of the most differentially expressed genes code for pairs of interacting proteins (except for the Nphs1/Nphs2 pair). This clearly demonstrates the power of our approach.

It has been suggested for endocrine and acinar pancreatic cells, as well as for hepatocytes, that they regress into an embryonic progenitor-like state as judged by the expression of progenitor marker genes.^{40–42} Although the gene expression pattern and the expression changes in cultured podocytes better fit to that of the renal vesicle than E13.5 podocytes (Supplementary Figures S4 and S5 online), cultured podocytes express neither markers of the renal vesicle nor even earlier markers of metanephric mesenchyme (Supplementary Figure S7 online).

Taken together, we have built podocyte PPI networks and demonstrate their utility in analyzing differential changes of the podocyte interactome. The PPI networks (PodNet, XPodNet, and GlobalNet) and the software for differential interactome analysis (ExprEssence)⁷ are implemented in Cytoscape, an open source platform.²⁰ Podocyte PPI networks and differential interactome analysis will become indispensable tools for the interpretation of the steadily increasing number of podocyte expression data. The public availability of our tools and podocyte PPI networks (www.PodNet.de and WikiPathways) ensures broad and immediate access by the research community.

MATERIALS AND METHODS

Generation of PPI networks

To set up PodNet, we examined the podocyte-related literature for the period from 1 January 2000 to 1 January 2011 by hand for reliable information on PPIs (co-immunoprecipitation and pull-down), as well as for gene/protein expression (*in situ* hybridization, immunofluorescence, immunohistochemistry, and immunoelectron microscopy) in the podocyte (for details, see Supplementary Methods online). Interactions from both human and mouse were considered to obtain a more comprehensive network, as restriction to PPIs directly proven to occur in mice results in a sparse PPI network, and as differences between mouse and human interactions are limited to a negligible number of proteins. Only mouse gene identifiers were used in the network, which allowed for direct integration of mouse gene expression data into the network (Supplementary Data File S1 and S2 online). By using the STRING database,¹⁹ we expanded PodNet to build XPodNet by adding nodes that interact with the nodes of PodNet and by adding interactions between all nodes. As a network without podocyte bias, we built GlobalNet featuring all mouse interactions from STRING (see Supplementary Methods online). In general, only interactions described as experimentally verified were taken from STRING. Enrichment of Gene Ontology terms was determined with the help of BiNGO plugin (version 2.44; <http://apps.cytoscape.org/apps/bingo>) for Cytoscape.

Identification of a podocyte core network

To isolate a subnetwork of XPodNet that contains as many essential nodes as possible, essential nodes and their interacting nodes were kept, whereas all other nodes were removed. Nodes were considered as ‘essential’ if human gene mutation, gene knockout in mice, or gene knockdown in zebrafish is associated with a podocyte phenotype (for example, foot process effacement). The subnetwork containing the highest number of essential nodes was selected, and those nodes interacting with essential nodes were deleted as long as the subnetwork did not fall apart.

Gene expression data

We used our own and publicly available gene expression data from Gene Expression Omnibus (GEO) based on Affymetrix Mouse Genome 430 2.0 Array (Santa Clara, CA) and Mouse Gene 1.0 ST Array platforms (Table 3). To extend publicly available gene expression data of cultured podocytes, we determined the transcriptome of our mouse podocyte cell line SVI in the same way as described previously.¹⁴ Podocytes were grown and differentiated according to the original protocol.¹⁷ The CEL files were processed with JMP Genomics 4.0 software (Cary, NC), using custom CDF files from Brainarray.org for Entrez Gene ID. Separately for each array platform, background correction was done using the robust multichip average method, followed by quantile normalization as provided by JMP Genomics.

Networks of most abundant interactions

To obtain a network of the most abundant interactions in the adult podocyte, we determined the most abundant interactions in Adult S-Pod and Adult G-Pod (Table 3) by applying the Law of Mass Action as described in Supplementary Methods online. For XPodNet, we took the subnetworks of the 50 most abundant interactions for both data sets (Supplementary Figure S1A and B online), keeping nonabundant interactions that connected proteins of most abundant interactions (gray edges). Thereafter, we built the intersection network (Figure 3a), which encompassed 19 common interactions. The same procedure was applied to GlobalNet, but with 75 most abundant interactions for both networks (Supplementary Figure S1C and D online), as for this number the intersection also yielded 19 shared most abundant interactions (Figure 3b).

PCA of gene expression data

We performed two PCA—one for each microarray platform. Input data for PCA were the mean log-transformed, robust multichip average background-corrected, and quantile-normalized expression values as provided as Lsmean values by JMP Genomics’ analysis of variance macro for each condition. The first two principal components are visualized.

Identification of most differentially regulated interactions

For both XPodNet and GlobalNet, we used ExprEssence⁷ (<http://sourceforge.net/projects/expronsense/>) (ExprEssWeb, a browser-based version of ExprEssence to map raw data on networks, will be available in due time (see presentation at http://www.ibima.med.uni-rostock.de/IBIMA/software_project/SoftwareProject.html)) for pairwise differential analyses of (1) developing versus adult podocytes (comparisons e, f, g, i, j, and k in Table 3) and (2) *in vivo* versus cultured podocytes (comparisons a, c, and d, in Table 3). Preprocessed (see section on Gene Expression Data) transcriptome data sets (see Table 3) were integrated into XPodNet

and GlobalNet. ExprEssence was applied without using variance data for the gene expression values. The link score is calculated from the relative expression changes of each interacting protein (for details, see Warsow *et al.*⁷). In brief, the link score increases if the expression of both interacting proteins increases; it decreases if the expression of both interacting proteins decreases; and it remains unchanged if the expression of the interacting proteins changes in opposite directions. The link score percentile thresholds were set as to give subnetworks with the desired number of remaining interactions.

In addition, we determined the most differentially regulated interactions in the set of the six pairwise comparisons of developing podocytes and in the set of the three pairwise comparisons of cultured podocytes. In each set, we ranked the interactions by their ExprEssence link scores, normalized the ranks by the number of links in the respective network (XPodNet or GlobalNet), and calculated the mean ranks. The 50 most differentially regulated interactions for both sets and both PPI networks are given in Supplementary Table S3 online. Starting with the highest rank, the 10 most differentially regulated interactions were selected (Table 4) for which the interaction in STRING could be verified in the literature and for which expression in adult podocytes could be confirmed using the Human Protein Atlas (Proteinatlas.org).

DISCLOSURE

All the authors declared no competing interests.

ACKNOWLEDGMENTS

This study was supported by grants of the German Federal Ministry of Education and Research within the PodoRePro project (BMBF, grant nos. 01GN0804, 01GN0901, and 01GN0805) to MJM, GF and KE.

SUPPLEMENTARY MATERIAL

Figure S1. Most abundant interactions for Adult G-Pod and Adult S-Pod in XPodNet and GlobalNet.

Figure S2. ExprEssence analyses of XPodNet.

Figure S3. ExprEssence analyses of GlobalNet.

Figure S4. Scatter plots of gene expression and changes of gene expression (1.0 ST platform).

Figure S5. Scatter plots of gene expression and changes of gene expression (430 2.0 platform).

Figure S6. Differentially expressed genes in XPodNet and GlobalNet.

Figure S7. Expression of marker genes of different stages in kidney and podocyte development.

Table S1. Overrepresented GO terms of molecular function in PodNet, XPodNet, and GlobalNet.

Table S2. Microarray data sets used for differential analysis.

Table S3. The 50 strongest differentially up- and downregulated interactions in culture and during development for XPodNet and GlobalNet.

Table S4. The 50 strongest differentially up- and downregulated genes during development and in culture.

Data File S1. Proteins contained in PodNet.

Data File S2. Protein-protein interactions in PodNet.

Supplementary material is linked to the online version of the paper at <http://www.nature.com/ki>

REFERENCES

- Pavenstadt H, Kriz W, Kretzler M. Cell biology of the glomerular podocyte. *Physiol Rev* 2003; **83**: 253–307.
- Wiggins RC. The spectrum of podocytopathies: a unifying view of glomerular diseases. *Kidney Int* 2007; **71**: 1205–1214.
- Keller BJ, Martini S, Sedor JR *et al.* A systems view of genetics in chronic kidney disease. *Kidney Int* 2012; **81**: 14–21.
- He JC, Chuang PY, Ma'ayan A *et al.* Systems biology of kidney diseases. *Kidney Int* 2012; **81**: 22–39.

5. Bhavnani SK, Eichinger F, Martini S et al. Network analysis of genes regulated in renal diseases: implications for a molecular-based classification. *BMC Bioinformatics* 2009; **10**(Suppl 9): S3.
6. Ideker T, Krogan NJ. Differential network biology. *Mol Syst Biol* 2012; **8**: 565.
7. Warsow G, Greber B, Falk SS et al. ExprEssence—revealing the essence of differential experimental data in the context of an interaction/regulation net-work. *BMC Syst Biol* 2010; **4**: 164.
8. Som A, Harder C, Greber B et al. The PluriNetWork: an electronic representation of the network underlying pluripotency in mouse, and its applications. *PLoS One* 2010; **5**: e15165.
9. He L, Sun Y, Takemoto M et al. The glomerular transcriptome and a predicted protein-protein interaction network. *J Am Soc Nephrol* 2008; **19**: 260–268.
10. Sun Y, He L, Takemoto M et al. Glomerular transcriptome changes associated with lipopolysaccharide-induced proteinuria. *Am J Nephrol* 2009; **29**: 558–570.
11. Akilesh S, Huber TB, Wu H et al. Podocytes use FcRn to clear IgG from the glomerular basement membrane. *Proc Natl Acad Sci USA* 2008; **105**: 967–972.
12. Bollee G, Flamant M, Schordan S et al. Epidermal growth factor receptor promotes glomerular injury and renal failure in rapidly progressive crescentic glomerulonephritis. *Nat Med* 2011; **17**: 1242–1250.
13. Brunskill EW, Georgas K, Rumballe B et al. Defining the molecular character of the developing and adult kidney podocyte. *PLoS One* 2011; **6**: e24640.
14. Kabgani N, Grigoleit T, Schulte K et al. Primary cultures of glomerular parietal epithelial cells or podocytes with proven origin. *PLoS One* 2012; **7**: e34907.
15. Takemoto M, He L, Norlin J et al. Large-scale identification of genes implicated in kidney glomerulus development and function. *EMBO J* 2006; **25**: 1160–1174.
16. Mundel P, Reiser J, Zuniga Mejia Borja A et al. Rearrangements of the cytoskeleton and cell contacts induce process formation during differentiation of conditionally immortalized mouse podocyte cell lines. *Exp Cell Res* 1997; **236**: 248–258.
17. Schiwek D, Endlich N, Holzman L et al. Stable expression of nephrin and localization to cell-cell contacts in novel murine podocyte cell lines. *Kidney Int* 2004; **66**: 91–101.
18. Koh GC, Porras P, Aranda B et al. Analyzing protein-protein interaction networks. *J Proteome Res* 2012; **11**: 2014–2031.
19. Szklarczyk D, Franceschini A, Kuhn M et al. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res* 2011; **39**: D561–D568.
20. Shannon P, Markiel A, Ozier O et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003; **13**: 2498–2504.
21. Koda R, Zhao L, Yaoita E et al. Novel expression of claudin-5 in glomerular podocytes. *Cell Tissue Res* 2011; **343**: 637–648.
22. Doyonnas R, Kershaw DB, Duhme C et al. Anuria, omphalocele, and perinatal lethality in mice lacking the CD34-related protein podocalyxin. *J Exp Med* 2001; **194**: 13–27.
23. Kirkin V, Lamark T, Sou YS et al. A role for NBR1 in autophagosomal degradation of ubiquitinylated substrates. *Mol Cell* 2009; **33**: 505–516.
24. Hartleben B, Godel M, Meyer-Schwesinger C et al. Autophagy influences glomerular disease susceptibility and maintains podocyte homeostasis in aging mice. *J Clin Invest* 2010; **120**: 1084–1096.
25. Pitera JE, Scambler PJ, Woolf AS. Fras1, a basement membrane-associated protein mutated in Fraser syndrome, mediates both the initiation of the mammalian kidney and the integrity of renal glomeruli. *Hum Mol Genet* 2008; **17**: 3953–3964.
26. Harita Y, Miyauchi N, Karasawa T et al. Altered expression of junctional adhesion molecule 4 in injured podocytes. *Am J Physiol Renal Physiol* 2006; **290**: F335–F344.
27. Yaoita E, Yao J, Yoshida Y et al. Up-regulation of connexin43 in glomerular podocytes in response to injury. *Am J Pathol* 2002; **161**: 1597–1606.
28. Niranjan T, Bielez B, Gruenwald A et al. The Notch pathway in podocytes plays a role in the development of glomerular disease. *Nat Med* 2008; **14**: 290–298.
29. Alexandropoulos K, Baltimore D. Coordinate activation of c-Src by SH3 and SH2-binding sites on a novel p130Cas-related protein, Sin. *Genes Dev* 1996; **10**: 1341–1355.
30. Marshall CB, Shankland SJ. Cell cycle regulatory proteins in podocyte health and disease. *Nephron Exp Nephrol* 2007; **106**: e51–e59.
31. Ohse T, Vaughan MR, Kopp JB et al. De novo expression of podocyte proteins in parietal epithelial cells during experimental glomerular disease. *Am J Physiol Renal Physiol* 2010; **298**: F702–F711.
32. Saleem MA, O'Hare MJ, Reiser J et al. A conditionally immortalized human podocyte cell line demonstrating nephrin and podocin expression. *J Am Soc Nephrol* 2002; **13**: 630–638.
33. Charbonnier S, Gallego O, Gavin AC. The social network of a cell: recent advances in interactome mapping. *Biotechnol Annu Rev* 2008; **14**: 1–28.
34. Ning K, Fermin D, Nesvizhskii AI. Comparative analysis of different label-free mass spectrometry based protein abundance estimates and their correlation with RNA-Seq gene expression data. *J Proteome Res* 2012; **11**: 2261–2271.
35. Schwahnhauser B, Busse D, Li N et al. Global quantification of mammalian gene expression control. *Nature* 2011; **473**: 337–342.
36. Zhu J, Attias O, Aoudjit L et al. p21-activated kinases regulate actin remodeling in glomerular podocytes. *Am J Physiol Renal Physiol* 2010; **298**: F951–F961.
37. Roselli S, Gribouval O, Boute N et al. Podocin localizes in the kidney to the slit diaphragm area. *Am J Pathol* 2002; **160**: 131–139.
38. Ruotsalainen V, Patrakka J, Tissari P et al. Role of nephrin in cell junction formation in human nephrogenesis. *Am J Pathol* 2000; **157**: 1905–1916.
39. Chittiprol S, Chen P, Petrovic-Djergovic D et al. Marker expression, behaviors, and responses vary in different lines of conditionally immortalized cultured podocytes. *Am J Physiol Renal Physiol* 2011; **301**: F660–F671.
40. Negi S, Jetha A, Aikin R et al. Analysis of beta-cell gene expression reveals inflammatory signaling and evidence of dedifferentiation following human islet isolation and culture. *PLoS One* 2012; **7**: e30415.
41. Pinho AV, Rooman I, Reichert M et al. Adult pancreatic acinar cells dedifferentiate to an embryonic progenitor phenotype with concomitant activation of a senescence programme that is present in chronic pancreatitis. *Gut* 2011; **60**: 958–966.
42. Chen Y, Wong PP, Sjeklocha L et al. Mature hepatocytes exhibit unexpected plasticity by direct dedifferentiation into liver progenitor cells in culture. *Hepatology* 2012; **55**: 563–574.

Differential Network Analysis Applied to Preoperative Breast Cancer Chemotherapy Response

Gregor Warsow^{1,2,3}, Stephan Struckmann¹, Claus Kerkhoff⁴, Toralf Reimer⁵, Nadja Engel^{6,*}, Georg Fuellen^{1,*}

1 Institute for Biostatistics and Informatics in Medicine and Ageing Research, University of Rostock, Rostock, Germany

2 Department of Mathematics and Informatics, University of Greifswald, Greifswald, Germany

3 Department of Anatomy and Cell Biology, University Medicine Greifswald, Greifswald, Germany

4 Department of Biomedical Sciences, School of Human Sciences, University of Osnabrueck, Germany

5 Department of Obstetrics and Gynecology, University of Rostock, Germany

6 Department of Cell Biology, University of Rostock, Rostock, Germany

* E-mail: nadja.engel-lutz@uni-rostock.de; fuellen@uni-rostock.de

Abstract

All processes in a living cell are organized by a complex functional network of interactions between genes, their products and other molecules of various origin. If their concerted actions are disturbed, this may have consequences for the cell and the organism, giving rise to the development of diseases such as breast cancer. Several strategies have been developed for treatment of breast cancer. One of them, neoadjuvant TFAC chemotherapy, is used in cases where application of preoperative systemic therapy is indicated. Estimating response to treatment allows or improves clinical decision-making and this, in turn, may be based on a good understanding of the underlying molecular mechanisms. Ever increasing amounts of high throughput data are available for integration into functional networks. In this study, we applied our software tool *ExprEssence* to identify specific mechanisms relevant for TFAC therapy response, thereby validating our method by recovering known mechanisms. Furthermore, we identified a mechanism that may further explain the synergism between paclitaxel and doxorubicin in TFAC treatment: Paclitaxel may attenuate MELK gene expression, resulting in lower levels of its target MYBL2, already associated with doxorubicin synergism in hepatocellular carcinoma cell lines. We tested our hypothesis in three breast cancer cell lines, confirming it in part. In particular, the predicted effect on MYBL2 could be validated, and a synergistic effect of paclitaxel and doxorubicin could be demonstrated in some cell lines.

Introduction

For the successful treatment of breast cancer, the most common type of cancer in women worldwide, knowledge of cancer-treatment responsiveness is most useful. Substantial progress was made in understanding disease mechanisms of breast cancer, but many questions are still unanswered. The rise of genome-scale gene expression profiling allowed for identification of biomarkers that help to further subcategorize known groups of breast cancer, among them luminal ($ER^+/HER2^-$), HER2-enriched ($HER2^+$) and triple-negative ($ER^-/PR^-/HER2^-$) types.

Profiling approaches were first based on the identification of single, differentially expressed genes or of gene sets (signatures). Nowadays, research follows an integrative approach utilizing protein/gene interaction networks, thereby reflecting that biological processes are performed by genes/proteins/molecules interacting with each other and not by single proteins individually. In terms of complexity, this approach goes beyond former analysis methods, as the number of genes in the human genome is surprisingly low (around 23,000 protein coding genes), but the number of interactions and dependencies between them allows for a great number of processes in the cell. The work presented here also attempts to model certain aspects of cell biology to better understand breast cancer treatment.

We use a protein/gene interaction network, into which large genome-scale datasets, assembled from more than 200 patients from various subtypes were integrated [1]. Patient collectives of this size enable unprecedented statistical power and robustness despite subgroup differences. We applied our previously published *ExprEssence* method [2] for the identification of altered protein/gene interactions that characterize the differences between the responders and non-responders to neoadjuvant TFAC therapy. We assume these differentially regulated interactions to be related or even critical for therapy outcome. Knowing about the differences between responders and non-responders may help to gain more detailed insights into both the progression of breast cancer and how it is affected by drugs, which is of high relevance for choosing individualized cancer treatment. Besides ours, there are several network-based approaches aiming to identify genes or proteins involved in the response to a treatment or external condition [3–5], including the pioneering work of Ideker et al. [6]. We compare the results of our method to two such methods, OptDis [5] and KeyPathwayMiner [7] investigating the same breast cancer dataset as the one explored by OptDis, and find that *ExprEssence* generates subnetworks more directly related to disease- and drug-related processes than both other methods.

Materials and Methods

Microarray data and generation of the condensed network

We used breast cancer-related gene expression data from the MicroArray Quality Control (MAQC)-II study ([1], GSE20194). The data were generated at the MD Anderson Cancer Center (MDACC, Houston TX, USA). In this study, transcriptome data as well as ER, PR and HER2 receptor status of 230 patients with newly diagnosed breast cancer were acquired by fine-needle aspiration before neoadjuvant chemotherapy with TFAC (a combination of paclitaxel (Taxol®), 5-fluorouracil, doxorubicin (Adriamycin®) and cyclophosphamide). Application of TFAC has become a combinatorial standard-therapy scheme since extension of the older FAC scheme by paclitaxel turned out to be beneficial for response [8]. The patients were classified as responders or non-responders after tumor resection. For each gene represented on the array, we pooled individual expression measures to obtain single expression values for all responders and all non-responders, respectively. The same data was used by Dao et al. for the application of their OptDis method, which generates subnetworks that are most suitable for distinction between two conditions (responder and non-responder in this case) [5]. In their study, the 230 cases were split up into a discovery and a validation group. After applying OptDis on both sets individually, they intersected the gene sets that made up the respective top-50 subnetworks for each group (discovery and validation). The overlap was a set of 39 genes, denoted as O39.

We integrated the gene expression data into a functional network based on the String database, version 9.0 [9]. The network contained all human interactions scoring at least 0.85 for experimental, database or textmining evidence channels. *ExprEssence* [2], a plugin for the Cytoscape platform [10], was then applied to the network. After integration of gene expression data into a network, *ExprEssence* identifies a subnetwork of those protein interactions that are most differential between the two compared states (responders versus non-responders in this case). *ExprEssence* link score thresholds are adjustable by the user and were set so that we obtained a number of genes comparable to the 39 genes from OptDis (the O39), resulting in 40 proteins (denoted as E40).

Cell culture and treatment conditions

Breast cancer cell lines MCF-7, BT-20 and SK-BR-3 and MCF-10A cells were obtained from American Type Culture Collection (ATCC, USA). MCF-7 and BT-20 were maintained in Dulbecco's modified Eagle's medium (Invitrogen, Germany) with 10 % fetal bovine serum (PAN Biotech GmbH, Germany) and 1% gentamycin (Ratiopharm, Germany). SKBR3 cell line was cultured in McCoy's 5a Medium (ATCC, USA) supplemented with 10 % fetal bovine serum (PAN Biotech GmbH, Germany) and 1% gentamycin (Ratiopharm, Germany). The non-tumorigenic control cell line MCF-10A was grown in Dulbecco's modified Eagle's medium Ham's F12 without phenol red (Invitrogen, Germany) containing 10 % horse serum (PAA Laboratories GmbH, Germany), the Mammary Epithelial Cell Growth Medium Supplement Pack (Promo Cell, Germany) including Bovine Pituitary Extract (0.004 ml/ml), Epidermal Growth Factor (recombinant human) 10 ng/ml, Insulin (recombinant human) 5 µg/ml, Hydrocortisone 0.5 µg/ml and 1% gentamycin (Ratiopharm, Germany). All cell lines were authenticated by morphology

and growth rate and were mycoplasma free. Prior treatment, all cell lines were seeded in 6-well plates and adapted to phenol-red-free Dulbecco's modified Eagle's medium (PAA Laboratories GmbH, Germany) with 10 % charcoal stripped fetal bovine serum (PAN Biotech GmbH, Germany) for 48 h (assay medium). Paclitaxel (T, Taxol; Ratiopharm, Germany) at a final concentration of 0.1 nM or 0.1 μ M, doxorubicin hydrochloride (A, Adriamycin; Sigma, Germany) at a final concentration of 1 nM or 1 μ M, or both were added to the cells for 24 or 48 h in fresh assay medium. As negative control the diluent EtOH (0.1 %) was used in the same manner.

Western blot

After treatment with T or/and A or rather with the control substance EtOH for at least 48 h, the cells were trypsinized, washed with PBS and lysed in ice-cold lysis buffer (Bio-Plex Cell Lysis Kit, Bio-Rad, USA). Cells were homogenized by brief sonification at 4 °C and centrifuged at 8,000 g for 1 min at 4 °C. Protein concentrations of supernatants were estimated by Bradford protein assay [11] so that equal amounts (10-20 μ g) of total soluble protein could be separated by Criterion TGX Stain-Free precast gels (Bio-Rad, Germany) and blotted on PVDF membranes. After SDS-PAGE, protein content per lane as well separation quality was additionally controlled with the Criterion Stain FreeTM gel imaging system (Bio-Rad, Germany). Protein transfer was carried out with a tank blotting system (Bio-Rad, Germany) and then, membranes were blocked with 5 % skim milk in Tris-buffered saline (TBS) and washed five times in TBS. For protein detection, primary antibodies (anti-rabbit anti-MYBL2: AP20207PU-N, Acris, USA; anti-mouse anti-PCNA, sc-56, Santa Cruz, USA; anti-mouse anti-Actin, sc-47778, Santa Cruz, USA) were incubated overnight at 4 °C followed by a labeling with a horseradish peroxidase (HPR)-conjugated secondary antibody (Dako, Glostrup, Denmark) for 1 h at room temperature. Protein signals were visualized by using SuperSignal West Femto Chemiluminescent Substrate (Pierce Biotechnology, Rockford, USA) for detection of peroxidase activity. Band intensity was analyzed densitometrically with the Molecular Imager ChemiDoc XRS and Image Lab 3.0.1 software (Bio-Rad, USA). Protein detection was repeated at a minimum of three times with individually prepared cell lysates from independent passaged cells.

MTS assay

Cells seeded in 96-well-plates in 100 μ L medium were treated with indicated compounds. After 48 h incubation at 37 °C in a 5 % CO₂ atmosphere, cells were assayed with 10 μ L MTS [3-(4,5 dimethylthiazol-2-yl)-5-(3-carboxymethoxyphenyl)-2-(4-sulfophenyl)-2H-tetrazolium] solution (Promega Corp., Madison, WI) for 1 h at 37 °C. The vehicle EtOH (0.1 %) was used in the same manner to serve as control. Colorimetric changes were measured at 492 nm and correction for background absorbance was done by measuring the absorbance of the compounds and MTS solution without the cells. Raw data were transferred to Microsoft Excel for analysis.

Live-Dead Assay

Live-Dead Assay was carried out following manufacture's instructions (PromoCell GmbH, Heidelberg, Germany). After treatment with the indicated compounds cells were washed with phosphate buffered saline to remove serum esterase activity and then treated with 200 μ L of Calcein AM/Ethidium homodimer-III (EthD-III) standard working solution. Cells were incubated at 37 °C for 1 h (Promo cell GmbH, Heidelberg, Germany). The percentage of stained live and dead cells was measured by a fluorescence multiplate reader (Tecan®M200, GmbH, Austria) at appropriate wavelengths; Calcein AM (Ex/Em ~495/~/515 nm), EthD-III (Ex/Em ~530/~/635 nm). The relative reference to the cell number was ensured by a simultaneous Hoechst staining. All obtained values were normalized with respect to the cell number.

Cell cycle measurement for proliferation analysis

To determine proliferation, cell cycle analysis was performed by flow cytometry [12]. The software FlowJo version 10.0.5 (Tree Star Inc., USA) was used to acquire data. A minimum of 15,000 ungated events were recorded. Double and clumps were excluded by gating on the DNA pulse width versus pulse area displays. For statistical analysis, the S-phase and G2/M-phase cells were defined as proliferative cells.

Results and Discussion

After integration of differential gene expression data of responders and non-responders into the STRING-based functional network, we applied *ExprEssence* as described in Materials and Methods in order to extract a condensed network of most strongly up- or downregulated interactions comparing responders to non-responders with respect to TFAC therapy. The resulting 40 proteins (denoted as E40) and their interactions are shown in the ExprEssence-condensed network in Figure 1. All interactions except one (KRT16–KRT7) are differentially regulated in a statistically significant way with Benjamini-Hochberg adjusted P-values below 0.05 for the two-tailed t-test.

We conducted an Ingenuity® IPA Functional Enrichment Analysis (using the Ingenuity database as of 01/21/2013) on our E40 proteins and the OptDis O39 proteins. In contrast to O39, our E40 gene set does not just feature tumor/cancer, apoptosis and cell cycle among the top enriched terms, but also many breast cancer terms, among them proliferation or cell cycle progression of breast cancer cell lines (Table 1). Therefore, compared to the O39 set, our E40 gene is more specific with respect to the biological mechanisms that distinguish responders from non-responders with regard to breast cancer treatment with TFAC.

The ExprEssence-condensed network

Liedtke et al. [13] report that triple-negative breast cancer is responding better to neoadjuvant chemotherapy compared to other types, especially ER-positive breast cancer. Thus, we expect a large proportion of TFAC responders to feature low expression of the ER receptor. This indeed reflects our results (Figure 1)

Enriched Functional Terms	
E40 Genes	O39 Genes
Carcinoma	Transactivation
Cell cycle progression	Cell movement
Chromosomal congression of chromosomes	Cell cycle progression
Cancer	Migration of cells
<i>Carcinoma in breast</i>	Proliferation of tumor cell lines
<i>Breast cancer</i>	Apoptosis of tumor cell lines
Digestive organ tumor	Proliferation of connective tissue cells
Proliferation of cells	Proliferation of cells
Proliferation of tumor cells	Necrosis
Plaque psoriasis	Hypoplasia
Cell movement	Proliferation of epithelial cells
<i>Proliferation of breast cancer cell lines</i>	Transcription
Amenorrhea	Cell death of tumor cell lines
Cell cycle progression of tumor cell lines	Differentiation of cells
Proliferation of cancer cells	Organismal death
Hyperplasia	Cell survival
Mitosis	Synthesis of DNA
Skin development	Transcription of RNA
Development of epidermis	Cell viability
<i>Cell cycle progression of breast cancer cell lines</i>	Binding of DNA
Cell viability of endometrial cells	Quantity of cells
Cloning of fibroblast cell lines	Development of carcinoma
Infanticide	Proliferation of fibroblasts
Mammary gland development	Development of tumor
Carcinoma in situ	Morphology of embryonic tissue

Tab. 1. The top 25 terms of Ingenuity Functional Enrichment Analysis for 40 genes found by ExprEssence (E40) and 39 genes found by OptDis (O39). Breast cancer related terms are written italic. Supp.Tab. 1 and 2 contain also P-values and the lists of the genes associated with the terms.

showing lower ESR1 (as well as AR and PGR) expression in the group of responders compared to non-responders. More generally, in our condensed network, these genes and their interactions describe some of the major differences between responders and non-responders. They build up parts of the subnetwork boxed in green, which contains interactions that are downregulated most strongly in responders compared to non-responders. In particular, this subnetwork includes the mutual stimulation between GATA3 and ESR1. This interaction has been hypothesized by Eeckhoute et al. [14] to promote breast cancer progression via a positive cross-regulatory loop. Despite its cancer promoting role, expression of GATA3 is indicative for good general prognosis [15], as it is strongly correlated with ESR1 expression [16] and such cancer cells can be treated with high rates of success with hormone therapy. Since TFAC therapy performs poorly in ER-positive breast cancer [13], however, the mutual stimulation between GATA3 and ESR1 implies that low GATA3 expression is in fact indicative for good response under the TFAC regimen. In summary, we successfully identified the downregulated interaction between GATA3 and ESR1 to be beneficial for good TFAC response, even though prognosis is worse in general for triple-negative breast cancer compared to hormone therapy-treatable receptor positive breast cancers.

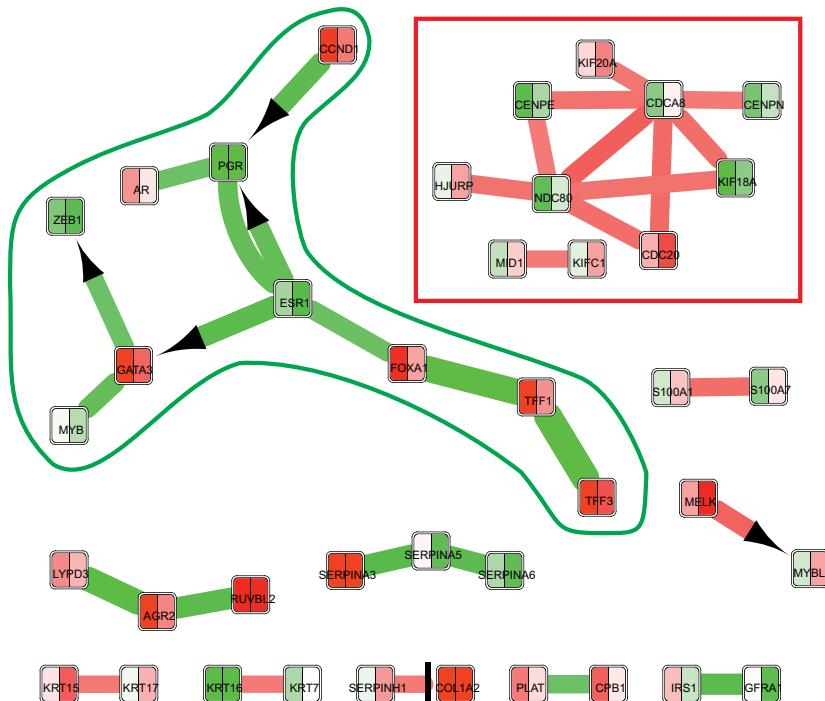


Fig. 1. *ExprEssence*-condensed network describing the 16 most and 16 least active interactions between 40 proteins (E40). For each protein, its mean expression level is visualized for non-responders (left) and responders (right) by color (green for low, white for intermediate, red for high expression). Interactions between proteins are represented by a line between the proteins. Stimulations are indicated by an arrow on the target, inhibitions by a t-bar. The up- (red) and down-regulation (green) of interactions are also color-coded. Full gene names can be found in Table 3.

Another protein in the subnetwork boxed in green, the pioneer transcription factor FOXA1, opens the chromatin and allows estrogen and anti-estrogens to bind the DNA [17]. FOXA1 expression is a signal for proliferation of ER-positive breast cancer and an indicator for good outcome of hormone therapy, as are ESR1 and GATA3 [18]. However, with our data, response to neoadjuvant TFAC treatment was examined. Since patients did not receive anti-estrogens, the proliferative effect of FOXA1 that mediates estrogen binding [19] implies that downregulation of the interaction between FOXA1 and ESR1 indicates good response under TFAC therapy. Lower FOXA1 gene expression has also been correlated with lower TFF1 protein levels and prevention of hormone-induced reentry into the cell cycle [20]. Further, the lower activity of the interaction between TFF1 and TFF3 in responders matches the hypothesis that high levels of the Trefoil factors promote both tumor growth and migration [21, 22]. Therefore, the downregulation of the interaction path FOXA1–TFF1–TFF3 indicates good response under TFAC therapy.

The subnetwork boxed in red in Figure 1 contains exclusively interactions between genes associated with cell cycle and mitosis. In contrast to the subnetwork boxed in green that we discussed above, all these interactions are more active in responders than in non-responders. Cells of responders are thus mitotically more active than non-responder cells, allowing the mitotic spindle poison paclitaxel to have

a stronger therapeutical effect [23].

Furthermore, we observed an upregulation of the interactions KRT7–KRT16 and KRT15–KRT17 in responders, where only the former is differentially regulated in a statistically significant way. We found no consistent biological interpretation for this observation.

The upregulation of SERPINH1 (also known as HSP47) may contribute to a good treatment response by its stimulating effect on procollagen secretion [24,25], which evokes a protective barrier around cancerous cells. Before cancer cells can migrate to distant sites, this obstacle has to be overcome and therefore the ability to metastasize is decreased in case of collagen secreting cells [26]. Some other serpins are also differentially regulated, but their relevance for breast cancer therapy is unclear.

In addition, we observe an upregulation of S100A1 and S100A7 in responders. Both genes belong to the S100 gene family, which is implicated in breast cancer and melanoma, metastasis, Alzheimer's disease, cardiomyopathy and other diseases. S100A7 was found to be expressed in high-grade ductal carcinoma *in situ* (DCIS), a key stage to invasive breast cancer [27]. S100A7 is associated with poor prognostic markers in DCIS and influences progression of breast carcinoma through its interaction with the c-Jun activation domain-binding protein 1 (Jab1) [28], thereby enhancing survival under conditions of cellular stress, such as anoikis [29]. Its downregulation was also shown to inhibit EGF-induced cell migration and invasion in ER-negative MDA-MB-468 cells [30,31]. Some recent studies, however, also identified tumor-suppressive effects of S100A7 in ER-positive breast cancer cells [32,33]. High expression of S100A1, as observed among the group of responders, leads to significant reduction of motility and invasion rates in cells expressing also high levels of metastasis-promoting S100A4 [34]. This observation suggests an antagonistic role of S100A1 against S100A4-mediated metastasis [35]. However, it remains unclear whether S100A1 also counteracts tumorigenic effects mediated by S100A7.

Another protein known to be involved in breast cancer tumorigenesis [36], IRS1, interacts with GFRA1. This interaction is part of a RET oncogene related pathway, which mediates IRS1 activation through GFRA-signalling proteins [37]. The lower activity of this growth-related oncogenetic pathway and lower IRS1 levels in responders may slow down breast cancer progression.

In concordance with high expression levels of AGR2 being associated with decreased survival [38], we also observe lower amounts of AGR2 in responders. The interaction between AGR2 and LYPD3 was considered a "viable target for oestrogen-responsive breast cancer intervention" by Fletcher et al. [39]. Moreover, according to Maslon et al. [40], RUVBL2 is "overproduced in a panel of primary breast cancer biopsy specimens" and a validated interactor of AGR2. Thus, we suggest that RUVBL2, AGR2 and LYPD3 are forming a module whose downregulation is beneficial and indicative of good prognosis in case of TFAC therapy. The interaction between PLAT and CPB1 is downregulated in responders, but we have no indication of its role based on the literature.

Finally, we identified the stimulation of MYBL2 by MELK to be upregulated in responders. At first glance, this finding is contradictory to the general association of high MYBL2 and MELK expression levels with aggressive tumor growth and poor outcome in breast cancer and other tumors [41–43]. Moreover, the proto-oncogene MYBL2 is known to allow cells to override growth inhibitory signals and is essential for S-phase entry [44–46]. However, the BioGraph Database [47] and the Comparative Toxicogenomics Database (CTD, [48]) suggest that MELK is related to susceptibility to TFAC therapy. Both databases

refer to analyses performed by Hess et al., where MELK has been observed to be significantly upregulated in responders to TFAC therapy [49], although no *direct* effects of paclitaxel on MELK were reported there. MELK is known to stimulate MYBL2 based on observations by Nakano et al., who found a downregulation of MYBL2 to be induced by MELK knockdown [50]. Thus, we decided to investigate the underlying stimulation, and its possible inhibition by paclitaxel, in more depth. In the literature we found indications that paclitaxel inhibits MELK via E2F transcription factors, and that MELK stimulates MYBL2 via ZPR9, implying that paclitaxel inhibits MYBL2 indirectly via MELK.

More specifically, paclitaxel has been shown to induce the cyclin inhibitor p21^{WAF1} in MCF-7 breast cancer cells [51], which leads to lower Cdk2 activity [52], resulting in less phosphorylation of the pocket proteins p107/p130 and persistent association of E2F transcription factors with p107/p130 [53]. Verlinden et al. found the MELK gene to carry E2F responsive elements in its promoter region [54]. Hence, paclitaxel-induced complexation of E2F transcription factors could lead to a downregulation of MELK gene expression. This could trigger less MYBL2 expression, since MELK has been shown to phosphorylate the zinc-finger-like protein ZPR9 [55], which, in turn enhances transcriptional activity of MYBL2 [50,56]. Moreover, Nakano et al. suggested that both ZPR9 and MYBL2 are transcriptionally regulated by MELK [50]. (Since ZPR9 is not represented in the data we used, it could not become a member of our E40 network.)

According to Calvisi et al. [57], low expression of MYBL2 is beneficial for chemotherapy response. More specifically, Calvisi et al. investigated hepatocellular carcinoma (HCC) cell lines with wildtype and mutated p53 and found MYBL2-inhibited HCC cells to be associated with reduced proliferation, increased DNA damage, and induction of apoptosis irrespective of p53 status. A p53 mutant status was correlated with higher levels of MYBL2 and advanced tumor stage of human breast cancer [58]. However, especially HCC cells with mutated p53, which are not able to arrest in the G₁ phase and therefore enter into mitosis with DNA heavily damaged by doxorubicin, show higher rates of apoptosis than p53 wildtype HCC cells. Therefore, Calvisi et al. concluded that MYBL2 inhibition could represent a valuable adjuvant for doxorubicin treatment against human hepatocellular carcinoma especially with mutated p53.

Taken together, paclitaxel may play an important role as a co-player of doxorubicin by repressing MELK expression, which in turn attenuates MYBL2 expression and hence allows for more efficient effects of doxorubicin. We will investigate this aspect in more detail in the section on MELK and MYBL2 as targets for TFAC therapy.

General patterns in the *ExprEssence*-condensed network

In summary, with respect to up- and downregulated links in the *ExprEssence*-condensed network, we observe that links that are *upregulated* in responders can be divided into two classes. The first class supports *tumorsuppressive* processes (indicating good prognosis irrespective of treatment, e.g. the protective collagen barrier built with the help of SERPINH1). The other class supports *pro-oncogenic* processes (e.g. cell cycle/mitosis related links). In this study, for all cases of the latter type, the process itself or at least one of the proteins involved in it is known as a target of TFAC. In turn, non-responders do not feature these specific targets for chemotherapy and therefore cannot benefit from the therapy as much.

Therefore, we suppose that upregulated pro-oncogenic processes are targets for therapy and hence a basis for TFAC response.

Downregulated links, on the other hand, cannot be associated closely to response to TFAC. Instead, if the downregulated links were upregulated, they would generally indicate worse response. Moreover, we observe that the subnetwork of downregulated links renders a collection of targets for other kinds of therapies. In particular, targets of anti-hormone therapy (e.g. ESR1 and AR) are part of the subnetwork boxed in green.

Summing up, we could show that, by utilizing an interaction network and gene expression data contrasting responders with non-responders of TFAC chemotherapy, *ExprEssence* yielded a subnetwork that can be incorporated into the cancer-related body of knowledge and, in addition, gives rise to new hypotheses with regard to the mechanistic workings of TFAC. One of these mechanisms will be investigated further in the following section.

MELK and MYBL2 as targets for TFAC therapy

MELK is expressed in several developing tissues, but it is also found in breast tumor-initiating cells, and is required for mammary tumor growth *in vivo* [59]. In our analysis (Figure 1), the stimulation of MYBL2 by MELK is upregulated in responders, despite high MYBL2 and MELK levels being associated with poor prognosis [41,42]. Therefore, we investigated this interaction in more detail using breast cancer cell lines. Our hypothesis was that paclitaxel (abbreviated by T for its trade name Taxol®) inhibits MELK gene expression, which leads to lower expression of MYBL2, suggesting that T may act as a co-player of doxorubicin (abbreviated by A for its trade name Adriamycin®) [57].

Probably due to very low levels of MELK protein in the breast cancer cell lines used in this study, we were not able to detect MELK using immunofluorescence and western blotting (Figure 2). This reflects that primarily tumor-initiating cells or stroma cells, which are not represented by the used cell cultures, express MELK [59]. Accordingly, MELK gene expression of the specimens investigated in this study indicates that they originate from freshly diagnosed breast cancer tissue which may, besides breast cancer cells, contain also tumor-initiating and stoma cells. In contrast, we observed high levels of MYBL2 especially in the breast cancer cell lines (Figures 2,3) and a decrease of MYBL2 protein levels after application of T and A both individually and in combination could be verified by Western blotting, as follows.

Four different cell line types were chosen to compare effects imputed to breast cancer subtype (Table 2). As a non-tumorigenic/normal breast-like control the cell line MCF-10A was selected. The cell line MCF-7 represents the most prevalent and most common breast cancer subtype (luminal, estrogen receptor (ER) and progesterone receptor (PR) positive). The highly invasive cell line BT-20 was used as a model for the triple negative type because neither ER, PR nor human epidermal growth factor receptor 2 (HER2) expressions is observed in BT-20. As HER2 positive cell type the cell line SKBR3 was used. Prior treatment with chemotherapeutic agents, all cells were adapted to phenol red free medium with charcoal treated serum to avoid cross stimulation with endogenous hormones like 17 β -estradiol. Final concentrations of paclitaxel and doxorubicin were selected on the basis of published IC50 values for both

substances [60].

For our experimental setup, we tested paclitaxel alone for 48 h (T), doxorubicin alone for 48 h (A), a combination (T + A) (48 h) and successive treatment so that paclitaxel was first given for 24 h and thereafter doxorubicin was added (T (24 h), A (24 h)). These various treatment combinations showed the direct influence of the single agents and also the combined effects on the protein expression level of MYBL2. In the left panel of Figure 3(a), the protein expression level of MYBL2 in the non-tumorigenic cell line MCF-10A after treatment with chemotherapeutic agents in comparison with vehicle control is given. The representative blot as well as the densitometric statistics of the three individual replicates shows that MYBL2 is expressed with no significant alterations in the non-tumorigenic cell line MCF-10A after treatment with the chemotherapeutics. As a marker for proliferative behavior, the Proliferating Cell Nuclear Antigen, commonly known as PCNA, was detected on the same blots. PCNA expression was significantly reduced after treatment with T and A alone as well as with the combined treatments. Only the lowest concentrations (0.1 nM T + 1 nM A) caused no proliferative alterations in comparison with control. This result reflects the strong inhibitory effect of T and A on proliferation of dividing cell populations by either stabilizing microtubules or by intercalating DNA. As loading control, the counter labeling with β -actin as a housekeeping protein and also the stain-free imaging of the SDS-PAGE separations for visual monitoring of the loaded total protein contents were utilized.

The first observation we gained from the western blotting experiments of the breast cancer cell lines (MCF-7, BT-20, SKBR3) was that these displayed significantly stronger expression levels of MYBL2 in the untreated state compared to the non-tumorigenic control (MCF-10A). On each blot, 10 μ g total soluble protein was transferred, making the expression levels of the cell lines comparable. The high expression levels of MYBL2 in the cancer cell lines render them as potential targets for treatment with the chemotherapeutic agents. In contrast to the non-tumorigenic cell line, MYBL2 expression levels of the breast cancer cell lines showed a distinct response to treatment with the chemotherapeutic agents (Figure 3). For MCF-7, a significant reduction of MYBL2 expression (by \sim 80 %) after treatment with the simultaneously given agents (T + A (48 h)) was observed. Furthermore, the exposure to 1 μ M A alone revealed a significant reduction of MYBL2 expression. The triple negative breast cancer cell line, BT-20, and the HER2 positive one, SKBR3, displayed a strong response after the simultaneous treatment with both chemotherapeutics (Figure 3(b)). The expression levels of MYBL2 in BT-20 cells decreased up to approximately 80 %. For SKBR3 cells, a reduction of 95 % for MYBL2 protein was reached. But in contrast to MCF-7, the single agents also influenced the protein contents of BT-20 and SKBR3. BT-20 cells showed a strong downregulation of MYBL2 protein after 0.1 μ M T or 1 μ M A exposure, similar to their combined application. The successive exposure with chemotherapeutic agents did not further enhance the protein repression in BT-20 cells. The influence on SKBR3 cells turned out to be somewhat different. Although the single chemotherapeutics paclitaxel and doxorubicin decreased MYBL2 expression significantly, the highest downregulation was reached after combined or successive treatment. In conclusion, the combined exposure of paclitaxel and doxorubicin (T + A (48 h)) revealed the strongest response on MYBL2 repression in the breast cancer cell lines (MCF-7, BT-20, SKBR3) while non-tumorigenic control cells (MCF-10A) were not affected. All three tested breast cancer cell lines were sensitive for the combined treatment of both chemotherapeutic adjuvants, as reflected by decreased PCNA

expression, except for the lowest treatment conditions in MCF-7 cells. However, SKBR3 and not BT-20 cells showed the most sensitive response to the combined chemotherapeutic treatment with paclitaxel and doxorubicin in concordance with repression of MYBL2 contents. Therefore, further investigations should be performed on MYBL2 as a marker for TFAC chemotherapy response in breast cancer cells, in particular in HER2-positive cells.

Currently, TFAC therapy is preferably used for triple-negative breast tumors [13]. Therefore, we also analyzed the cytotoxic potential of the combined treatment of paclitaxel (T) and doxorubicin (A) in the triple-negative cell line BT-20 in comparison to the non-tumorigenic control (MCF-10A) (Supp. Figure 1). Towards this end, we decided to use three independent cytotoxic measurement methods (MTS assay, Live-Dead-assay and cell cycle measurements for proliferation and apoptosis determination). The MTS assay reflects the influence on the metabolic viability of the cells, while the live-dead-test directly provides information about the induction of apoptosis. Viability was significantly lowered after treatment with A alone and in combination with T in both cell lines (Fig. 3A). In contrast, the live-dead staining revealed a significant higher level of apoptotic BT-20 cells after exposure to $1 \mu\text{M}$ A and the combined treatment with T (Supp. Figure 1 B,D).

Finally, we analyzed the cell cycle phases by flow cytometry since paclitaxel stabilizes microtubules and induces a G2-phase arrest. As expected, T alone induced an arrest in the G2-phase, leading to higher proliferation rates (G2/M + S phase) in both cell lines. Treatment with A or the combined exposure of both agents showed a significant decrease of the proliferative phases in BT-20 cells, while MCF-10A proliferation was stimulated. This effect on MCF-10A cells is not unusual, since, in an epithelial tissue, apoptosis is often compensated for by increased proliferation rates to maintain the tissue structure. These three cytotoxic assays confirm the postulated effects of T and A on the BT-20 cell line, a representative of the triple-negative breast cancer subtype. Furthermore, these results validate the MYBL2 western blotting experiments, demonstrating that the combined exposure to T and A leads to the strongest effects.

We can conclude that high expression levels of MYBL2 are associated with response to TFAC treatment, which should be verified in further experiments by the investigation of human tissue material. The results of the bioinformatics analysis are consistent with cell biological results concerning the downregulation of MYBL2 protein induced by TFAC treatment, rendering MYBL2 as a potential breast cancer marker for a successful TFAC therapy.

Conclusions

We have applied *ExprEssence*, a software tool for extraction of condition-related interactions from a functional interaction network, to preoperative breast cancer chemotherapy response. We identified interactions that were already known to be related to TFAC therapy response. Further, we proposed a putative response-related mechanism via MELK and MYBL2 which has not been taken into account yet for assessment of chances of response. We performed experiments with cell lines representing various breast cancer subtypes to test our hypothesis that paclitaxel acts synergistically with doxorubicin via suppression of MELK, which in turn attenuates MYBL2 gene expression, known to be advantageous for

Subtype	Markers	Prevalence	Cell line
Luminal	ER ⁺ and/or PR ⁺ , HER2 ⁻ , low Ki67	42-59 %	MCF-7
Triple negative	ER ⁻ , PR ⁻ , HER2 ⁻ , cytokeratin 5/6 ⁺	14-20 %	BT-20
HER2 ⁺	ER ⁻ , PR ⁻ , HER2 ⁺	7-12 %	SKBR3
Non-tumorigenic/basal-like/ normal breast-like	ER ^{+/-} and/or PR ^{+/-} , HER2 ⁻ ,	—	MCF-10A

Tab. 2. Selected breast cancer subtypes with their most common marker profile, their overall prevalence and a representative human cell line with these molecular features. This table was compiled from different sources [61–64]. ER: Estrogen receptor; PR: Progesterone receptor; HER2: human epidermal growth factor receptor 2; +: positive; -: negative.

TFAC response. Though, probably due to low amounts, we were not able to detect MELK protein, we could demonstrate attenuated MYBL2 protein levels in TFAC treated cells and a synergism of paclitaxel and doxorubicin.

Our method relies on a network of gene/protein interactions onto which gene expression data is mapped. A disadvantage of starting with a known network is that we are not able to discover novel interactions and hence false negatives may arise. Also, gene expression data may not reflect post-translational modifications such as phosphorylations. Nevertheless, we can generate valuable hypotheses based on highlighting some interactions as particularly relevant. These may be false positives, since network interactions are context-dependent events, and gene expression data may give false evidence in cases where changes of gene expression are irrelevant. Thus, the highlighted interactions we found may not give a complete picture, and they need to be validated experimentally.

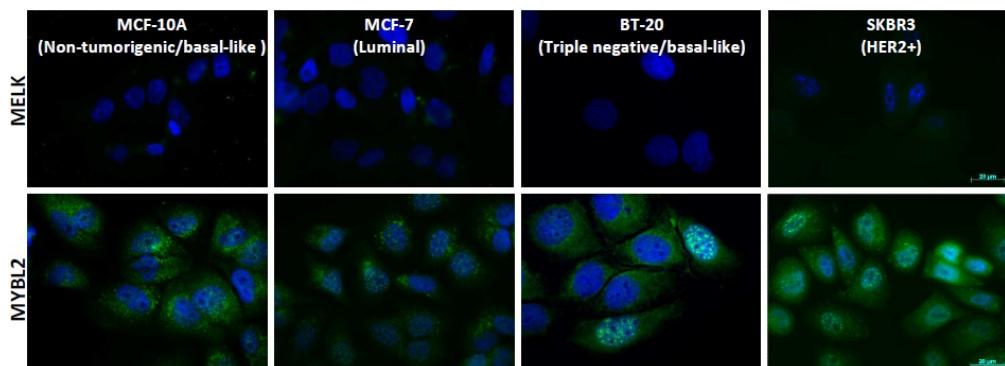
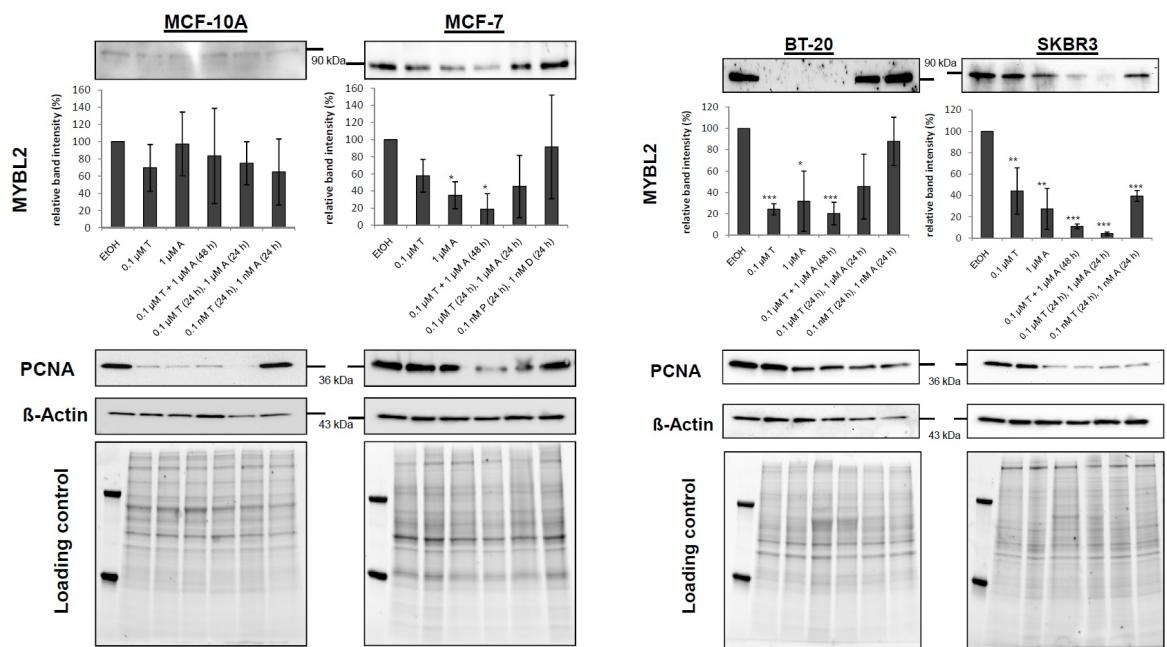


Fig. 2. Expression levels of MELK and MYBL2 protein in the non-tumorigenic cell line MCF-10A in contrast to the breast cancer cell lines MCF-7, BT-20 and SKBR3 detected by immunofluorescence. Note that MELK protein levels were below detection threshold while MYBL2 protein was abundant in all cell lines. The strongest MYBL2 signal was reached in the cell line SKBR3. MELK and MYBL2 protein: green; cell nuclei: blue.



(a) Non-tumorigenic cell line MCF-10A and the breast cancer cell line MCF-7.

(b) Estrogen receptor negative breast cancer cell line BT-20 and the HER2 positive breast cancer cell line SKBR3.

Fig. 3. Expression analysis of MYBL2 protein after treatment with paclitaxel (Taxol, T) and doxorubicin (Adriamycin, A) in several cell lines by Western blotting. Single treatment with T or A for 48 h (T (48 h); A (48 h)), combined treatment for 48 h (T + A (48 h)) or successive treatment for each for 24 h (T (24 h), A (24 h)) was applied. Quantification of western blotting results was carried out with individual passaged cells for at least three times. Representative western blots were displayed on the top the graphs. Proliferative alterations were detected against Proliferating Cell Nuclear Antigen (PCNA). Loading controls were labeling of the house keeping protein β-actin and stain-free imaging of the SDS-PAGEs prior blotting procedure. Mean \pm SD values ($n = 3$). * : $p < 0.05$; ** : $p < 0.01$; *** : $p < 0.001$ as compared to control treatment (unpaired t test).

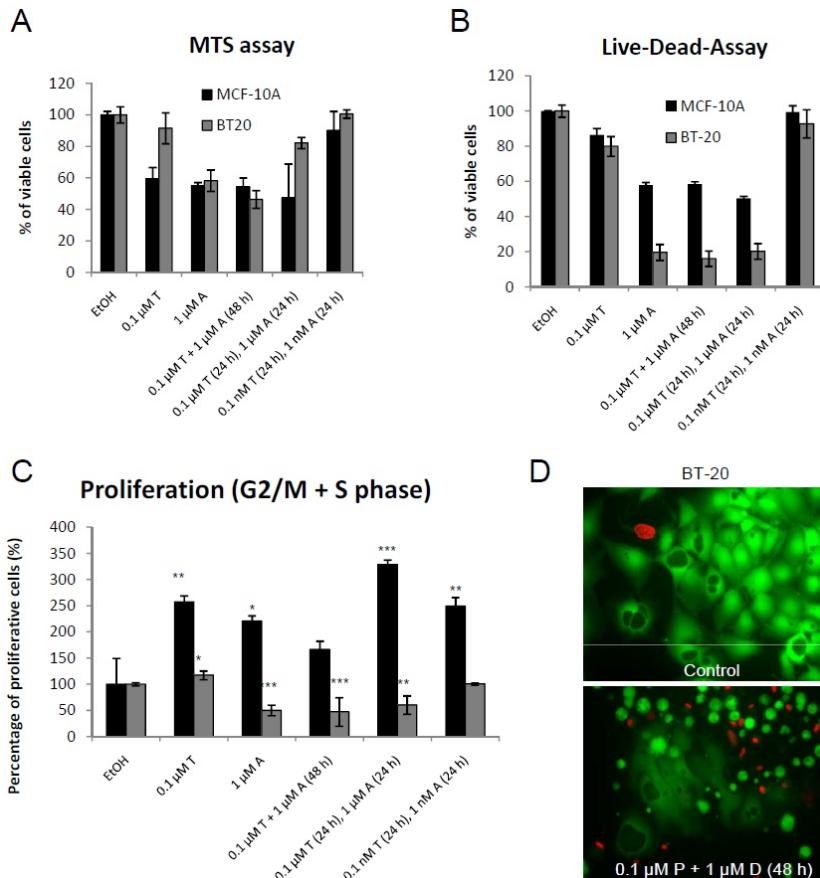


Fig. 4. Cytotoxic activity on non-tumorigenic control cell line MCF-10A (black bar) and triple negative breast cancer cell line BT-20 (grey bar) after treatment with paclitaxel (P) and doxorubicin (D) was calculated by three individual assays: MTS (A), Live-Dead (B, D) and Cell cycle analysis (C). In each measurement the control treatment with 0.1 % EtOH was set to 100 % to validate the results after exposure to the compounds. All measurements were repeated at a minimum of three replicates. Fluorescence pictures of live (green) and dead (red) stained cells were taken with a fluorescence microscope (Axio Scope. A1, Carl Zeiss, Germany). Mean \pm SD values ($n = 3$). * : $p < 0.05$; ** : $p < 0.01$; *** : $p < 0.001$ as compared to control treatment (unpaired t test).

Acknowledgments

We would like to thank Hermann Ragg, Bielefeld, for advice on Serpins, and Dirk Repsilber, FBN Dummerstorf, Germany, for his help with the IPA analysis.

References

1. Popovici V, Chen W, Gallas BG, Hatzis C, Shi W, et al. (2010) Effect of training-sample size and classification difficulty on the accuracy of genomic predictors. *Breast Cancer Res* 12: R5.
2. Warsow G, Greber B, Falk SSI, Harder C, Siatkowski M, et al. (2010) Expressence—revealing the essence of differential experimental data in the context of an interaction/regulation net-work. *BMC Syst Biol* 4: 164.
3. Cabusora L, Sutton E, Fulmer A, Forst CV (2005) Differential network expression during drug and stress response. *Bioinformatics* 21: 2898–2905.
4. Nacu S, Critchley-Thorne R, Lee P, Holmes S (2007) Gene expression network analysis and applications to immunology. *Bioinformatics* 23: 850–858.
5. Dao P, Wang K, Collins C, Ester M, Lapuk A, et al. (2011) Optimally discriminative subnetwork markers predict response to chemotherapy. *Bioinformatics* 27: i205–i213.
6. Ideker T, Ozier O, Schwikowski B, Siegel AF (2002) Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* 18 Suppl 1: S233–S240.
7. Baumbach J, Friedrich T, Kötzting T, Krohmer A, Müller J, et al. (2012) Efficient algorithms for extracting biological key pathways with global constraints. In: Proceedings of the fourteenth international conference on Genetic and evolutionary computation conference. ACM, pp. 169–176. URL <http://dl.acm.org/citation.cfm?id=2330188>.
8. Buzdar AU, Singletary SE, Valero V, Booser DJ, Ibrahim NK, et al. (2002) Evaluation of paclitaxel in adjuvant chemotherapy for patients with operable breast cancer: preliminary data of a prospective randomized trial. *Clin Cancer Res* 8: 1073–1079.
9. Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, et al. (2011) The string database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res* 39: D561–D568.
10. Saito R, Smoot ME, Ono K, Ruscheinski J, Wang PL, et al. (2012) A travel guide to cytoscape plugins. *Nat Methods* 9: 1069–1076.
11. Bradford MM (1976) A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal Biochem* 72: 248–254.

12. Engel N, Liseck J, Piechulla B, Nebe B (2012) Metabolic profiling reveals sphingosine-1-phosphate kinase 2 and lyase as key targets of (phyto-) estrogen action in the breast cancer cell line mcf-7 and not in mcf-12a. *PLoS One* 7: e47833.
13. Liedtke C, Mazouni C, Hess KR, Andr F, Tordai A, et al. (2008) Response to neoadjuvant therapy and long-term survival in patients with triple-negative breast cancer. *J Clin Oncol* 26: 1275–1281.
14. Eeckhoute J, Keeton EK, Lupien M, Krum SA, Carroll JS, et al. (2007) Positive cross-regulatory loop ties gata-3 to estrogen receptor alpha expression in breast cancer. *Cancer Res* 67: 6477–6483.
15. Chou J, Provost S, Werb Z (2010) Gata3 in development and cancer differentiation: cells gata have it! *J Cell Physiol* 222: 42–49.
16. Abba MC, Nunez MI, Colussi AG, Croce MV, Segal-Eiras A, et al. (2006) Gata3 protein as a muc1 transcriptional regulator in breast cancer cells. *Breast Cancer Res* 8: R64.
17. Taube JH, Allton K, Duncan SA, Shen L, Barton MC (2010) Foxa1 functions as a pioneer transcription factor at transposable elements to activate afp during differentiation of embryonic stem cells. *J Biol Chem* 285: 16135–16144.
18. Jacquemier J, Charafe-Jauffret E, Monville F, Esterni B, Extra JM, et al. (2009) Association of gata3, p53, ki67 status and vascular peritumoral invasion are strongly prognostic in luminal breast cancer. *Breast Cancer Res* 11: R23.
19. Eeckhoute J, Carroll JS, Geistlinger TR, Torres-Arzayus MI, Brown M (2006) A cell-type-specific transcriptional network required for estrogen regulation of cyclin d1 and cell cycle progression in breast cancer. *Genes Dev* 20: 2513–2526.
20. Laganire J, Deblois G, Lefebvre C, Bataille AR, Robert F, et al. (2005) From the cover: Location analysis of estrogen receptor alpha target promoters reveals that foxa1 defines a domain of the estrogen response. *Proc Natl Acad Sci U S A* 102: 11651–11656.
21. Westley BR, May FEB (2006) Identification of steroid hormone-regulated genes in breast cancer. *Methods Mol Med* 120: 363–388.
22. Dignass A, Lynch-Devaney K, Kindon H, Thim L, Podolsky DK (1994) Trefoil peptides promote epithelial migration through a transforming growth factor beta-independent pathway. *J Clin Invest* 94: 376–383.
23. Long BH, Fairchild CR (1994) Paclitaxel inhibits progression of mitotic cells to g1 phase by interference with spindle formation without affecting other microtubule functions during anaphase and telephase. *Cancer Res* 54: 4355–4361.
24. Hendershot LM, Bulleid NJ (2000) Protein-specific chaperones: the role of hsp47 begins to gel. *Curr Biol* 10: R912–R915.

25. Ragg H (2007) The role of serpins in the surveillance of the secretory pathway. *Cell Mol Life Sci* 64: 2763–2770.
26. Fields GB (1991) A model for interstitial collagen catabolism by mammalian collagenases. *J Theor Biol* 153: 585–602.
27. Enerbck C, Porter DA, Seth P, Sgroi D, Gaudet J, et al. (2002) Psoriasin expression in mammary epithelial cells in vitro and in vivo. *Cancer Res* 62: 43–47.
28. Emberley ED, Alowami S, Snell L, Murphy LC, Watson PH (2004) S100a7 (psoriasin) expression is associated with aggressive features and alteration of jab1 in ductal carcinoma in situ of the breast. *Breast Cancer Res* 6: R308–R315.
29. Emberley ED, Niu Y, Curtis L, Troup S, Mandal SK, et al. (2005) The s100a7-c-jun activation domain binding protein 1 pathway enhances prosurvival pathways in breast cancer. *Cancer Res* 65: 5696–5702.
30. Paruchuri V, Prasad A, McHugh K, Bhat HK, Polyak K, et al. (2008) S100a7-downregulation inhibits epidermal growth factor-induced signaling in breast cancer cells and blocks osteoclast formation. *PLoS One* 3: e1741.
31. Armstrong DK, Kaufmann SH, Ottaviano YL, Furuya Y, Buckley JA, et al. (1994) Epidermal growth factor-mediated apoptosis of mda-mb-468 human breast cancer cells. *Cancer Res* 54: 5280–5283.
32. Deol YS, Nasser MW, Yu L, Zou X, Ganju RK (2011) Tumor-suppressive effects of psoriasin (s100a7) are mediated through the -catenin/t cell factor 4 protein pathway in estrogen receptor-positive breast cancer cells. *J Biol Chem* 286: 44845–44854.
33. Yu SE, Jang YK (2012) The histone demethylase lsd1 is required for estrogen-dependent s100a7 gene expression in human breast cancer cells. *Biochem Biophys Res Commun* 427: 336–342.
34. Zhang S, Wang G, Liu D, Bao Z, Fernig DG, et al. (2005) The c-terminal region of s100a4 is important for its metastasis-inducing properties. *Oncogene* 24: 4401–4411.
35. Wang G, Zhang S, Fernig DG, Martin-Fernandez M, Rudland PS, et al. (2005) Mutually antagonistic actions of s100a4 and s100a1 on normal and metastatic phenotypes. *Oncogene* 24: 1445–1454.
36. Shaw LM (2011) The insulin receptor substrate (irs) proteins: at the intersection of metabolism and cancer. *Cell Cycle* 10: 1750–1756.
37. Morandi A, Plaza-Menacho I, Isacke CM (2011) Ret in breast cancer: functional and therapeutic implications. *Trends Mol Med* 17: 149–157.
38. Brychtova V, Vojtesek B, Hrstka R (2011) Anterior gradient 2: a novel player in tumor cell biology. *Cancer Lett* 304: 1–7.

39. Fletcher GC, Patel S, Tyson K, Adam PJ, Schenker M, et al. (2003) hag-2 and hag-3, human homologues of genes involved in differentiation, are associated with oestrogen receptor-positive breast tumours and interact with metastasis gene c4.4a and dystroglycan. *Br J Cancer* 88: 579–585.
40. Maslon MM, Hrstka R, Vojtesek B, Hupp TR (2010) A divergent substrate-binding loop within the pro-oncogenic protein anterior gradient-2 forms a docking site for reptin. *J Mol Biol* 404: 418–438.
41. Thorner AR, Hoadley KA, Parker JS, Winkel S, Millikan RC, et al. (2009) In vitro and in vivo analysis of b-myb in basal-like breast cancer. *Oncogene* 28: 742–751.
42. Pickard MR, Green AR, Ellis IO, Caldas C, Hedge VL, et al. (2009) Dysregulated expression of fau and melk is associated with poor prognosis in breast cancer. *Breast Cancer Res* 11: R60.
43. Raschell G, Cesi V, Amendola R, Negroni A, Tanno B, et al. (1999) Expression of b-myb in neuroblastoma tumors is a poor prognostic factor independent from mycn amplification. *Cancer Res* 59: 3365–3368.
44. Joaquin M, Watson RJ (2003) The cell cycle-regulated b-myb transcription factor overcomes cyclin-dependent kinase inhibitory activity of p57(kip2) by interacting with its cyclin-binding domain. *J Biol Chem* 278: 44255–44264.
45. Lin D, Fiscella M, O'Connor PM, Jackman J, Chen M, et al. (1994) Constitutive expression of b-myb can bypass p53-induced waf1/cip1-mediated g1 arrest. *Proc Natl Acad Sci U S A* 91: 10079–10083.
46. Sala A, Casella I, Bellon T, Calabretta B, Watson RJ, et al. (1996) B-myb promotes s phase and is a downstream target of the negative regulator p107 in human cells. *J Biol Chem* 271: 9363–9367.
47. Liekens AML, De Knijf J, Daelemans W, Goethals B, De Rijk P, et al. (2011) Biograph: unsupervised biomedical knowledge discovery via automated hypothesis generation. *Genome Biol* 12: R57.
48. Davis AP, Murphy CG, Johnson R, Lay JM, Lennon-Hopkins K, et al. (2013) The comparative toxicogenomics database: update 2013. *Nucleic Acids Res* 41: D1104–D1114.
49. Hess KR, Anderson K, Symmans WF, Valero V, Ibrahim N, et al. (2006) Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. *J Clin Oncol* 24: 4236–4244.
50. Nakano I, Paucar AA, Bajpai R, Dougherty JD, Zewail A, et al. (2005) Maternal embryonic leucine zipper kinase (melk) regulates multipotent neural progenitor proliferation. *J Cell Biol* 170: 413–427.
51. Blagosklonny MV, Schulte TW, Nguyen P, Mimnaugh EG, Trepel J, et al. (1995) Taxol induction of p21waf1 and p53 requires c-raf-1. *Cancer Res* 55: 4623–4626.

52. Brugarolas J, Bronson RT, Jacks T (1998) p21 is a critical cdk2 regulator essential for proliferation control in rb-deficient cells. *J Cell Biol* 141: 503–514.
53. Vantieghem K (2006) Photoproduction of Vitamin D3 & Activation into 1a, 25-dihydroxyvitamin D3 in Human Epidermal Keratinocytes, Dermal Fibroblasts & Other Cells. Number 371 in *Acta Biomedica Lovaniensia*. Leuven University Press, 117 pp.
54. Verlinden L, Eelen G, Beullens I, Van Camp M, Van Hummelen P, et al. (2005) Characterization of the condensin component cnap1 and protein kinase melk as novel e2f target genes down-regulated by 1,25-dihydroxyvitamin d3. *J Biol Chem* 280: 37319–37330.
55. Seong HA, Gil M, Kim KT, Kim SJ, Ha H (2002) Phosphorylation of a novel zinc-finger-like protein, zpr9, by murine protein serine/threonine kinase 38 (mpk38). *Biochem J* 361: 597–604.
56. Seong HA, Kim KT, Ha H (2003) Enhancement of b-myb transcriptional activity by zpr9, a novel zinc finger protein. *J Biol Chem* 278: 9655–9662.
57. Calvisi DF, Simile MM, Ladu S, Frau M, Evert M, et al. (2011) Activation of v-myb avian myeloblastosis viral oncogene homolog-like2 (mybl2)-lin9 complex contributes to human hepatocarcinogenesis and identifies a subset of hepatocellular carcinoma with mutant p53. *Hepatology* 53: 1226–1236.
58. Mannefeld M, Klassen E, Gaubatz S (2009) B-myb is required for recovery from the dna damage-induced g2 checkpoint in p53 mutant cells. *Cancer Res* 69: 4073–4080.
59. Hebbard LW, Maurer J, Miller A, Lesperance J, Hassell J, et al. (2010) Maternal embryonic leucine zipper kinase is upregulated and required in mammary tumor-initiating cells in vivo. *Cancer Res* 70: 8863–8873.
60. Tegze B, Szllsi Z, Haltrich I, Pnzvlt Z, Tth Z, et al. (2012) Parallel evolution under chemotherapy pressure in 29 breast cancer cell lines results in dissimilar mechanisms of resistance. *PLoS One* 7: e30804.
61. Carey LA, Perou CM, Livasy CA, Dressler LG, Cowan D, et al. (2006) Race, breast cancer subtypes, and survival in the carolina breast cancer study. *JAMA* 295: 2492–2502.
62. Hannemann J, Kristel P, van Tinteren H, Bontenbal M, van Hoesel QGCM, et al. (2006) Molecular subtypes of breast cancer and amplification of topoisomerase ii alpha: predictive role in dose intensive adjuvant chemotherapy. *Br J Cancer* 95: 1334–1341.
63. Lund MJ, Butler EN, Hair BY, Ward KC, Andrews JH, et al. (2010) Age/race differences in her2 testing and in incidence rates for breast cancer triple subtypes: a population-based study and first report. *Cancer* 116: 2549–2559.
64. Yang XR, Chang-Claude J, Goode EL, Couch FJ, Nevanlinna H, et al. (2011) Associations of breast cancer risk factors with tumor subtypes: a pooled analysis from the breast cancer association consortium studies. *J Natl Cancer Inst* 103: 250–263.

Supplement

Supp. Figure 1. Cytotoxic activity on non-tumorigenic control cell line MCF-10A (black bar) and triple negative breast cancer cell line BT-20 (grey bar) after treatment with paclitaxel (T) and doxorubicin (A) was calculated by three individual assays: MTS (A), Live-Dead (B, D) and Cell cycle analysis (C). In each measurement the control treatment with 0.1 % EtOH was set to 100 % to validate the results after exposure to the compounds. All measurements were repeated at a minimum of three replicates. Fluorescence pictures of live (green) and dead (red) stained cells were taken with a fluorescence microscope (Axio Scope. A1, Carl Zeiss, Germany). Mean \pm SD values ($n = 3$). * : $p < 0.05$; ** : $p < 0.01$; *** : $p < 0.001$ as compared to control treatment (unpaired t test).

Functions Annotation	P-value	Molecules
Carcinoma	2.00E-07	AGR2, AR, CCND1, CDC20, CDCA8, COL1A2, ESR1, FOXA1, GATA3, GFRA1, IRS1, KIF20A, KIFC1, KRT16, KRT7, MELK, MYB, MYBL2, PGR, PLAT, S100A1, SERPINA3, SERPINA5, SERPINA6, TFF1, TFF3, ZEB1
Cell cycle progression	2.00E-07	AR, CCND1, CDC20, CDCA8, CENPE, ESR1, FOXA1, GATA3, HJURP, IRS1, KIFC1, MELK, MYB, MYBL2, NDC80, PGR, SERPINA5, ZEB1
Chromosomal congression of chromosomes	1.58E-06	CENPE, KIF18A, KIFC1, NDC80
Cancer	1.58E-06	AGR2, AR, CCND1, CDC20, CDCA8, COL1A2, ESR1, FOXA1, GATA3, GFRA1, IRS1, KIF20A, KIFC1, KRT15, KRT16, KRT17, KRT7, MELK, MYB, MYBL2, NDC80, PGR, PLAT, S100A1, SERPINA3, SERPINA5, SERPINA6, TFF1, TFF3, ZEB1
Carcinoma in breast	1.58E-06	AGR2, AR, CCND1, COL1A2, ESR1, GATA3, GFRA1, MYBL2, PGR, SERPINA5, TFF1
Breast cancer	1.58E-06	AGR2, AR, CCND1, CDC20, COL1A2, ESR1, FOXA1, GATA3, GFRA1, KRT15, KRT17, MYB, MYBL2, PGR, SERPINA5, TFF1
Digestive organ tumor	1.89E-05	AR, CCND1, CDC20, CDCA8, COL1A2, ESR1, FOXA1, IRS1, KIF20A, KIFC1, KRT7, MELK, MYB, PGR, PLAT, SERPINA3, SERPINA6, TFF1, TFF3
Proliferation of cells	3.49E-05	AGR2, AR, CCND1, CDC20, CDCA8, COL1A2, ESR1, FOXA1, GATA3, GFRA1, IRS1, KIF20A, KRT16, KRT17, MELK, MYB, MYBL2, PGR, PLAT, S100A1, S100A7, SERPINA5, SERPINH1, TFF1, TFF3, ZEB1
Proliferation of tumor cells	6.25E-05	AR, CCND1, ESR1, GFRA1, IRS1, MYB, MYBL2, PLAT, S100A7, TFF3
Plaque psoriasis	6.25E-05	CCND1, GATA3, KRT15, KRT16, KRT17, S100A7

Continued on next page

Supp. Table 1 – continued from previous page

Functions Annotation	P-value	Molecules
Cell movement	7.72E-05	AGR2, AR, CCND1, ESR1, FOXA1, GATA3, GFRA1, IRS1, KRT16, LYPD3, MYB, PLAT, S100A1, S100A7, SERPINA3, SERPINA5, TFF1, TFF3, ZEB1
Proliferation of breast cancer cell lines	1.00E-04	AR, CCND1, ESR1, IRS1, MYB, PGR, S100A1, S100A7, ZEB1
Amenorrhea	2.46E-04	AR, PGR, SERPINA6
Cell cycle progression of tumor cell lines	2.65E-04	AR, CCND1, ESR1, FOXA1, MELK, PGR, ZEB1
Proliferation of cancer cells	2.98E-04	ESR1, GFRA1, IRS1, MYB, MYBL2, PLAT, S100A7, TFF3
Hyperplasia	3.32E-04	AR, CCND1, ESR1, GATA3, IRS1, KRT16, KRT17, MYB, PGR, S100A7
Mitosis	3.98E-04	AR, CCND1, CDC20, CDCA8, CENPE, IRS1, KIFC1, MYBL2, NDC80
Skin development	4.33E-04	CCND1, COL1A2, GFRA1, KRT15, KRT16, KRT17, S100A7
Development of epidermis	4.92E-04	CCND1, GFRA1, KRT15, KRT16, KRT17, S100A7
Cell cycle progression of breast cancer cell lines	5.04E-04	CCND1, ESR1, MELK, PGR
Cell viability of endometrial cells	5.06E-04	ESR1, PGR
Cloning of fibroblast cell lines	5.06E-04	CCND1, MYBL2
Infanticide	5.06E-04	ESR1, PGR
Mammary gland development	5.09E-04	AR, CCND1, ESR1, GATA3, IRS1, PGR
Carcinoma in situ	5.09E-04	AR, ESR1, GATA3, KRT16, KRT7, TFF3

Supp. Table 1. Top 25 terms of Ingenuity Functional Enrichment Analysis for E40 gene set including associated genes and P-values (corrected for multiple testing using BenjaminiHochberg correction).

Functions Annotation	P-value	Molecules
Transactivation	2.31E-12	AR, EP300, ESR1, GADD45G, GATA3, IGFBP4, MAPK3, MDM2, MED1, MED31, NCOA3, PRKACA, RARA, RET, SMAD3, SRC, STUB1, UBE2I, ZBTB16
Cell movement	4.89E-11	AGR2, AR, C4B (includes others), CST3, ESR1, EVL, GATA3, GFRA1, IGF1R, IGF2, IGFBP4, IL6ST, IRS1, MAPK3, MED1, NCOA3, PRKACA, RABEP1, RARA, RET, SEMA6A, SHC1, SMAD3, SRC, TSC2, UBE2I, ZBTB16
Cell cycle progression	2.37E-10	AR, EP300, ESR1, GADD45G, GATA3, IGF1R, IGF2, IRS1, MAPK3, MDM2, NCOA3, PRKACA, RARA, RET, SHC1, SMAD3, SRC, TSC2, UBE2I, YWHAB, ZBTB16
Migration of cells	2.44E-10	AR, C4B (includes others), CST3, ESR1, EVL, GATA3, GFRA1, IGF1R, IGF2, IGFBP4, IL6ST, IRS1, MAPK3, NCOA3, PRKACA, RABEP1, RARA, RET, SEMA6A, SHC1, SMAD3, SRC, TSC2, UBE2I, ZBTB16
Proliferation of tumor cell lines	1.96E-09	AR, EP300, ESR1, GADD45G, IGF1R, IGF2, IGFBP4, IL6ST, IRS1, MAPK3, MDM2, MED1, NCOA3, PRKACA, RARA, RET, SHC1, SMAD3, SRC, STUB1, UBE2I, ZBTB16

Continued on next page

Supp. Table 2 – continued from previous page

Functions Annotation	P-value	Molecules
Apoptosis of tumor cell lines	3.95E-09	AR, EP300, ESR1, GADD45G, IGF1R, IGF2, IGFBP4, IL6ST, MAPK3, MDM2, NCOA3, PRKACA, RARA, RET, SHC1, SMAD3, SRC, STUB1, TSC2, ZBTB16
Proliferation of connective tissue cells	5.15E-09	AR, ESR1, IGF1R, IGF2, IGFBP4, IL6ST, IRS1, MAPK3, MDM2, MED1, RARA, RET, SMAD3, SRC, TSC2
Proliferation of cells	5.15E-09	AGR2, AR, ASH2L, CALM1 (includes others), CST3, EP300, ESR1, GADD45G, GATA3, GFRA1, IGF1R, IGF2, IGFBP4, IL6ST, IRS1, MAPK3, MDM2, MED1, NCOA3, PRKACA, RABEP1, RARA, RET, SEMA6A, SHC1, SMAD3, SRC, STUB1, TSC2, UBE2I, ZBTB16
Necrosis	5.15E-09	AGR2, AR, CST3, EP300, ESR1, GADD45G, GATA3, GFRA1, IGF1R, IGF2, IGFBP4, IL6ST, IRS1, MAPK3, MDM2, MED1, NCOA3, PRKACA, RARA, RET, SHC1, SMAD3, SRC, STUB1, TSC2, YWHAB, ZBTB16
Hypoplasia	5.77E-09	AR, ESR1, IGF1R, IGF2, IL6ST, MDM2, MED1, RARA, RET, SHC1, SMAD3, SRC, STUB1, TSC2
Proliferation of epithelial cells	5.77E-09	AGR2, AR, EP300, ESR1, IGF1R, IGF2, IGFBP4, MED1, RARA, RET, SMAD3, TSC2, ZBTB16
Transcription	6.11E-09	AP1G2, AR, ASH2L, EP300, ESR1, GADD45G, GATA3, IGF2, IL6ST, MAPK3, MDM2, MED1, NCOA3, PRKACA, RARA, RET, SHC1, SMAD3, SMAD9, SRC, STUB1, UBE2I, YWHAB, ZBTB16
Cell death of tumor cell lines	6.78E-09	AGR2, AR, EP300, ESR1, GADD45G, IGF1R, IGF2, IGFBP4, IL6ST, MAPK3, MDM2, NCOA3, PRKACA, RARA, RET, SHC1, SMAD3, SRC, STUB1, TSC2, ZBTB16
Differentiation of cells	8.30E-09	AR, EP300, ESR1, GADD45G, GATA3, GFRA1, IGF1R, IGF2, IL6ST, IRS1, MAPK3, MDM2, MED1, NCOA3, PRKACA, RARA, RET, SHC1, SMAD3, SMAD9, SRC, TSC2, UBE2I, ZBTB16
Organismal death	1.29E-08	AR, C4B (includes others), CST3, EP300, ESR1, GATA3, IGF1R, IGF2, IL6ST, MAPK3, MDM2, MED1, NCOA3, PRKACA, RARA, RET, SHC1, SMAD3, SMAD9, SRC, STUB1, TSC2
Cell survival	1.29E-08	AR, EP300, ESR1, GATA3, GFRA1, IGF1R, IGF2, IL6ST, IRS1, MAPK3, MDM2, MED1, PRKACA, RARA, RET, SHC1, SMAD3, SRC, TSC2, UBE2I
Synthesis of DNA	1.32E-08	AR, EP300, ESR1, IGF1R, IGF2, IGFBP4, IL6ST, IRS1, MDM2, PRKACA, SHC1, SRC, TSC2
Transcription of RNA	2.48E-08	AR, ASH2L, EP300, ESR1, GADD45G, GATA3, IGF2, IL6ST, MAPK3, MDM2, MED1, NCOA3, PRKACA, RARA, RET, SHC1, SMAD3, SMAD9, SRC, STUB1, UBE2I, YWHAB, ZBTB16
Cell viability	2.62E-08	AR, EP300, ESR1, GATA3, GFRA1, IGF1R, IGF2, IL6ST, MAPK3, MDM2, MED1, PRKACA, RARA, RET, SHC1, SMAD3, SRC, TSC2, UBE2I
Binding of DNA	2.80E-08	AR, CALM1 (includes others), EP300, ESR1, GATA3, IL6ST, MAPK3, MDM2, MED1, PRKACA, RARA, SMAD3, SRC, YWHAB

Continued on next page

Supp. Table 2 – continued from previous page

Functions Annotation	P-value	Molecules
Quantity of cells	3.81E-08	AGR2, AR, C4B (includes others), EP300, ESR1, GADD45G, GATA3, GFRA1, IGF1R, IGF2, IL6ST, IRS1, MAPK3, MDM2, MED1, RARA, RET, SHC1, SMAD3, SRC, TSC2, ZBTB16
Development of carcinoma	3.81E-08	CST3, NCOA3, RARA, RET, SMAD3, SRC, TSC2
Proliferation of fibroblasts	4.76E-08	ESR1, IGF1R, IGF2, IGFBP4, IL6ST, MAPK3, MDM2, RARA, SMAD3, SRC, TSC2
Development of tumor	4.76E-08	CST3, ESR1, IGF1R, IL6ST, NCOA3, RARA, RET, SMAD3, SRC, TSC2, ZBTB16
Morphology of embryonic tissue	5.03E-08	AR, EP300, ESR1, GATA3, GFRA1, IL6ST, MDM2, PRKACA, RARA, RET, SHC1, SMAD3, SMAD9, UBE2I

Supp. Table 2. Top 25 terms of Ingenuity Functional Enrichment Analysis for O39 gene set including associated genes and P-values (corrected for multiple testing using BenjaminiHochberg correction).

Gene Symbol	Gene Name
AGR2	anterior gradient 2 homolog (Xenopus laevis)
AR	androgen receptor
CCND1	cyclin D1
CDC20	cell division cycle 20
CDCA8	cell division cycle associated 8
CENPE	centromere protein E, 312kDa
CENPN	centromere protein N
COL1A2	collagen, type I, alpha 2
CPB1	carboxypeptidase B1 (tissue)
ESR1	estrogen receptor 1
FOXA1	forkhead box A1
GATA3	GATA binding protein 3
GFRA1	GDNF family receptor alpha 1
HJURP	Holliday junction recognition protein
IRS1	insulin receptor substrate 1
KIF18A	kinesin family member 18A
KIF20A	kinesin family member 20A
KIFC1	kinesin family member C1
KRT15	keratin 15
KRT16	keratin 16
KRT17	keratin 17
KRT7	keratin 7

Continued on next page

Supp. Table 3 – continued from previous page

Gene Symbol	Gene Name
LYPD3	LY6/PLAUR domain containing 3
MELK	maternal embryonic leucine zipper kinase
MID1	midline 1 (Opitz/BBB syndrome)
MYB	v-myb myeloblastosis viral oncogene homolog (avian)
MYBL2	v-myb myeloblastosis viral oncogene homolog (avian)-like 2
NDC80	NDC80 kinetochore complex component
PGR	progesterone receptor
PLAT	plasminogen activator, tissue
RUVBL2	RuvB-like 2 (<i>E. coli</i>)
S100A1	S100 calcium binding protein A1
S100A7	S100 calcium binding protein A7
SERPINA3	serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 3
SERPINA5	serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 5
SERPINA6	serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 6

Supp. Table 3. Full gene names for the gene symbols in Figure 1.

Teil III

Appendix

7 Hintergrundinformationen zu den Berechnungen

7.1 Details zur LinkScore-Berechnung durch ExprEssence

Der von *ExprEssence* berechnete LinkScore schätzt die Änderung der Stärke (Häufigkeit) einer Interaktion zweier Proteine beim Übergang von einer Bedingung in eine andere. Die Interaktion zwischen zwei Proteinen A und B wird hier vereinfachend als bimolekulare Elementarreaktion zwischen A und B betrachtet. Der Stoßtheorie folgend, müssen A und B aufeinandertreffen, um miteinander in Interaktion zu treten. Die Rate r dieser Elementarreaktion, somit also die Rate der Interaktion (im Folgenden als Interaktionsstärke, Auftretenshäufigkeit oder Aktivität bezeichnet), ist gemäß der Stoßtheorie proportional zum Produkt der Konzentrationen der an der Reaktion teilnehmenden Moleküle:

$$r = k \cdot [A] \cdot [B], k: \text{Stoßfaktor.} \quad (7.1)$$

Je ein Protein A und B werden für eine Interaktion (die Komplexbildung AB) dem Pool der einzeln vorliegenden Proteine entnommen und diesem bei Reversibilität der Bindung zwischen A und B nach „Bedingung“ der Interaktion wieder zugeführt. Es bildet sich ein Gleichgewicht zwischen den im Komplex und in freier Form vorliegenden Proteinen aus, das je nach Größe des Stoßfaktors - oder äquivalent - der Affinität der interagierenden Proteine zueinander, eher auf der Seite der einzelnen Proteine oder des interagierenden Proteinkomplexes liegt. Dieses Gleichgewicht lässt sich durch das Massenwirkungsgesetz beschreiben, das die Proportionalität der Konzentration des Komplexes AB zum Produkt der Konzentrationen von A und B ausdrückt:

$$[AB] = K \cdot [A] \cdot [B], K: \text{Gleichgewichtskonstante.} \quad (7.2)$$

Der Stoßfaktor sowie die Gleichgewichtskonstante sind in der Regel für Interaktionen zwischen verschiedenen Proteinen unterschiedlich.

Wie in Abschnitt 2.1 legitimiert, werden im Folgenden Genexpressionswerte direkt als Proteinmengen interpretiert. Des Weiteren wird von nun an direkt mit den Genexpressionswerten anstatt mit Konzentrationen von Proteinmengen gearbeitet. Die Unterscheidung ist unerheblich, da die Überführung von Proteinmengen in Konzentrationen durch Multiplikation mit einem für alle Proteine konstanten Faktor ω umgesetzt werden kann, was zu einem geänderten Faktor k aus Formel 7.1 führt und - wie in Formel 7.5 gezeigt

wird - dieser Faktor für die Berechnung des LinkScores vernachlässigbar ist.

$$\begin{aligned}
 \text{Sei } [A] &= \omega \cdot E_A. \text{ Dann gilt} & (7.3) \\
 r &= k \cdot (\omega \cdot E_A) \cdot (\omega \cdot E_B) \\
 &= k \cdot \omega^2 \cdot (E_A) \cdot (E_B) \\
 &= k' \cdot (E_A) \cdot (E_B), \text{ mit} \\
 k' &= k \cdot \omega^2.
 \end{aligned}$$

Im Folgenden wird mit zur Basis 2 logarithmierten Interaktionsstärken gearbeitet. Dies bringt den Vorteil mit sich, dass die ebenfalls log2-transformierten Genexpressionsdaten direkt genutzt werden können. Somit wird im weiteren Verlauf mit Summen und Differenzen normalverteilter Variablen gearbeitet, welche wiederum normalverteilt sind. Aus der Produktbildung der nichtlogarithmierten Expressionswerte von A und B gemäß Formel 7.1 wird somit die Summation der logarithmierten Werte. Im Sinne des besseren Verständnisses beinhaltet die Bezeichnung (I) für die Interaktionsstärke bereits die log2-Transformation. Es gilt dann:

$$\begin{aligned}
 I &= ld(r) = ld(k' \cdot (E_A) \cdot (E_B)) & (7.4) \\
 &= ld(k') + ld(E_A) + ld(E_B) \\
 &= ld(k') + E^A + E^B,
 \end{aligned}$$

wobei $ld(\cdot)$ den Logarithmus zur Basis 2 beschreibe und E^A und E^B die zur Basis 2 logarithmierten Genexpressionswerte der Proteine A und B seien.

Der LinkScore ist ein Maß für die Änderung der Interaktionsstärke beim Übergang von einer Bedingung in eine andere. Er berechnet sich als die Differenz der Interaktionsstärken zwischen der ersten und der zweiten Bedingung:

$$\begin{aligned}
 LS_{1 \rightarrow 2} &= I_2 - I_1 & (7.5) \\
 &= (ld(k') + E_2^A + E_2^B) - (ld(k') + E_1^A + E_1^B) \\
 &= (E_2^A + E_2^B) - (E_1^A + E_1^B),
 \end{aligned}$$

wobei $1 \rightarrow 2$ den Vergleich von der ersten zu der zweiten Bedingung beschreibe. Wie zu erkennen ist, entfallen die Summanden $ld(k')$ bei Bildung der Differenz. Mit diesem Wissen kann die Berechnung der Interaktionsstärke aus Formel 7.4 wie folgt vereinfacht werden:

$$I = E^A + E^B. \quad (7.6)$$

Die LinkScores sind direkt interpretierbar: Ein LinkScore von +1 entspricht einer Verdopplung, ein Wert von +2 einer Vervierfachung etc. der Interaktionsstärke von der ersten zur zweiten Bedingung. Entsprechend beschreibt -1 die Halbierung und -2 die Reduktion der Interaktionsstärke auf ein Viertel.

Nachdem die LinkScore-Berechnung für zwei physisch interagierende Proteine hergeleitet wurde, sollen nun auch genregulative Interaktionen (Stimulation, Inhibition) betrachtet werden. Hier übt ein Protein Einfluss auf die Expression des Gens eines anderen Proteins aus.

Ist dieser Einfluss positiv - beispielsweise bei der Stimulation/Aktivierung eines Ziel-Gens Z durch einen Transkriptionsfaktor T - wird man diese Regulation als aktiv bezeichnen, ihr also einen großen Wert für die Interaktionsstärke zusprechen, wenn die Expressionswerte von T und Z hoch sind. Dies ist zwar kein Beleg dafür, dass die Expression von Z tatsächlich durch T induziert wurde, jedoch ist es hierfür ein Indiz, da sich die Expressionswerte wie für eine Stimulation erwartet verhalten. Sind die Expressionswerte hingegen niedrig, wird die Interaktion als nicht-aktiv angesehen und die Interaktionsstärke sollte gering sein. Diese Bedingungen werden durch die Formel 7.6 zur Berechnung der Interaktionsstärke bei physischen Interaktionen erfüllt. Die LinkScore-Berechnung bleibt folglich ebenso bestehen, da der LinkScore auch im positiv-regulativen Fall den Unterschied der Interaktionsstärke/Regulationsstärke zwischen den betrachteten Bedingungen beschreiben soll. Den vom Betrag her größten LinkScore hat eine positiv-regulative Interaktion dann, wenn die Expressionswerte von T und Z unter der einen Bedingung niedrig und unter der anderen Bedingung hoch sind. Wird nur entweder T oder Z exprimiert, ist die Interaktionsstärke intermediär und kann keinen so großen Beitrag mehr zu einem extremen LinkScore leisten. Die Interaktion wird unter diesen Umständen nicht zu den am stärksten differentiell regulierten Interaktionen zählen. Dies ist auch erwünscht, da in dem beschriebenen Fall kein Indiz für das Vorhandensein der Interaktion existiert. Eine Inhibition ist aktiv, und soll dann einen großen Wert für die Interaktionsstärke erhalten, wenn große Mengen des Repressors R vorhanden sind, das Ziel-Gen Z jedoch nicht exprimiert wird. Sie ist hingegen nicht aktiv, wenn kein R vorhanden ist und Z in großen Mengen exprimiert wird. Um dies abzubilden, wird die Stärke einer Inhibition daher wie dargestellt berechnet:

$$I^i = E^R - E^Z. \quad (7.7)$$

7.1. DETAILS ZUR LINKSCORE-BERECHNUNG DURCH EXPRESSION

Der zugehörige LinkScore (LS^i) ist

$$\begin{aligned} LS_{1 \rightarrow 2}^i &= I_2^i - I_1^i \\ &= (E_2^R - E_2^Z) - (E_1^R - E_1^Z), \end{aligned} \tag{7.8}$$

wobei i kennzeichnet, dass es sich um eine *inhibitorische* Interaktion handelt. Auch hier kann im Fall von nicht zum Interaktionstyp passenden Expressionsdaten (Repressor und Ziel-Gen sind beide exprimiert oder sie sind beide nicht exprimiert) kein extremer LinkScore erreicht werden.

7.2 Details zur Berechnung der Varianz der Interaktionsstärke

Es gilt gemäß Formeln 7.6 und 7.7:

$$\text{Var}(I) = \text{Var}(E_A + E_B) \quad (7.9)$$

$$= \text{Var}(E_A) + \text{Var}(E_B) + 2\text{Cov}(E_A, E_B) \text{ und}$$

$$\text{Var}(I^i) = \text{Var}(E_R - E_Z) \quad (7.10)$$

$$= \text{Var}(E_R) + \text{Var}(E_Z) + 2\text{Cov}(E_R, E_Z).$$

Die Varianz der Interaktionsstärke berechnet sich demnach immer gleich – unabhängig von der Qualität der Interaktion (physische, positiv- oder negativ-regulative Interaktion). Die Varianzdaten der Genexpressionswerte stehen meist zur Verfügung. In der Regel sind jedoch die Kovarianzmatrizen nicht verfügbar. Eine vereinfachende Annahme, dass E_A mit E_B oder E_R mit E_Z statistisch nicht korreliert wären und der Kovarianz-Term deswegen entfallen könnte, ist gerade aufgrund der durch die Interaktion beschriebenen Wechselwirkung nicht zulässig. Daher soll kurz diskutiert werden, wann der Kovarianz-Term dennoch entfallen kann und was die Folge ist. Die Kovarianz kann unter Anwendung der Cauchy-Schwarzschen Ungleichung durch die Standardabweichungen nach oben abgeschätzt werden:

$$|\text{Cov}(E_A, E_B)| \leq \sqrt{\text{Var}(E_A)} \cdot \sqrt{\text{Var}(E_B)} \quad (7.11)$$

Die Gleichheit wird genau dann erreicht, wenn der Betrag des Korrelationskoeffizienten 1 ist. Dies ergibt sich direkt aus der Definition des Korrelationskoeffizienten ϱ :

$$\varrho(E_A, E_B) = \frac{\text{Cov}(E_A, E_B)}{\sqrt{\text{Var}(E_A)} \cdot \sqrt{\text{Var}(E_B)}} \quad (7.12)$$

Gilt die Gleichheit nicht, wird die Kovarianz der Genexpressionswerte und somit die Varianz der Interaktion bei Anwendung der Ungleichung überschätzt. Dies führt zu einer kleineren Teststatistik (Formel 2.6) und in Folge zu einer möglichen Nichtablehnung der Nullhypothese. Diese Abschätzung der Kovarianz führt demnach zu einem größeren Fehler 2. Art (β -Fehler), wodurch differentiell regulierte Interaktionen häufiger ausgeblichen dargestellt werden. Im Umkehrschluss führt das Weglassen des Kovarianz-Terms zu einem größeren Fehler 1. Art (α -Fehler) ohne den Fehler 2. Art zu erhöhen. Somit bleiben maximal regulierte Interaktionen häufiger vollwertig dargestellt (nicht blass), obwohl die Änderung tatsächlich nicht signifikant ist.

7.2. DETAILS ZUR BERECHNUNG DER VARIANZ DER INTERAKTIONSSTÄRKE

Insgesamt ist bei der Anwendung von *ExprEssence* ein erhöhter Fehler 1. Art einem 2. Art vorzuziehen, da bei letzterem der Erkenntnisgewinn durch das Ausbleichen der betroffenen Interaktionen unnötig stark eingeschränkt wird. Diese Argumentation berücksichtigt, dass *ExprEssence* ein Tool zur Generierung von Hypothesen zur mechanistischen Erklärung von Unterschieden verschiedener Zustände ist, und folglich Aussagen, die aus *ExprEssence*-Subnetzwerken abgeleitet werden, unbedingt einer experimentellen Prüfung zu unterziehen sind.

7.3 Schnittmengenbildung ExprEssence-kondensierter Netzwerke mit ExprEsSector

ExprEsSector (***ExprEssence Intersector***) ist ein Programm, das aus mehreren ($n \in \mathbb{N}$) verschiedenen *ExprEssence*-kondensierten Netzwerken ($CN_i, i = \{1, \dots, n\}$, condensed networks) diejenigen Interaktionen identifiziert, die *gemeinsam* als differentiell reguliert identifiziert wurden. Sei \widetilde{CN} die Menge der ausgewählten *ExprEssence*-Netzwerke:

$$\widetilde{CN} = \{CN_1, \dots, CN_n\}. \quad (7.13)$$

ExprEsSector bildet das Schnittmengennetzwerk CN_{\cap_n} der n *ExprEssence*-kondensierten Netzwerke:

$$CN_{\cap_n} = (V_{\cap_n}, E_{\cap_n}), \quad (7.14)$$

wobei

$$V_{\cap_n} = \bigcap_{i \in \{1, \dots, n\}} V_{CN_i} = \{v | \forall V_{CN} \in \{V_{CN_1}, \dots, V_{CN_n}\} : v \in V_{CN}\} \quad (7.15)$$

und

$$E_{\cap_n} = \bigcap_{i \in \{1, \dots, n\}} E_{CN_i} = \{e | \forall E_{CN} \in \{E_{CN_1}, \dots, E_{CN_n}\} : e \in E_{CN}\}. \quad (7.16)$$

$V_{CN_{(i)}}$ beschreibe die Vertextmenge und $E_{CN_{(i)}}$ die Kantenmenge des *ExprEssence*-Subnetzwerks $CN_{(i)}$.

Anstatt die kondensierten Netzwerke direkt miteinander zu schneiden, können auch vom Nutzer ausgewählte *ExprEssence*-Netzwerke $CN_{u_1}, \dots, CN_{u_k}$ zu einem gemeinsamen Netzwerk vereinigt werden, um dieses anschließend mit anderen Vereinigungsnetzwerken oder unveränderten *ExprEssence*-Netzwerken zu schneiden:

$$CN_{\bigcup_{u_1, \dots, u_k}} = (V_{\bigcup_{u_1, \dots, u_k}}, E_{\bigcup_{u_1, \dots, u_k}}), \quad (7.17)$$

wobei

$$V_{\bigcup_{u_1, \dots, u_k}} = \bigcup_{i \in \{u_1, \dots, u_k\}} V_{CN_i} = \{v | \exists V_{CN} \in \{V_{CN_{u_1}}, \dots, V_{CN_{u_k}}\} : v \in V_{CN}\} \quad (7.18)$$

und

$$E_{\bigcup_{u_1, \dots, u_k}} = \bigcup_{i \in \{u_1, \dots, u_k\}} E_{CN_i} = \{e \mid \exists E_{CN} \in \{E_{CN_{u_1}}, \dots, E_{CN_{u_k}}\} : e \in E_{CN}\}. \quad (7.19)$$

Für das so erstellte und zu schneidende Vereinigungsnetzwerk gelte

$$CN_{\bigcup_{u_1, \dots, u_k}} \in \widetilde{CN}, \quad (7.20)$$

wohingegen von den unveränderten *ExprEssence*-Netzwerken $CN_{u_1}, \dots, CN_{u_k}$ nur diejenigen in \widetilde{CN} vertreten seien, die vom Nutzer nochmals separat als zu schneidend definiert wurden.

Die Anforderung des Vorhandenseins aller Knoten und Kanten in *jedem* zu schneidenden Netzwerk kann auf eine beliebige kleinere Anzahl von Netzwerken $m \in \{1, \dots, n-1\}$ relaxiert werden, sodass das Fehlen eines Knotens oder einer Kante in bis zu $q = n-m$ Netzwerken nicht zur Elimination aus dem Schnittmengennetzwerk führt:

$$CN_{\cap_n^m} = (V_{\cap_n^m}, E_{\cap_n^m}), \quad (7.21)$$

wobei

$$V_{\cap_n^m} = \{v \mid \exists \widetilde{CN}^* \subset \widetilde{CN}, |\widetilde{CN}^*| \geq m \ \forall CN \in \widetilde{CN}^* : v \in V_{CN}\} \quad (7.22)$$

und

$$E_{\cap_n^m} = \{e \mid \exists \widetilde{CN}^* \subset \widetilde{CN}, |\widetilde{CN}^*| \geq m \ \forall CN \in \widetilde{CN}^* : e \in E_{CN}\}. \quad (7.23)$$

Wie leicht erkennbar ist, ist unbedeutend, in welchen q der n ausgewählten *ExprEssence*-Netzwerke eine Kante oder ein Knoten nicht vorhanden ist.

Literaturverzeichnis

- [1] Johnson DS, Mortazavi A, Myers RM, Wold B (2007) Genome-wide mapping of in vivo protein-dna interactions. *Science* 316: 1497–1502.
- [2] Aparicio O, Geisberg JV, Struhl K (2004) Chromatin immunoprecipitation for determining the association of proteins with specific genomic sequences in vivo. *Curr Protoc Cell Biol Chapter 17: Unit 17.7.*
- [3] Barabási AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. *Nat Rev Genet* 5: 101–113.
- [4] Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, et al. (2009) Human protein reference database—2009 update. *Nucleic Acids Res* 37: D767–D772.
- [5] Chatr-Aryamontri A, Breitkreutz BJ, Heinicke S, Boucher L, Winter A, et al. (2013) The biogrid interaction database: 2013 update. *Nucleic Acids Res* 41: D816–D823.
- [6] Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, et al. (2013) String v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res* 41: D808–D815.
- [7] Kanehisa M, Goto S (2000) Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 28: 27–30.
- [8] Li C, Liakata M, Rebholz-Schuhmann D (2013) Biological network extraction from scientific literature: state of the art and challenges. *Brief Bioinform* .
- [9] Wermter J, Tomanek K, Hahn U (2009) High-performance gene name normalization with geno. *Bioinformatics* 25: 815–821.
- [10] Greene CS, Troyanskaya OG (2011) Pilgrm: an interactive data-driven discovery platform for expert biologists. *Nucleic Acids Res* 39: W368–W374.
- [11] Gardner TS, di Bernardo D, Lorenz D, Collins JJ (2003) Inferring genetic networks and identifying compound mode of action via expression profiling. *Science* 301: 102–105.

- [12] Bonneau R, Reiss DJ, Shannon P, Facciotti M, Hood L, et al. (2006) The inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome Biol* 7: R36.
- [13] Marbach D, Mattiussi C, Floreano D (2009) Replaying the evolutionary tape: biomimetic reverse engineering of gene networks. *Ann N Y Acad Sci* 1158: 234–245.
- [14] D'haeseleer P, Liang S, Somogyi R (2000) Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics* 16: 707–726.
- [15] Galagan JE, Minch K, Peterson M, Lyubetskaya A, Azizi E, et al. (2013) The mycobacterium tuberculosis regulatory network and hypoxia. *Nature* 499: 178–183.
- [16] Marbach D, Prill RJ, Schaffter T, Mattiussi C, Floreano D, et al. (2010) Revealing strengths and weaknesses of methods for gene network inference. *Proc Natl Acad Sci U S A* 107: 6286–6291.
- [17] Warsow G, Endlich N, Schordan E, Schordan S, Chilukoti RK, et al. (2013) Podnet, a protein-protein interaction network of the podocyte. *Kidney Int* 84: 104–115.
- [18] Goh KI, Cusick ME, Valle D, Childs B, Vidal M, et al. (2007) The human disease network. *Proc Natl Acad Sci U S A* 104: 8685–8690.
- [19] Barrenas F, Chavali S, Holme P, Mobini R, Benson M (2009) Network properties of complex human disease genes identified through genome-wide association studies. *PLoS One* 4: e8090.
- [20] Sun PG, Gao L, Han S (2011) Prediction of human disease-related gene clusters by clustering analysis. *Int J Biol Sci* 7: 61–73.
- [21] Bauer-Mehren A, Bundschus M, Rautschka M, Mayer MA, Sanz F, et al. (2011) Gene-disease network analysis reveals functional modules in mendelian, complex and environmental diseases. *PLoS One* 6: e20284.
- [22] Gevaert O, De Smet F, Timmerman D, Moreau Y, De Moor B (2006) Predicting the prognosis of breast cancer by integrating clinical and microarray data with bayesian networks. *Bioinformatics* 22: e184–e190.

- [23] Chuang HY, Lee E, Liu YT, Lee D, Ideker T (2007) Network-based classification of breast cancer metastasis. *Mol Syst Biol* 3: 140.
- [24] Chen R, Mias GI, Li-Pook-Than J, Jiang L, Lam HYK, et al. (2012) Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell* 148: 1293–1307.
- [25] Li-Pook-Than J, Snyder M (2013) ipop goes the world: integrated personalized omics profiling and the road toward improved health care. *Chem Biol* 20: 660–666.
- [26] Magger O, Waldman YY, Ruppin E, Sharan R (2012) Enhancing the prioritization of disease-causing genes through tissue specific protein interaction networks. *PLoS Computational Biology* 8: e1002690.
- [27] Przytycka TM, Singh M, Slonim DK (2010) Toward the dynamic interactome: it's about time. *Brief Bioinform* 11: 15–29.
- [28] Chen B, Fan W, Liu J, Wu FX (2013) Identifying protein complexes and functional modules—from static ppi networks to dynamic ppi networks. *Brief Bioinform* .
- [29] Thorne T, Stumpf MPH (2012) Inference of temporally varying bayesian networks. *Bioinformatics* 28: 3298–3305.
- [30] Oates CJ, Mukherjee S (2012) Network inference and biological dynamics. *Ann Appl Stat* 6: 1209–1235.
- [31] Mutz KO, Heilkenbrinker A, Lönne M, Walter JG, Stahl F (2013) Transcriptome analysis using next-generation sequencing. *Curr Opin Biotechnol* 24: 22–30.
- [32] Kogenaru S, Qing Y, Guo Y, Wang N (2012) Rna-seq and microarray complement each other in transcriptome profiling. *BMC Genomics* 13: 629.
- [33] Schillert A (2010) Statistische Qualitätssicherung von Hochdurchsatz-Genotypisierungsdaten. Ph.D. thesis, Universität zu Lübeck.
- [34] Carvalho B, Bengtsson H, Speed TP, Irizarry RA (2007) Exploration, normalization, and genotype calls of high-density oligonucleotide snp array data. *Biostatistics* 8: 485–499.
- [35] Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T (2011) Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* 27: 431–432.

- [36] Vogel C, Marcotte EM (2012) Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat Rev Genet* 13: 227–232.
- [37] Greenbaum D, Colangelo C, Williams K, Gerstein M (2003) Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biol* 4: 117.
- [38] Warsow G, Greber B, Falk SSI, Harder C, Siatkowski M, et al. (2010) Expressence—revealing the essence of differential experimental data in the context of an interaction/regulation net-work. *BMC Syst Biol* 4: 164.
- [39] Welch BL (1938) The significance of the difference between two means when the population variances are unequal. *Biometrika* 29: 350–362.
- [40] Welch BL (1947) The generalization of Student's problem when several different population variances are involved. *Biometrika* 34: 28–35.
- [41] Amaratunga D (2004) Exploration and analysis of DNA microarray and protein array data, volume 446. Wiley. com.
- [42] Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* 98: 5116–5121.
- [43] Broberg P (2003) Statistical methods for ranking differentially expressed genes. *Genome Biol* 4: R41.
- [44] Bortz J, Schuster C (2010) Statistik für Human- und Sozialwissenschaftler. Lehrbuch mit Online-Materialien. Springer DE.
- [45] Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B (Methodological)* : 289–300.
- [46] Nakao, Bono, Kawashima, Kamiya, Sato, et al. (1999) Genome-scale gene expression analysis and pathway reconstruction in kegg. *Genome Inform Ser Workshop Genome Inform* 10: 94–103.
- [47] Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, et al. (2006) Gene prioritization through genomic data fusion. *Nat Biotechnol* 24: 537–544.

- [48] Köhler S, Bauer S, Horn D, Robinson PN (2008) Walking the interactome for prioritization of candidate disease genes. *Am J Hum Genet* 82: 949–958.
- [49] Schwikowski B, Uetz P, Fields S (2000) A network of protein-protein interactions in yeast. *Nat Biotechnol* 18: 1257–1261.
- [50] Chua HN, Sung WK, Wong L (2006) Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. *Bioinformatics* 22: 1623–1630.
- [51] Ideker T, Ozier O, Schwikowski B, Siegel AF (2002) Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* 18 Suppl 1: S233–S240.
- [52] Dao P, Wang K, Collins C, Ester M, Lapuk A, et al. (2011) Optimally discriminative subnetwork markers predict response to chemotherapy. *Bioinformatics* 27: i205–i213.
- [53] Baumbach J, Friedrich T, Kötzing T, Krohmer A, Müller J, et al. (2012) Efficient algorithms for extracting biological key pathways with global constraints. In: Proceedings of the fourteenth international conference on Genetic and evolutionary computation conference. ACM, pp. 169–176. URL <http://dl.acm.org/citation.cfm?id=2330188>.
- [54] Pereira-Leal JB, Enright AJ, Ouzounis CA (2004) Detection of functional modules from protein interaction networks. *Proteins* 54: 49–57.
- [55] Guo Z, Wang L, Li Y, Gong X, Yao C, et al. (2007) Edge-based scoring and searching method for identifying condition-responsive protein-protein interaction subnetwork. *Bioinformatics* 23: 2121–2128.
- [56] Dittrich MT, Klau GW, Rosenwald A, Dandekar T, Müller T (2008) Identifying functional modules in protein-protein interaction networks: an integrated exact approach. *Bioinformatics* 24: i223–i231.
- [57] Rajagopalan D, Agarwal P (2005) Inferring pathways from gene lists using a literature-derived network of biological relationships. *Bioinformatics* 21: 788–793.

- [58] Wu X, Harrison SH, Chen JY (2009) Pattern discovery in breast cancer specific protein interaction network. *Summit on Translat Bioinforma* 2009: 1–5.
- [59] Ichimura K, Kurihara H, Sakai T (2003) Actin filament organization of foot processes in rat podocytes. *J Histochem Cytochem* 51: 1589–1600.
- [60] Pavenstädt H, Kriz W, Kretzler M (2003) Cell biology of the glomerular podocyte. *Physiol Rev* 83: 253–307.
- [61] Endlich N, Kress KR, Reiser J, Uttenweiler D, Kriz W, et al. (2001) Podocytes respond to mechanical stress in vitro. *J Am Soc Nephrol* 12: 413–422.
- [62] Endlich N, Endlich K (2012) The challenge and response of podocytes to glomerular hypertension. *Semin Nephrol* 32: 327–341.
- [63] Shankland SJ (2006) The podocyte's response to injury: role in proteinuria and glomerulosclerosis. *Kidney Int* 69: 2131–2147.
- [64] Asanuma K, Kim K, Oh J, Giardino L, Chabanis S, et al. (2005) Synaptopodin regulates the actin-bundling activity of alpha-actinin in an isoform-specific manner. *J Clin Invest* 115: 1188–1198.
- [65] Endlich N, Schordan E, Cohen CD, Kretzler M, Lewko B, et al. (2009) Palladin is a dynamic actin-associated protein in podocytes. *Kidney Int* 75: 214–226.
- [66] Poulsom R, Little MH (2009) Parietal epithelial cells regenerate podocytes. *J Am Soc Nephrol* 20: 231–233.
- [67] Ronconi E, Sagrinati C, Angelotti ML, Lazzeri E, Mazzinghi B, et al. (2009) Regeneration of glomerular podocytes by human renal progenitors. *J Am Soc Nephrol* 20: 322–332.
- [68] Zhang J, Pippin JW, Kroft RD, Naito S, Liu ZH, et al. (2013) Podocyte repopulation by renal progenitor cells following glucocorticoids treatment in experimental fsgs. *Am J Physiol Renal Physiol* 304: F1375–F1389.
- [69] Pippin JW, Sparks MA, Glenn ST, Buitrago S, Coffman TM, et al. (2013) Cells of renin lineage are progenitors of podocytes and parietal epithelial cells in experimental glomerular disease. *Am J Pathol* 183: 542–557.

- [70] Kelder T, Pico AR, Hanspers K, van Iersel MP, Evelo C, et al. (2009) Mining biological pathways using wikipathways web services. *PLoS One* 4: e6447.
- [71] Giardine B, Borg J, Higgs DR, Peterson KR, Philipsen S, et al. (2011) Systematic documentation and analysis of human genetic variation in hemoglobinopathies using the microattribution approach. *Nat Genet* 43: 295–301.
- [72] Williams AJ, Harland L, Groth P, Pettifer S, Chichester C, et al. (2012) Open phacts: semantic interoperability for drug discovery. *Drug Discov Today* 17: 1188–1198.
- [73] Przulj N, Corneil DG, Jurisica I (2006) Efficient estimation of graphlet frequency distributions in protein-protein interaction networks. *Bioinformatics* 22: 974–980.
- [74] Przulj N (2007) Biological network comparison using graphlet degree distribution. *Bioinformatics* 23: e177–e183.
- [75] Milenkovi? T, Przulj N (2008) Uncovering biological network function via graphlet degree signatures. *Cancer Inform* 6: 257–273.
- [76] Ashburn TT, Thor KB (2004) Drug repositioning: identifying and developing new uses for existing drugs. *Nat Rev Drug Discov* 3: 673–683.
- [77] von Eichborn J, Murgueitio MS, Dunkel M, Koerner S, Bourne PE, et al. (2011) Promiscuous: a database for network-based drug-repositioning. *Nucleic Acids Res* 39: D1060–D1066.
- [78] Kuhn M, Szklarczyk D, Franceschini A, von Mering C, Jensen LJ, et al. (2012) Stitch 3: zooming in on protein-chemical interactions. *Nucleic Acids Res* 40: D876–D880.