Die Integration von Multiskalen- und Multi-Omik-Daten zur Erforschung von Wirt-Pathogen-Interaktionen am Beispiel von pathogenen Pilzen

Dissertation

zur Erlangung des akademischen Grades

***„doctor rerum naturalium“* (Dr. rer. nat.)**

vorgelegt dem Rat der Fakultät für Biowissenschaften

der Friedrich-Schiller-Universität Jena

von

M. Sc. Theresia Conrad

geboren am 17.04.1991 in Jena

Gutachter

apl. Prof. Dr. Reinhard Guthke, Hans-Knöll-Institut Jena

PD Dr. Kerstin Voigt, Hans-Knöll-Institut Jena

Prof. Dr. Christoph Kaleta, Christian-Albrechts-Universität zu Kiel

Verteidigung

03.02.2020, Jena

„Remember to look up at the stars and not down at your feet. Try to make sense about what you see and wonder about what makes the universe exist. Be curious.“

Stephen Hawking

Danksagung

Zu Beginn möchte ich mich ganz herzlich bei all jenen Personen bedanken, die maßgeblich zum Gelingen der vorliegenden Dissertation beigetragen haben.

An allererster Stelle sind das meine Betreuer Prof. Dr. Ilse Jacobsen, Prof. Dr. Reinhard Guthke und Dr. Jörg Linde, durch die diese Arbeit ermöglicht wurde. Egal ob bioinformatischer, biologischer oder organisatorischer Natur – mit ihren wertvollen Hinweisen und motivierenden Worten haben sie mich bei jeder zu meisternden Herausforderung tatkräftig unterstützt. Vor allem dir, Reinhard, möchte ich ganz besonders dafür danken, dass du mich auch während deines wohlverdienten Ruhestands und neben all deinen anderen Projekten stets mit Rat und Tat begleitet hast!

Ein großes Dankeschön geht auch an die Graduiertenschule *Jena School for Microbial Communication* (JSMC), die mich in ihr umfangreiches Exzellenzförderprogramm aufgenommen hat. Zusammen mit der Jenaer Graduierten-Akademie hat sie mit ihrem vielfältigen Kurs- und Weiterbildungsangebot erheblich zur Förderung meiner akademischen Laufbahn beigetragen.

Außerdem möchte ich mich ganz herzlich bei all meinen Kooperationspartnern für die angenehme und konstruktive Zusammenarbeit bedanken. Dazu gehören unter anderem die Arbeitsgruppen Mikrobielle Immunologie, Molekulare und Angewandte Mikrobiologie, und Mikrobielle Pathogenitätsmechanismen des Hans-Knöll-Instituts (HKI) sowie die Mitarbeiter des Lehrstuhls Bioinformatik an der Friedrich-Schiller-Universität Jena und meine ehemaligen Kollegen von Host Septomics. Ein besonderer Dank gilt Dr. Sebastian Henkel von der BioControl Jena GmbH, der mich durch viele anregende Diskussionen und kritische Betrachtungsweisen in meiner Arbeit sehr unterstützt und vorangebracht hat. Auch Agata Kilar *(Silesian University of Technology,* Gliwice, Polen), deren Bachelor- und Masterarbeit ich während meiner

Zeit am HKI betreuen durfte, möchte ich für ihre großartige und engagierte Arbeit danken.

Natürlich möchte ich mich ebenso bei all den (ehemaligen) Kollegen der beiden Gruppen Systembiologie/Bioinformatik und Angewandte Systembiologie für die tolle Arbeitsatmosphäre und ihre stete Unterstützung und Motivation bedanken. Dieser Dank geht besonders an Dr. Patricia Sieber, Maria Prauße, Franziska Hörhold, Dr. Thomas Wolf und Dr. Sebastian Vlaic. Sowohl eure fachliche Unterstützung als auch die gemeinsamen Unternehmungen, die Spaziergänge zur Mensa und die vielen unvergesslichen Kaffeepausen haben für mich die Zeit am HKI zu einer ganz besonderen gemacht.

Zuletzt – aber deswegen nicht weniger wichtig – gilt ein ganz großes, liebes Dankeschön meiner Familie und meinen Freunden. Insbesondere meinen Eltern und Christine Zech ist es zu verdanken, dass diese Arbeit entstanden ist. Obwohl in der Ferne sind sie die unsichtbaren, aber doch nicht wegzudenkenden Zahnräder im großen Getriebe „Dissertation“. Ob mit finanzieller Unterstützung oder moralischer, mit Schokolade, Cocktail oder Käffchen, mit aufmunternden Worten oder einfach nur mit einer Umarmung – ohne euren Rückhalt und euren Glauben an mich hätte ich es nicht so weit gebracht. Ich danke euch!

Zusammenfassung

Die stetige Entwicklung und Verbesserung neuer wissenschaftlicher Messmethoden haben dazu geführt, dass uns eine Fülle an Daten aus heterogenen Quellen zur Verfügung steht. Dazu zählen unter anderem Daten unterschiedlicher zeitlicher und struktureller Skalen wie die verschiedenen Omik-Ebenen. Durch die Integration solcher sogenannten Multiskalen- und Multi-Omik-Daten ist es möglich, ein umfassendes Verständnis für die Komplexität und Dynamik biologischer Systeme und deren Prozesse zu entwickeln. Jedoch stellt die Integration aufgrund der biologisch und methodisch bedingten Datenheterogenität eine wohlbekannte Herausforderung in den Biowissenschaften dar.

Ziel der vorliegenden Dissertation war es, unter Zuhilfenahme verschiedener computergestützter Integrationsansätze neue Erkenntnisse im Bereich der Infektionsbiologie bezüglich der Wirt-Pathogen-Interaktionen zu erlangen. Der Fokus lag dabei auf pathogenen Pilzen, die eine Vielzahl lokaler und systemischer Infektionen hervorrufen können. Anhand von aktuellen Forschungsbeispielen wurden einerseits einige bereits gut etablierte Analysemethoden von Multiskalen- und Multi-Omik-Daten aufgezeigt. Andererseits wurde der neu entwickelte ModuleDiscoverer-Algorithmus zur Identifikation von regulatorischen Modulen in Protein-Protein-Interaktionsnetzwerken vorgestellt. Es konnte gezeigt werden, dass ModuleDiscoverer die Integration von Multi-Omik-Daten effektiv unterstützt und auch die Detektion potenzieller Schlüsselfaktoren gestattet, deren Identifizierung über andere klassische Ansätze nicht möglich ist.

Mit dieser Dissertation konnte ein tieferer Einblick in die komplexen Zusammenhänge und Dynamiken biologischer Systeme erhalten und so ein wichtiger Beitrag zur Erforschung von Wirt-Pathogen-Interaktionen im Kontext pathogener Pilze geleistet werden. Die Komplexität der Interaktionen, das nur eingeschränkt über

Datenbanken zur Verfügung stehende Wissen und die methodischen Grenzen der bioinformatischen Werkzeuge zeigen allerdings auch den Bedarf an weiterführenden Forschungsarbeiten, um ein umfassendes Verständnis der Komplexität biologischer Systeme zu ermöglichen.

Abstract

The ongoing development and improvement of novel measurement techniques for scientific research result in a huge amount of available data coming from heterogeneous sources. Amongst others, these sources comprise diverse temporal and spatial scales including different omics levels. The integration of such multiscale and multi-omics data enables a comprehensive understanding of the complexity and dynamics of biological systems and their processes. However, due to the biologically and methodically induced data heterogeneity, the integration process is a well-known challenge in nowadays life science.

Applying several computational integration approaches, the present doctoral thesis aimed at gaining new insights into the field of infection biology regarding host-pathogen interactions. In this context, the focus was on fungal pathogens causing a variety of local and systemic infections. Based on current examples of research, on the one hand, several well-established approaches for the analysis of multiscale and multi-omics data have been presented. On the other hand, the novel ModuleDiscoverer approach was introduced to identify regulatory modules in protein-protein interaction networks. It has been shown that ModuleDiscoverer effectively supports the integration of multi-omics data and, in addition, allows the detection of potential key factors that cannot be detected by other classical approaches.

This thesis provides deeper insights into the complex relationships and dynamics of biological systems and, thus, represents an important contribution to the investigation of host-pathogen interactions. Due to the interactions' complexity and the limitations of the currently available knowledge databases as well as the bioinformatic tools, further research is necessary to gain a comprehensive understanding of the complexity of biological systems.

Inhaltsverzeichnis

Abkürzungen

AD	*average distance*
ADM	*average distance between means*
AGNES	*agglomerative nesting*
A. fumigatus	*Aspergillus fumigatus*
APN	*average proportion of non-overlap*
C. albicans	*Candida albicans*
C. glabrata	*Candida glabrata*
C. parapsilosis	*Candida parapsilosis*
C. tropicalis	*Candida tropicalis*
CLARA	*clustering large applications*
DAMPs	*damage-associated molecular patterns*
DIANA	*divisive analysis*
DNA	Desoxyribonukleinsäure
et al.	*et alii, et aliae*
FOM	*figure of merit*
GO	*gene ontology*
GRM	gesamtregulatorisches Modul
GRN	genregulatorisches Netzwerk
GSEA	*gene set enrichment analysis*
M. tuberculosis	*Mycobacterium tuberculosis*
mRNA	*messenger* RNA
PAM	*partitioning around medoids*
PAMPs	*pathogen-associated molecular patterns*
PPIN	Protein-Protein-Interaktionsnetzwerk(e)

RNA	Ribonukleinsäure
SOTA	*self-organization tree algorithm*
WPI	Wirt-Pathogen-Interaktion
WPPI	Wirt-Pilzpathogen-Interaktion

1 Einleitung

1.1 Multiskalen- und Multi-Omik-Daten im Kreislauf der Systembiologie

Von ganzen Populationen bis hin zur einzelnen Zelle – biologische Systeme lassen sich in allen Bereichen des Lebens finden. Sie bestehen aus einer Vielzahl von Komponenten, die miteinander auf verschiedenen Ebenen interagieren. Um die Untersuchung dieser komplexen biologischen Systeme zu ermöglichen, können sie als ein Zusammenspiel einzelner Ebenen unterschiedlicher zeitlicher und struktureller Größenordnungen (Skalen) beschrieben werden [Castiglione *et al.,* 2014]. Ein Beispiel aus dem molekularen Bereich sind sogenannte Omik-Ebenen. Diese werden durch die Gesamtheit ihrer Komponenten (Gene, Proteine, Metabolite etc.) einer strukturellen Skala (Zelle, Gewebe, Organ, Organismus etc.) charakterisiert. Auf einer zeitlichen Skala (Sekunden, Minuten, Tage etc.) lässt sich die Dynamik ihrer Interaktionen beschreiben. Von Multiskalen oder Multi-Omik ist die Rede, wenn mehrere Skalen bzw. Omik-Ebenen betrachtet werden, die mit dem gleichen biologischen System assoziiert sind [Schleicher *et al.,* 2016].

Die vorliegende Arbeit ist der integrierten Analyse von Multiskalen- und Multi-Omik-Daten gewidmet. Unter Zuhilfenahme sowohl experimenteller als auch computergestützter Ansätze kann ein umfassendes Verständnis der Komplexität und Dynamik biologischer Systeme und deren Prozesse erlangt werden. Noch bis vor einigen Jahren wurden die einzelnen Komponenten als jeweils separate Einheiten untersucht (z. B. ein einzelnes Gen oder Protein). Neue Omik-Technologien ermöglichen es, mehrere Komponenten einer bestimmten Omik-Ebene gleichzeitig zu analysieren und sie so als Teil eines komplexen Netzwerks zu betrachten [De Keersmaecker *et al.,* 2006]. Die Ergebnisse solcher Betrachtungen liefern neue Hypothesen, die wichtige Anhalts-

punkte für weitere experimentelle Untersuchungen darstellen. Dieser Ablauf von experimentellen Untersuchungen, Datenanalyse, Hypothesengenerierung und neuen experimentellen Untersuchungen zur Überprüfung der Hypothesen wird auch als Kreislauf der Systembiologie bezeichnet (Abbildung 1.1) [Butcher *et al.*, 2004].

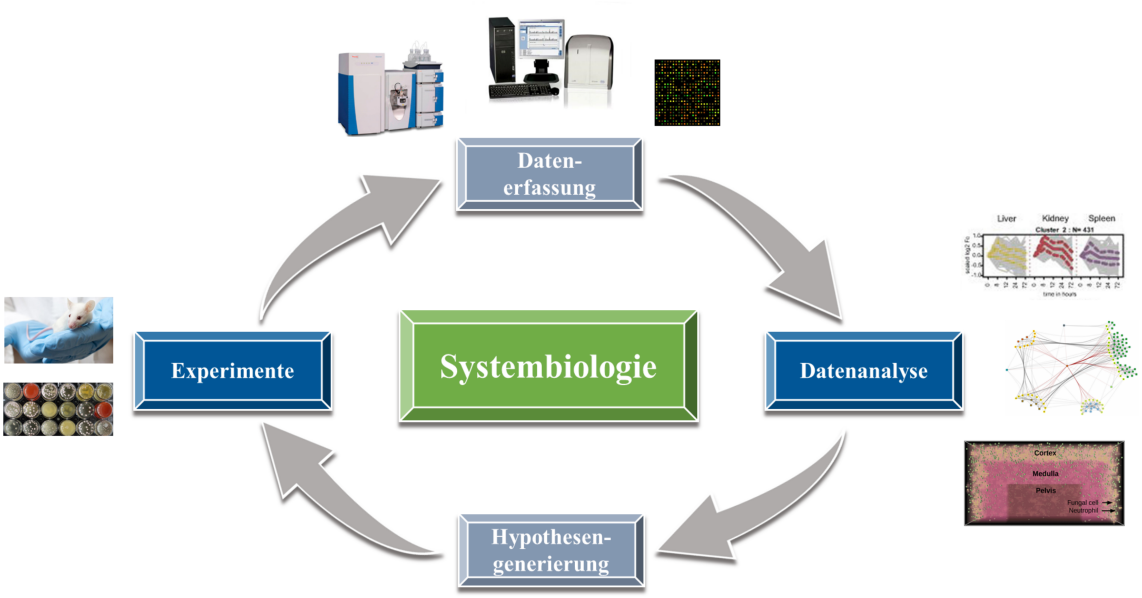


Abbildung 1.1: Kreislauf der Systembiologie (modifiziert nach [1])

Das Aufkommen und die Weiterentwicklung der modernen Omik-Technologien sowie die damit verbundene Zunahme wissenschaftlicher Erkenntnisse haben dazu geführt, dass immer komplexere wissenschaftliche Studien durchgeführt und enorme Datenmengen produziert werden. Allein in den letzten 20 Jahren hat sich die Anzahl von Omik-Daten-basierten Studien um ein Vielfaches gesteigert (Abbildung 1.2). Für die Bewältigung all dieser nun verfügbaren Daten ist die Entwicklung neuer Analysemethoden von entscheidender Bedeutung. Die Verarbeitung der enthaltenen

[1] *http://www.scinexx.de/wissen-aktuell-20400-2016-07-18.html*, Stand vom 07.08.2018;
http://www.directindustry.de/prod/thermo-scientific-scientific-instruments-and-aut/product-7217-1713268.html, Stand vom 07.08.2018;
https://upload.wikimedia.org/wikipedia/commons/f/f2/Cdnaarray.jpg, Stand vom 07.08.2018;
https://www.management-krankenhaus.de/products/labor-diagnostik/next-generation-dna-sequenzierer-eroeffnet-grosser-wissenschaftsgemeinde-d, Stand vom 07.08.2018;
[Hebecker *et al.*, 2016];
[Conrad *et al.*, 2018]

Informationen entsprechend ihrer Qualität und Relevanz leistet einen wesentlichen Beitrag für das tiefere Verständnis der Komplexität eines biologischen Systems.

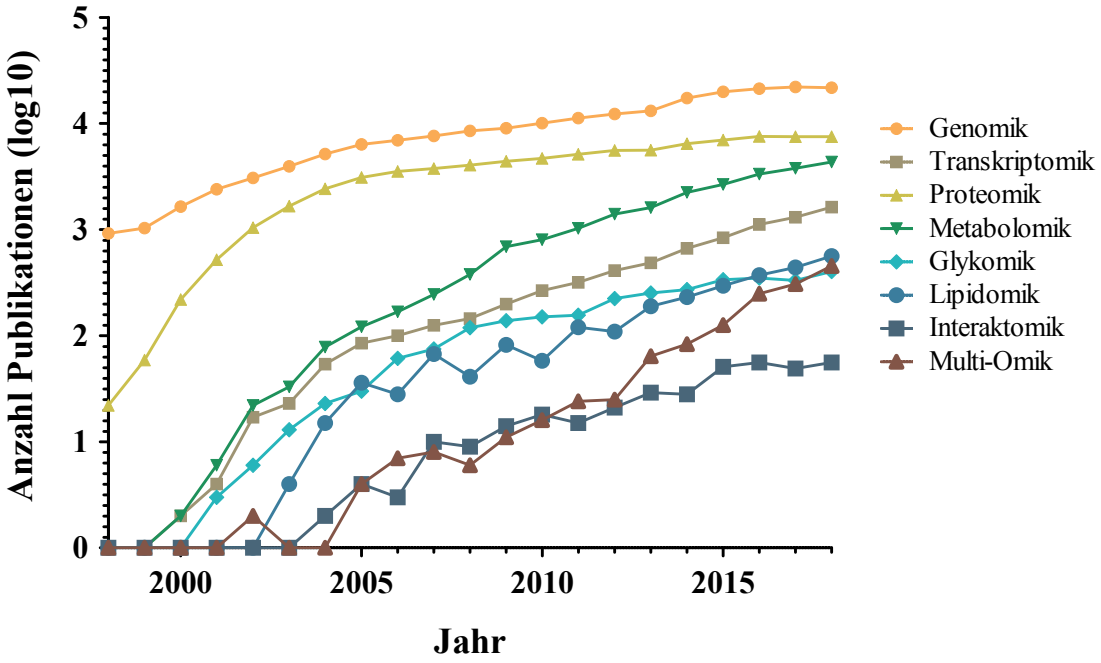


Abbildung 1.2: Publikationen Omik-Daten-basierter Studien der letzten 20 Jahre (basierend auf PubMed-Informationen [2])

1.2 Die Bedeutung von Pilzinfektionen

Ob im Wasser, in der Luft, im Boden oder sogar im Menschen selbst – überall in der Umwelt befinden sich Mikroorganismen, von denen einige für den Menschen „nützlich“ sind, aber andere eine Vielzahl von Krankheiten und Schädigungen auslösen können. Letztere werden als Pathogene bezeichnet [Pirofski *et al.,* 2012]. Dazu zählen Viren sowie pathogene Bakterien und Pilze. Schätzungsweise gibt es etwa 1400 bekannte Spezies von humanpathogenen Erregern, die unter den unterschiedlichsten Bedingungen wachsen und sich Umgebungsänderungen mithilfe von diversen Überlebens- und Infektionsstrategien gezielt anpassen können [*„Microbiology by*

[2] *https://www.ncbi.nlm.nih.gov/pubmed/,* Stand vom 01.05.2019

numbers“, 2011]. Der Ausgang einer Infektion mit Erregern hängt oftmals vom Immunstatus der betroffenen Individuen ab. Problematisch wird es vor allem dann, wenn die Erreger auf immunsupprimierte Individuen treffen. Deren Immunsystem ist unter anderem durch Vorerkrankungen oder immunsupprimierende Therapien bereits geschwächt und somit anfälliger gegenüber Eindringlingen [Heinekamp *et al.,* 2015]. Mortalitätsraten über 50 % wurden beispielsweise mit invasiven Pilzinfektionen (Mykosen) assoziiert. Weltweit sterben jährlich etwa 1.5 Millionen Menschen durch invasive Mykosen. Über 90 % dieser Todesfälle gehen auf die Gattungen *Candida, Aspergillus, Cryptococcus* und *Pneumocystis* zurück. Dabei zählen *Candida albicans* und *Aspergillus fumigatus* in den Industrienationen Europas und Nordamerikas zu den bedeutendsten humanpathogenen Pilzen, die eine Vielzahl lokaler und systemischer Infektionen hervorrufen [Brown *et al.,* 2012].

1.2.1 *Candida albicans*

C. albicans ist der vermutlich am häufigsten vorkommende humanpathogene Pilz. Hierbei handelt es sich um einen meist harmlosen kommensalen Organismus, der im Magen-Darm-Trakt sowie in der oralen und vaginalen Schleimhaut vieler gesunder Individuen zu finden ist [Kim *et al.,* 2011]. Doch in immunsupprimierten Individuen kann *C. albicans* zu lebensbedrohlichen Infektionen führen. Dazu zählen sowohl Infektionen des Blutkreislaufs, Candidämie genannt, als auch die Kolonisierung von inneren Organen als Folge der Candidämie (disseminierte Candidose). Die Sterblichkeitsrate solcher Erkrankungen liegt zwischen 30 % und 50 % [Kim *et al.,* 2011]. Die Virulenz eines Pathogens, also das Maß seiner krankmachenden Wirkung, wird maßgeblich über seine Virulenzfaktoren bestimmt. Zu den bisher identifizierten Virulenzfaktoren von *C. albicans* gehört beispielsweise die Fähigkeit, in den drei verschiedenen Morphologien Hefe, Pseudohyphe und echte Hyphe aufzutreten, wachsen und zwischen ihnen wechseln zu können. Besonders in ihrer Hyphenform kann *C. albicans* zu schwerwiegenden Infektionen führen [Jacobsen *et al.,* 2017]. Weitere Virulenzmerkmale sind unter anderem die Produktion von Molekülen speziell für die Invasion des Wirts, die Evasion des Immunsystems des Wirts, die Bildung von Biofilmen und Toxinen oder die Fähigkeit, sich essentielle Nährstoffe und Spurenelemente (z. B. Metalle) vom Wirt zu beschaffen [Krüger *et al.,* 2015].

1.2.2 *Aspergillus fumigatus*

A. fumigatus ist ein im Erdreich und Kompost befindlicher saprotropher Schimmelpilz, der komplexes organisches Material abbaut [Krüger *et al.,* 2015]. Seine Sporen, auch Konidien genannt, werden über die Luft übertragen und stellen die eigentlichen Infektionserreger dar. Aufgrund ihrer geringen Größe können sie leicht vom Wirt eingeatmet werden. Daher sind vor allem die Atemorgane das häufigste primär betroffene Organsystem. Wie *C. albicans* stellt der Pilz für gesunde Individuen im Regelfall keine Bedrohung dar. Ist das Immunsystem jedoch geschwächt, kann *A. fumigatus* verschiedenste Krankheitsformen auslösen – angefangen mit der lokal begrenzten allergischen Reaktion bis hin zur disseminierten Infektion. Invasive Aspergillose kann zu Sterblichkeitsraten von 50 % bis nahezu 100 % führen [Brown *et al.,* 2012]. Zu den bekanntesten Virulenzfaktoren von *A. fumigatus* gehört unter anderem die Beschaffung von essentiellen Nährstoffen und Metallen [Krüger *et al.,* 2015]. Darüber hinaus hat der Pilz spezielle Schutzmaßnahmen entwickelt, um der Erkennung durch die wirtseigenen Immunzellen zu entgehen. Solche Immunevasionsmechanismen betreffen beispielsweise die Erkennung von Konidien oder die Regulierung der Phagozytose [Heinekamp *et al.,* 2015].

1.3 Wirt-Pathogen-Interaktionen am Beispiel pathogener Pilze

Wirt-Pathogen-Interaktionen (WPI) beschreiben das Zusammenspiel und die Wechselwirkungen zwischen dem eindringenden Erreger und dem jeweiligen Wirt. Im vorangegangenen Kapitel wurde auf die Bedeutung der Pilze als Erreger eingegangen. Dementsprechend liegt der Fokus dieser Dissertation auch auf Interaktionen zwischen Wirt und pathogenen Pilzen, im Folgenden WPPI (Wirt-Pilzpathogen-Interaktionen) genannt. Solche WPPI finden auf unterschiedlichen zeitlichen und strukturellen Skalen statt und involvieren Mechanismen, die entweder von der Wirts- oder von der Pilzseite initiiert werden.

1.3.1 Wirtsinitiierte Wirt-Pilzpathogen-Interaktionen

Einer der wichtigsten Ausgangspunkte für WPPI ist das Immunsystem des Wirts. Dazu zählen sowohl die Prozesse der angeborenen unspezifischen als auch die der erworbenen spezifischen Immunantwort. Beide sind maßgeblich für die Erkennung, Kontrolle und Eliminierung der eingedrungenen Pathogene verantwortlich [Murphy *et al.,* 2014]. Die Initiierung der Immunantwort erfolgt meist über die Erkennung hoch konservierter Erregerstrukturen, den so genannten *pathogen-associated molecular patterns* (PAMPs) oder *damage-associated molecular patterns* (DAMPs). Letztere werden unter anderem durch infektionsinduzierte oder traumataverursachte Zellschädigungen freigesetzt. DAMPs und PAMPs können von den Immunzellrezeptoren (z. B. *pattern recognition receptors)* des angeborenen Immunsystems detektiert werden. Durch die dadurch hervorgerufene Aktivierung verschiedener zellulärer Signalkaskaden werden weitere Immunzellen rekrutiert. Ebenfalls aktivierte Transkriptionsfaktoren tragen zu einer gezielten Transkription, Synthese und Sekretion von bestimmten Entzündungsmediatoren, wie Zytokinen, Chemokinen oder Lipidmediatoren, bei. So wird, entweder direkt oder indirekt, auch die erworbene Immunantwort initiiert [Krüger *et al.,* 2015].

1.3.2 Pilzinitiierte Wirt-Pilzpathogen-Interaktionen

Pathogene Pilze haben verschiedene Virulenzstrategien entwickelt, um der Immunverteidigung des Wirts entgegenzuwirken. Sie sekretieren z. B. bestimmte Effektorproteine, die mit wirtseigenen Enzymen, zellulären Rezeptoren oder anderen biologischen Makromolekülen interagieren. Auf diese Weise können sie in den Wirtsmetabolismus eingreifen und ein erleichtertes Eindringen und Überleben im Wirt ermöglichen. Um sich vor der Erkennung durch die Immunzellen zu schützen, hat beispielsweise *A. fumigatus* eine Abschirmungsstrategie entwickelt. Die Zellwand der Konidien besteht unter anderem aus β-1,3,-Glucanen, Galactomannanen und Chitinen, die von den Immunzellrezeptoren als PAMPs erkannt werden. Durch die Ummantelung der Konidien mit 1,8-Dihydroxynaphthalin-Melanin und dem hydrophoben Protein RodA werden sie immunologisch inert (inaktiv). Andere Pilzstrategien zielen auf die Beschaffung bestimmter Nährstoffe ab, die für das Überleben und die Vermehrung benötigt werden. Darüber hinaus rufen Pathogene durch ihr Eindringen schwerwiegende Schäden im wirtseigenen Gewebe hervor. Diese können über Organschädigung zum

Organversagen und Tod des Wirts führen [Krüger *et al.,* 2015].

1.4 Heterogenität von Wirt-Pilzpathogen-Interaktionsdaten

1.4.1 Biologische Prozesse auf verschiedenen strukturellen Skalen

WPPI sind hochkomplexe biologische Prozesse, die sich über verschiedene strukturell voneinander getrennte Skalen erstrecken (Abbildung 1.3). Eine dieser Skalen ist die Wirt- oder Pilzpathogenpopulation selbst, deren Wachstums- und Interaktionsdynamiken sich in Abhängigkeit von ihrer Umgebung untersuchen lassen. Diese Populationsskala ist eng mit der Organismusskala verknüpft, auf der ein einzelnes Individuum entsprechend seines Phänotyps und Verhaltens analysiert werden kann. Die Organskala beschreibt zum einen die spezifischen Gegebenheiten der einzelnen Organe, zum anderen aber auch die Art und Weise, wie Organe miteinander in Verbindung stehen und gegenseitig Signale austauschen. Organe setzen sich wiederum aus verschiedenen Geweben und einer Vielzahl von Zellen zusammen, die ebenfalls als eigene Skalen betrachtet werden können. Die „unteren“ Skalen bilden die zelluläre und die molekulare Skala, die die Aktivitäten und biologischen Prozesse einzelner Zellen beschreiben. Um Informationen aus den verschiedenen Skalen gewinnen zu können, werden unter anderem bildgebende Verfahren angewendet, mikrobiologische Kulturen angelegt oder Körperflüssigkeiten und Gewebeproben im Labor untersucht. Speziell für die molekulare Skala sind Knockout-Verfahren, Perturbationsexperimente und die Anwendung von Hochdurchsatztechnologien zur Beobachtung auf verschiedenen Omik-Ebenen etabliert [Castiglione *et al.,* 2014; Schleicher *et al.,* 2016].

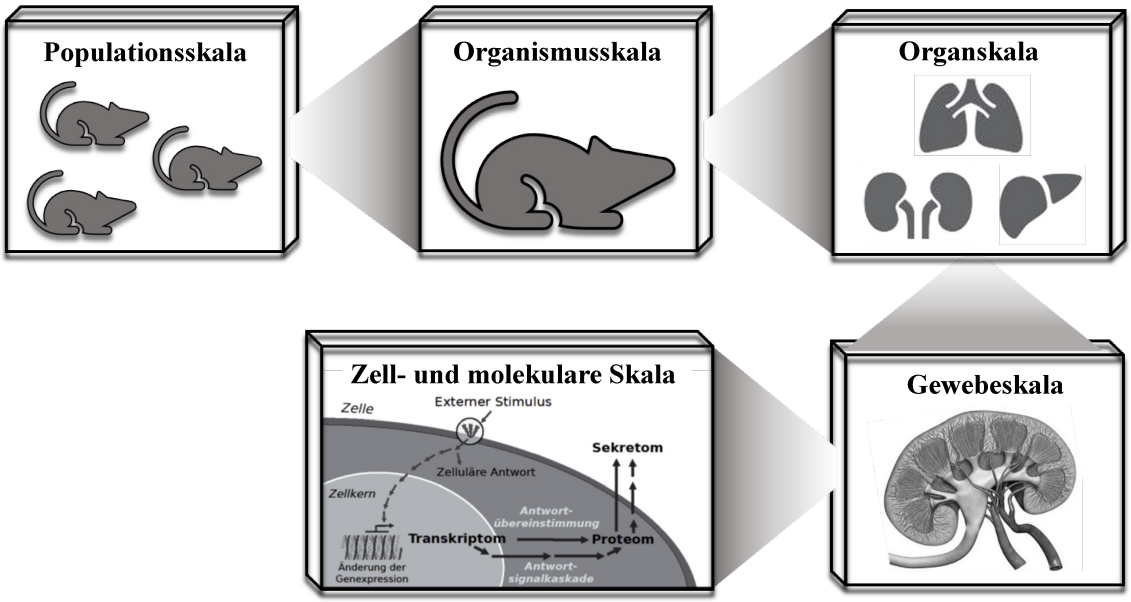


Abbildung 1.3: Strukturell voneinander getrennte Skalen im biologischen System (modifiziert nach [3])

Wie bereits in Kapitel 1.1 erwähnt, haben sich in den letzten zwei Jahrzehnten die Anzahl von Omik-Daten-basierten Studien und in diesem Zusammenhang auch die Nachfrage nach geeigneten Analysemethoden erheblich gesteigert. Aus diesem Grund sind die molekulare Skala und die damit verbundenen Omik-Ebenen für die vorliegende Dissertation von besonderer Bedeutung. Zu den bekanntesten Omik-Ebenen gehören das Genom, Transkriptom, Proteom, Metabolom und Interaktom. Das Genom beschreibt die Sequenz der Desoxyribonukleinsäure (DNA). Das Transkriptom enthält Informationen über die Anwesenheit und die relative Menge von Ribonukleinsäure (RNA)-Transkripten zu einem bestimmten Zeitpunkt. Das Proteom beschreibt das Ergebnis der RNA-Translation in Proteine bzw. die Menge aller Proteine z. B. einer Zelle. Das Metabolom repräsentiert die Gesamtheit der niedermolekularen Verbindungen, das Interaktom das komplette Netzwerk der molekularen Interaktionen [Joyce *et al.,* 2006].

Neben den strukturellen Skalen spielt auch die zeitliche Komponente von WPPI

[3] *http://media.gettyimages.com/vectors/human-organ-icons-acme-series-vector-id648899432?s=612x612,* Stand vom 09.08.2018;
http://www.paradisi.de/images_artikel/1/13977_0.jpg, Stand vom 01.10.2016

bzw. biologischen Prozessen eine wichtige Rolle. Während regulatorische Interaktionen auf der molekularen Skala innerhalb von Sekunden oder wenigen Minuten stattfinden können, umfassen die daraus hervorgehenden Effekte auf Zell-, Gewebe-, Organ- oder Ganzkörperskala Zeitspannen von wenigen Stunden bis mehreren Tagen [Schleicher *et al.,* 2016].

1.4.2 Experimentdesign und Messung von Omik-Daten

In experimentellen Studien spielt das zugrunde liegende Design eine wichtige Rolle für die Vergleichbarkeit der resultierenden Ergebnisse. So kann beispielsweise die Wahl der Kontrollen, die Anzahl der Replikate oder die Wahl der Messzeitpunkte zu Heterogenität der Daten führen. Auch die jeweils angewendeten Methoden zur Datenerhebung stellen eine große Herausforderung für die Datenintegration dar. So haben sich beispielsweise für die Untersuchung des Transkriptoms Microarray- oder *Next Generation Sequencing*-Ansätze (z. B. RNA-Seq) etabliert. Für Proteommessungen werden unter anderem zweidimensionale Gelelektrophoresen, Western Blots oder Massenspektrometrie verwendet. Die unterschiedlichen Ansätze und auch deren methodische Grenzen wirken sich maßgeblich auf die Quantität und die Qualität der Ergebnisse aus. Eine direkte Vergleichbarkeit der erhaltenen Daten ist somit meist nicht möglich [Hasin *et al.,* 2017]. Hierauf wird in Kapitel 3 detaillierter eingegangen.

1.4.3 Notwendigkeit der Datenintegration

Jede Omik-Technologie für sich ermöglicht es, einzelne Bereiche des biologischen Gesamtsystems näher zu betrachten und daraus neue Hypothesen für weitere Forschungsansätze zu entwickeln. Allerdings steigt durch die Fokussierung auf einzelne Aspekte nicht automatisch auch das Verständnis für die Struktur und Dynamik des Gesamtsystems. Da die globale Organisation auf den Interaktionen der einzelnen Ebenen und Komponenten basiert, können bereits kleine Änderungen in den „untersten“ Ebenen enorme Auswirkungen auf das Gesamtverhalten des biologischen Systems haben [Castiglione *et al.,* 2014]. Um diese Interebenenkomplexität zu verstehen, sollten möglichst viele Informationen aus den zur Verfügung stehenden Daten extrahiert und miteinander in Relation gesetzt werden. Die Schwierigkeit liegt darin, die zuvor beschriebene biologisch und methodisch bedingte Datenhete-

rogenität zu berücksichtigen und die Analysen entsprechend darauf auszurichten. Ein Vorteil ist, dass mit der integrierten Analyse einer Vielzahl von Datensätzen die biologische Variabilität und das experimentell bedingte Messrauschen innerhalb der einzelnen Datensätze erkannt und dann zumindest teilweise eliminiert werden kann (z. B. durch sogenannte globale Normalisierungsmethoden). So können auch jene Datensätze in die Analysen einbezogen werden, die sonst aufgrund übermäßigen Rauschens nur bedingt nutzbar sind [De Keersmaecker *et al.,* 2006; Hasin *et al.,* 2017].

1.5 Methoden zur Multi-Omik-Datenanalyse

Die in den letzten Jahrzehnten entwickelten Methoden zur Einzel- und Multiskalendatenanalyse lassen sich wie folgt klassifizieren: Globale oder fragestellungsspezifische Analysen, *Bottom-Up*- oder *Top-Down*-Ansätze, überwachtes oder unüberwachtes Lernen, sequentielle oder simultane Analysen. Diese Klassen werden im Folgenden nach dem Review von De Keersmaecker *et al.* aus dem Jahr 2006 näher beschrieben.

Globale Analysen beziehen alle verfügbaren Daten mit ein, ohne dabei einen Fokus auf z. B. einen bestimmten Signalweg zu legen. Sie eignen sich besonders für die Untersuchung globaler, d. h. skalenübergreifender Muster, oder um eine ganzheitliche Sicht auf das Verhalten eines Organismus zu erhalten. Ein intuitiver und weit verbreiteter Ansatz zur Aufdeckung von Gemeinsamkeiten zwischen den Datensätzen verschiedener Skalen ist der Vergleich ihrer jeweiligen Komponenten und gegebenenfalls deren Regulation. *Gene ontology* (GO)-Term- oder Signalweganalysen können Rückschlüsse darauf liefern, welche biologischen Prozesse oder Signalwege mit den Datensätzen korreliert oder assoziiert sind [Ashburner *et al.,* 2000]. Die fragengesteuerte Analyse bezieht sich dagegen auf Daten und Informationen, die einen ganz bestimmten Prozess betreffen [De Keersmaecker *et al.,* 2006].

Bei *Top-Down*-Ansätzen werden große Datenmengen (z. B. Hochdurchsatz-Omik-Daten) eines biologischen Systems auf konkrete Informationen für ein spezielles Anwendungsbeispiel heruntergebrochen. Bei *Bottom-Up*-Ansätzen verhält es sich genau andersherum. Hier werden Detailinformationen zusammengefasst, sodass all-

gemeinere Aussagen zu einem System getroffen werden können. Am Beispiel von Netzwerkrekonstruktionen starten *Bottom-Up*-Ansätze mit einem detaillierten, vorwissensbasierten oder hypothetischen Netzwerk, um so das Verhalten des biologischen Systems als Ganzes vorhersagen zu können. Dagegen zielen *Top-Down*-Ansätze auf die hypothesenfreie Rekonstruktion von einzelnen Interaktionen ab. Diese beruht auf allen verfügbaren Daten, die global und somit ohne Berücksichtigung von Vorwissen oder Hypothesen gemessen wurden [De Keersmaecker *et al.,* 2006].

Ziel des überwachten Lernens ist die Abschätzung einer mathematischen Funktion auf Grundlage eines Trainingsdatensatzes mit Paaren von Ein- und Ausgabevariablen. Diese Funktion soll in der Lage sein, für neue Eingabevariablen die zugehörigen Ausgabevariablen bestmöglich vorherzusagen. Klassifikations- und Regressionsprobleme sind typische Anwendungsbeispiele für das überwachte Lernen. Beim unüberwachten Lernen wird nicht zwischen Ein- und Ausgabevariablen unterschieden. Hier soll die zugrunde liegende Struktur oder Verteilung der vorliegenden Daten analysiert werden, um auf diese Weise mehr über die Daten selbst in Erfahrung zu bringen. Beispiele für unüberwachte Lernmethoden sind die Hauptkomponentenanalyse oder diverse Clustering-Methoden [De Keersmaecker *et al.,* 2006].

Bei einer sequentiellen Analyse wird eine Datenquelle nach der anderen genutzt, bei einer simultanen erfolgt die Nutzung gleichzeitig. So kann etwa das Clustering von Daten und die Suche nach überrepräsentierten Motiven in den jeweiligen Clustern entweder gleichzeitig oder nacheinander erfolgen. Die Entscheidung darüber, welcher Ansatz der geeignetere ist, hängt unter anderem von den jeweiligen Qualitätsanforderungen, aber auch von den unterschiedlichen Berechnungskomplexitäten und verfügbaren Rechnerressourcen ab [De Keersmaecker *et al.,* 2006].

Im Folgenden werden einige der Methoden vorgestellt, die auch in den Manuskripten des Hauptteils Anwendung finden.

1.5.1 Clustering-basierte Ansätze

In der heutigen Forschung hat sich das Clustering als ein bewährter Ansatz in der WPI-Datenanalyse erwiesen. Clustering ist eine Technik des unüberwachten Lernens und

gruppiert Komponenten basierend auf Ähnlichkeits- bzw. Distanzmaßen (oder auch Kriterien des Zusammenhangs, der Kompaktheit bzw. der Separierbarkeit). Komponenten innerhalb eines Clusters weisen dabei eine große Ähnlichkeit (z. B. Korrelation) und geringe Distanz zueinander auf, Komponenten verschiedener Cluster eine größtmögliche Distanz und geringe Ähnlichkeit. Damit ist es möglich, jene Komponenten zu identifizieren, die eine ähnliche Funktion oder ein gemeinsames Profil (z. B. in der Genexpression) aufweisen [Oyelade *et al.,* 2016]. So können auch solche Signalwege oder Komponenten detektiert werden, die bisher noch nicht mit der jeweils zugrunde liegenden Fragestellung in Verbindung gebracht wurden. Oyelade *et al.* beschreiben in ihrem Review aus dem Jahr 2016 einige der gebräuchlichsten Clustering-Methoden, auf die in den folgenden Absätzen etwas näher eingegangen werden soll.

Grundlegend kann zwischen partiellem und komplettem Clustering mit entweder klar voneinander abgegrenzten oder aber überlappenden Clustern unterschieden werden. Im Gegensatz zum kompletten Clustering müssen beim partiellen nicht alle zu betrachtenden Komponenten tatsächlich auf die einzelnen Cluster verteilt werden. Bei klar voneinander abgegrenzten Clustern kann jede Komponente genau einem Cluster zugeordnet werden. Entsprechend erlaubt das überlappende Clustering die Zuordnung zu mehreren Clustern. Eine der ältesten Ansätze ist das hierarchische Clustern, bei dem die vorliegenden Daten in hierarchisch geordnete Cluster aufgeteilt werden. Dabei wird entweder nach einem *Bottom-Up-* oder einem *Top-Down*-Verfahren vorgegangen. Bei dem *Bottom-Up*-Ansatz des agglomerativen Clusterns (*agglomerative nesting,* kurz AGNES) stellt jedes Objekt des vorliegenden Datensatzes zunächst ein eigenes Cluster dar. Auf Grundlage von minimaler Distanz bzw. der größten Ähnlichkeit werden die einzelnen Cluster dann schrittweise zu immer größeren zusammengefasst, bis nur noch ein einziges großes Cluster vorliegt. Ein *Top-Down*-Ansatz ist das divisive Clustern, bei dem zunächst alle Komponenten einem einzigen großen Cluster angehören und dann nach und nach entsprechend ihrer Distanzen in kleinere Cluster aufgesplittet werden. Beispiele hierfür sind die Methoden DIANA *(divisive analysis)* und SOTA *(self-organization tree algorithm)* [Oyelade *et al.,* 2016].

Eine weitere große Gruppe von Clustering-Methoden nutzt den Partitionierungsansatz. Bei diesem wird bereits zu Beginn eine bestimmte Anzahl von resultierenden Clustern festgelegt. Über die Verschiebung der Clusterzentren und der Minimierung

einer Fehlerfunktion werden dann die jeweiligen Komponentenzugehörigkeiten bestimmt. Im Unterschied zu hierarchischem Clustern, bei dem eine einmal festgelegte Zugehörigkeit nicht mehr verändert wird, ist hier ein Wechsel möglich. Der einfachste und weit verbreitetste Partitionierungsalgorithmus ist k-Means, der auf der zufälligen Auswahl der initialen Clusterzentren basiert. Neben k-Means gibt es z. B. auch das k-Medoids Clustering (PAM, *partitioning around medoids).* In PAM wird jedes Cluster durch eines seiner Elemente, das sogenannte Medoid, repräsentiert. Dabei wird das Medoid so gewählt, dass die durchschnittliche Distanz zwischen ihm und allen anderen Elementen des Clusters minimal ist. Eine Erweiterung von PAM ist CLARA *(clustering large applications).* CLARA kann auch für große Datensätze genutzt werden, da jeweils nur ein kleiner, repräsentativer Teil der tatsächlichen Daten für das Clustering verwendet wird. Neben den eben genannten gibt es noch viele weitere Methoden wie model-, dichte-, grid-basierte oder kombinierte Verfahren, auf die hier aber nicht näher eingegangen werden soll [Oyelade *et al.,* 2016].

Ein Beispiel aus der aktuellen Wirt-Pathogen-Forschung ist die Studie von Roy *et al.* aus dem Jahr 2018. Sie befasst sich mit den Interaktionen zwischen murinen Makrophagen und *Mycobacterium tuberculosis* während einer Infektion. Dabei liegt das Hauptaugenmerk auf der zeitlichen Dynamik des Makrophagentranskriptoms sowie dem Einfluss von Interferon-γ und den Zytokinen IL-4 und IL-13. Mithilfe des k-Means-Clustering konnten Gruppen von Genen identifiziert werden, die die gleichen Genexpressionsmuster im Zeitverlauf aufweisen. Das erhaltene Genexpressionsprofil der infizierten Makrophagen konnte zu einem besseren Verständnis der Wirt-Pathogen-Dynamiken während einer *M. tuberculosis*-Infektion beitragen und gleichzeitig wichtige Hinweise auf potenzielle transkriptionelle Biomarker einer solchen liefern [Roy *et al.,* 2018].

1.5.2 Interaktionsnetzwerkbasierte Ansätze

Die Zelle als biologisches System besteht aus einer Vielzahl unterschiedlicher Komponenten (Gene, Proteine, Metabolite etc.), die miteinander interagieren. Mit ihrer Hilfe kann die Zelle umweltbedingte Änderungen wahrnehmen, die jeweiligen Signale verarbeiten, weiterleiten und schließlich dynamisch auf die Änderungen reagieren. All diese Komponenten und deren direkte (z. B. physikalische Verbindungen) und

indirekte Interaktionen (z. B. regulatorische Funktionen) spannen ein umfangreiches Netzwerk auf [De Keersmaecker *et al.,* 2006]. Graphisch wird ein solches Netzwerk typischerweise über Knoten, die die Komponenten repräsentieren, und gerichtete oder ungerichtete Kanten dargestellt, die die Interaktionen zwischen ihnen widerspiegeln. Weitverbreitet sind vor allem genregulatorische Netzwerke (GRN) und Protein-Protein-Interaktionsnetzwerke (PPIN). Die Problematik besteht darin, dass solche Netzwerke selten in ihrer kompletten Struktur bekannt sind. Häufig fehlt Teilwissen über alle tatsächlich beteiligten Komponenten und/oder deren Interaktionen. Durch die Rekonstruktion von Netzwerken gelingt es, Zusammenhänge innerhalb des realen biologischen Netzwerks besser zu verstehen und mathematische oder Computermodelle zu entwickeln, die ein bestimmtes Verhalten simulieren oder vorhersagen können.

Netzwerkinferenz

In einem GRN werden Gene und deren direkte oder indirekte Beeinflussung untereinander dargestellt. Bei einer indirekten Interaktion kann beispielsweise ein Gen einen Transkriptionsfaktor kodieren, der an die Promoterregion eines anderen Gens binden und somit dessen Regulation beeinflussen kann. Auch Signalkaskaden oder ganze Stoffwechselwege können durch Kanten repräsentiert werden. Netzwerkinferenz wird genutzt, um solche, teilweise sehr komplexen, Interaktionen zwischen den Genen vorherzusagen. Sie setzt sich zusammen aus der Identifikation von potenziellen Regulatoren, der Vorhersage von Zielgenen und der Vorhersage der jeweiligen Art der Interaktion. Dafür spielen vor allem Genexpressionsintensitäten, Zeitreihendaten und die Verwendung von Vorwissen eine wichtige Rolle. Je nach Größe und Dynamik des Netzwerks, der Richtung der Kanten und der Verwendung von Vorwissen wurde eine Vielzahl von Methoden entwickelt, um die Netzwerkinferenz zu ermöglichen. Darunter finden sich beispielsweise Methoden wie Boolesche Modellierung, probabilistische Modellierung, informationstheoretische Methoden, Regression oder Optimierung [Linde *et al.,* 2015].

Ein Beispiel für die Rekonstruktion von großen (z. B. genomweiten), gerichteten Netzwerken ist der ExTILAR-Algorithmus [Vlaic *et al.,* 2012]. Er kombiniert dynamische Modellierungsansätze mit linearen Regressionsmethoden, um dynamische Netzwerke auf der Basis von gewöhnlichen Differentialgleichungen zu inferrieren.

ExTILAR konnte unter anderem bereits erfolgreich auf Genexpressionsdaten von Maushepatozyten angewendet werden, die auf einem Wechsel von nährstoffarmem zu nährstoffreichem Kulturmedium beruhen. Auf Grundlage dieser Daten und zusätzlichem Vorwissen war es möglich, ein Transkriptionsfaktornetzwerk zu generieren, das die biologischen Hauptprozesse in den Hepatozyten widerspiegelt. Darüber hinaus konnten bisher unbekannte Interaktionen zwischen den zwei Transkriptionsfaktoren Tgif1 und Atf3 identifiziert werden, die Hinweise auf deren Rolle in weiteren biologischen Prozessen geben [Vlaic *et al.,* 2012]. Erste Ansätze zur Netzwerkinferenz sind auch zur Erforschung von WPPI etabliert [Guthke *et al.,* 2016]. Auch hier konnten neue Interaktionen vorhergesagt werden, die anschließend experimentell bestätigt wurden [Altwasser *et al.,* 2015; Linde *et al.,* 2010; Linde *et al.,* 2012].

Netzwerkmodule

Ist ein Netzwerk durch vorhandenes Vorwissen oder Inferenz bereits größtenteils bekannt, spiegelt es die komplexen strukturellen und funktionellen Zusammenhänge seiner Komponenten wider. Um diese detaillierter untersuchen zu können, werden zusammenhängende Subnetzwerke aus den größeren Interaktionsnetzwerken extrahiert. Solche Subnetzwerke enthalten Gruppen von eng miteinander in Verbindung stehenden Komponenten und werden als Module bezeichnet [Lin *et al.,* 2015]. Die Zerlegung eines Netzwerks in seine modulare Struktur hat auch den Vorteil, dass weitere Subnetzwerke integriert werden können, die von anderen Omik-Datentypen stammen. So ist es möglich, in ein bereits bestehendes PPIN, basierend auf Proteomdaten, ein GRN, basierend auf Transkriptomdaten, einzubeziehen. Auf diese Weise werden nicht nur die einzelnen molekularen Ebenen wie Transkriptom oder Proteom analysiert, sondern eine Verbindung zwischen beiden geschaffen, sodass eine zusammenhängende Betrachtung möglich ist [Joyce *et al.,* 2006].

Module können anhand der zugrunde liegenden Annahmen kategorisiert werden (Abbildung 1.4). So gibt es beispielsweise funktionelle Module, deren Komponenten mit spezifischen biologischen Funktionen assoziiert sind. Andere Module beziehen sich dagegen auf die topologische Struktur des Netzwerks, wie sie beispielsweise durch maximale Cliquen charakterisiert wird. Eine Clique ist eine Gruppe von Komponenten, wobei jedes Paar von Komponenten miteinander durch eine Kante verbunden ist. Sie ist maximal, wenn sie nicht Teil einer noch größeren Clique ist.

Auch Mischformen von Modulen sind denkbar. Sogenannte regulatorische Module berücksichtigen zum einen die jeweilige Cliquenstruktur des Netzwerks, basieren zum anderen aber auch auf den Expressions- bzw. Regulationswerten ihrer Komponenten. Somit beschreiben sie co-regulierte Komponenten, die eine gemeinsame biologische Funktion aufweisen [Vlaic *et al.*, 2012].

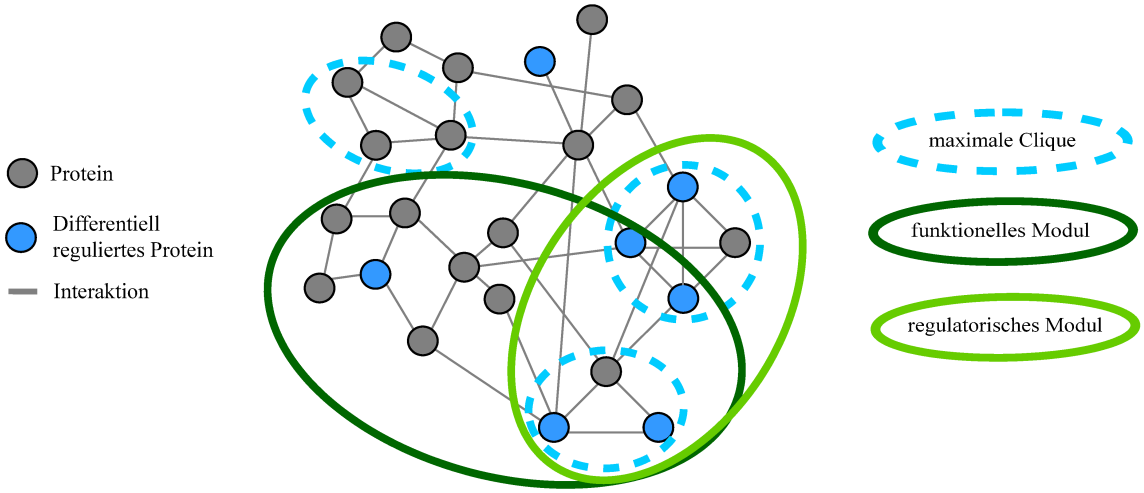


Abbildung 1.4: Potenzielle Netzwerkmodule eines PPIN

Ein Beispiel für eine moduldetektierende Software ist KeyPathwayMiner. Sie wurde im Jahr 2011 von Alcaraz *et al.* entwickelt und seither regelmäßig erweitert, um die Extraktion und Visualisierung von maximal zusammenhängenden Subnetzwerken aus umfangreicheren Interaktionsnetzwerken zu ermöglichen. Auf der Grundlage von Vorwissen und experimentellen Regulationsdaten erlaubt KeyPathwayMiner die Identifizierung von funktionellen Modulen und regulatorisch aktiven Signalwegen. Dabei ist KeyPathwayMiner nicht nur für Daten einer einzelnen molekularen Ebene geeignet, sondern auch auf die Kombination multipler Omik-Datensätze anwendbar [Alcaraz *et al.*, 2014].

1.5.3 Modellierung

Eine weitere Möglichkeit für die Integration von Multiskalen-WPI-Daten ist die computergestützte Modellierung. Dabei wird die Komplexität der WPI gezielt reduziert, um mithilfe eines vereinfachten Modells wichtige Schlüsseleigenschaften identifizieren zu können. Die dafür entwickelten Methoden lassen sich in zeitunabhängige,

zeitkontinuierliche und zeitdiskrete Gruppen einteilen. Bei den zeitunabhängigen Methoden wird der Zeitfaktor bei der Modellierung vernachlässigt. Beispiele sind die bedingungsbasierte Modellierung sowie Spieltheorieansätze. Da jedoch oftmals die WPI-Dynamik über eine bestimmte Zeitspanne hinweg untersucht werden soll, kann die Zeit abhängig von der jeweiligen Fragestellung entweder als kontinuierliche oder diskrete Größe in das Modell integriert werden. Gewöhnliche Differentialgleichungen für die zeitkontinuierliche Modellierung sind hier eine weitverbreitete Methode. Kommt zusätzlich zum Zeitfaktor auch der räumliche Aspekt hinzu, eignen sich beispielsweise partielle Differentialgleichungen. Agentenbasierte, zustandsbasierte, zellulär-automatenbasierte, Boolesche und probabilistische Ansätze sind hingegen Beispiele für die zeitdiskrete Modellierung [Schleicher *et al.,* 2016].

Ein aktuelles Beispiel für die Modellierung von WPPI ist das von Dühring *et al.* entwickelte spieltheoretische Modell aus dem Jahr 2017. Der biologische Hintergrund ist hierbei die wirtsinitiierte Aufnahme eingedrungener *C. albicans*-Zellen durch die Makrophagen (auch Phagozytose genannt). Ziel des Modells ist es, die verschiedenen (Überlebens-) Strategien von Makrophagen und *C. albicans* nach der Phagozytose zu beschreiben und in Abhängigkeit davon Nash-Gleichgewichte (Lösungen des Spiels) zu bestimmen. Da die Nash-Gleichgewichte direkte Konsequenzen der Modellparametrisierung sind, lassen sich daraus verschiedene biologische Szenarien ableiten. Dabei wird unter Zuhilfenahme dynamischer Optimierung die Populationsdynamik der Makrophagen-*C. albicans*-Interaktionen untersucht, um die optimale Kontrolle der Pilze durch die Makrophagen zu erreichen [Dühring *et al.,* 2017].

Für die Integration von Multiskalen-WPI-Daten hat sich in den vergangenen Jahren die Generierung von Hybridmodellen – also die Kombination verschiedener Ansätze – als vorteilhaft erwiesen. Blickensdorf *et al.* haben beispielsweise ein Modell entwickelt, das sowohl die agentenbasierte Modellierung als auch partielle Differentialgleichungen miteinander vereint. Hintergrund des Modells ist, dass oftmals Mäuse als Modellorganismen für die Untersuchung von WPPI genutzt werden. Um allerdings Rückschlüsse auf Infektionen im Menschen ziehen zu können, müssen die unterschiedlichen physiologischen Faktoren von Mensch und Maus sowie die verschiedenen Inhalations- bzw. Infektionsdosen berücksichtigt werden. Die Simulierung einer *A. fumigatus*-Infektion in einer humanen und in einer murinen Lunge ermöglicht die vergleichende

Quantifizierung der Infektionsbekämpfung in den beiden Wirten. Mithilfe des entwickelten Hybridmodells erhielten Blickensdorf *et al.* Hinweise darauf, dass die Beseitigung der *A. fumigatus*-Konidien in der Maus effizienter erfolgt als im Menschen [Blickensdorf *et al.,* 2019].

1.6 Zielstellung

In den letzten Jahren wurden immer neue Ansätze entwickelt, um die Integration von Daten aus heterogenen Quellen, d. h. aus verschiedenen zeitlichen und strukturellen Skalen, zu ermöglichen. Allerdings bezieht sich deren Anwendung hinsichtlich der Untersuchung von WPI zumeist auf Interaktionen zwischen Wirt und bakteriellen Erregern. Dagegen steht die Erforschung von Interaktionen zwischen Wirt und pathogenen Pilzen unter Zuhilfenahme solcher Integrationsansätze derzeit noch am Anfang. Daher behandelt die vorliegende Dissertation die Integration von Multiskalen- und Multi-Omik-Daten, um neue Erkenntnisse im Bereich der WPPI zu erlangen. Es wird auf Beispiele der aktuellen infektionsbiologischen Forschung eingegangen, an denen ich im Rahmen der Dissertation mitgewirkt habe. Anhand dieser Beispiele werden die Vorteile und die Wichtigkeit sowie bewährte Methoden der Integration von Multiskalen- und Multi-Omik-Daten aufgezeigt. Dabei werden nicht nur einige gängige Methoden zur Analyse dieser Daten vorgestellt, sondern auch der neu entwickelte ModuleDiscoverer-Algorithmus zur Identifikation von regulatorischen Modulen.

2 Manuskripte

2.1 Manuskript 1: „Dual-species transcriptional profiling during systemic candidiasis reveals organ-specific host-pathogen interactions“

Status
Veröffentlicht im November 2016

Literaturangabe
Hebecker, B.*, Vlaic, S.*, Conrad, T., Bauer, M., Brunke, S., Kapitan, M., Linde, J., Hube, B., Jacobsen, I. D. (2016). Dual-species transcriptional profiling during systemic candidiasis reveals organ-specific host-pathogen interactions. *Scientific Reports,* 6. https://doi.org/10.1038/srep36055

Übersicht
Im ersten Manuskript dieser Dissertation wird das Projekt von Hebecker und Vlaic *et al.* vorgestellt. In diesem sollten die WPPI in mit *C. albicans*-infizierten Mäusen auf Basis von Multiskalendaten (d. h. Transkriptomdaten verschiedener Organe und Zeitpunkte) untersucht werden. Mithilfe von klassischen Methoden (z. B. Komponentenvergleich oder Signalweganalyse), Clustering-Verfahren sowie der Inferenz eines Interspezies-GRN wurden die Daten analysiert und integriert. So war es möglich, umfassende Genexpressionsprofile für Wirt und Pilz zu erstellen und als WPPI in Zusammenhang zu bringen. Die Studie hat gezeigt, dass die Antwort des Wirts und die Anpassung des Pilzes während einer disseminierten Candidose auf dem Transkriptomlevel sowohl organspezifisch als auch zeitabhängig sind. So

* geteilte Erstautorenschaft

kommt es beispielsweise in der Niere zu einer späteren Aktivierung der angeborenen Abwehrmechanismen, dafür aber auch zu einer stärkeren proinflammatorischen Antwort als im Vergleich zur Leber. Die verschiedenartigen Umgebungen von Leber und Niere tragen auch zu einer unterschiedlichen Aktivität von zellwand- und zelloberflächenmodifizierenden Enzymen bei. Die dadurch hervorgerufenen strukturellen Veränderungen können die Interaktionen von Immunzellen beeinflussen und somit zu einem spezifischen Infektionsverlauf in den jeweiligen Organen beitragen.

Beiträge
JID und HeB konzipierten und planten die Experimente. HeB war für die Mausversuche, die RNA-Isolation und die *C. albicans*-Microarrays zuständig. Zytokin- und Phagozytose-*Assays* wurden durch HeB und KM durchgeführt. VS analysierte die Genexpressionsdaten der Wirtsseite. Die Cluster- und Inferenznetzwerkanalysen wurden von VS und CT durchgeführt. CT, HeB und LJ waren für die Genexpressionsanalyse der Pathogenseite zuständig. BS und HeB führten eine GSEA durch. HeB, HuB, BM und JID interpretierten die Daten. HeB und JID schrieben das Manuskript, wobei alle Autoren an der Überprüfung und Überarbeitung des Manuskripts beteiligt waren.

www.nature.com/scientificreports

SCIENTIFIC REPORTS

OPEN

Dual-species transcriptional profiling during systemic candidiasis reveals organ-specific host-pathogen interactions

Received: 19 July 2016
Accepted: 11 October 2016
Published: 03 November 2016

Betty Hebecker[1,2,*,†], Sebastian Vlaic[3,4,5,*], Theresia Conrad[4], Michael Bauer[2,6], Sascha Brunke[7], Mario Kapitan[1], Jörg Linde[4], Bernhard Hube[2,7,8] & Ilse D. Jacobsen[1,2,8]

Candida albicans **is a common cause of life-threatening fungal bloodstream infections. In the murine model of systemic candidiasis, the kidney is the primary target organ while the fungal load declines over time in liver and spleen. To better understand these organ-specific differences in host-pathogen interaction, we performed gene expression profiling of murine kidney, liver and spleen and determined the fungal transcriptome in liver and kidney. We observed a delayed transcriptional immune response accompanied by late induction of fungal stress response genes in the kidneys. In contrast, early upregulation of the proinflammatory response in the liver was associated with a fungal transcriptome resembling response to phagocytosis, suggesting that phagocytes contribute significantly to fungal control in the liver. Notably, *C. albicans* hypha-associated genes were upregulated in the absence of visible filamentation in the liver, indicating an uncoupling of gene expression and morphology and a morphology-independent effect by hypha-associated genes in this organ. Consistently, integration of host and pathogen transcriptional data in an inter-species gene regulatory network indicated connections of *C. albicans* cell wall remodelling and metabolism to the organ-specific immune responses.**

Candida albicans is the most common cause of life-threatening fungal bloodstream and disseminated infections. The crude mortality of disseminated infections caused by *Candida* is >50%, higher than for bacterial blood stream infection[1]. The murine model of hematogenously disseminated candidiasis is most commonly used to study systemic candidiasis and to evaluate efficacy of antifungal therapy. Following intravenous infection, *C. albicans* initially infects almost all organs; however, while the fungal load increases over time in kidneys, it declines in liver and spleen[2,3]. Moreover, *C. albicans* is able to form filaments – a hallmark of pathogenicity - in the kidney within two hours after intravenous infection, whereas no hypha formation is observed in liver and spleen[4]. Thus, the kidney appears to be the prime target organ for disseminated candidiasis in mice. In contrast, disseminated candidiasis in humans affects kidneys but also commonly leads to infection of liver and spleen[5].

The relative susceptibility of the kidneys in murine candidiasis has been linked to the specific immunological setup of this organ: Neutrophils and macrophages are present in higher numbers in liver and spleen than in the kidneys of naive animals[2]. In addition, in comparison to spleen and liver, leukocytes are recruited later to *C. albicans*-infected kidneys[2]. Early accumulation of neutrophils has a protective effect and mononuclear

[1]Research Group Microbial Immunology, Leibniz Institute for Natural Product Research and Infection Biology (Hans Knöll Institute), Jena, Germany. [2]Center for Sepsis Control and Care, Jena University Hospital, Jena, Germany. [3]Department of General, Visceral and Vascular Surgery, Experimental Transplantation Surgery, Jena University Hospital, Jena, Germany. [4]Research Group Systems Biology/Bioinformatics, Leibniz Institute for Natural Product Research and Infection Biology (Hans Knöll Institute), Jena, Germany. [5]Department of Bioinformatics, Friedrich-Schiller-University Jena, Germany. [6]Department of Anaesthesiology and Intensive Care Therapy, Jena University Hospital, Jena, Germany. [7]Department of Microbial Pathogenicity Mechanisms, Leibniz Institute for Natural Product Research and Infection Biology (Hans Knöll Institute), Jena, Germany. [8]Friedrich-Schiller-University, Jena, Germany. [†]Present address: Aberdeen Fungal Group, Institute of Medical Sciences, University of Aberdeen, Aberdeen, UK. [*]These authors contributed equally to this work. Correspondence and requests for materials should be addressed to I.D.J. (email: ilse.jacobsen@leibniz-hki.de)

phagocytes can directly kill *C. albicans in vitro* and *in vivo*[2,6,7]. As kidney-resident macrophages and inflammatory monocytes are important for fungal clearance in the kidney, the delay in the early phagocytic response in the kidney seems to be critical for progression of infection[6,7]. However, at later stages of infection continued neutrophil recruitment induced by locally produced cytokines and chemokines in the kidney directly contributes to immunopathology[3]. Thus, the immune response, fungal clearance and pathological consequences vary significantly in different organs during systemic *C. albicans* infection in the murine model.

C. albicans expresses a variety of virulence factors that contribute to pathogenesis, including the morphological transition between yeast and hypha, the expression of hypha-associated virulence factors and metabolic flexibility[6,8,9]. In addition to morphogenesis, the *C. albicans* cell wall structure is strongly influenced by environmental factors, such as available carbon sources or environmental stresses, leading to rapid cell wall remodelling processes *in vitro* and *in vivo* that affect interaction with immune cells[10–12]. Furthermore, *C. albicans* rapidly adapts to the changing environments encountered during infection by changes in gene expression, translation, and post-translational modifications[13,14].

Thus, both the host and the fungus respond rapidly to the changing conditions during systemic candidiasis. Such changes can be monitored on the transcriptional level by *in vivo* gene expression profiling to estimate the functional adjustments during invasive growth of *C. albicans*. Recently, Xu *et al.* analysed 479 *C. albicans* and 46 mouse genes from kidney samples using the NanoString technology[15]. Although less than 2% of the *Candida* genome was covered, a clear induction of *C. albicans* hypha-associated as well as zinc and iron responsive genes was observed. Similarly, RNASeq of *C. albicans* cells infecting the mouse kidney revealed regulation of iron metabolism, hypha formation, and metabolic adaptation during infection[16].

Even though these studies significantly contributed to our understanding of disseminated candidiasis, they focused on the kidney as the main target organ. However, given the different immune responses in different organs, it is likely that organ-specific host-pathogen interactions occur and contribute to the different course of infection in murine organs. Understanding these differences may be crucial in understanding pathogenicity of life-threatening systemic *C. albicans* infections. Therefore, we performed a time-course transcriptional analysis of liver, spleen, and kidney samples from mice infected intravenously with *C. albicans*. Our results demonstrate not only a delayed immune response in the kidney compared to liver and spleen, but also qualitative differences in the responses of these organs. Microarray analyses of *C. albicans* cells likewise showed organ-specific adaptation processes. This includes an increased expression of hypha-associated genes in the liver in the absence of visible filamentation, suggesting that gene expression and morphogenesis is uncoupled under these specific conditions *in vivo*.

Results and Discussion

Murine spleen, liver, and kidney show different kinetics of transcriptional responses to systemic candidiasis. Transcriptional profiling is a powerful tool to elucidate host-pathogen interactions during infection[17,18] that has been successfully applied to the murine model of disseminated candidiasis[15,16,19–21]. However, previous studies focused on the kidney as the main target organ. Given the differing progression of infection in kidney, liver, and spleen in mice and the frequent involvement of liver and spleen in human disseminated candidiasis, we performed a transcriptional profiling of these three organs over the time course of infection in mice.

Principal component analysis (PCA) of the murine transcriptional data sets showed that 54% of the variance (PC1 and PC2) among all transcripts in the data sets could be attributed to organ-specific expression differences (Fig. 1A). Temporal changes also contributed to variance (7%) and biological replicates from individual organs clustered according to time point (Fig. 1A). PCA reflects general expression trends in the data which are due to alterations in the expression of the majority of genes represented on the microarray. When specifically analyzing differentially expressed genes (DEGs), organ-specific kinetics of transcriptional changes became evident: Most of the changes in gene expression in liver and spleen were observed already 24 h p.i., whereas in kidney the number of DEGs continuously increased over time (Fig. 1B). We thus performed cluster analysis to categorize the genes according to organ expression kinetics and analysed these clusters for enriched GO terms. All organs responded to infection by regulation of genes associated with immune processes and stress response (Suppl. Fig. 1A, Suppl. data 1 and 2) and, more specifically, upregulation of genes associated with the response to interferon β and cytokine biosynthesis (represented by cluster 5, Suppl. Fig. 1B). However, the expression kinetics differed between organs: As shown in Fig. 2A, several genes in TLR and NLR pathways were upregulated from 8 h to 24 h p.i. in the liver, while in the kidney sizable induction of these pathways was detected only after 24 h and further increased towards 72 h p.i. At 72 h p.i., complement activation was significantly induced in the liver. In the spleen, especially genes in T- and B-cell receptor signalling pathways were enriched; however, the involved genes were mainly downregulated compared to the control. The subset of genes displayed in Fig. 2B highlights the differences in the expression kinetics of genes associated with inflammation: Expression of most of these genes in the liver increased strongly early after infection with a tendency to return to control sample expression levels at later time points, whereas in kidney samples the largest changes were detected 72 h p.i.

Thus, while all organs responded to infection by downregulation of genes associated with organ-specific functions and upregulation of immune processes, the kinetics and type of immune responses differed.

Organ-specific transcriptional changes in systemic candidiasis reflect activation of local immune responses and organ damage. As discussed above, the transcriptional changes in the kidney were characterised by late upregulation of proinflammatory pathways, consistent with the reported delayed recruitment of immune cells in this organ[2]. Notable exceptions were the chemokine CXCL1 (also known as keratinocyte cell–derived chemokine [KC]) and the intercellular adhesion molecule ICAM-1, which is important for leukocyte transmigration. These genes were highly upregulated after 8 h in the kidney (Fig. 2B). As we analysed the transcriptome of whole organs, the precise cellular source of early immune transcripts is not clear. It has, however, been shown that CXCL1 is mainly produced by resident tissue macrophages via TLR4 signalling through

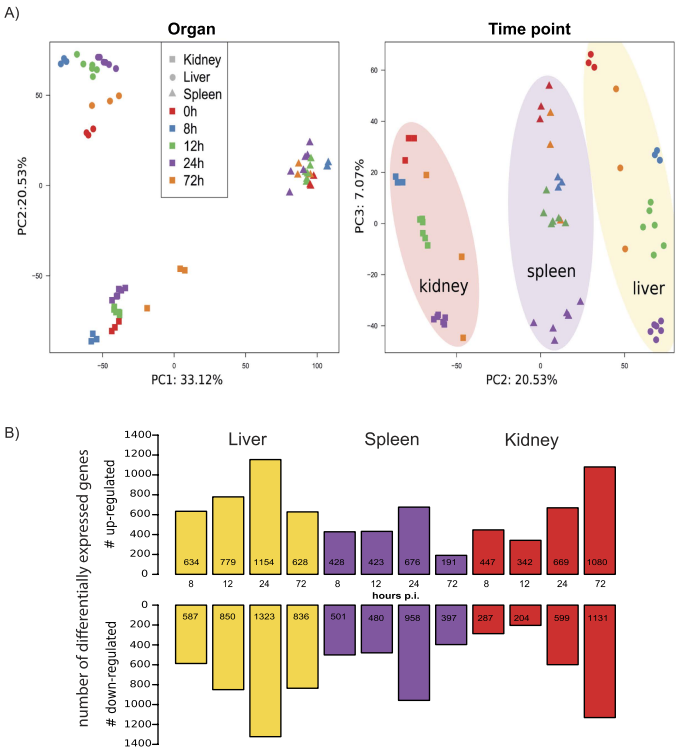


Figure 1. Principle component (PC) analysis of mouse expression data (A) and numbers of differentially expressed genes (DEGs) (B). (**A**) Left: Approximately 54% of the variation between data sets was captured by the first two PCs and coincided with the organ the sample was obtained from. Right: The third PC captured about 7% of the variance in the data and correlated with the temporal effects in the data set. (**B**) Numbers of genes differentially regulated ($p < 0.05$) in kidneys, liver and spleen during *C. albicans* infection as compared to the PBS control.

MyD88[22]. In comparison to liver and spleen, the number of resident immune cells in the kidney is low[2] and even though transcript abundance of CXCL1 increased in the kidney 8 h p.i., protein levels are known to increase more steeply only after 24 h[3]. This suggests that the potential of resident renal macrophages to produce this chemokine is not sufficient to drive early neutrophil recruitment at high levels. The progressive upregulation of proinflammatory pathways in the kidney over the course of infection can be explained partially by the increasing or persistent fungal burden which leads to ongoing immune stimulation. Fungal growth is furthermore associated with progressive renal damage[3,23], reflected in our data set by (i) the upregulation of genes involved in wound healing and (ii) the downregulation of genes in "renal system processes", "transport processes", and "ion homeostasis" observed at later time points (Suppl. Fig. 1).

Organ-specific immune responses included the induction of the acute phase response in the liver (Suppl. Fig. 1A) and the TGF-β pathway and genes associated with lymphocyte activation and leukocyte proliferation in the kidney (Suppl. Fig. 1A). While the latter were upregulated at late time points in the kidney, their expression was continuously decreasing until 24 h in the spleen, with a tendency to return to basal levels towards 72 h p.i. As a secondary lymphoid organ, main functions of the spleen are associated with leukocyte activation and proliferation. Therefore, transiently reduced expression of these genes could be interpreted as reduced expression of genes with organ-specific functions. The transient nature of these alterations might thus reflect the successful control

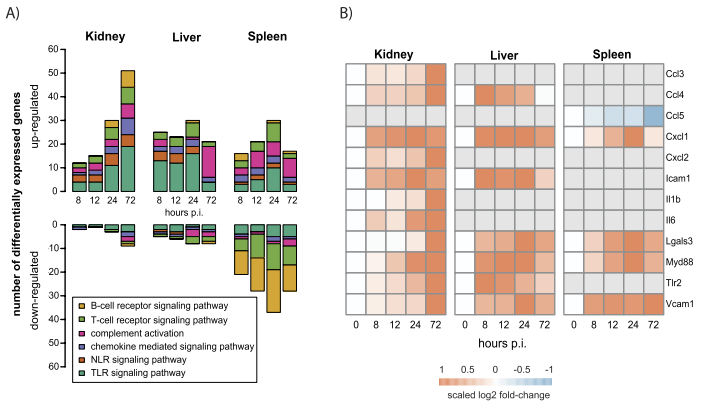


Figure 2. Gene expression changes in immune response pathways after systemic *C. albicans* infection. (**A**) Number of differentially regulated genes in distinct immune pathways. The org.Mm.eg.db package for R was used to identify differentially expressed genes associated with the pathways "TLR signaling pathway" (GO:0002224), "NLR signaling pathway" (GO:0035872), "chemokine mediated signaling pathway" (GO:0070098), "complement activation" (GO:0006956), "T-cell receptor signaling pathway" (GO:0050852) and "B-cell receptor signaling pathway" (GO:0050853) including associated child-terms, respectively. (**B**) Gene expression of selected differentially expressed immune genes. For each organ, log2 fold-changes were gene-wise maximum scaled between −1 and 1 (scaled log2 FC). Measurements of probes mapping to the same gene were mean averaged.

of fungal growth in the spleen. Downregulation was also observed in the liver for genes involved in liver function ("monocarboxylic acid metabolism" and "lipid homeostasis", cluster 9, Suppl. Fig. 1). While these genes also showed a trend to return to basal levels at 72 h p.i., it was not as pronounced as for the spleen, indicating that the transcriptional response of this organ was continuing even though infection is controlled to a similar degree in liver and spleen. Furthermore, although the kinetics of fungal burden and immune cell recruitment in liver and spleen are similar during systemic candidiasis, the transcriptional immune response of these organs differed significantly: Components of the innate immune response, such as TLR and NLR signalling pathways, were induced to a higher extend in the liver. The lack of strong induction of genes associated with proinflammatory responses is consistent with the comparatively low and transient production of proinflammatory cytokines in the spleen during systemic candidiasis[3]. We find it noteworthy, however, that systemic responses were clearly induced in the liver, e.g. the acute phase response, and that genes involved in complement activation were upregulated especially after 72 h p.i. This might indicate the role of the liver for the systemic immune response during ongoing candidiasis independent of local pathogen control, and might explain why expression of genes involved in metabolic liver function (cluster 9, Suppl. Fig. 1) did not return to steady-state levels by 72 h p.

In summary, all organs showed downregulation of genes associated with organ function, likely as a result of organ impairment and/or as consequence of upregulation of genes dedicated to immune responses. Consistent with the development of fungal burden, these changes were transient in spleen and liver but increased over time in the kidney.

Transcriptional profiles of *C. albicans* indicate organ specific fungal adaptation in kidney and liver. One of the largest challenges for *in vivo* fungal transcription analysis is the relative abundance of host transcripts hampering the recovery of sufficient amounts of fungal RNA[17]. We used an enrichment protocol based on sequential lysis of host and fungal cells on ice. Determination of genome wide transcription levels by microarrays and of selected genes (*TSP1*, *HSP90*, *HSP104*) by quantitative RT-PCR (q RT-PCR) revealed approximately 10% variation between RNA isolated from spiked murine organs using our enrichment protocol and RNA isolated from the spiking culture by standard methods[24]. We were able to isolate sufficient amounts of high-quality fungal RNA to perform microarray analysis from *C. albicans* cells retrieved from the liver 8 h p.i., the kidneys 12 h p.i., and the kidneys 24 h p.i., enabling us to detect 6569 genes in the arrays. The amounts recovered at other time points were sufficient only for q RT-PCR.

The lower amounts of fungal RNA obtained from the kidney at 8 h p.i. or the liver at later time points are possibly due to lower amounts of total fungal biomass and RNA at these time points in the organs. We furthermore detected no cross-hybridization of uninfected kidney samples with the *C. albicans* microarrays, indicating

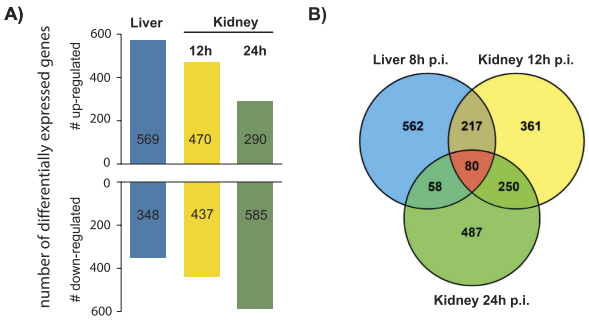


Figure 3. *C. albicans* genes differentially expressed during systemic candidiasis in mice. (**A**) Number of differentially expressed *C. albicans* genes in liver and kidney samples 8 h, 12 h and 24 h p.i. relative to common reference (log phase SC5314 grown in YPD at 37 °C). (**B**) Venn diagram representing the overlap of *C. albicans* DEGs in liver 8 h, kidney 12 h and kidney 24 h p.i. Within all three samples, 80 genes were differentially expressed in all samples, whereas most of the genes were differentially regulated at specific time points or organs.

that the detected transcripts originate from *C. albicans*. The microarray data were verified by qPCR analyses of selected genes (Suppl. Table 1).

Compared to YPD culture (control), 917 *C. albicans* genes were differentially expressed in liver, 908 in kidney 12 h p.i., and 875 in kidney 24 h p.i., with 80 genes differentially expressed in all organs (Fig. 3, Suppl. data 3). These included infection-associated, highly upregulated individual genes, such as *HWP1*, *ECE1, ALS1, ALS3, HYR1,* and *SOD5* (Table 1). Furthermore, GO categories related to cell wall and cell surface were significantly enriched in positively regulated genes of all samples (Fig. 4, Suppl. data 4), indicating remodelling of the cell wall and reflecting the upregulation of hypha-associated genes involved in adhesion such as *ALS1*, *ALS3*, *DIF1,* and *HWP1* (Table 1).

While this demonstrates common regulation of a set of virulence-associated genes, most DEGs displayed organ- or time point-specific regulation (Fig. 3B). Some of the organ-specific differences might be due to the differences in the time points analysed (liver 8 h p.i. vs. kidney 12 h and 24 h p.i.), which is a limitation of this study; however, the higher similarity among gene expression profiles at two time points in the kidney compared to the liver (Suppl. Fig. 2A) still suggests organ-specific adaptations.

The transcriptional response of *C. albicans* during growth in the kidney is characterized by iron acquisition and metabolic adaptation. To elucidate which processes were affected by the DEGs, we performed GO term enrichment analysis and gene set enrichment analysis (GSEA) using published lists of *in vivo* and *in vitro* regulated *C. albicans* genes. Specific expression patterns deduced from GO term enrichment analysis included upregulation of "iron assimilation" in kidney samples at both time points, supported by GSEA results matching gene regulation in the kidney with iron homeostasis[25] and consistent with two recently published *in vivo* transcriptional profiles of *C. albicans* cells infecting the mouse kidney[15,16]. This correlates well with local iron sequestration shown in kidney lesions[26] and the upregulation of lactoferrin and haptoglobin by the host (Suppl. data 1), supporting the concept of nutritional immunity as a host defence mechanism[27].

During early kidney infection, according to GO terms *C. albicans* furthermore upregulated translational processes (e.g. "ribosome biogenesis", Fig. 4), which was supported by GSEA (ribosomal proteins in a set of putative targets of the transcription factor Fhl1[28]). After 24 h in the kidney, the transcriptional profile of *C. albicans* indicated a response to starvation (Fig. 4), possibly as a result of rapid growth and hypha formation. In contrast, metabolic genes differentially expressed in the liver indicated catabolic processes and carbohydrate transport (Fig. 4). Some of the organ-specific metabolic adaptations of *C. albicans* could be a response to organ-specific nutrient supply. For example, while the glycerol biosynthetic gene *RHR2* showed decreased expression in kidney, possibly reflecting the presence of the renal osmoprotectant glycero-phosphocholine[15], *RHR2* expression was increased in the liver (Suppl. data 3). Indeed, it has been shown that the ability of *C. albicans* to utilize external glycero-phosphocholine is important for virulence in murine systemic candidiasis[29] and that *RHR2* is essential for proliferation of *C. albicans* in the liver of intraperitoneally infected mice[30].

Inference modelling identified three inter-species pathogen-host gene-regulatory networks linking fungal metabolism with the host immune response. Metabolic adaptation of *C. albicans* might also be influenced by immune reactions, especially restriction of nutrient availability upon phagocytosis[13,31]. To determine which fungal and host activities may directly influence each other, we inferred an inter-species pathogen-host gene-regulatory network[32] using the ExTILAR algorithm[33]. This method integrates transcriptional data (DEGs of both species clustered by k-means; Suppl. Fig. 2B,C) with existing knowledge

orf	Name	Liver 8 h	Kidney 12 h	Kidney 24 h
orf19.3374	*ECE1*	10.535	4.930	7.899
orf19.2060	*SOD5*	8.370	4.242	4.074
orf19.5741	*ALS1*	6.613	2.065	2.277
orf19.6037	*ASM3*	5.801	2.351	2.414
orf19.1321	*HWP1*	5.765	4.585	3.079
orf19.2942	*DIP5*	4.992	1.723	1.530
orf19.6993	*GAP2*	3.758	−1.404	−1.459
orf19.4975	*HYR1*	3.708	1.906	2.603
orf19.2355	*ALS3*	3.657	2.471	2.412
orf19.85	*GPX2*	3.203	1.483	1.783
orf19.5760	*IHD1*	3.180	1.850	1.870
orf19.6367	*SSB1*	2.925	1.975	1.637
orf19.711		2.802	3.277	1.592
orf19.7469	*ARG1*	2.762	1.222	1.276
orf19.7114	*CSA1*	2.669	8.687	9.024
orf19.5916		2.599	1.439	−1.435
orf19.4456	*GAP4*	2.314	1.896	2.056
orf19.7084	*DFI1*	2.286	1.371	1.479
orf19.6844	*ICL1*	2.268	1.562	1.637
orf19.2608	*ADH5*	2.075	−2.394	−1.873
orf19.6837	*FMA1*	2.047	1.376	−1.635
orf19.3384		2.040	1.273	1.526
orf19.3829	*PHR1*	2.014	1.542	1.604
orf19.5170	*ENA21*	1.931	1.478	2.167

Table 1. List of fungal genes significantly upregulated by *C. albicans* in the liver 8 h p.i. Values indicate absolute log2-FC. Note, that *ADH5* and *GAP2* were significantly upregulated in the liver 8 h p.i., but downregulated in the kidney at both time points.

on transcription factor-dependent regulation of the genes to infer a regulatory network between DEGs of host (Suppl. data 5) and pathogen (Suppl. data 6). The resulting network showed three disconnected sub-networks each composed of mouse and *Candida* clusters.

The first network (Fig. 5A) was composed of a murine cluster containing downregulated genes involved in tissue homeostasis, remodelling and renal function, and thereby reflecting parts of the renal response to infection. This host cluster was both regulated by and regulating a cluster of upregulated fungal genes affecting iron homeostasis and biofilm formation (as observed in the kidney). This *C. albicans* cluster in turn was connected to the second *C. albicans* cluster, comprising downregulated genes involved in amino acid metabolism, similar to the observed changes in the liver. Thus, this sub-network suggests a link between host tissue function on the one side and fungal iron homeostasis and amino acid metabolism on the other, and reflects the organ-specific fungal transcription. Amino acid biosynthesis is affected by iron starvation in *S. cerevisiae* in a complex feedback loop involving transcription factors and biosynthesis gene containing iron-sulfur clusters[34]. Cysteine is important for the formation of iron-sulfur clusters and the *C. albicans* key iron regulator Sfu1 contains a cysteine-rich central domain that is typical for GATA-type transcription factors[35]. Cysteine might thus possibly link Sfu1 and amino acid biosynthesis *via* iron-sulfur clusters; this is however highly speculative. Alternatively, tissue damage caused by invasive fungal growth and the subsequent host response might significantly affect nutrient availability for and metabolism of the fungus.

In the second network (Fig. 5B) the central *C. albicans* cluster contained genes related to metabolic processes that were upregulated in the liver but downregulated in the kidney. This cluster was predicted to negatively affect the expression of genes in both, another fungal and a host gene cluster. In the host cluster, genes were associated with immune system processes and included genes such as IL6, CXCL10 and Interferon- and TNF-receptors, which were upregulated only in the kidney. The second *C. albicans* cluster also showed organ-specific regulation, with increased expression in the kidney and decreased expression in the liver. This cluster was significantly enriched for terms associated with the cell surface and included genes like *CDC12*, *CHS3*, and *MYO2*, that influence filamentation and cell wall composition and thereby could affect interaction with the immune system[36]. Moreover, binding sites for the transcription factor *BCR1*, which regulates biofilm formation and cell-surface associated genes[37], was over-represented in the promoter region of genes in this cluster. Biofilm-associated genes were also found to be enriched in the kidney by GO term analysis (Fig. 4). Taken together, this sub-network indicates an organ-specific regulatory relationship between host immune responses, fungal metabolism and *C. albicans* morphology.

Similarly, the *C. albicans* clusters within the third network (Fig. 5C) also contained genes with organ-specific transcription patterns and functions for metabolic adaptation, while two of the four connected murine clusters (#1 and #3) were enriched in genes with functions for the immune response and organ-specific transcription.

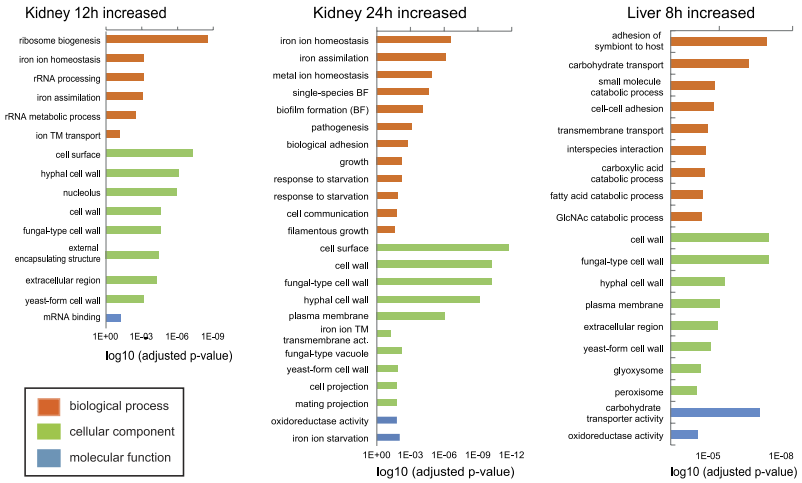


Figure 4. GO terms enriched in positively regulated differentially expressed genes of *C. albicans* during systemic candidiasis. The most significantly regulated ontologies were determined by the adjusted p-value (BP - biological process (orange); CC - cellular component (green), MF - molecular function (blue)). A full list of enriched GO terms is provided in Suppl. data 4.

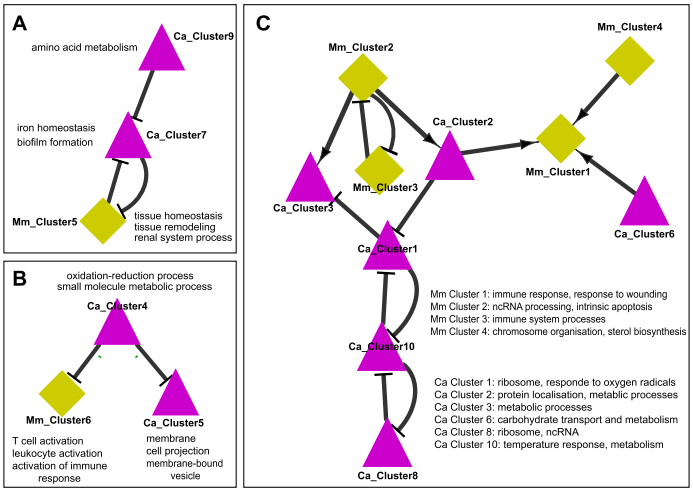


Figure 5. Inferred inter-species pathogen-host gene-regulatory networks. Nodes represent clusters of co-regulated genes for *M. musculus* (diamond) and *C. albicans* (triangle). Edges between two clusters represent directed regulatory interactions that are either positive (arrow) or negative (blunt end). A full list of genes and enriched GO terms of the clusters is provided in Suppl. Data 5 (mouse) and 6 (*C. albicans*).

	Liver 8 h		Liver 12 h	Kidney 12 h		Liver 24 h	Kidney 24 h	
Gene name	qPCR	Microarray	qPCR	qPCR	Microarray	qPCR	qPCR	Microarray
ALS3	80.6	1.9	84.8	89.2	*1.1*	139.3	62.4	*1.2*
DCK1	1.8	*1.6*	1.3	0.9	*1.2*	1.2	0.8	1.5
ECE1	793.5	10.5	629.3	3061.2	4.9	1817.4	3654.9	7.9
HGT2	150.3	5.3	288.5	71.2	*1.4*	194.6	127.4	*1.5*
HWP1	292.2	5.8	128.2	201.4	4.6	81.6	172.0	3.1
IHD1	17.2	3.2	9.2	10.1	1.9	9.2	13.9	1.9
RBT1	1.1	*1.0*	9.6	7	*0.9*	2.0	7.2	*0.9*
orf19.2457	3.7	1.8	5.1	2.7	1.4	4.3	3.3	1.4

Table 2. Gene expression of the core filamentation response determined by qPCR and microarray. Values in italic were not detected as significant differentially expressed by microarray analysis. Overall, gene expression quantified by qPCR (fold change determined by $2^{-\Delta\Delta Ct}$ method) and microarray (absolute log2-FC), respectively, show the same trend in regulation.

Thus, all deduced networks support an organ-specific interplay between fungal metabolism and immune processes during infection. Immune reactions likely affect the availability of nutrients, e.g. upon phagocytosis and by iron sequestration[13,27] but metabolic adaptations of *C. albicans* in turn can affect the interaction of the fungus with the immune system[38].

Expression of hypha-associated genes in the absence of filamentation in the liver. This link between fungal metabolism and interaction with immune cells is likely to be mediated by changes of the fungal surface and cell wall[38]. Genes involved in biofilm formation were enriched at later time points in the kidney, consistent with observations by others[15,16]. GSEA furthermore identified a significant overlap of the *in vivo* transcriptome 24 h p.i. with genes differentially regulated in *C. albicans ccr4*Δ/Δ and *sit4*Δ/Δ[39,40], both involved in cell wall maintenance and filamentation. The cell wall composition of *C. albicans* is also linked to morphogenesis, a central virulence attribute of *C. albicans*, and filamentation occurs early upon systemic infection of the murine kidney[3]. Not surprisingly, filamentation-associated genes were among the most strongly upregulated *Candida* genes in the kidney (Tables 1 and 2). In the liver, however, filamentation does not occur after intravenous infection[2,41] (and own unpublished data). Thus, we were surprised to observe induction of hypha-associated genes (HAGs) and core filamentation response genes[42] in the liver (Tables 1 and 2). To confirm the results obtained by microarrays, we performed qPCR on our RNA samples to analyse the expression levels of the six genes of the core filamentation response[42]. With the exception of *DCK1*, all core filamentation response genes showed a moderate to strong induction in both organs at all time-points under investigation and, with the exception of *RBT1*, similar expression levels in all samples (Table 2; Suppl. Fig. 3).

Expression of hypha-associated genes by *C. albicans* yeast cells in the liver could indicate cell wall remodelling. In addition to the core filamentation response genes, several genes encoding cell wall remodelling factors were specifically induced in *C. albicans* in the liver, e.g. *CRH11* (cell wall transglycosylase), *MP65* (cell surface mannoprotein), *PHR1, BGL2* and *PGA4* (glucanosyltransferases), and the chitinase gene *CHT1*.

It is known that the *C. albicans* cell wall is highly dynamic. *In vitro*, the exposure of *C. albicans* to different stresses that are encountered following phagocytosis can lead to rapid architectural changes and structural realignments of the cell wall[43]. Changes observed *in vivo* include altered exposure of chitin and β-glucan[11,12]. Both morphology and cell wall composition can influence the interaction with immune cells. For example, murine macrophages phagocytose *C. albicans* yeast cells more efficiently than hyphae[44] and preferentially ingest O- and N-linked mannan-deficient mutants[14]. Cell wall glycosylation also affects phagosome maturation in macrophages[45]. We thus hypothesized that the induction of HAGs in yeast cells, as observed in the liver, could influence the interaction of the fungus with macrophages. To test this hypothesis, we inoculated log-phase yeast cells grown in YPD (0 min) into RPMI at 37 °C. In this condition, increased expression of *ALS3*, *ECE1*, *HWP1*, and *IHD1* was detected by qPCR as early as 15 min after induction (data not shown), even though first germ tube formation only became visible after 45 min (Suppl. Fig. 4A). These *C. albicans* yeasts were then used in a phagocytosis assay with human monocyte-derived macrophages. We observed a significant increase in the phagocytosis rate of MDMs challenged with RPMI-induced yeast cells (Suppl. Fig. 4B,C) and alterations in the release of cytokines by macrophages stimulated with thimerosal-killed cells (Suppl. Fig. 4D). While this suggests that the induction of the filamentation program is indeed associated with changes that influence the interaction with immune cells, this hypothesis requires further investigation.

It has, however, been well established that both phagocytosis and cytokine production depend on the interaction of the fungal surface and components of the cell wall with host cell pattern recognition receptors. As the cell wall of *C. albicans* responds to numerous environmental cues[45], it is plausible that the yeast cells found in the liver differ from both, cells grown *in vitro* under yeast-favouring conditions and the yeasts grown under hypha-inducing conditions in our *in vitro* experiment. Morphology-independent expression of HAGs has also been observed during growth in the intestinal tract[46], suggesting that hypha-like modification of the yeast surface can occur in specific host niches, and GSEA revealed a significant overlap of DEGs in the liver with genes

upregulated in the caecum[46]. It is however difficult to predict to which extent the interaction with immune cells in the liver might be influenced by the observed changes of the transcriptome. Given that HAGs were upregulated in the liver 8 h p.i., after prolonged exposure to a likely hypha-inducing environment, it nonetheless appears striking that visible hypha were absent in this organ – especially as they readily form in the kidney and *in vitro* within this time span. Possibly, phagocytosis of *C. albicans* yeasts in the liver exposes the fungus to an environment that prevents hyphal growth.

***C. albicans* likely faces phagocytosis early after infection of the liver.** According to GSEA, the greatest overlap of analysed transcriptomes was observed between *C. albicans* DEGs specific for liver and the transcriptional response to phagocytosis by bone marrow-derived macrophages (Suppl. data 7)[47]. The upregulation of catabolic and transport processes by *C. albicans* in the liver (Fig. 4) furthermore resembled previously observed transcriptional changes upon phagocytosis[13,31]. Stress responsive genes that were upregulated following phagocytosis by PMNs, e.g. *SOD5*, *YHB1*, *GPX2*, *ASR1*, and *CAT1*, were also significantly upregulated by *C. albicans* cells in the liver (Suppl. data 8). In the kidney, in contrast, *SOD5* and *SSB1* were upregulated while other core stress response genes showed decreased expression (e.g. *HSP70*, *HSP90* and *CCP1*; Suppl. data 8). All of the stress-responsive genes with increased expression in the kidney 24 h p.i. are also known to be induced upon oxidative stress[48], possibly indicating interaction with immigrating phagocytes at later stages of infection in the kidney. These organ-specific differential expression kinetics of stress-associated fungal genes might reflect the higher abundance of immune cells in the liver. Residential immune cells likely induce an early profound stress response in *C. albicans*, which we were able to detect in the liver 8 h p.i. It has been shown that interaction between *C. albicans* and resident macrophages also occurs in the kidney within 2 h p.i.[49]; however, this interaction does not inhibit fungal filamentation and could not be detected on day 1 p.i. The low transcription of fungal stress genes in the kidney 12 h p.i. observed by us further supports that the interaction with resident phagocytes in the kidney is only transient and that additional phagocytes need to be recruited upon infection to this organ[2,49]. Thus, low transcription of fungal stress genes in the kidney at the 12 h time point, followed by delayed induction (compared to liver infection) could reflect a fatal later onset of interaction with recruited immune cells in the kidney, whereas early phagocytosis by resident or rapidly recruited immune cells may contribute significantly to the control of *C. albicans* in the liver.

As discussed above, phagocytosis of *Candida* yeasts in the liver possibly exposes the fungus to an environment that prevents hyphal growth. Even though *C. albicans* filaments within macrophages *in vitro*[31], recent findings in a zebrafish model[50] suggest that some macrophages can inhibit filamentation *in vivo*[51] and *C. albicans* DEGs in the zebrafish model showed a significant overlap with the transcriptome in the liver. Furthermore, phagocytosis by PMNs prevents *C. albicans* filamentation[13,52]. It therefore appears plausible that *C. albicans* responds to the hepatic environment by induction of a hypha-associated program while physical formation of hypha is counteracted by the activity of phagocytes. This hypothesis is supported by the observation that combined neutropenia and C5 deficiency in mice led to the development of foci of fungal invasion in liver and spleen. This was not observed in neutropenic C5-sufficient mouse strains, indicating that in the absence of neutrophils other phagocytes control *Candida* in the liver[53]. However, further studies are needed to determine the type of immune cells and exact mechanisms that prevent hypha formation and facilitate fungal clearance in the liver.

Comparability of gene expression profiles across different studies. One very surprising result of the GSEA was that we did not observe a significant overlap with transcriptional data from *C. albicans* in the murine data published previously[15,16,21]. One explanation could be the approach used for our GSEA analysis, in which we combined data from both time points for the kidney to compare this data set to the DEGs in the liver. This initial comparison, designed to detect inter-organ differences, might partially explain why no significant enrichment was observed with other *in vivo* transcription data sets from individual organs referenced to pre-infection samples.

To analyse the comparability of our and published data sets in more detail, we performed a direct comparison (Suppl. Data 9), revealing a 31–44% overlap of DEGs identified in the different studies, with 73–90% of the overlapping DEGs displayed similar trends in expression (up- or downregulated, respectively). While this overlap may appear to be low, it should be interpreted considering that technical differences between the different studies influence the fold-change gene expression and thereby the DEGs identified depending on the cut off. For example, the NanoString technology used by Xu *et al.*[15] is highly sensitive and might thus detect low-expression DEGs missed in our study, but is also limited to the set of genes included in the methodological set up. Similarly, RNASeq after specific enrichment procedures, as performed by Amorim-Vaz *et al.*[16], is likely to be superior in sensitivity compared to microarrays. Amorim-Vaz *et al.* furthermore analysed time points different to our study, which will affect the direct comparison. Further technical differences hampering a meaningful comparison are the different control conditions chosen for *C. albicans in vitro* and the fungal strain used[19].

While all these factors influence the DEGs identified in the different studies, it should be highlighted that distinct pathways (generally identified by GO analysis) were found to be induced in the kidney in all studies. This includes the upregulation of iron acquisition systems, filamentation and cell wall modification as well as specific virulence factors. This demonstrates that these pathways can be robustly identified in different settings and by different technologies and thus underlines their importance for the infection process. It also shows that pathway analyses (via GO-analysis or other bioinformatical approaches) are able to produce more robust and often more informative results than individual analysis of selected genes, for which expression levels might be strongly influenced by the chosen method.

In summary, our study clearly shows that at the transcriptional level both host responses and fungal adaptation during disseminated candidiasis are organ-specific. In the kidney, late onset of innate defence mechanisms, likely due to the comparatively low number of resident immune cells and slow recruitment of additional

effector cells, facilitates fungal proliferation accompanied by filament formation, upregulation of iron acquisition mechanisms, and metabolic adaptation. Failure to control fungal growth likely drives the observed exacerbated induction of proinflammatory responses, thereby contributing to immunopathology. In contrast, innate immune factors are quickly upregulated in the liver and the fungal response indicates possible phagocytosis that might help to explain the noteworthy lack of filamentation in this organ. The distinct environments in the different organs likely drive the observed differential expression of cell wall and cell surface modifying enzymes. This in turn may lead to structural alterations that can affect interaction with immune cells and thus possibly contribute to the specific course of infections in different organs.

Material and Methods

Ethics statement. All animal experiments were performed in accordance with the German animal protection law and were approved by the responsible Federal State authority (Thüringer Landesamt für Lebensmittelsicherheit und Verbraucherschutz) ethics committee (beratende Komission nach §15 Abs. 1 Tierschutzgesetz; permit no. 03–009/13). The animals were cared for in accordance with the European Convention for the Protection of Vertebrate Animals Used for Experimental and Other Scientific Purposes.

Strains and culture condition. *Candida albicans* SC5314 and the GFP expressing strain M137 (Fradin *et al.*, 2005) were maintained as glycerol stocks at −80 °C. For experiments, single colonies obtained from YPD agar (1% w/v peptone, 1% w/v yeast extract, 2% w/v glucose, 2% w/v agar) were grown overnight in liquid YPD (without agar) at 30 °C or in RPMI 1640 at 37 °C in a shaking incubator.

Murine infection model. The infection was performed as described previously[54]. Briefly, SC5314 cells from liquid YPD overnight cultures were washed twice with ice-cold PBS and adjusted to the desired concentration. The infection dose was confirmed by plating. On Day 0, three mice per time point were infected via the lateral tail vein with *C. albicans* ranging from 2.5×10^4 to 6.25×10^6 cfu/g body weight. Female BALB/c mice of 10–12 weeks were used for infection. The remaining infection solution was centrifuged and the pellet was snap frozen in liquid nitrogen and stored at −80 °C for later RNA extraction.

After infection, the health status of the mice was examined twice a day by a veterinarian, and surface temperature and body weight were recorded daily. Mice were humanely sacrificed 8 h, 12 h, 24 h, and 72 h post infection. Immediately after euthanasia, kidneys, spleen, and liver were removed aseptically, rinsed with sterile PBS and snap frozen in liquid nitrogen. PBS mock-infected animals served as controls.

RNA isolation. We designed an RNA isolation protocol for isolation of both fungal and murine RNA from the same sample based on step-wise isolation of host and fungal RNA (Suppl. Fig. 5). First, organs were aseptically homogenized in RLT buffer (Qiagen) with 1% β-Mercaptoethanol (β-ME) on ice water using an Ika T10 basic UltraTurrax homogenizer. Then, homogenates were centrifuged at 3,000 g at 4 °C for 3 min. The supernatant was used for mouse tissue RNA isolation with the RNeasy Mini Kit (Qiagen) as described by the manufacturer. The remaining pellets were vortexed on a mini-beadbeater (Precellys) for 5 sec at 5000 m/s in 1 ml RLT buffer (Qiagen) with 1% β-ME. After centrifugation at 4 °C, fungal RNA was isolated from the remaining pellets as previously described[24]. The RNA quality was determined using a Bioanalyzer (Agilent Inc.), the quantity was measured with a Nanodrop ND1000 (Peqlab).

Gene expression profiling. Genome-wide gene expression of *C. albicans* was analysed with *C. albicans*-specific microarrays (ClinEuroDiag). The red channel represents hybridization with Cy5-labeled *C. albicans* RNA from experimental mouse sepsis while the green channel always shows hybridization with Cy3-labeled RNA from *C. albicans* yeast cells growing logarithmically in standard YPD medium at 37 °C with shaking (common control). *C. albicans* RNA labelling, microarray hybridization, and scanning were performed as previously described[24].

Genome-wide gene expression profiling of mouse samples was performed using the MouseRef-8 Expression and MouseWG6 v2.0 BeadChips (Illumina) according to the manufacturer's instruction. From the RNA isolation, 200 ng total RNA with a Bioanalyzer RIN greater than seven were used for amplification prior to chip hybridization. Samples were analysed using the iScan platform (Illumina) measuring the variation of expression rate of > 42,000 transcripts. All microarray data are MIAME compliant and raw data have been deposited at GEO (GSE83682). The gene expression of liver, spleen, and kidney tissue 8 h, 12 h, 24 h, and 72 h after intravenous infection was compared to control samples of mock infected mice 24 h after PBS injection.

Expression data analyses. All analyses were performed in R using packages provided by Bioconductor 2.26[55]. *C. albicans* two-colour microarray data were pre-processed using the limma package[56]. "Printtiplowess" normalisation was used on each array separately to correct for spatial effects or cross-hybridization. Array spots corresponding to the same gene were summarized using the duplicated correlation function of limma. Normalization of the arrays was performed using between-array quantile normalization. Log2 fold-changes (log2-FC) were calculated between *in vivo* samples and the common reference using limma. Genes with Benjamini-Yekutieli corrected p-value < 0.05 were considered as differentially expressed.

Mouse data was annotated using appropriate Illumina manifest files. For each platform raw data were pre-processed independently using the lumi package[57] for R, including background correction (bgAdjust) and subsequent data normalization using variance stabilization transformation and quantile normalization. Detection calls for each probe were performed at default parameters settings and probes detected as absent in all samples were removed. Subsequently, the illuminaMousev2.db package for R was used for re-annotation and probe filtering. We retained only probes with a quality grade of "good" or "perfect". The normalized and pre-processed data set for the two different microarray platforms was integrated into a combined data set by probe ID matching and subsequent between array loess normalization (affy package 1.44; ref. 58) with the "span" parameter set to

0.1. Quality of the integration procedure was verified by analysis of the expression values of known housekeeping genes and identified stable reference genes before and after loess normalisation. Differential expression was assessed using limma with a Benjamini-Yekutieli corrected p-value below 0.05 and an absolute log2-FC of 0.5 for each organ and time-point. To compare expression values of the control samples between organs we included the within-donor correlation for samples derived from the same animal estimated using the duplicate Correlation function provided by limma.

GO-term analysis of *C. albicans* expression data was performed using FungiFun2[59]. The results contain GO-categories with Fisher's exact test determined, Benjamini-Hochberg corrected p-values below 0.05.

Cluster analysis. Clustering of differentially expressed genes was performed using the clValid package[60] for R, allowing direct comparison of the results from various clustering methods. For our analysis, we compared the results of 5 clustering algorithms (*Hierarchical clustering*, *k-means*, *SOTA*[61], *Diana* and *Clara*[62]) for 3 to 15 clusters. The resulting scores of each validation measure were scaled between 0 and 1 and transformed such that a score of 1 represents the best result. Subsequently, all scores were combined using the average of the mean internal score and the mean stability score. Finally, we selected the clustering method and the number of clusters according to the maximal overall score. Subsequent gene ontology enrichment analyses for the genes of each cluster were performed using the GOstats package[63] excluding all terms with less than 10 or more than 500 genes.

Gene Set Enrichment Analysis (GSEA). For gene set enrichment analysis (GSEA) a gene list was produced from (i) the combined DEGs of *C. albicans* in kidney 12 h and 24 h and (ii) liver 8 h p.i., which were imported as "phenotypes" to be compared into the GSEA software v2.2.0 (Broad Institute)[64]. Analysis parameters were as follows: norm, meandiv; scoring_scheme, weighted; Metric:Diff_of_Classes; set_min, 15; nperm, 1000; set_max, 5000. Gene sets with an FDR < 25% were considered as significant enriched. This set was used for a gene set enrichment analysis (GSEA) based on the recently published list of regulated genes compiled from the literature[16] expanded with another recently performed *in vivo* transcriptional profile of *C. albicans*[15] and the gene list "GSEA_Nantel_2012" kindly supplied to CGD by Andre Nantel (www.candidagenome.org/download/community/GSEA_Nantel_2012/).

Inference of pathogen-host gene-regulatory network. The ExTILAR algorithm[33] was used for the inference of an inter-species pathogen-host gene-regulatory network (GRN) based on prior knowledge about transcription factor binding sites (TFBSs) and about genes known to affect the activity of transcription factors. oPOSSUM 3.0[65] was used to detect over-represented TFBSs for each cluster. For clusters formed by the genes of *M. musculus* we restricted the promoter region of the target genes between 2,000 base pairs (bp) upstream and 0 bp downstream of the transcription start site. TFs corresponding to TFBSs with a Z-score of 10 or more and a Fisher-score of 7 or more (default values) were considered as potential regulators of the respective cluster. For *C. albicans*, the *S. cerevisiae* tool required the mapping of ORF-IDs to gene symbols of the *Saccharomyces* Genome Database (SGD). oPOSSUM was then used at default values restricting the number of background genes to 5,000 randomly selected. TFs corresponding to TFBSs with a Z-score of 10 or more and a Fisher score of 5 or more were considered as potential regulators of the respective cluster.

Based on the temporal mean cluster expression profile and the respective standard deviation (SD) as well as the TF-to-gene information obtained from oPOSSUM we inferred 10,000 networks. For each inference, we sampled expression data from a normal distribution using the mean and SD of the respective cluster profile/time point. ExTILAR was applied at default parameters including stepwise forward selection for the optimization of the structure-template. The final stable consensus network was derived by selection of edges between clusters that were inferred in more than 50 percent of all networks. Using the stable consensus network structure as a template we applied an ordinary least squares approach for parameter estimation of the final network model.

Real time qPCR. For microarray validation and expression analysis of core filamentation response genes, qRT-PCR was performed using the my-Budget 5 × EvaGreen QPCR Mix II (Bio&Sell) in a C1000TM Thermal Cycler (BioRad) according to manufacture recommendations. Gene specific primers are listed in Suppl. Table 2. Relative gene expression levels were determined by the $2^{\Delta\Delta Ct}$ method with *ACT1* and *EFB1* as housekeeping genes. RNA isolated from the *C. albicans* infection solution served as reference. Murine RNA from mock-infected mice served as negative control.

References

1. Perlroth, J., Choi, B. & Spellberg, B. Nosocomial fungal infections: epidemiology, diagnosis, and treatment. *Med. Mycol.* **45,** 321–346 (2007).
2. Lionakis, M. S., Lim, J. K., Lee, C.-C. R. & Murphy, P. M. Organ-specific innate immune responses in a mouse model of invasive candidiasis. *J. Innate Immun.* **3,** 180–199 (2011).
3. MacCallum, D. M., Castillo, L., Brown, A. J. P., Gow, N. A. R. & Odds, F. C. Early-expressed chemokines predict kidney immunopathology in experimental disseminated *Candida albicans* infections. *PloS One* **4,** e6420 (2009).
4. Lionakis, M. S. & Netea, M. G. *Candida* and host determinants of susceptibility to invasive candidiasis. *PLoS Pathog.* **9,** e1003079 (2013).
5. Lewis, R. E. *et al.* Epidemiology and sites of involvement of invasive fungal infections in patients with haematological malignancies: a 20-year autopsy study. *Mycoses* **56,** 638–645 (2013).
6. Brown, G. D. Innate antifungal immunity: the key role of phagocytes. *Annu. Rev. Immunol.* **29,** 1–21 (2011).
7. Ngo, L. Y. *et al.* Inflammatory monocytes mediate early and organ-specific innate defense during systemic candidiasis. *J. Infect. Dis.* **209,** 109–119 (2014).
8. Moyes, D. L. *et al.* Candidalysin is a fungal peptide toxin critical for mucosal infection. *Nature* **532,** 64–68 (2016).
9. Sudbery, P. E. Growth of *Candida albicans* hyphae. *Nat. Rev. Microbiol.* **9,** 737–748 (2011).
10. Ene, I. V. *et al.* Host carbon sources modulate cell wall architecture, drug resistance and virulence in a fungal pathogen. *Cell. Microbiol.* **14,** 1319–1335 (2012).

11. Hopke, A. *et al.* Neutrophil Attack Triggers Extracellular Trap-Dependent *Candida* Cell Wall Remodeling and Altered Immune Recognition. *PLoS Pathog.* **12,** e1005644 (2016).
12. Marakalala, M. J. *et al.* Differential adaptation of *Candida albicans in vivo* modulates immune recognition by dectin-1. *PLoS Pathog.* **9,** e1003315 (2013).
13. Fradin, C. *et al.* Granulocytes govern the transcriptional response, morphology and proliferation of *Candida albicans* in human blood. *Mol. Microbiol.* **56,** 397–415 (2005).
14. Lewis, L. E. *et al.* Stage specific assessment of *Candida albicans* phagocytosis by macrophages identifies cell wall composition and morphogenesis as key determinants. *PLoS Pathog.* **8,** e1002578 (2012).
15. Xu, W. *et al.* Activation and alliance of regulatory pathways in *C. albicans* during mammalian infection. *PLoS Biol.* **13,** e1002076 (2015).
16. Amorim-Vaz, S. *et al.* RNA Enrichment Method for Quantitative Transcriptional Analysis of Pathogens *In Vivo* Applied to the Fungus *Candida albicans*. *mBio* **6,** e00942–915 (2015).
17. Allert, S., Brunke, S. & Hube, B. *In Vivo* Transcriptional Profiling of Human Pathogenic Fungi during Infection: Reflecting the Real Life? *PLoS Pathog.* **12,** e1005471 (2016).
18. Weber, M. *et al.* Hepatic induction of cholesterol biosynthesis reflects a remote adaptive response to pneumococcal pneumonia. *FASEB J. Off. Publ. Fed. Am. Soc. Exp. Biol.* **26,** 2424–2436 (2012).
19. Walker, L. A. *et al.* Genome-wide analysis of *Candida albicans* gene expression patterns during infection of the mammalian kidney. *Fungal Genet. Biol. FG B* **46,** 210–219 (2009).
20. MacCallum, D. M. Massive induction of innate immune response to *Candida albicans* in the kidney in a murine intravenous challenge model. *Fems Yeast Res.* **9,** 1111–1122 (2009).
21. Andes, D., Lepak, A., Pitula, A., Marchillo, K. & Clark, J. A simple approach for estimating gene expression in *Candida albicans* directly from a systemic infection site. *J. Infect. Dis.* **192,** 893–900 (2005).
22. De Filippo, K., Henderson, R. B., Laschinger, M. & Hogg, N. Neutrophil chemokines KC and macrophage-inflammatory protein-2 are newly synthesized by tissue macrophages using distinct TLR signaling pathways. *J. Immunol. Baltim. Md 1950* **180,** 4308–4315 (2008).
23. Majer, O. *et al.* Type I interferons promote fatal immunopathology by regulating inflammatory monocytes and neutrophils during *Candida* infections. *PLoS Pathog.* **8,** e1002811 (2012).
24. Wächtler, B., Wilson, D., Haedicke, K., Dalle, F. & Hube, B. From Attachment to Damage: Defined Genes of *Candida albicans* Mediate Adhesion, Invasion and Damage during Interaction with Oral Epithelial Cells. *PLoS One* **6,** e17046 (2011).
25. Chen, C., Pande, K., French, S. D., Tuch, B. B. & Noble, S. M. An iron homeostasis regulatory circuit with reciprocal roles in *Candida albicans* commensalism and pathogenesis. *Cell Host Microbe* **10,** 118–135 (2011).
26. Potrykus, J. *et al.* Fungal iron availability during deep seated candidiasis is defined by a complex interplay involving systemic and local events. *PLoS Pathog.* **9,** e1003676 (2013).
27. Crawford, A. & Wilson, D. Essential metals at the host-pathogen interface: nutritional immunity and micronutrient assimilation by human fungal pathogens. *FEMS Yeast Res.* **15** (2015).
28. Lavoie, H. *et al.* Evolutionary tinkering with conserved components of a transcriptional regulatory network. *PLoS Biol.* **8,** e1000329 (2010).
29. Bishop, A. C. *et al.* Glycerophosphocholine utilization by *Candida albicans*: role of the Git3 transporter in virulence. *J. Biol. Chem.* **288,** 33939–33952 (2013).
30. Desai, J. V. *et al.* Coordination of *Candida albicans* Invasion and Infection Functions by Phosphoglycerol Phosphatase Rhr2. *Pathog. Basel Switz.* **4,** 573–589 (2015).
31. Lorenz, M. C., Bender, J. A. & Fink, G. R. Transcriptional response of *Candida albicans* upon internalization by macrophages. *Eukaryot. Cell* **3,** 1076–1087 (2004).
32. Schulze, S., Schleicher, J., Guthke, R. & Linde, J. How to Predict Molecular Interactions between Species? *Front. Microbiol.* **7,** 442 (2016).
33. Vlaic, S. *et al.* The extended TILAR approach: a novel tool for dynamic modeling of the transcription factor network regulating the adaption to *in vitro* cultivation of murine hepatocytes. *BMC Syst. Biol.* **6,** 147 (2012).
34. Philpott, C. C., Leidgens, S. & Frey, A. G. Metabolic remodeling in iron-deficient fungi. *Biochim. Biophys. Acta* **1823,** 1509–1520 (2012).
35. Lan, C.-Y. *et al.* Regulatory networks affected by iron availability in *Candida albicans*. *Mol. Microbiol.* **53,** 1451–1469 (2004).
36. Warenda, A. J. & Konopka, J. B. Septin function in *Candida albicans* morphogenesis. *Mol. Biol. Cell* **13,** 2732–2746 (2002).
37. Nobile, C. J. *et al.* A recently evolved transcriptional network controls biofilm development in *Candida albicans*. *Cell* **148,** 126–138 (2012).
38. Brown, A. J. P., Brown, G. D., Netea, M. G. & Gow, N. A. R. Metabolism impacts upon *Candida* immunogenicity and pathogenicity at multiple levels. *Trends Microbiol.* **22,** 614–622 (2014).
39. Dagley, M. J. *et al.* Cell wall integrity is linked to mitochondria and phospholipid homeostasis in *Candida albicans* through the activity of the post-transcriptional regulator Ccr4-Pop2. *Mol. Microbiol.* **79,** 968–989 (2011).
40. Lee, C.-M., Nantel, A., Jiang, L., Whiteway, M. & Shen, S.-H. The serine/threonine protein phosphatase SIT4 modulates yeast-to-hypha morphogenesis and virulence in *Candida albicans*. *Mol. Microbiol.* **51,** 691–709 (2004).
41. Rozell, B., Ljungdahl, P. O. & Martínez, P. Host-pathogen interactions and the pathological consequences of acute systemic *Candida albicans* infections in mice. *Curr. Drug Targets* **7,** 483–494 (2006).
42. Martin, R. *et al.* A core filamentation response network in *Candida albicans* is restricted to eight genes. *PloS One* **8,** e58613 (2013).
43. Ene, I. V. *et al.* Cell Wall Remodeling Enzymes Modulate Fungal Cell Wall Elasticity and Osmotic Stress Resistance. *mBio* **6,** e00986 (2015).
44. Keppler-Ross, S., Douglas, L., Konopka, J. B. & Dean, N. Recognition of yeast by murine macrophages requires mannan but not glucan. *Eukaryot. Cell* **9,** 1776–1787 (2010).
45. Bain, J. M. *et al.* *Candida albicans* hypha formation and mannan masking of β-glucan inhibit macrophage phagosome maturation. *mBio* **5,** e01874 (2014).
46. Rosenbach, A., Dignard, D., Pierce, J. V., Whiteway, M. & Kumamoto, C. A. Adaptations of *Candida albicans* for growth in the mammalian intestinal tract. *Eukaryot. Cell* **9,** 1075–1086 (2010).
47. Marcil, A. *et al.* Analysis of PRA1 and its relationship to *Candida albicans*- macrophage interactions. *Infect. Immun.* **76,** 4345–4358 (2008).
48. Enjalbert, B., MacCallum, D. M., Odds, F. C. & Brown, A. J. P. Niche-specific activation of the oxidative stress response by the pathogenic fungus *Candida albicans*. *Infect. Immun.* **75,** 2143–2151 (2007).
49. Lionakis, M. S. *et al.* CX3CR1-dependent renal macrophage survival promotes *Candida* control and host survival. *J. Clin. Invest.* **123,** 5035–5051 (2013).
50. Chen, Y. Y. *et al.* Dynamic transcript profiling of *Candida albicans* infection in zebrafish: a pathogen-host interaction study. *PloS One* **8,** e72483 (2013).
51. Gilbert, A. S., Wheeler, R. T. & May, R. C. Fungal Pathogens: Survival and Replication within Macrophages. *Cold Spring Harb. Perspect. Med.* **5,** a019661 (2015).

52. Rubin-Bejerano, I., Fraser, I., Grisafi, P. & Fink, G. R. Phagocytosis by neutrophils induces an amino acid deprivation response in *Saccharomyces cerevisiae* and *Candida albicans*. *Proc. Natl. Acad. Sci. USA* **100,** 11007–11012 (2003).
53. Fulurija, A., Ashman, R. B. & Papadimitriou, J. M. Neutrophil depletion increases susceptibility to systemic and vaginal candidiasis in mice, and reveals differences between brain and kidney in mechanisms of host resistance. *Microbiol. Read. Engl.* **142** (Pt 12), 3487–3496 (1996).
54. Jacobsen, I. D., Lüttich, A., Kurzai, O., Hube, B. & Brock, M. *In vivo* imaging of disseminated murine *Candida albicans* infection reveals unexpected host sites of fungal persistence during antifungal therapy. *J. Antimicrob. Chemother.* **69,** 2785–2796 (2014).
55. Huber, W. *et al.* Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Methods* **12,** 115–121 (2015).
56. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43,** e47 (2015).
57. Du, P., Kibbe, W. A. & Lin, S. M. lumi: a pipeline for processing Illumina microarray. *Bioinforma. Oxf. Engl.* **24,** 1547–1548 (2008).
58. Gautier, L., Cope, L., Bolstad, B. M. & Irizarry, R. A. affy–analysis of Affymetrix GeneChip data at the probe level. *Bioinforma. Oxf. Engl.* **20,** 307–315 (2004).
59. Priebe, S., Kreisel, C., Horn, F., Guthke, R. & Linde, J. FungiFun2: a comprehensive online resource for systematic analysis of gene lists from fungal species. *Bioinforma. Oxf. Engl.* **31,** 445–446 (2015).
60. Hartigan, J. A. *Clustering algorithms*. (Wiley, 1975).
61. Herrero, J., Valencia, A. & Dopazo, J. A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinforma. Oxf. Engl.* **17,** 126–136 (2001).
62. Kaufman, L. & Rousseeuw, P. J. *Finding Groups in Data: An Introduction to Cluster Analysis*. (Wiley, 2005).
63. Falcon, S. & Gentleman, R. Using GOstats to test gene lists for GO term association. *Bioinforma. Oxf. Engl.* **23,** 257–258 (2007).
64. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102,** 15545–15550 (2005).
65. Kwon, A. T., Arenillas, D. J., Worsley Hunt, R. & Wasserman, W. W. oPOSSUM-3: advanced analysis of regulatory motif over-representation across genes or ChIP-Seq datasets. *G3 Bethesda Md* **2,** 987–1002 (2012).

Acknowledgements

We thank Christine Dunker for providing material to optimize the RNA isolation method and Volha Skrahina for qRT-PCR primers. For experimental assistance we thank Markus Bläss, Melanie Polke, Birgit Weber, Nadja Jablonowski, and Stephanie Wisgott. This work was supported by the German Federal Ministry of Education and Health (BMBF) Germany, (FKZ 01EO1002, Integrated Research and Treatment Center, Center for Sepsis Control and Care (CSCC), the Jena School for Microbial Communication (JSMC) and by the Hans Knöll Institute Jena. IDJ, MB and JL are supported by the Deutsche Forschungsgemeinschaft CRC/Transregio 124 'Pathogenic fungi and their human host: Networks of interaction' subprojects C5 (IDJ, MB) and INF (JL). SV is supported by the German Federal Ministry of Education and Research (BMBF, 'InfectControl,' subproject FINAR).

Author Contributions

I.D.J. and B.H. designed the experiments. B.H. performed murine work, RNA isolation, and *C. albicans* microarrays. Cytokine and phagocytosis assays were performed by B.H. and M.K. S.V. performed differential murine gene expression analysis. Cluster analysis and inference network analysis were performed by S.V. and T.C. *C. albicans* gene expression analysis was performed by T.C., B.H. and J.L. S.B. and B.H. performed GSEA. B.H., B.H., M.B. and I.D.J. interpreted the data. B.H. and I.D.J. wrote the paper; all authors reviewed the manuscript.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Hebecker, B. *et al.* Dual-species transcriptional profiling during systemic candidiasis reveals organ-specific host-pathogen interactions. *Sci. Rep.* **6**, 36055; doi: 10.1038/srep36055 (2016).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

2.2 Manuskript 2: „ModuleDiscoverer: Identification of regulatory modules in protein-protein interaction networks“

Status

Veröffentlicht im Januar 2018

Literaturangabe

Vlaic, S., Conrad, T., Tokarski-Schnelle, C., Gustafsson, M., Dahmen, U., Guthke, R., Schuster, S. (2018). ModuleDiscoverer: Identification of regulatory modules in protein-protein interaction networks. *Scientific Reports,* 8(1). https://doi.org/10.1038/s41598-017-18370-2

Übersicht

Besonderes Augenmerk dieser Dissertation liegt auf der Entwicklung und Anwendung des moduldetektierenden Algorithmus ModuleDiscoverer. Mit seinem heuristischen Ansatz stellt er ein nützliches Werkzeug für die Identifikation von regulatorischen Modulen in großen, gesamtgenomischen PPIN und Hochdurchsatzgenexpressionsdaten dar. Seine Anwendbarkeit und Effektivität bei der Datenanalyse einzelner Ebenen konnte auf Grundlage von *Rattus norvegicus*-Genexpressionsdaten bezüglich einer diätinduzierten, nicht-alkoholischen Fettleberhepatitis demonstriert werden. Die identifizierten Module sind stabil, biologisch relevant und reflektieren die bereits in anderen Studien getätigten klinischen und histologischen Beobachtungen bezüglich der nicht-alkoholischen Fettleberhepatitis. Die von ModuleDiscoverer erzeugten Ergebnisse sind vergleichbar mit denen von anderen moduldetektierenden Ansätzen wie DEGAS, MATISSE oder KeyPathwayMiner. Allerdings liefert das mit ModuleDiscoverer erzeugte regulatorische Modul signifikant mit der nicht-alkoholischen Fettleberhepatitis assoziierte Einzelnukleotidpolymorphismen. Diese konnten weder über die anderen drei Algorithmen noch durch die Analyse von ausschließlich differenziell regulierten Genen detektiert werden. Die Ergebnisse des Manuskripts lassen vermuten, dass ModuleDiscoverer auch für andere Organismen und weitere Omik-Daten anwendbar ist.

Beiträge

VS konzipierte die Studie und führte zusammen mit CT und TSC die Analysen durch. VS, CT und TSC interpretierten die Ergebnisse und schrieben das Manuskript. GM, DU, GR und SS unterstützten die Interpretation der Ergebnisse und die Erstellung des Manuskripts. Alle Autoren waren am Überprüfungs- und Überarbeitungsprozess beteiligt.

www.nature.com/scientificreports

SCIENTIFIC REPORTS

OPEN

ModuleDiscoverer: Identification of regulatory modules in protein-protein interaction networks

Received: 25 August 2017
Accepted: 6 December 2017
Published online: 11 January 2018

Sebastian Vlaic[1,2], Theresia Conrad[1], Christian Tokarski-Schnelle[1,3], Mika Gustafsson[4], Uta Dahmen[3], Reinhard Guthke[1] & Stefan Schuster[2]

The identification of disease-associated modules based on protein-protein interaction networks (PPINs) and gene expression data has provided new insights into the mechanistic nature of diverse diseases. However, their identification is hampered by the detection of protein communities within large-scale, whole-genome PPINs. A presented successful strategy detects a PPIN's community structure based on the maximal clique enumeration problem (MCE), which is a non-deterministic polynomial time-hard problem. This renders the approach computationally challenging for large PPINs implying the need for new strategies. We present ModuleDiscoverer, a novel approach for the identification of regulatory modules from PPINs and gene expression data. Following the MCE-based approach, ModuleDiscoverer uses a randomization heuristic-based approximation of the community structure. Given a PPIN of *Rattus norvegicus* and public gene expression data, we identify the regulatory module underlying a rodent model of non-alcoholic steatohepatitis (NASH), a severe form of non-alcoholic fatty liver disease (NAFLD). The module is validated using single-nucleotide polymorphism (SNP) data from independent genome-wide association studies and gene enrichment tests. Based on gene enrichment tests, we find that ModuleDiscoverer performs comparably to three existing module-detecting algorithms. However, only our NASH-module is significantly enriched with genes linked to NAFLD-associated SNPs. ModuleDiscoverer is available at http://www.hki-jena.de/index.php/0/2/490 (Others/ModuleDiscoverer).

Structural analysis of intracellular molecular networks has attracted ample interest over several decades[1]. This includes cellular networks such as protein interaction maps[2], metabolic networks[3,4] transcriptional regulation maps[5], signal transduction networks[6,7] as well as functional association networks[8]. Recent advances in the field of network medicine have focused on the identification of disease-associated modules within the organism-specific interactome[9]. The interactome captures interactions between all molecules of a cell[10] and is represented by a graph composed of nodes denoting cellular molecules that are connected by edges representing interactions between them. Within the interactome, modules are sub-graphs that can be linked to phenotypes such as diseases or traits. Up to date, the identification of disease-associated modules has been applied mostly based on protein-protein interaction networks (PPINs) of *Homo sapiens*. They have been successfully identified for, e.g., asthma[11], inflammatory and malignant diseases[12], obesity and type-2-diabetes (among others)[13] as well as different subtypes of breast cancer[14–16], providing new in-depth insights into the underlying molecular mechanisms of the respective disease. For example, biomarker identification for the classification of 402 breast tumor samples into their respective subtype was successfully performed based on subtype-specific protein signaling networks[15]. Furthermore, the same study highlighted that strongly connected genes (i.e., hub genes) present in either subtype-specific network are valid drug targets for the respective subtype.

There are three fundamental assumptions underlying the identification of disease modules[17] (Fig. 1). Firstly, entities forming dense clusters within the interactome (topological modules) are involved in similar biological functions (functional modules). Secondly, molecules associated to the same disease, such as disease-associated proteins, tend to be located in close proximity within the network, which defines the disease module. Thirdly,

[1]Leibniz Institute for Natural Product Research and Infection Biology - Hans-Knöll-Institute, Systems Biology and Bioinformatics, Jena, 07745, Germany. [2]Friedrich-Schiller-University, Department of Bioinformatics, Jena, 07743, Germany. [3]University Hospital Jena, Friedrich-Schiller-University, General, Visceral and Vascular Surgery, Jena, 07749, Germany. [4]Linköping University, Bioinformatics, Department of Physics, Chemistry and Biology, Linköping, 581 83, Sweden. Correspondence and requests for materials should be addressed to S.V. (email: Sebastian.Vlaic@leibniz-hki.de)

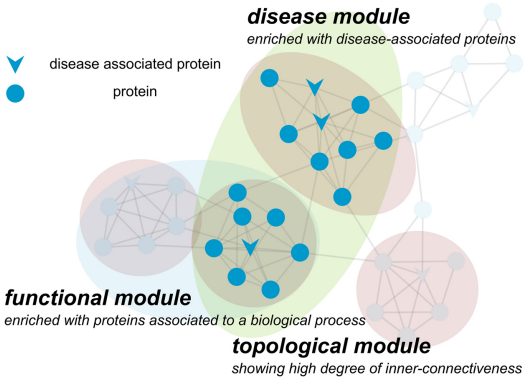


Figure 1. The concept of disease modules exemplified using a sample PPIN. One or more topological modules (highlighted red) contain proteins involved in similar biological processes forming functional modules (highlighted blue). A disease module (highlighted green) is a sub-network of proteins enriched with disease-relevant proteins, e.g., known disease-associated proteins.

disease modules and functional modules overlap. Thus, a disease relates to the breakdown of one or more connected functional modules.

A variety of approaches have been presented specifically for the identification of disease modules. They can be categorized into two different groups. On the one hand, there are algorithms that make use of known disease-associated molecules or genetic loci, the known interactome as well as some association function for the identification of disease modules and/or new disease-associated molecules[18–22]. For example, the disease module detection (DIAMOnD) algorithm[20] utilizes known disease-associated proteins (seed proteins) to identify proteins (DIAMOnD proteins) significantly connected to seed proteins. Iterative application of the algorithm results in a growing disease module with a ranked list of DIAMOnD proteins, i.e., candidate disease-associated proteins. On the other hand, there are algorithms that identify disease modules as well as disease-associated molecules *'ab initio'* based on the projection of omics data onto the interactome in conjunction with a community structure detecting algorithm[12,13,23]. Like topological modules, communities are groups of proteins with higher within-edge density compared to the edge density connecting them[24]. For example, the approach presented by Barrenäs *et al.*[13] identifies protein communities by decomposition of the human PPIN into sub-graphs of maximal cliques. A clique is a sub-graph of the PPIN, where each pair of proteins is connected by an edge. A maximal clique is a clique that is not part of a larger clique. The regulatory module is then formed by the union of all maximal cliques that are significantly enriched with disease-associated-proteins, e.g., differentially expressed genes.

The idea of disease modules can obviously be generalized towards the detection of regulatory modules underlying an arbitrary phenotype of any organism. This can be of high interest, e.g., for the molecular characterization of animal models of diverse human diseases. This includes animal models of infectious diseases such as fungal infections with *Candida albicans* and *Aspergillus fumigatus*[25], animal models of inflammation[26], asthma[27] as well as metabolic diseases such as fatty liver disease (FLD)[28]. Since animal models reflect only certain aspects of the human disease phenotype[29], identification of the underlying regulatory module can provide additional information regarding the functional context in which such models are valid. A variety of algorithms for the identification of such phenotype (or condition)-specific modules in PPINs have been published[30]. Like the MCE-based approach by Barrenäs *et al.*[13], so called 'module cover approaches' (see Batra *et al.*[31]) such as MATISSE[32], DEGAS[33] and KeyPathwayMiner[34] consider the detection of differential gene expression as a separate pre-processing step and can handle proteins in the PPIN with missing expression information. In contrast to the MCE-based approach, these algorithms avoid assumptions about the community structure. In turn, they introduce additional parameters controlling, e.g., the allowed noise in the network structure (DEGAS and KeyPathwayMiner) or the module size (MATISSE), or introduce additional assumptions such as the expected fraction of similarly expressed genes in the regulatory module (MATISSE). The optimization problem underlying these approaches is non-deterministic polynomial time (NP)-hard (see Batra *et al.*[31], Ulitsky *et al.*[32] and Eblen *et al.*[35], respectively). Thus, application of any of these algorithms to large-scale PPINs becomes computationally challenging. While heuristics were presented for DEGAS, KeyPathwayMiner and MATISSE, an efficient heuristic following the idea of the MCE-based approach is missing.

We present ModuleDiscoverer, a new approach to the *ab initio* identification of regulatory modules. ModuleDiscoverer is a heuristic that, based on the idea of the MCE-based approach, approximates the PPIN's underlying community structure by iterative enumeration of cliques starting from random seed proteins in the

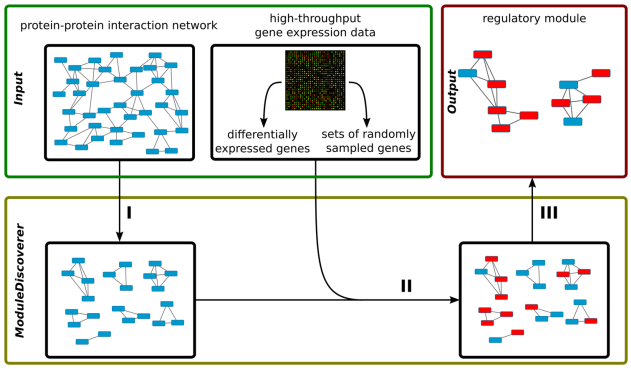


Figure 2. Given a PPIN and gene expression data (Input), the algorithm works in three steps. Step I) The community structure underlying the PPIN is approximated by the identification of protein cliques. Step II) Identification of cliques significantly enriched with DEGs. Step III) Assembly of the regulatory module based on the union of significantly enriched cliques.

network. We identify the regulatory module underlying a diet-induced rat model of non-alcoholic steatohepatits (NASH), the severe form of the non-alcoholic fatty liver disease (NAFLD). The identified NASH-regulatory module is then validated using NAFLD-associated single nucleotide polymorphism (SNP) data from independent genome-wide association studies (GWASs) as well as gene enrichment tests based on known gene-to-disease relations. We compare our results to those derived from DEGAS, MATISSE and KeyPathwayMiner. Finally, we show that our NASH-module reflects histological and clinical parameters as reported by Baumgardner *et al.*[36], who first introduced the animal model.

Results

ModuleDiscoverer: detection of regulatory modules. The detection of regulatory modules is divided into three steps I–III (Fig. 2). Starting with a PPIN (Fig. 2, Input) the algorithm first approximates the underlying community structure by iterative enumeration of protein cliques from random seed proteins in the network (Fig. 2, I). Next, DEGs obtained from high-throughput gene expression data in conjunction with sets of randomly sampled genes (Fig. 2, Input) are used to calculate a p-value for each clique (Fig. 2,II). Finally, significantly enriched cliques are assembled (Fig. 2,III) resulting in the identified regulatory module (Fig. 2, Output).

Step I: Approximation of the PPIN's community structure. Approximation of the community structure underlying the PPIN (Fig. 2, I) is composed of three phases: transformation, identification and extension. In brief, the PPIN is transformed into a graph with labeled nodes and edges (Fig. 3A,B). Starting from one or more random seed nodes, the algorithm then identifies minimal cliques of size three (Fig. 3C,E). Finally, all minimal cliques are stepwise extended competing for nodes in the network until no clique can be extended further (Fig. 3F).

The number of seed nodes used defines two strategies for the enumeration of cliques, the single-seed and the multi-seed approach. Notably, there are advantages as well as disadvantages for both strategies (Supplementary File S1). The single-seed approach identifies cliques using only one seed node in the PPIN. This is suitable for the identification of regulatory modules that are comparable to the results of current, MCE-based algorithms. However, in dense regions of highly overlapping cliques, the single-seed approach favors the enumeration of large maximal cliques. Consequently, proteins that are part of only small cliques can be missed. In contrast, the use of two or more seed nodes (the multi-seed approach), which compete for nodes during the enumeration of cliques, leads to a breakdown of large maximal cliques. While this increases the probability for proteins contained in small cliques only to become part of the final regulatory module, it also leads to an inflation of the regulatory module with proteins not associated to DEGs. Concluding, the multi-seed based regulatory modules can be seen as a comprehensive extension to the single-seed based regulatory modules. In the following example we will illustrate our approach showing one iteration of ModuleDiscoverer using three seed proteins ($p4$, $p6$ and $p9$).

Phase 1 of Step I: Transformation of the PPIN into a labeled graph: Figure 3(A) shows a PPIN as provided by databases such as STRING[37]. It consists of 10 nodes representing the proteins $p1$ to $p10$ and 26 connecting edges. These edges refer to prior-knowledge interactions between connected proteins. First, the network is transformed into an undirected labeled graph $G(V, E)$ (Fig. 3B). The graph G consists of 10 vertices $V(G) = \{v_1, \ldots, v_{10}\}$ and 26 edges $E(G) = \{e_1, \ldots, e_{26}\}$. Each vertex is labeled with one protein ($p1$–$p10$). Notably, a vertex can be labeled with more than one protein. In such case, the proteins in the label form a clique in the PPIN (e.g., vertex $p1$, $p2$, $p4$ in Fig. 3D). Two vertices v_x and v_y (with $x, y \in 1, \ldots, 10$ and $x \neq y$) are connected by an edge if there is at least one

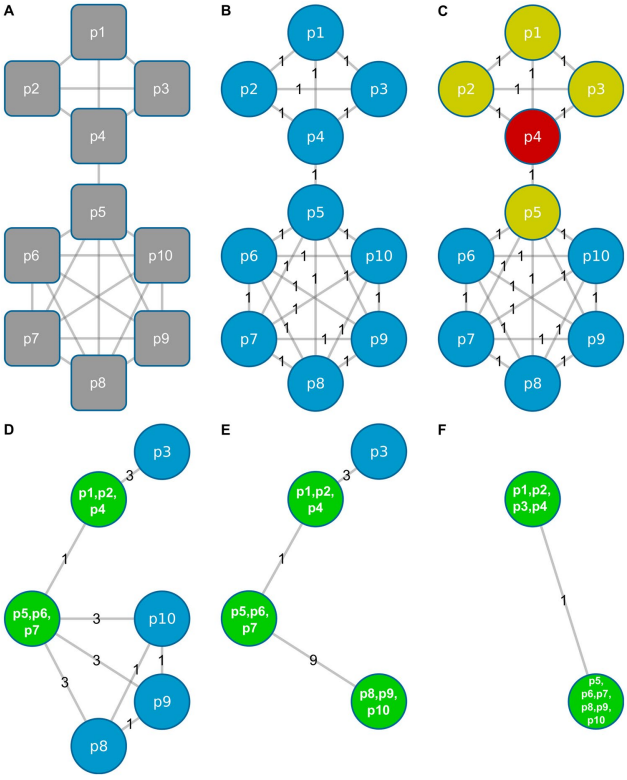


Figure 3. Clique enumeration using ModuleDiscoverer. (**A**) Sample PPIN with 10 proteins and 26 known relations. (**B**) Representation of the PPIN as an undirected labeled graph with each vertex representing one of the proteins in (**A**). The edge weight denotes for the number of existing relations between its connecting nodes. (**C**–**F**) Red vertices denote for seed nodes. Yellow vertices are first neighbors of seed nodes. Green vertices represent cliques. Their label represents clique forming proteins.

known relation in the PPIN between the proteins represented by v_x as well as the proteins represented by v_y. The weight of the edge connecting v_x and v_y denotes for the number of relations between the proteins represented by v_x and the proteins represented by v_y. Initially, all edges have weight 1.

Phase 2 of Step I: Identification of minimal cliques of size three: Starting with randomly selected seed proteins, the algorithm first identifies minimal cliques of size three. A seed is dropped if it is not part of a minimal clique. In Fig. 3(C), we start with *p*4 (colored red) as a seed and search for any minimal clique of size three by exploring its neighbors (colored yellow) as well as their neighbors. The order in which vertices are explored is random. In our example, the first clique identified is formed by *p*1, *p*2 and *p*4 and the corresponding vertices are merged into the vertex *p*1, *p*2, *p*4 (Fig. 3D). Next, the weights of the edges are updated. In our example (Fig. 3D), the edge between *p*1, *p*2, *p*4 and *p*3 is now weighted 3, since the proteins *p*1, *p*2 and *p*4 are all connected to protein *p*3 (Fig. 2A). The edge's weight connecting *p*1, *p*2, *p*4 with *p*5 remains 1, since only *p*4 is connected to *p*5. Following the same strategy, the minimal clique *p*5, *p*6, *p*7 is identified starting from the seed *p*6 (Fig. 3D) while the seed *p*9 is merged with *p*8 and *p*10 into *p*8, *p*9, *p*10 (Fig. 3E). All edge weights are updated accordingly.

Phase 3 of Step I: Extension of all minimal cliques: All minimal cliques of size three (Fig. 3E; green) are now iteratively extended in random order until they cannot be enlarged further. Once a node becomes part of a clique,

it cannot become part of another clique, i.e., cliques compete for nodes in the graph. Starting from Fig. 3(E), $p1$, $p2$, $p4$ is processed first. $p1$, $p2$, $p4$ is connected to $p3$ by an edge of weight 3. Thus, all proteins $p1$, $p2$ and $p4$ are connected to $p3$ (Fig. 3A). Therefore, both vertices can be merged to form the new vertex $p1$, $p2$, $p3$, $p4$ (Fig. 3F). Next, the clique represented by $p5$, $p6$, $p7$ is processed. The edge connecting $p5$, $p6$, $p7$ with $p8$, $p9$, $p10$ has a weight of 9. This indicates that all proteins of $p5$, $p6$, $p7$ are connected with all proteins of $p8$, $p9$, $p10$. Therefore, both vertices are merged to form $p5$, $p6$, $p7$, $p8$, $p9$, $p10$ (Fig. 3F). Finally, no clique can be enlarged any further. The algorithm terminates reporting two cliques, i.e., the clique formed by the proteins $p1$, …, $p4$ as well as the clique formed by the proteins $p5$, …, $p10$.

Phases 1–3 of step I of the algorithm are repeated for n iterations with random seed proteins in each iteration until the set of obtained cliques sufficiently approximates the community structure underlying the PPIN.

Step II: Identification of significantly enriched cliques. In step II (Fig. 2II) all enumerated cliques are tested for their enrichment with phenotype-associated proteins, e.g., proteins corresponding to DEGs from high-throughput gene expression data (Fig. 2, Input). The p-value for each clique is calculated using a permutation-based test[38]. In detail, for a gene expression platform measuring N genes, with $D \in N$ being the set of DEGs, the gene sets B are created, each containing $|D|$ genes sampled from N. For each clique in C, the p-value $p_{i,D}$ of clique c_i ($i = 1, \ldots, |C|$) is calculated using the one-sided Fisher's exact test. Accordingly, the p-value $p_{i,b}$ of clique c_i is calculated for each gene set b in B. The final p-value p_i^* is then calculated according to equation 1.

$$p_i^* = \frac{|\forall B\colon p_{i,b} \leq p_{i,D}|}{|B|} \tag{1}$$

Step III: Assembly of the regulatory module. Based on an user-defined p-value cutoff we filter significantly enriched cliques. Since cliques can overlap in their proteins, the union of all significantly enriched cliques (Fig. 2III) results in a large regulatory module (Fig. 2, Output). This module summarizes biological processes and molecular mechanisms underlying the respective phenotype.

Reproducibility of regulatory modules. ModuleDiscoverer is a heuristic that approximates the underlying community structure. Since the exact solution is unknown, quality of the approximation cannot be assessed directly. Instead, we can test if additional iterations of the algorithm, i.e., the enumeration of more cliques, has a qualitative impact on the regulatory module in terms of additional nodes and edges. To this end, non-parametric bootstrapping sampling (with replacement) is applied to assess reproducibility of the regulatory module. Based on the results of n iterations of ModuleDiscoverer, we create bootstrap samples of n iterations and identify the respective regulatory modules. Pairwise comparison of the regulatory modules in terms of shared edges and nodes then provides a distance between the two regulatory modules. The median of all distances divided by the average number of nodes and edges reflects the stability of the regulatory module. See Supplementary File S1 section 1.4 for details.

ModuleDiscoverer: application to biological data. To demonstrate the application of ModuleDiscoverer we used the PPIN of *R. norvegicus* in conjunction with gene expression data of a rat model of diet-induced NASH for the identification of a NASH-regulatory module. The results will be presented in three sections: (i) processing of the PPIN (Fig. 2, I), (ii) identification of significantly enriched cliques based on high-throughput expression data (Fig. 2, II) and iii), assembly of the regulatory module based on the union of all significantly enriched cliques (Fig. 2, III). Finally, the NASH-regulatory module will be analyzed and validated.

Processing of the PPIN. The PPIN of *R. norvegicus* (STRING, version 10) was filtered for high-confidence relations with a score >0.7. This retained 15,436 proteins connected by 474,395 relations. Next, we used the single-seed approach of ModuleDiscoverer to enumerate maximal cliques using 2,000,000 iterations. This identified 1,494,126 maximal cliques in total, enclosing 185,178 unique maximal cliques. Additionally, we applied ModuleDiscoverer with 1,020,000 iterations using the multi-seed approach with 25 seed proteins per iteration. This resulted in 18,807,344 cliques in total enclosing 2,269,022 unique cliques.

Identification of significantly enriched cliques. Based on the expression data, we identified 286 DEGs (p-value < 0.05) out of 4,590 EntrezGeneID-annotated genes on the microarray platform (Supplementary File F2). 10,000 data sets were created sampling 286 random genes out of 4,590 genes in the statistical background. Finally, genes of all data sets were translated into EnsemblProteinIDs using the R-package *org.Rn.eg.db*.

P-value calculation according to equation 1 was performed for each clique satisfying the following two properties. First, at least one protein in the clique is associated to a DEG. Second, at least half of the proteins in the clique are associated to genes in the statistical background. For the p-value cutoff 0.01 we identified 696 significantly enriched cliques for the single-seed approach and 5,386 significantly enriched cliques for the multi-seed approach. Notably, permutation-based calculated p-values were similar to p-values calculated using the one-sided Fisher's exact test (Supplementary Figure F1).

Assembly and analysis of the regulatory module. The single-seed regulatory modules contains five disconnected sub-networks composed of 311 proteins connected by 3,180 relations. 175 of the 311 proteins are associated to background genes and 60 are associated to DEGs. Similar, the regulatory module of the multi-seed approach contains five sub-networks composed of 415 proteins and 4,975 relations in total (Fig. 4). 210 of these 415 proteins are associated with background genes and 67 proteins are associated with DEGs. Both of the regulatory modules are

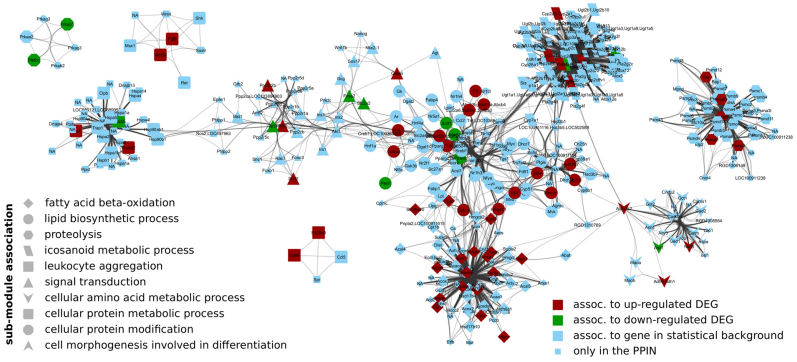


Figure 4. The identified NASH-regulatory module. Nodes (proteins) are labeled with the official gene symbol. Their membership in a sub-module is shape-coded.

significantly enriched ($p < 10^{-4}$) with proteins associated to DEGs. Based on 100 bootstrap samples we found that both regulatory modules are reproducible with an average variability of less than 5% (Supplementary Figure F2). Furthermore, we investigated the robustness of the modules to changes in the edge score cutoff of the PPIN, i.e., the robustness of the algorithm to noise in the PPIN. We found that both regulatory modules are composed of a reproducible set of core proteins (Supplementary File S1), which contribute to a strong similarity among these regulatory modules compared with the similarity to regulatory modules identified with other algorithms. Apart from a single edge, the multi-seed regulatory module encloses the single-seed regulatory module. Thus, we decided to focus on the multi-seed regulatory module as an extension to the single-seed regulatory module.

Next, we identified pathways significantly enriched with proteins for the regulatory module shown in Fig. 4. The results (Supplementary File S3) highlighted NASH-relevant pathways such as fatty acid degradation and elongation, PPAR signaling pathway[39], arachidonic acid metabolism[40], the metabolism of diverse amino acids[41] as well as insulin signaling pathway[42,43]. Identification of sub-modules based on the edge-betweenness centrality measure[44] in the network revealed 10 sub-modules. These sub-modules are sparsely connected with each other but densely connected within themselves. In Fig. 4, the sub-module membership of each protein (and thus its associated biological process) is shape-coded. We performed an enrichment analysis for the proteins of each sub-module to identify its potential biological functions (Supplementary File S4).

We found that the most central sub-module (Fig. 4, circles) is associated with the lipid biosynthetic process. For example, the KEGG PPAR-signaling pathway is significantly enriched with proteins from the module. This pathway plays a key-role in the development of FLD by regulating the beta-oxidation of fatty acids, the activation of anti-inflammatory pathways and the interaction with insulin signaling[45]. In agreement with these findings, the sub-module is directly connected to sub-modules associated to fatty acid beta-oxidation (diamonds), icosanoid-metabolic processes (parallelogram) and cellular signal transduction such as the insulin signaling pathway (triangles). Another directly connected sub-module is associated to the metabolism of cellular amino acids (V-shaped) such as alanine, aspartate and glutamate metabolism as well as phenylalanine, tyrosine and tryptophan metabolism.

Another two sub-modules are associated to proteolysis (hexagons) and the metabolism of cellular proteins (round rectangle) with the latter being directly connected to the sub-module associated with signal transduction (triangles). The connection between cellular protein metabolic processes such as the response to unfolded proteins (Supplementary File S4, sub-module 8) and NAFLD as well as NASH has been studied extensively and is reviewed in[46].

Detection of regulatory modules using module cover approaches. We compared the identified NASH-regulatory module with the regulatory modules identified by three 'module cover algorithms' (see Batra *et al.*[31]), namely MATISSE, DEGAS and KeyPathwayMiner (see methods for details).

The identified modules were compared based on EnsemblProteinIDs and results are summarized in Table 1. We found that DEGAS produced the smallest module composed of 42 proteins, followed by KeyPathwayMiner with 100 proteins. The modules produced by MATISSE (314) and ModuleDiscoverer (single-seed: 311; multi-seed 415) are similar in size. With app. 24%, the modules of MATISSE and KeyPathwayMiner show the highest overlap with the set of proteins associated to all DEGs, followed by ModuleDiscoverer (app. 9%) and DEGAS (app. 2%). The regulatory module by MATISSE overlaps with the modules of ModuleDiscoverer and KeyPathwayMiner to about 22%–26%. The module of KeyPathwayMiner overlaps with the modules of ModuleDiscoverer by app. 13%–16%.

	DRPs	MD-SS	MD-MS	DEGAS	MATISSE	KPM
DRPs	410	9.08%	8.84%	2.26%	23.55%	23.79%
MD-SS		311	74.49%	3.22%	22.79%	15.77%
MD-MS			415	2.47%	21.50%	13.44%
DEGAS				42	3.19%	8.40%
MATISSE					314	26.22%
KPM						100

Table 1. Node-wise overlap between identified regulatory modules of DEGAS, MATISSE, KeyPathwayMiner (KPM), ModuleDiscoverer single-seed (MD-SS) and multi-seed (MD-MS) as well as the set of DEG-associated proteins, i.e., differentially regulated proteins (DRPs). The overlap (given in %) is defined as fraction of the intersection of the module's nodes from the union of the module's nodes. The diagonal of the matrix contains the total number of proteins in the module.

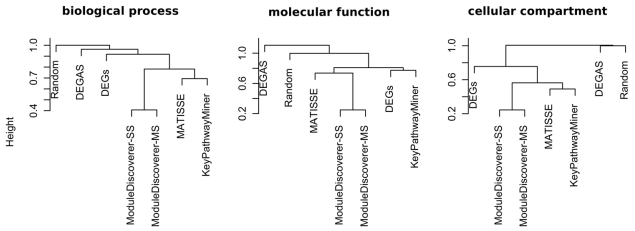


Figure 5. Similarity of modules given by the correlation-based distance measure of ranked lists of significantly enriched GO-terms. The height corresponds to the correlation-based distance (see methods), where values <1 denote for a positive average correlation.

Thus, modules produced by ModuleDiscoverer are more related to the modules produced by MATISSE compared to KeyPathwayMiner.

Next, we were interested in the module's mutual agreement regarding the underlying biology. Hierarchical clustering was used to visualize the correlation-based distance measure (see methods) between regulatory modules obtained from lists of significantly enriched GeneOntology (GO)-terms. Figure 5 outlines the results for the ontologies biological process (BP), molecular function (MF) and cellular compartment (CC). Compared to random lists of GO-terms (Fig. 5, Random), KeyPathwayMiner, MATISSE and ModuleDiscoverer show a positive average correlation for all three ontologies. For BP and CC (Fig. 5, left and right) the regulatory modules of KeyPathwayMiner and MATISSE show a higher agreement in the derived GO-term lists compared to ModuleDiscoverer. With respect to MF (Fig. 5, middle), the GO-term list of the KeyPathwayMiner module shows a high correlation with the GO-term list derived from the set of DEGs. The GO-term list of the MATISSE module are correlated with the GO-term lists of both ModuleDiscoverer modules. Overall, GO-term lists derived from the modules of MATISSE, KeyPathwayMiner as well as ModuleDiscoverer show a positive average correlation with the GO-term lists derived from the set of DEGs.

Literature validation of the regulatory module. We corroborated both NASH-modules (single-seed and multi-seed) using curated disease-to-SNP associations (see methods). Disease-to-SNP associations are based on DNA-sequence information. Thus, they can be considered independent from the gene expression data used to identify the module. In contrast to the set of DEGs as well as the set of proteins captured by the modules identified using DEGAS, MATISSE or KeyPathwayMiner, we found that both NASH-modules are significantly enriched (p-value <0.05) with genes associated to NAFLD-relevant SNPs (Supplementary File S5).

Next, we performed a gene enrichment analysis using a list of curated disease-to-gene associations (see methods). The results are outlined in Fig. 6. Both of our NASH-modules show significantly enriched FLD-associated diseases such as obesity, (non-insulin dependent) diabetes mellitus type-2, liver carcinoma and insulin resistance. Notably, for the set of DEGs almost all of these disease-terms (with the exception of 'Fatty liver') show a slight, but non-significant enrichment (p-value ≥ 0.05). Compared to ModuleDiscoverer, the modules produced by KeyPathwayMiner and MATISSE show increasing similarity to the results of ModuleDiscoverer.

Discussion

We have presented ModuleDiscoverer, an algorithm for the identification of regulatory modules based on large-scale, whole-genome PPINs and high-throughput gene expression data. To show applicability of the

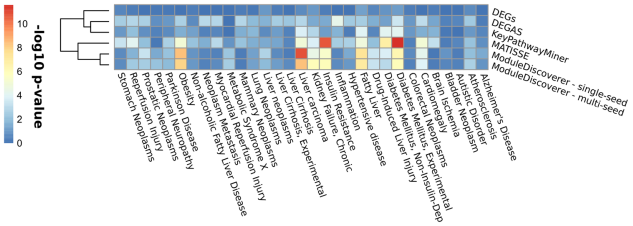


Figure 6. Enrichment of FLD-related diseases with proteins of modules produced by ModuleDiscoverer (single-seed and multi-seed), DEGAS, KeyPathwayMiner and MATISSE as well as the set of DEGs. Higher values equal lower p-values.

algorithm, we identified a non-alcoholic steatohepatitis (NASH)-regulatory module for which we relied on the STRING resource only. STRING integrates information from a variety of resources, such as primary interaction databases, algorithms for interaction prediction, pathway databases, text-mining and knowledge transfer based on orthology. Reported relations are thus based on known physical interaction as well as associative information. To ensure quality of the relations, we selected a high cutoff (>0.7) for the combined edge score. Additionally, we found that a small increase/decrease of the selected cutoff has no substantial effect on our results. To further assure robustness of the identified regulatory modules, a comparison of the modules based on different PPINs should be considered. If working with human data, for example, our algorithm could be applied to the human signaling network provided by the Wang Lab[47]. If there is no comparative PPIN or even no PPIN at all for the organism of interest, a yet to explore alternative might be the use of whole-genome gene regulatory networks (GRNs). Algorithms such as presented in Altwasser *et al.*[48] are based on mathematical models that combine expression data and prior-knowledge interaction data. In such GRNs, relations denote for functional relationships between genes/proteins acting in common biological contexts, which equals networks derived from STRING[37]. This corresponds to the idea of regulatory modules as shown in Fig. 1.

We compared the ModuleDiscoverer-identified NASH-modules to the modules detected by DEGAS, KeyPathwayMiner and MATISSE. Based on the comparison of rank-transformed lists of significantly enriched GO-terms, the DEGAS-, KeyPathwayMiner-, MATISSE- and ModuleDiscoverer-produced modules as well as the set of DEGs correlate in their underlying biology. Interestingly, the module by MATISSE (followed by KeyPathwayMiner) overlaps most with the ModuleDiscoverer-identified module. This can be explained by the methodology underlying the algorithms. KeyPathwayMiner identifies connected sub-networks of proteins associated to DEGs. Exception nodes, i.e., nodes not associated to DEGs, are included as 'bridges' to identify the overall maximal connected sub-network. Thus, modules by KeyPathwayMiner are always centered around proteins associated to DEGs. In contrast, MATISSE calculates weights for the PPIN's edges based on a probabilistic model estimating the similarity between proteins given the underlying expression data. Proteins without expression information do not contribute to the score during the module finding process. Thus, MATISSE-identified modules contain also peripheral exception nodes. This relates to the 'guild-by-association' principle of ModuleDiscoverer, which includes an exception gene in the module if a significant amount of measured genes in its direct neighborhood, i.e., the set of genes that form the maximal clique, is associated to a DEG. In contrast to MATISSE however, the clique assumption by ModuleDiscoverer naturally limits the number of exception nodes to those that are part of the clique. In consequence, we cannot state the best performing algorithm since the results strongly depend on the underlying assumptions. However, based on the validation, we found that only the ModuleDiscoverer-identified NASH-modules contain a significant number of proteins associated to NAFLD-relevant SNPs.

We find that the identified NASH-module (Fig. 4) reflects the experimental clinical and histological observations by Baumgardner *et al.* For example, the NASH-module highlights the disease-term 'Obesity' as significantly enriched with proteins of the module (Fig. 5). In agreement, Baumgardner *et al.*[36] observed a significant increase in body weight in the treatment group compared to control ($p \leq 0.05$). Moreover, they reported a significant increase in fat mass as percentage of body weight between treatment and control reflecting adiposity. Additionally, serum leptin levels were observed to be significantly increased in the treatment group. The serum leptin level is a marker that positively correlates with obesity[49]. Other significantly enriched disease terms include 'Insulin Resistance', 'Diabetes Mellitus Type-2' and 'Diabetes Mellitus, Experimental'. Baumgardner *et al.*[36] reported significantly increased serum insulin concentrations compared to control rats that were overfed with a high-fat 5% corn oil diet at (220 $kcal*kg^{-3/4}*day^{-1}$~17%) for 21 days. They concluded that this observation points towards hyperinsulinemia, which can be due to insulin resistance and is often associated with type-2 diabetes. Finally, we found the disease-term 'Fatty Liver' significantly enriched in proteins of the module. Baumgardner *et al.*[36] reported that histological examination of the liver samples showed steatosis, macrophage infiltration and focal necrosis in the treatment samples. This was accompanied by significantly elevated serum alanine aminotransferase (ALT) levels and significantly increased serum and liver triglyceride concentrations. Notably though, other inflammation-associated scores such as hepatocellular ballooning and lobular inflammation/necrosis

were reported to be elevated but not statistically significant. This could explain the non-significantly enriched disease-terms such as 'Inflammation' and 'Liver Cirrhosis'.

To further evaluate our algorithm, we used a small sub-network of the high-confidence PPIN of *R. norvegicus* (Supplementary File S1). We showed that the single-seed approach as well as the multi-seed approach work well in principle and highlighted their advantages as well as disadvantages. In summary, in cases where large-scale, genome-wide PPINs cannot be processed by MCE-solving algorithms, i.e., the regulatory module based on the exact solution cannot be determined, the use of ModuleDiscoverer becomes inevitable. In such situations, the regulatory module of the single-seed and the multi-seed approach should be identified. While single-seed-based regulatory module is more consistent with results of MCE-based approaches, the multi-seed regulatory module will extend the single-seed based regulatory module with proteins that may have been missed due to a PPIN structure of highly overlapping maximal cliques.

Conclusion

We presented ModuleDiscoverer, a heuristic approach for the identification of regulatory modules in large-scale, whole-genome PPINs. The application of ModuleDiscoverer becomes favorable with increasing size and density of PPINs. Compared to the MCE-based approach, we demonstrated that ModuleDiscoverer identifies modules that can be identical (single-seed approach) or even more comprehensive (multi-seed approach). We applied our algorithm to experimental data for the identification of the regulatory module underlying a rat model of diet-induced NASH. The identified NASH-regulatory module is stable, biologically relevant, reflects experimental observations on the clinical and histological level and is comparable to the results of three published module detection algorithms. In contrast to any of the modules identified by these algorithms or the set of DEGs alone, our NASH-module is significantly enriched with NAFLD-associated SNPs derived from independent GWASs. Altogether, we consider ModuleDiscoverer a valuable tool in the identification of regulatory modules based on large-scale, whole-genome PPINs and high-throughput gene expression data.

Methods

Microarray data, pre-processing and differential gene expression analysis. Affymetrix microarray gene expression data of a rodent model of diet-induced NASH published by Baumgardner *et al.*[36] was downloaded from Gene Omnibus Express[50] (GSE8253). In brief, the animal model was obtained by overfeeding rodents with a high-fat diet based on 70% corn oil at moderate caloric excess ($220\,kcal{*}kg^{-3/4}{*}day^{-1}$~17%) for 21 days via total enteral nutrition (TEN)[36]. They compared the treatment group against a control group of rats fed a diet based on 5% corn oil at normal caloric levels ($187\,kcal{*}kg^{-3/4}{*}day^{-1}$) for 21 days via TEN. Gene expression in each experimental group was measured using three microarrays.

Affymetrix Rat Genome U34 arrays were annotated with custom chip definition files from Brainarray version 15[51]. Raw data was pre-processed using RMA[52]. Differential gene expression was assessed using *limma*[53] with a p-value <0.05 (Supplementary File S2).

SNP-gene-disease and gene-disease association data. Disease-to-SNP relations as well as curated disease-to-gene associations for *H. sapiens* were obtained from DisGeNET[54]. All text-mining based disease-to-SNP associations were removed. Furthermore, we removed all associations involving genes without an orthologue in *R. norvegicus*. Orthology information was obtained from the RGD[55]. For the disease-to-gene associations we created a disease network similar to Goh *et al.*[56]. In this network, two diseases (nodes) are connected if they share ≥ 10 genes. Selecting the first neighbors of the terms 'Fatty Liver' and 'Non-alcoholic Fatty Liver Disease' yielded a list of 31 NAFLD-relevant diseases.

Algorithms for phenotype-specific module identification. We tested three different phenotype-specific module identification algorithms named MATISSE[32], DEGAS[33] and KeyPathwayMiner[34]. MATISSE and DEGAS are implemented in the MATISSE toolbox[57]. For KeyPathwayMiner we downloaded the stand-alone application (version 4.0)[58]. For all algorithms, the high-confidence interactome of *R. norvegicus* from STRING was converted to *sif*-format. EntrezGeneID-based gene identifiers of the microarray were converted to EnsemblProteinIDs using the *org.Rn.eg.db* database.

Matisse. Matisse aims at the identification of connected components (connected sub-networks) composed of nodes associated with genes of high similarity, e.g., genes with similar expression profiles. MATISSE starts from small, high-scoring groups of proteins (as defined by a probabilistic model estimating the similarity between genes). These seed groups are step-wise modified (extended, reduced, exchanged or merged) until the overall score is maximized. We applied MATISSE to expression data of all six samples (three control, three case) for all DEGs. Starting from seed protein groups with minimal/maximal size of 5/50, MATISSE was run to identify regulatory modules with minimal/maximal size of 5/100. Pearson correlation was used to assess similarity between gene expression patterns (default parameter settings). A total of four regulatory modules was identified, which we combined into a single regulatory module for further analysis.

Degas. Degas aims at the identification of minimal (k, l)-components (connected sub-networks) where at least k genes are differentially expressed in all but l cases. The algorithm was run using expression data of all six samples (three control, three case) for the full set of genes available on the microarray. The CUPS heuristic was used to identify all regulatory modules with at least $k = 40$ genes differentially expressed (p-value <0.05) in all but $l = 1$ case. k was optimized automatically within a range of 10 and 50 using k-steps of 10 (default parameter settings). The algorithm identified one regulatory module, which was used for further analysis.

KeyPathwayMiner. KeyPathwayMiner identifies maximal (k, l)-components (connected sub-networks) with at most k genes that are not differentially expressed in all but l cases. The algorithm was applied using the full set of genes available in the data. Instead of using expression data for all six samples we provided an indicator flag (0/1) to mark differentially expressed genes (1). The algorithm identified regulatory modules containing a maximum of $k = 2$ genes, which are not differentially expressed ($l = 0$) using the INES strategy. The best-scoring module was selected for further analysis.

Comparing modules based on lists of GO-term. The distance between regulatory modules from different algorithms was estimated based on the correlation of ranked lists of significantly enriched GO-terms. For each identified regulatory module we performed a gene enrichment analysis using GOstats with the *org.Rn.eg.db* package. P-values $>= 0.05$ were set to 1 and p-value-ordered GO-term lists were rank-transformed. Indices corresponding to ties were ordered at random. The ranking was repeated 1,000 times. Spearman's rank correlation coefficient was calculated for each repeat. The final correlation between the GO-term lists of two methods was averaged over all 1,000 repeats. We defined the distance as 1 minus the correlation coefficient.

References

1. Albert, R. Scale-free networks in cell biology. *J Cell Sci.* **118**, 4947–4957 (2005).
2. Uetz, P. *et al.* A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**, 623–627 (2000).
3. Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N. & Barabási, A. L. Hierarchical organization of modularity in metabolic networks. *Science* **297**, 1551–1555 (2002).
4. Schuster, S., Pfeiffer, T., Moldenhauer, F., Koch, I. & Dandekar, T. Exploring the pathway structure of metabolism: decomposition into subnetworks and application to *Mycoplasma pneumoniae*. *Bioinformatics (Oxford, England)* **18**, 351–361 (2002).
5. Lee, T. I. *et al.* Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science.* **298**, 799–804 (2002).
6. Fu, C., Li, J. & Wang, E. Signaling network analysis of ubiquitin-mediated proteins suggests correlations between the 26S proteasome and tumor progression. *Mol Biosyst.* **5**, 1809–1816 (2009).
7. Ma'ayan, A. *et al.* Formation of regulatory patterns during signal propagation in a mammalian cellular network. *Science.* **309**, 1078–1083 (2005).
8. Tong, A. H. Y. *et al.* Global mapping of the yeast genetic interaction network. *Science* **303**, 808–813 (2004).
9. Ivanov, P. C., Liu, K. K. L. & Bartsch, R. P. Focus on the emerging new fields of network physiology and network medicine. *New J. Phys.* **18**, 100201 (2016).
10. Sanchez, C. *et al.* Grasping at molecular interactions and genetic networks in *Drosophila melanogaster* using flynets, an internet database. *Nucleic Acids Res.* **27**, 89–94 (1999).
11. Sharma, A. *et al.* A disease module in the interactome explains disease heterogeneity, drug response and captures novel pathways and genes in asthma. *Hum Mol Genet.* **24**, 3005–3020 (2015).
12. Gustafsson, M. *et al.* Integrated genomic and prospective clinical studies show the importance of modular pleiotropy for disease susceptibility, diagnosis and treatment. *Genome Med.* **6**, 17 (2014).
13. Barrenäs, F. *et al.* Highly interconnected genes in disease-specific networks are enriched for disease-associated polymorphisms. *Genome Biol.* **13**, R46 (2012).
14. Li, J. *et al.* Identification of high-quality cancer prognostic markers and metastasis network modules. *Nat Commun.* **1**, 34 (2010).
15. Zaman, N. *et al.* Signaling network assessment of mutations and copy number variations predict breast cancer subtype-specific drug targets. *Cell Rep.* **5**, 216–23 (2013).
16. McGee, S. R., Tibiche, C., Trifiro, M. & Wang, E. Network analysis reveals a signaling regulatory loop in the PIK3CA-mutated breast. *Cancer Predicting Survival Outcome. Genomics Proteomics Bioinformatics* **15**, 121–129 (2017).
17. Barabási, A. L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. *Nat Rev Genet.* **12**, 56–68 (2011).
18. Ahn, Y. Y., Bagrow, J. P. & Lehmann, S. Link communities reveal multiscale complexity in networks. *Nature* **466**, 761–764 (2010).
19. George, R. A. *et al.* Analysis of protein sequence and interaction data for candidate disease gene prediction. *Nucleic Acids Res.* **34**, e130 (2006).
20. Ghiassian, S. D., Menche, J. & Barabási, A. L. A disease module detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. *PLoS Comput Biol.* **11**, e1004120 (2015).
21. Köhler, S., Bauer, S., Horn, D. & Robinson, P. N. Walking the interactome for prioritization of candidate disease genes. *Am J Hum Genet.* **82**, 949–958 (2008).
22. Oti, M., Snel, B., Huynen, M. A. & Brunner, H. G. Predicting disease genes using protein–protein interactions. *J Med Genet.* **43**, 691–698 (2006).
23. Zhang, X., Gao, L., Liu, Z. P. & Chen, L. Identifying module biomarker in type 2 diabetes mellitus by discriminative area of functional activity. *BMC Bioinformatics* **16**, 92 (2015).
24. Fortunato, S. Community detection in graphs. *Physics Reports* **486**, 75–174 (2010).
25. Hardwood, C. G. & Rao, R. P. Host pathogen relations: exploring animal models for fungal pathogens. *Pathogens* **3**, 549–562 (2014).
26. Webb, D. R. Animal models of human disease: inflammation. *Biochem Pharmacol.* **87**, 121–130 (2014).
27. Mullane, K. & Williams, M. Animal models of asthma: reprise or reboot? *Biochem Pharmacol.* **87**, 131–139 (2014).
28. Imajo, K. *et al.* Rodent models of nonalcoholic fatty liver disease/nonalcoholic steatohepatitis. *Int J Mol Sci.* **14**, 21833–21857 (2013).
29. McGonigle, P. & Ruggerie, B. Animal models of human disease: challenges in enabling translation. *Biochem Pharmacol.* **87**, 162–171 (2014).
30. Mitra, K., Carvunis, A. R., Ramesh, S. K. & Ideker, T. Integrative approaches for finding modular structure in biological networks. *Nat Rev Genet.* **14**, 719–732 (2013).
31. Batra, R. *et al.* On the performance of de novo pathway enrichment. *npj Systems Biology and Application.* **3**, 1 (2017).
32. Ulitsky, I. & Shamir, R. Identification of functional modules using network topology and high-throughput data. *BMC Systems Biology* **1**, 8 (2007).
33. Ulitsky, I., Krishnamurthy, K., Karp, R. M. & Shamir, R. DEGAS: DeNovoDiscovery of dysregulated pathways in human diseases. *PLoS ONE* **5**, e13367 (2010).
34. Alcaraz, N., Kücük, H., Weile, J., Wipat, A. & Baumbach, J. KeyPathwayMiner: detecting case-specific biological pathways using expression data. *Internet Mathematics.* **7**, 299–313 (2011).
35. Eblen, J., Phillips, C. A., Rogers, G. L. & Langston, M. A. The maximum clique enumeration problem: algorithms, applications, and implementations. *BMC Bioinformatics* **13**, S5 (2012).
36. Baumgardner, J. N., Shankar, K., Hennings, L., Badger, T. M. & Ronis, M. J. A new model for nonalcoholic steatohepatitis in the rat utilizing total enteral nutrition to overfeed a high-polyunsaturated fat diet. *Am J Physiol Gastrointest Liver Physiol.* **294**, G27–G38 (2008).

37. Szklarczyk, D. *et al*. Stringv10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **43**, D447–D452 (2015).
38. Ge, Y., Dudoit, S. & Speed, T. P. Resampling-based multiple testing for microarray data analysis. *TEST* **12**, 1–44 (2003).
39. Souza-Mello, V. Peroxisome proliferator-activated receptors as targets to treat non-alcoholic fatty liver disease. *World J Hepatol.* **7**, 1012–1019 (2015).
40. Loomba, R., Quehenberger, O., Armando, A. & Dennis, E. A. Polyunsaturated fatty acid metabolites as novel lipidomic biomarkers for noninvasive diagnosis of nonalcoholic steatohepatitis. *J Lipid Res.* **56**, 185–192 (2015).
41. Cheng, S. *et al*. Metabolite profiling identifies pathways associated with metabolic risk in humans. *Circulation* **125**, 2222–2231 (2012).
42. Chitturi, S. *et al*. Nash and insulin resistance: Insulin hypersecretion and specific association with the insulin resistance syndrome. *Hepatology* **35**, 373–379 (2002).
43. Nassir, F. & Ibdah, J. A. Role of mitochondria in nonalcoholic fatty liver disease. *Int J Mol Sci* **15**, 8713–8742 (2014).
44. Newman, M. E. J. & Girvan, M. Finding and evaluating community structures in networks. *Physical Review E* **69**, 026113 (2004).
45. Pawlak, M., Lefebvre, P. & Staels, B. Molecular mechanism of pparα action and its impact on lipid metabolism, inflammation and fibrosis in non-alcoholic fatty liver disease. *J Hepatol.* **62**, 720–733 (2015).
46. Henkel, A. & Green, R. M. The unfolded protein response in fatty liver disease. *Semin Liver Dis.* **33**, 321–329 (2013).
47. Wang, E. Cancer Systems Biology and Bioinformatics. http://www.cancer-systemsbiology.org/data-software, (accessed 11.2017)
48. Altwasser, R., Linde, J., Buyko, E., Hahn, U. & Guthke, R. Genome-wide scale-free network inference for *Candida albicans*. *Front Microbiol.* **3**, 51 (2012).
49. Al Maskari, M. Y. & Aln, A. A. Correlation between serum leptin levels, body mass index and obesity in omanis. *Sultan Qaboos Univ Med J.* **6**, 27–31 (2006).
50. Barrett, T. *et al*. NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.* **41**, D991–D995 (2013).
51. Dai, M. *et al*. Evolving gene/transcript definitions significantly alter the interpretation of genechip data. *Nucleic Acids Res.* **33**, e175 (2005).
52. Irizarry, R. A. *et al*. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics.* **4**, 249–264 (2003).
53. Ritchie, M. E. *et al*. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
54. Piñero, J. *et al*. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database (Oxford)* **2015**, bav028 (2015).
55. Shimoyama, M. *et al*. The rat genome database 2015: genomic, phenotypic and environmental variations and disease. *Nucleic Acids Res.* **43**, D743–D750 (2015).
56. Goh, K. I. *et al*. The human disease network. *Proc Natl Acad Sci USA* **104**, 8685–8690 (2007).
57. Samir, R. MATISSE - identifying modules using gene expression and interaction networks. http://acgt.cs.tau.ac.il/matisse/, (accessed 05. 2017).
58. Baumbach, J., Alcaraz, N., Pauling, J. & List, M. KeyPathwayMiner. https://keypathwayminer.compbio.sdu.dk/keypathwayminer/, (accessed 05. 2017).

Acknowledgements

We thank Dr. Jens Schumacher (Institute of Stochastics) and Stefan Lang (Institute for Bioinformatics) from the Friedrich-Schiller-University Jena as well as Dr. Michael Weber from the Hans-Knöll-Institute Jena for helpful discussions. This work was financially supported by the Interdisciplinary Center for Clinical Research - IZKF Jena (J50) as well as the DFG within the Transregio 124 (FungiNet, projects B1 and INF) and the Jena School for Microbial Communication (JSMC).

Author Contributions

S.V. designed the study and performed the analysis together with T.C. as well as C.T.S. S.V., T.C. and C.T.S. interpreted the results and wrote the manuscript. M.G., U.D., R.G. and S.S. aided in interpretation of the results and contributed to writing the manuscript. All authors reviewed the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-017-18370-2.

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

2.3 Manuskript 3: „Module-detection approaches for the integration of multilevel omics data highlight the comprehensive response of *Aspergillus fumigatus* to caspofungin“

Status
Veröffentlicht im Oktober 2018

Literaturangabe
Conrad, T., Kniemeyer, O., Henkel, S., Krüger, T., Mattern, D. J., Valiante, V., Guthke, R., Jacobsen, I. D., Brakhage, A. A., Vlaic, S., Linde, J. (2018). Module-detection approaches for the integration of multilevel omics data highlight the comprehensive response of *Aspergillus fumigatus* to caspofungin. *BMC Systems Biology,* 12(1). https://doi.org/10.1186/s12918-018-0620-8

Übersicht
In der Studie von Conrad *et al.* konnte die Vermutung bestätigt werden, dass ModuleDiscoverer auch auf andere Organismen und Omik-Daten anwendbar ist. Es wurde der Frage nachgegangen, inwieweit die Integration von Multi-Omik-Daten über moduldetektierende Ansätze, Vorteile bringt gegenüber der Analyse von einzelnen Omik-Datensätzen und der Anwendung klassischer Ansätze, wie dem Komponentenvergleich. Zur Beantwortung dieser Frage wurde die Stressantwort von *A. fumigatus* auf das antifungale Medikament Caspofungin untersucht und zu verschiedenen Zeitpunkten Daten bezüglich der Transkriptom-, Proteom- und Sekretomantwort des Pilzes erhoben. Die von ModuleDiscoverer generierten regulatorischen Module weisen eine wesentlich höhere Übereinstimmung zwischen den verschiedenen Omik-Ebenen und Zeitpunkten auf als die Ergebnisse des klassischen Ansatzes. Die zusätzlichen strukturellen Informationen der regulatorischen Module ermöglichten außerdem ein Clustering der Modulkomponenten auf der Grundlage der Netzwerktopologie. So konnten biologische Prozesse identifiziert werden, die signifikant mit der Stressantwort des pathogenen Pilzes assoziiert sind, wie beispielsweise die Regulation der Kinaseaktivität, Transportmechanismen oder der Aminosäuremetablismus. Darüber

hinaus wurden potenzielle Schlüsselfaktoren der Immunantwort bzw. der WPPI identifiziert, die zu medikamentinduzierten Nebenwirkungen beitragen können.

Beiträge

CT, HSG, KT, LJ und VS führten die Datenanalysen durch. KO, MDJ und VV waren für die Durchführung der Experimente zuständig. CT, KO, HSG, GR, JID, BAA, VS und LJ interpretierten die Ergebnisse. CT, KO, KT, VS und LJ schrieben das Manuskript. Alle Autoren waren an der Überprüfung und Überarbeitung des finalen Manuskripts beteiligt.

Conrad *et al. BMC Systems Biology* (2018) 12:88
https://doi.org/10.1186/s12918-018-0620-8

BMC Systems Biology

RESEARCH ARTICLE Open Access

Module-detection approaches for the integration of multilevel omics data highlight the comprehensive response of *Aspergillus fumigatus* to caspofungin

T. Conrad[1*], O. Kniemeyer[2], S. G. Henkel[3], T. Krüger[2], D. J. Mattern[2,9], V. Valiante[4], R. Guthke[1], I. D. Jacobsen[5,6], A. A. Brakhage[2,6], S. Vlaic[1] and J. Linde[7,8]

Abstract

Background: Omics data provide deep insights into overall biological processes of organisms. However, integration of data from different molecular levels such as transcriptomics and proteomics, still remains challenging. Analyzing lists of differentially abundant molecules from diverse molecular levels often results in a small overlap mainly due to different regulatory mechanisms, temporal scales, and/or inherent properties of measurement methods. Module-detecting algorithms identifying sets of closely related proteins from protein-protein interaction networks (PPINs) are promising approaches for a better data integration.

Results: Here, we made use of transcriptome, proteome and secretome data from the human pathogenic fungus *Aspergillus fumigatus* challenged with the antifungal drug caspofungin. Caspofungin targets the fungal cell wall which leads to a compensatory stress response. We analyzed the omics data using two different approaches: First, we applied a simple, classical approach by comparing lists of differentially expressed genes (DEGs), differentially synthesized proteins (DSyPs) and differentially secreted proteins (DSePs); second, we used a recently published module-detecting approach, ModuleDiscoverer, to identify regulatory modules from PPINs in conjunction with the experimental data. Our results demonstrate that regulatory modules show a notably higher overlap between the different molecular levels and time points than the classical approach. The additional structural information provided by regulatory modules allows for topological analyses. As a result, we detected a significant association of omics data with distinct biological processes such as regulation of kinase activity, transport mechanisms or amino acid metabolism. We also found a previously unreported increased production of the secondary metabolite fumagillin by *A. fumigatus* upon exposure to caspofungin. Furthermore, a topology-based analysis of potential key factors contributing to drug-caused side effects identified the highly conserved protein polyubiquitin as a central regulator. Interestingly, polyubiquitin UbiD neither belonged to the groups of DEGs, DSyPs nor DSePs but most likely strongly influenced their levels.

Conclusion: Module-detecting approaches support the effective integration of multilevel omics data and provide a deep insight into complex biological relationships connecting these levels. They facilitate the identification of potential key players in the organism's stress response which cannot be detected by commonly used approaches comparing lists of differentially abundant molecules.

Keywords: Multilevel, Omics, Protein-protein interaction network, Module, *Aspergillus fumigatus*, Caspofungin, Stress response, ModuleDiscoverer

* Correspondence: Theresia.Conrad@leibniz-hki.de
[1]Systems Biology/Bioinformatics, Leibniz Institute for Natural Product Research and Infection Biology – Hans Knöll Institute, Jena, Germany
Full list of author information is available at the end of the article

Background

The permanent growth in the development and improvement of new measurement techniques have led to a wealth of data from heterogeneous sources. The integration of all available data obtained from diverse studies has the potential to provide a more comprehensive and deeper understanding of the studied subject [1–3]. One example is the investigation of an organism's response to an external stimulus at different molecular levels. Large-scale studies at molecular levels like transcriptomics, proteomics, lipidomics or metabolomics can be summarized by the term 'omics levels'. These omics levels are linked to each other and are considered in their entirety. They describe the overall biological processes which occur in the analyzed organism. Potential links can be characterized by level-shared ('overlapping') components (such as genes or proteins) or the participation of components of different molecular levels in level-shared pathways.

As widely reported, the integration and analysis of data from multiple levels measured with diverse techniques at different time points are challenging. In an intuitive and commonly used approach ('simple approach'), the analysis of several sets of omics data is based on the comparison of lists of differentially expressed genes (DEGs) and differentially synthesized proteins (DSyPs) identified in experimental datasets. However, the use of only DEGs and DSyPs is threshold-dependent and usually incomplete due to experimental limitations. For example, the use of liquid chromatography-mass spectrometry (LC-MS/MS)-based shotgun proteomics analysis for the identification of DSyPs is usually limited in the quantification of low abundant proteins due to the large dynamic range of protein abundances that needs to be covered [4, 5]. Other approaches, including diverse pathway enrichment analyses, assign both differentially and non-differentially expressed genes or their synthesized proteins to specific pathways which are part of biological processes. The level of activity of such pathways can be estimated by taking into account measurements of changes in gene expression or protein synthesis. However, as these approaches are based on pre-defined lists of pathways, they exclude unknown pathways which may also have important functions [6]. Over the last decades, the analysis of protein-protein interaction networks (PPINs) has become a useful approach [7]. By identifying direct (physical) contacts and indirect interactions (e.g., via regulatory cascades) between two or more proteins, PPINs point to structural and functional relationships between their nodes [8]. Several de novo network enrichment approaches were developed to extract connected sub-networks from larger interaction networks. Such sub-networks containing sets of closely related proteins are defined as modules [9]. There are many examples in the literature demonstrating the usefulness of modules in research data interpretation. For instance, Stuart et al. analyzed genetic modules to detect co-expressed genes that are involved in similar biological processes [10], while Trevino et al. [11] have shown the usefulness of investigating inter-module connectivity to identify molecular cross-talk between normal prostate epithelial and prostate carcinoma cells.

Another very interesting application of modules is the identification of prognostic or drug response biomarkers [12]. In this context, modules also show their potential for the characterization of drug-caused side effects occurring in addition to effects on the intended primary drug target. Wang et al. [13] demonstrated that major contributing factors of such side effects can be investigated by considering the primary drug target and its local network structure.

Several categories of modules have been described until now (Fig. 1). Examples are topological modules composed of proteins showing a high degree of inner-connectiveness or functional modules that contain proteins associated to specific biological functions [14, 15]. So-called regulatory modules are defined as sets of co-expressed genes which share a common function [16]. Popular methods for the detection of regulatory modules are: DEGAS [17], MATISSE [15], KeyPathwayMiner [18] and ModuleDiscoverer [19]. Among them, the recently published ModuleDiscoverer (MD) includes a heuristic that approximates the PPIN's underlying community structure based on maximal cliques. While a community defines a group of proteins featuring a higher within-edge density in comparison to the edge density connecting them, a clique represents a set of proteins with edges between each pair of them. A clique is maximal if no node (e.g., protein) exists which extends that clique. MD was shown to be very efficient in the detection of regulatory modules for gene expression data in the context of animal models of non-alcoholic fatty liver disease [19].

In this study, we applied the simple approach (SA), the recently published module-detection approach MD as well as KeyPathwayMiner to experimental data of different molecular levels, measurement techniques and time points. As a case study, we analyzed the molecular response of the human pathogenic fungus *Aspergillus fumigatus* to the antifungal drug caspofungin. *A. fumigatus* causes local and systemic infections in immunocompromised individuals [20]. One therapeutic approach is the use of the lipopeptide caspofungin of the group of echinocandins. Caspofungin specifically targets the fungal cell wall by inhibiting the synthesis of the polysaccharide β-(1,3)-D-glucan [21]. Fungal cells respond to caspofungin by the adaption of gene expression and, consequently, protein biosynthesis and secretion of molecules [22]. Therefore, we analyzed the transcriptomic, proteomic and

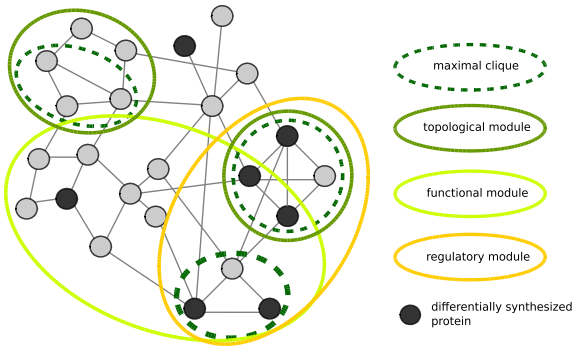


Fig. 1 Module categories. Exemplarily selected categories of modules within protein-protein interaction networks. Proteins are represented by circles, interactions by edges

secretomic response of *A. fumigatus* to caspofungin at several time points to gain a deeper understanding of the overall molecular response of this fungus to this drug.

We demonstrated the so far untested capacity of MD to integrate multilevel omics data and showed that this level of integration is not achievable using SA. Moreover, module-detecting approaches facilitate the identification of potential key players in the organism's stress response which are not detectable by commonly used approaches comparing lists of differentially abundant molecules.

Methods

Omics data and data processing

Data analyses were performed in R version 3.4.1 using packages provided by Bioconductor [23].

Strain and culture conditions

Mycelia of the *Aspergillus fumigatus* strain CEA17 Δ*akuB* [24] were pre-cultured for 16 h in *Aspergillus* minimal medium (AMM, [25]) containing 50 mM glucose and 70 mM $NaNO_3$ and then stressed with a sub-inhibitory concentration of caspofungin (100 ng/ml) as described in Altwasser et al. [26]. Liquid cultures were inoculated with 1×10^6 conidia/ml and cultivated at 37 °C with shaking at 200 rpm. Samples for analyzing the transcriptomic, proteomic and secretomic response of the fungus were taken at the indicated time points after treatment. Secreted proteins were precipitated overnight from culture supernatants as described below.

Transcriptome data

RNA extraction, cDNA library construction and RNA-Seq analysis by Illumina next-generation sequencing of samples taken at 0 h, 0.5 h, 1 h, 4 h and 8 h after caspofungin treatment were performed as described in [26]. Likewise, data were pre-processed as described in [26]. Genes were annotated by identifiers provided by the *Aspergillus* Genome Database (AspGD, as of September 2015 [27]). In addition, identifiers provided by the Central *Aspergillus* Data Repository (CADRE) [28] were obtained using the package *biomaRt* [29] provided by Bioconductor as of February 2017. For each time point, expression values were compared to the control sample taken at 0 h. Only those genes with an absolute log2 Fold Change (log2FC) value greater 1 and a False Discovery Rate (FDR) corrected *p*-value below 0.05 were considered to be differentially expressed.

Proteome and secretome data

Samples for proteome analysis were taken at 0 h, 4 h and 8 h after treatment. The mycelium was collected by filtering through Miracloth (Merck Millipore), subsequently washed with water and snap frozen with liquid nitrogen. Sample preparation of the mycelium for the proteome analysis was performed as previously described [30]. Samples for secretome analysis were taken at 0 h and 8 h after treatment and prepared as follows: Cell free-filtered supernatant of AMM medium from *A. fumigatus* cultures was precipitated by trichloroacetic acid (TCA) at 15% (*w*/*v*) final concentration (4 °C, overnight). Precipitates were washed with acetone and resolubilized in trifluoroethanol (TFE) mixed 1:1 with 100 mM triethylammonium bicarbonate (TEAB). Samples containing 100 µg of total protein (in 100 µl) were reduced with 50 mM tris(2-carboxyethyl)phosphine (TCEP) for 1 h at 55 °C and subsequently cysteine thiols were alkylated with 12.5 mM iodoacetamide for 30 min at room temperature. Proteins were digested at 37 °C for 18 h with trypsin+LysC mix (Promega) at 1:25 protease:-protein ratio. Proteome samples were labeled with tandem mass tags (TMT) 6plex and secretome samples

were labeled with isobaric tags for relative and absolute quantification (iTRAQ) 4plex according to the manufacturer's protocols.

LC-MS/MS analysis was performed as previously described [30] with the following modifications: Eluents A (0.1% *v*/v formic acid in H_2O) and B (0.1% v/v formic acid in 90/10 ACN/H_2O *v*/v) were mixed for 10 h gradient elution: 0–4 min at 4% B, 15 min at 5.5% B, 30 min at 6.5%, 220 min at 12.5% B, 300 min at 17% B, 400 min at 26% B, 450 min at 35% B, 475 min at 42% B, 490 min at 51% B, 500 min at 60% B, 515–529 min at 96% B, 530–600 min at 4% B. Precursor ions were monitored at m/z 300–1500, $R = 140$ k (FWHM), 3e6 AGC (automatic gain control) target, and 120 maximum injection time (maxIT). Top ten precursor ions (0.8 Da isolation width; z = 2–5) underwent data-dependent higher-energy collisional dissociation (HCD) fragmentation at normalized collision energy (NCE) 36% using N_2 gas. Dynamic exclusion was set to 40 s. MS^2 spectra were monitored at $R = 17.5$ k (FWHM), 2e5 AGC target, and 120 maxIT. The fixed first mass was set to m/z 110 to match the iTRAQ reporter ions (m/z 114–117).

Database searches were performed by Proteome Discoverer (PD) 1.4 (Thermo Fisher Scientific, Dreieich, Germany) using the AspGD protein database of *A. fumigatus* Af293 [31] and the algorithms of MASCOT 2.4.1 (Matrix Science, UK), SEQUEST HT (integral search engine of PD 1.4), and MS Amanda 1.0. Two missed cleavages were allowed for tryptic digestion. The precursor mass tolerance and the integration tolerance (most confident centroid) were set to 5 ppm and the MS2 tolerance to 0.02 Da. Static modifications were carbamidomethylation of cysteine and either TMT6plex (proteome) or iTRAQ4plex (secretome) at lysine residues and the peptide N-terminus. Dynamic modifications were oxidation of methionine and either TMT6plex of threonine or iTRAQ4plex of tyrosine. Percolator and a reverse decoy database were used for *q*-value validation of the spectral matches ($\Delta cn < 0.05$). At least two peptides per protein and a strict target FDR < 1% were required for confident protein hits. The significance threshold for differential protein abundances for TMT and iTRAQ experiments was set to factor 1.5.

With the aid of the *biomaRt* package, proteins were annotated using identifiers provided by AspGD as of September 2015 and CADRE as of February 2017.

Chemical analysis of secondary metabolites

For quantification of fumagillin, fungal cultures were extracted and run on a LC-MS system consisting of an HPLC, UltiMate 3000 binary RSLC with photo diode array detector (Thermo Fisher Scientific, Dreieich, Germany) and the mass spectrometer (LTQ XL Linear Ion Trap from Thermo Fisher Scientific, Dreieich, Germany) with an electrospray ion source as described in Jöhnk et al. [32]. Data were obtained from three biological replicates and three technical replicates. A standard curve (1000, 500, 250, 125 and 62.5 μg/mL) using an authentic fumagillin standard (Abcam, United Kingdom) was calculated. The Xcalibur Quan Browser software (Thermo Fisher Scientific, Dreieich, Germany) was used to calculate the amounts of fumagillin.

Application of module-detecting approaches

A high-confidence (score > 0.7) PPIN of *A. fumigatus* strain A1163 was downloaded from STRING version 10 [33]. Both the PPIN and the pre-processed omics data were taken as input for the module-detecting approaches. Thereby, protein identifier annotations provided by CADRE were used.

ModuleDiscoverer

In order to apply MD for transcriptome data, the background contains all known *A. fumigatus* proteins described in AspGD. Analyzing proteome and secretome data, all proteins detected via LC-MS/MS were taken as background. The single-seed MD algorithm was applied to the input data as described by Vlaic et al. [19]. In brief, maximal cliques were identified using only one seed node in the PPIN. Cliques were tested for their enrichment with DEGs/DSyPs/DSePs using a permutation-based test as described in Vlaic et al. [19]. Cliques with a *p*-value < 0.01 were considered significantly enriched. Based on the union of these significantly enriched cliques, the regulatory module was assembled.

For the integration of different omics datasets, all regulatory modules were merged by forming the union of all nodes and edges. The resulting union regulatory module is defined as 'overall regulatory module' (ORM). Sub-modules with a number of nodes < 10 were not considered. Cytoscape version 3.2.1 [34] was used to visualize and analyze regulatory modules, for example, regarding their nodes' degree and betweenness centrality.

KeyPathwayMiner

KeyPathwayMiner (KPM) detects maximal connected sub-networks. In these sub-networks, all but a specific number *K* components are DEGs, DSyPs or DSePs in all but at most a specific number *L* cases [18]. In this study, cases are defined as the available time points. In a first analysis (I), KPM was applied to each single experimental dataset to receive one module for each time point of the respective molecular level. In the single-level analysis (II), the modules for each molecular level over all time points were identified. A third analysis (III) directly combined all of the experimental datasets to get the overall regulatory module. For the KPM input, one

matrix for each time point (I) or molecular level ((II) and (III)) were generated consisting of information about the components' regulation at the respective time points. For (II) and (III), only those components were considered that were DEGs/DSyPs/DSePs in at least one of the time points of the respective molecular level. With these matrices, the *A. fumigatus* PPIN and with the aid of KeyPathwayMiner Cytoscape App [18], sub-networks were computed using following settings: Ant colony optimization meta heuristic (ACO) as search algorithm, individual node exceptions (INEs) as search strategy, maximum of exception nodes $K = 2$. For (I) and (II), the maximal case exception parameter was set to $L = 0$. For the multilevel omics analysis (III), the logical connector of the different levels was set to the logical 'OR' and L was set to $L1 = 3$ (transcriptome data), $L2 = 1$ (proteome data) and $L3 = 0$ (secretome data). These L values were based on the number of time points of the respective molecular level. The assumption was that the considered component is a DEG/DSyP/DSeP in at least one measured time point. For instance, as four measured transcriptome time points were available, a gene was allowed to be not differentially expressed in maximal three out of four time points. The top ten best-scoring sub-networks were selected for further analysis. A KPM regulatory module describes the union of these top ten sub-networks of the respectively considered datasets.

Comparison of the simple approach and a module-detecting approach

Overlap of components

The overlap (percentage value) is defined as fraction of the intersection of the respective datasets from the union of the datasets. For the simple approach (SA), the overlap of different molecular levels was analyzed by comparing lists of DEGs, DSyPs and DSePs at the considered time points. For the module-detecting approach, the overlap of all components of the respective regulatory modules was considered.

In addition to the comparison of percentage values of overlapping components, a more objective measurement based on a permutation-based test was considered. Considering all known *A. fumigatus* proteins (N) described in AspGD, $D \in N$ is a set of components (DEGs, DSyPs or DSePs) for each of the molecular levels. In $I = 100{,}000$ iterations, datasets B were created where each set consists of $|D|$ components sampled from N. In every iteration, the overlap P of the molecular levels was calculated based on the generated datasets for transcriptome, proteome and secretome. The p-value was calculated by dividing the number of iterations in which $P \geq O$, where O represents the overlap received by SA or MD, and the total number of iterations I.

Correlation of the components' regulation

All components detected in at least one of the transcriptomic and one of the proteomic time points were considered for correlation analyses. The distance between results obtained for different molecular levels and time points was estimated based on the correlation of ranked lists of the components' absolute gene expression or protein synthesis regulation values (absolute log2FCs). Lists of ordered, absolute regulation values were rank-transformed. Indices corresponding to ties (equal values) were randomly ordered. Spearman's rank correlation coefficient r was calculated. The ranking was repeated 1000 times. Over all repeats, the final correlation between the regulation lists was averaged. The distance d is defined as $d = 1 - r$.

Generalized topological overlap

The ORM was clustered via the generalized topological overlap measure (GTOM) as described in [35]. Matrix $T^{[m]} = [t_{ij}^{[m]}]$ is called the m-th order GTOM matrix and includes the overlap of nodes reachable from the nodes i and j within m steps:

$$t_{ij}^{[m]} = \frac{|N_m(i) \cap N_m(j)| + a_{ij} + I_{i=j}}{min\{|N_m(i)|, |N_m(j)|\} + 1 - a_{ij}}$$

$A = [a_{ij}]$ is defined as adjacency matrix, $N_m(i)$ as the set of neighbors of i, the Identity matrix $I_{i=j}$ equals 1 if $i = j$ and zero else, $|\cdot|$ denotes the number of elements (cardinality) in its argument j. The clustering was performed for second-order connections. With the aid of the *hclust* function *(method = average),* a dendrogram based on all distances between proteins were generated. A cutoff of 0.65 was chosen to receive the clusters. R packages *RcolorBrewer* [36] *a*nd *WGCNA* [37] were applied for coloring the single clusters.

Enrichment analysis (functional annotation of biological processes)

Gene Ontology (GO) terms were applied for functional annotation concerning biological processes. Gene (gene product) terms of *A. fumigatus* were retrieved from AspGD as of October 2017. In particular, GO information about the Af293 strain was extracted and imported into R and was transformed into custom annotation objects by packages *AnnotationDbi* [38] and *GSEABase* [39] (each of version 1.38.2 as part of Bioconductor package collection version 3.5). In addition, the packages *GO.db* [40], *GOstats* [41] as well as the helper function *GSEAGOHyperGParams* of package *Category* [42] were applied for the enrichment analysis. For SA, all *A. fumigatus* proteins described in AspGD were taken as background. For the MD approach, all proteins which are part of the PPIN downloaded from STRING, were taken as background. GO terms composed of at least two

members, associated with at least two components and leading to *p*-values below 0.05 were considered as significantly enriched.

Results

Data overview

We used experimental omics data of a *A. fumigatus* study that investigated the stress response to the antifungal drug caspofungin at different molecular levels (transcriptome, proteome, secretome) including different time points. Figure 2a provides an overview of the available datasets including all genes and proteins detected by RNA-Seq and LC-MS/MS. Over all considered time points, 9881 genes were measured for the transcriptomic response, 3858 proteins for the proteomic response and 1110 proteins for the secretome. Filtering the data for DEGs, DSyPs and DSePs resulted in 1058 DEGs (498 upregulated (↑), 560 downregulated (↓)) at 0.5 h, 1237 DEGs (876 ↑, 361 ↓) at 1 h, 1322 DEGs (784 ↑, 538 ↓) at 4 h and 1068 DEGs (600 ↑, 468 ↓) at 8 h after caspofungin treatment. In the proteome, 230 DSyPs (88 ↑, 142 ↓) were identified at 4 h after treatment, and 204 DSyPs (114 ↑, 90 ↓) at the 8 h time point. 136 DSePs (118 ↑, 18 ↓) were detected for the secretome at 8 h after treatment (Fig. 2b). Complete lists of DEGs, DSyPs and DSePs are provided in the Additional file 1.

Overlap of datasets of the different molecular levels

We started to analyze the molecular level overlap by comparing all measured genes or proteins (hereafter called 'components') independently of their differential regulation and time points. This comparison showed that the overlap of all three molecular levels amounted to 10.5% (Fig. 2a). Applying SA and MD to the experimental data (Fig. 3), this level overlap accounted for 0.5% (SA) and 6.1% (MD). Considering only two out of three molecular levels (including data of all considered time points, respectively), both approaches resulted in the highest overlap for the proteome/secretome comparison (11.2% SA, 21.4% MD). This observation was not surprising as the secreted proteins are also included in the global proteome. We found that MD provided an up to 12-fold higher overlap than SA.

A further analysis of overlapping components considered a more objective measurement based on a permutation-based test. In 100,000 iterations, random datasets for transcriptome, proteome and secretome were generated and the overlap of all three datasets was calculated. The median-value of all 100,000 random overlaps equaled 3. Thus, the level overlap accounted for 0.1%. For the SA-obtained overlap (11 components or 0.5% as presented in Fig. 3), we calculated a p-value = 2.8×10^{-4} which is statistically significant in comparison to random overlaps. In contrast, the MD-received overlap (58 components or 6.1% as presented in Fig. 3) resulted in the smaller p-value = 1.0×10^{-5}. Comparing the overlap percentage values, SA produced 5-fold and MD even 61-fold higher overlap values than random overlaps. The comparison of the SA- and MD-received overlap values resulted in the above-mentioned 12-fold higher values for MD.

Estimation of the best match of transcriptomic and proteomic time points

The selection of measured time points was based on the following assumption: The expression of a gene and the synthesis of its corresponding protein do not occur at the same time since they are consecutive processes. Thus, changes in the transcriptional regulation are also reflected in the differential synthesis of proteins at the proteomic level but most likely at a later time point. Therefore, different time points at the transcriptomic and proteomic level were selected to consider the delay between transcription and translation during the fungal response. Hence, we analyzed our results regarding best matches of level- and time point-dependent sub-responses.

We tested two approaches for estimating the best transcriptome-proteome time point match: Comparison of components, and correlation of the components'

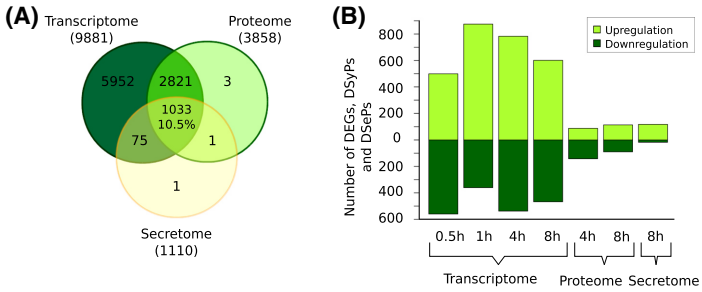


Fig. 2 Overview of the available datasets. **a** Number and overlap of all measured genes and proteins. **b** Number of differentially expressed genes (DEGs), differentially synthesized proteins (DSyPs) and differentially secreted proteins (DSePs) in all available experimental datasets

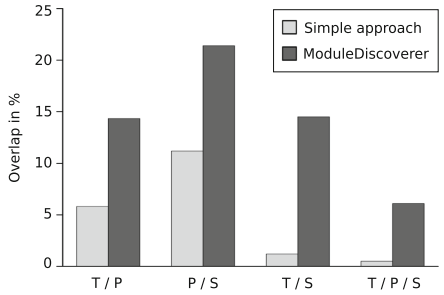


Fig. 3 Overlap of molecular levels. Overlap of transcriptome (T), proteome (P) and secretome (S) regarding their components (genes or proteins)

regulation. The first estimation approach aimed at analyzing overlapping components in the transcriptome and proteome which can be observed, for instance, as transcripts and their synthesized proteins. For the second estimation approach, the correlation of the components' regulation was calculated based on absolute gene expression or protein synthesis regulation values. This approach represents the regulation of response pathways which not necessarily contain overlapping components but also other genes or proteins contributing to these pathways. Therefore, in this approach not only the overlapping components were analyzed but also components which are part from at least one of the respectively compared transcriptome and proteome time points. This leads to a higher number of considered components.

Starting with the comparison of components (Fig. 4a), both SA and MD demonstrated the best match for the transcriptomic response 1 h and the proteomic response 4 h after caspofungin treatment (5.3% SA, 16.5% MD). While SA resulted in the best match of transcriptome at 8 h with proteome at 8 h (7.3%), MD showed the best match with transcriptome at 4 h (16.8%). Consequently, for both time point comparisons, MD-produced results indicated a delay of 3–4 h between the different sub-responses. Taking into account also the correlation of components' regulation, Fig. 4b shows that similar to the previous analyses MD provided a better, i.e., here lower, distance for MD values than for SA. Oppositely to SA, the MD results confirmed the best time point match of transcriptome at earlier time point (1 h) and proteome at the later one (8 h) (Fig. 4b), similarly to the aforementioned comparison of components (Fig. 4a).

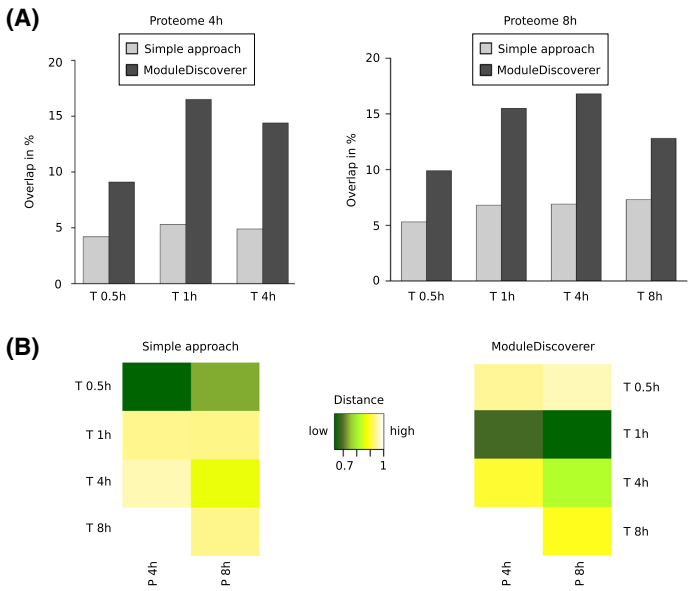


Fig. 4 Transcriptome-proteome time point match. Estimation of the best time point match for transcriptome (T) and proteome (P) time points regarding **a** comparison of components and **b** correlation of the components' regulation. Distance is defined as 1 minus correlation coefficient

The lowest distances were observed for the proteome at 8 h and transcriptome at 1 h (Fig. 4b, highlighted in dark green), followed by the proteome at 4 h and transcriptome at 1 h (Fig. 4b, dark green). These findings were also in agreement with the highest and second highest overlap values in Fig. 4a. Together with the observation that both approaches showed very high distance values (yellow and light yellow) between the same transcriptome and proteome time points, our results support the assumption of a time delay between level-dependent sub-responses (transcription and translation). Tendencies in the coherence of time points and an estimation of the resulting time delay between molecular levels may be helpful for further wet-lab studies regarding time- and cost-saving by focusing on the most relevant time points.

Another observation can be made by comparing the respective results of the two estimation approaches: There is a tendency that the correlation-based approach resulted in best matches for earlier transcriptome time points than the overlap-based approach. This observation may be based on the activation of stress response pathways induced by the fungus shortly after the caspofungin treatment. As such response pathways could involve components from both molecular levels transcriptome and proteome, we assume that the actual regulation of response pathways represented by the correlation-based approach already starts before the main translation process of potentially involved components occurs (represented by the overlap-based approach).

Integration of multilevel omics data

Analysis of the overall fungal response to caspofungin

All regulatory modules of each molecular level and time point identified by MD (Table 1 and Additional file 2: Table S1) can be considered to be part of the overall fungal response to caspofungin. Forming the union of them, the resulting overall regulatory module (ORM) was composed of five sub-modules including 894 components (Fig. 5). For a focused enrichment analysis based on the ORM's underlying topology, we performed a generalized topological overlap measurement regarding the main sub-module 1. Figure 5 represents the ORM with its five sub-modules and the 15 clusters of sub-module 1, where the cluster membership of each protein is color-coded. An overview of all components of the ORM including sub-modules and clusters is available in Additional file 2: Table S2). GO term enrichment analyses showed that the clusters were significantly enriched with distinct biological functions (see Additional file 2: Tables S3–21) for a list of all significantly associated biological processes of each cluster and the remaining sub-modules). Examples of such processes are protein phosphorylation and response to oxidative stress (cluster 2, Additional file 2: Table S4), actin filament-based process (cluster 3, Additional file 2: Table S5), regulation of kinase activity (cluster 5, Additional file 2: Table S7), amino acid metabolic processes (cluster 6, Additional file 2: Table S8 and cluster 9, Additional file 2: Table S11), (1,3)-alpha-D-glucan biosynthesis (cluster 7, Additional file 2: Table S9), secondary and lipid metabolic process (cluster 12, Additional file 2: Table S14 and cluster 13, Additional file 2: Table S15) or transport mechanisms (cluster 15, Additional file 2: Table S17 and sub-module 5, Additional file 2: Table S21).

Table 1 Regulatory modules generated by ModuleDiscoverer

Underlying experimental dataset	Number of nodes (components)	Number of edges (interactions)
Transcriptome 0.5 h	511	2967
Transcriptome 1 h	256	1336
Transcriptome 4 h	313	1604
Transcriptome 8 h	256	1208
Proteome 4 h	147	845
Proteome 8 h	124	520
Secretome 8 h	293	2413
Overall regulatory module	894	6111

Number of nodes (representing gene or protein components) and edges (representing interactions between the components) of the regulatory modules received by ModuleDiscoverer

Polyubiquitin and CBF/NF-Y family transcription factor as potential key factors contributing to the caspofungin-induced response

To investigate potential key factors in the fungal response contributing to, e.g., caspofungin-caused side effects, we analyzed the underlying topological network structure of the ORM. We took into account the network node-associated degree (number of edges connected to the node) and betweenness centrality (number of shortest paths that go through each node) [13]. We identified the node representing polyubiquitin UbiD with the fifth highest degree (Table 2) and the third highest betweenness centrality (Table 3). It was furthermore the only node that could be found in the top ten lists of both measures. Ubiquitin is a highly conserved 76-residue protein which can be found in all eukaryotic organisms [43]. In *Saccharomyces cerevisiae*, the orthologous gene UBI4, one out of four ubiquitin genes in yeast, was shown to be essential for resistance to different stresses including high temperatures and starvation [44].

In addition to this topology-based approach, we also applied an approach focused on transcription factors. Transcription factors play an important role in regulating the compensatory stress response to drugs. However, in many cases, it is difficult to measure transcription factors' activity since they are often constitutively expressed and/or activated posttranscriptionally. Therefore, we scanned the ORM for transcription factors connected to DEG-associated

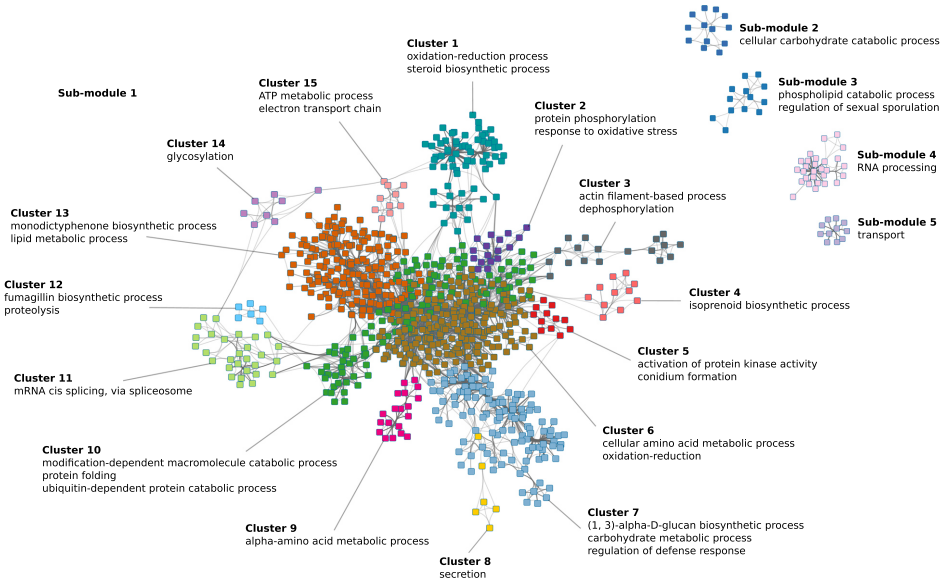


Fig. 5 Overall regulatory module representing the response of *A. fumigatus* to caspofungin. The overall regulatory module identified by ModuleDiscoverer is composed of five sub-modules including 894 components (see Additional file 2: Table S2). Clusters with exemplarily selected significantly enriched biological processes are color-coded

proteins, DSyPs or DSePs (Table 4). Among them, we detected the CBF/NF-Y family transcription factor. It shows similarities to DNA polymerase epsilon subunit *DPB4* of *S. cerevisiae* and *Schizosaccharomyces pombe*.

Both polyubiquitin and the CBF/NF-Y family transcription factor were detected in all transcriptome and, in case of the CBF/NF-Y family transcription factor, proteome time points but neither as DEG nor as DSyP. Figure 6 represents these two nodes and their respective first neighbors (including DEGs, DSyPs or DSePs) within the ORM.

The investigation of potential key factors in the drug-induced response, like polyubiquitin and CBF/NF-Y family transcription factor, may help to better understand the position and dynamics of drug targets and associated proteins in the interaction network and can potentially contribute to increase the safety of drugs.

Caspofungin induces increased production of the secondary metabolite fumagillin

As described above, the ORM contained two clusters, cluster 12 and 13, which included several enzymes that are involved in the biosynthesis of secondary metabolites. In particular, transcripts and their corresponding proteins of the antimicrobial agent fumagillin biosynthesis gene cluster (11 out of 15 cluster genes) showed increased levels after exposure of *A. fumigatus* to caspofungin. To verify whether caspofungin triggers the production of this meroterpenoid, we extracted *A. fumigatus* cultures exposed for 8 h to caspofungin (100 ng/ml) and control cultures with ethyl acetate and determined the fumagillin concentration by LC-MS. In cultures without caspofungin the concentration of fumagillin was 67.3 ± 21.7 μg/ml, while in cultures with caspofungin the concentration increased by 3-fold to 208.1 ± 63.8 μg/ml (Fig. 7). The level of other secondary metabolites such as pseurotin A stayed almost unchanged (Additional file 3).

Comparison of ModuleDiscoverer- and KeyPathwayMiner-generated regulatory modules

To estimate the comprehensiveness of MD-generated regulatory modules, we applied another available module-detecting approach, KeyPathwayMiner (KPM), to our experimental datasets and compared the identified regulatory modules with those identified by MD (Table 5).

Table 5 shows the numbers of components of the KPM-produced regulatory modules for each time point and the overall regulatory module in comparison with those based on MD. Exemplarily, the comparison showed that the ORM received by MD contains a 1.5-fold higher number of

Table 2 Nodes of the overall regulatory module with highest degree

CADRE-IDs	AspGD-IDs	Protein names	Degree	BC	log2FC						
					T 0.5 h	T 1 h	T 4 h	T 8 h	P 4 h	P 8 h	S 8 h
CADAFUBP00004294	AFUB_043760	Fatty acid synthase beta subunit, putative	146	0.126	−1.159	−0.628	− 0.693	− 0.828	− 0.006	− 0.136	0.534
CADAFUBP00004295	AFUB_043770	Fatty acid synthase alpha subunit FasA, putative	142	0.082	−1.008	−0.517	− 0.678	− 0.726	−0.038	− 0.105	1.448
CADAFUBP00002402	AFUB_024590	Acetyl-CoA carboxylase	124	0.122	−1.697	−0.812	−1.285	− 1.380	− 0.456	−0.628	1.518
CADAFUBP00004404	AFUB_044900	Nonribosomal peptide synthase SidE	114	0.048	1.077	1.918	1.613	1.229	NA	NA	NA
CADAFUBP00006564	AFUB_067450	Polyubiquitin UbiD/Ubi4, putative	111	0.396	−0.762	0.238	−0.248	−0.688	NA	NA	NA
CADAFUBP00007473	AFUB_076690	ATP citrate lyase, subunit 1, putative	98	0.037	−1.636	−0.705	−0.698	− 0.584	−0.474	− 0.521	0.683
CADAFUBP00007537	AFUB_077330	Bifunctional pyrimidine biosynthesis protein (PyrABCN), putative	82	0.073	−0.438	−0.468	− 0.214	−0.974	− 0.206	−0.340	1.586
CADAFUBP00000761	AFUB_007730	Glutamate synthase Glt1, putative	74	0.035	−2.168	−0.525	−0.119	− 0.483	−0.255	− 0.234	1.135
CADAFUBP00001006	AFUB_010250	Succinyl-CoA synthetase, alpha subunit, putative	72	0.021	−0.497	−0.273	0.167	−0.011	0.234	0.069	NA
CADAFUBP00003062	AFUB_031240	Sulfite reductase, putative	68	0.047	−0.900	−0.774	− 0.103	−0.598	− 0.016	−0.022	1.5

Top ten nodes of the overall regulatory module showing the highest degree and additional information regarding their betweenness centrality (BC) and gene- or protein-associated log2 Fold Change (log2FC) measured for the transcriptomic (T), proteomic (P) and secretomic (S) fungal response to caspofungin at all time points, respectively

Table 3 Nodes of the overall regulatory module with highest betweenness centrality

CADRE-IDs	AspGD-IDs	Protein names	Degree	BC	log2FC						
					T 0.5 h	T 1 h	T 4 h	T 8 h	P 4 h	P 8 h	S 8 h
CADAFUBP00007914	AFUB_081260	Peptidyl-arginine deiminase domain protein	4	0.6	0.696	3.125	2.328	1.647	NA	NA	NA
CADAFUBP00001626	AFUB_016580	Long-chain-fatty-acid-CoA ligase, putative	11	0.405	−1.395	−1.605	−0.406	−0.750	NA	NA	NA
CADAFUBP00006564	AFUB_067450	Polyubiquitin UbiD/Ubi4, putative	111	0.396	−0.762	0.238	−0.248	−0.688	NA	NA	NA
CADAFUBP00008739	AFUB_089890	Mandelate racemase/muconate lactonizing enzyme family protein	10	0.304	1.791	0.546	0.247	0.055	NA	NA	NA
CADAFUBP00003378	AFUB_034540	Lysophospholipase 3	5	0.303	0.906	1.357	1.185	1.357	0.706	0.937	1.513
CADAFUBP00002707	AFUB_027690	Lysophospholipase	10	0.273	−1.454	−0.217	−0.854	−1.700	0.387	0.192	NA
CADAFUBP00006379	AFUB_065540	Patatin-like phospholipase domain-containing protein	10	0.273	−0.639	− 0.570	−0.931	−1.460	NA	NA	NA
CADAFUBP00008747	AFUB_089980	Ribosome biogenesis protein (Rrs1), putative	26	0.259	1.045	−0.096	−0.154	0.413	0.185	0.045	NA
CADAFUBP00004062	AFUB_041460	Plasma membrane ATPase	4	0.25	−1.535	−0.670	−1.002	−1.152	0.041	−0.023	1.910
CADAFUBP00005096	AFUB_052070	Plasma membrane ATPase	4	0.25	−0.378	− 0.299	0.441	0.463	NA	NA	NA
CADAFUBP00000491	AFUB_004970	Alcohol dehydrogenase, zinc-containing, putative	3	0.167	−1.712	−1.178	−1.125	−0.941	−0.759	−0.454	0.914

Top ten nodes of the overall regulatory module showing the highest betweenness centrality (BC) and additional information regarding their node degree and gene- or protein-associated log2 Fold Change (log2FC) measured for the transcriptomic (T), proteomic (P) and secretomic (S) fungal response to caspofungin at all time points, respectively

Table 4 Transcription factors within the overall regulatory module

CADRE-IDs	AspGD-IDs	Protein names	log2FC						
			T 0.5 h	T 1 h	T 4 h	T 8 h	P 4 h	P 8 h	S 8 h
CADAFUBP00000978	AFUB_009970	CBF/NF-Y family transcription factor, putative	0.230	0.447	−0.124	0.002	−0.034	−0.150	NA
CADAFUBP00001789	AFUB_018340	HLH transcription factor, putative	−1.441	−0.236	0.128	−0.355	NA	NA	NA
CADAFUBP00003751	AFUB_038290	Zinc knuckle transcription factor/splicing factor MSL5/ZFM1, putative	0.366	0.062	0.111	0.143	−0.368	−0.322	3.379
CADAFUBP00004232	AFUB_043140	Transcription elongation factor SPT6, putative	−0.369	−0.235	− 0.177	−0.563	0.143	0.188	NA
CADAFUBP00005084	AFUB_051950	PHD transcription factor (Rum1), putative	−1.069	0.036	0.089	−0.780	−0.071	− 0.042	NA
CADAFUBP00007653	AFUB_078520	Stress response regulator/HFS transcription factor, putative	0.317	−0.271	0.034	−0.102	−0.130	0.072	NA
CADAFUBP00001318	AFUB_013400	TFIIH complex helicase Rad3, putative	0.154	0.034	−0.055	− 0.226	−0.032	0.222	NA
CADAFUBP00003811	AFUB_038920	Ccr4-Not transcription complex subunit (NOT1), putative	−0.786	−0.076	0.038	−0.834	− 0.100	−0.041	NA

Transcription factors detected in the overall regulatory module and their log2 Fold Change (log2FC) measured for the transcriptomic (T), proteomic (P) and secretomic (S) fungal response to caspofungin at all time points, respectively

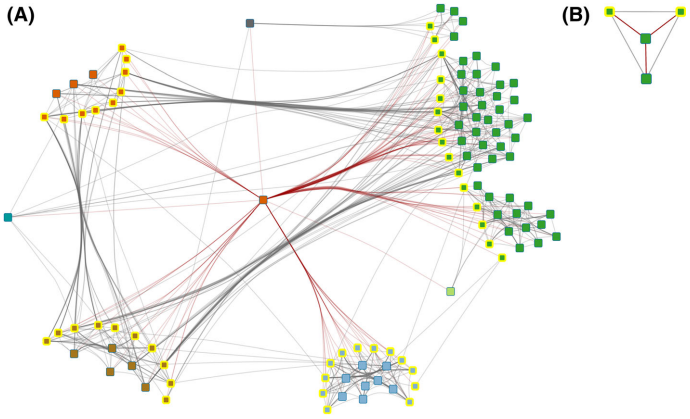


Fig. 6 Potential key factors within the overall regulatory module contributing to the caspofungin-caused fungal response. **a** Polyubiquitin and **b** CBF/NF-Y family transcription factor (centrally arranged, respectively) and their first neighbors in the overall regulatory module. DEG-associated proteins, DSyPs and DSePs are highlighted with a yellow border

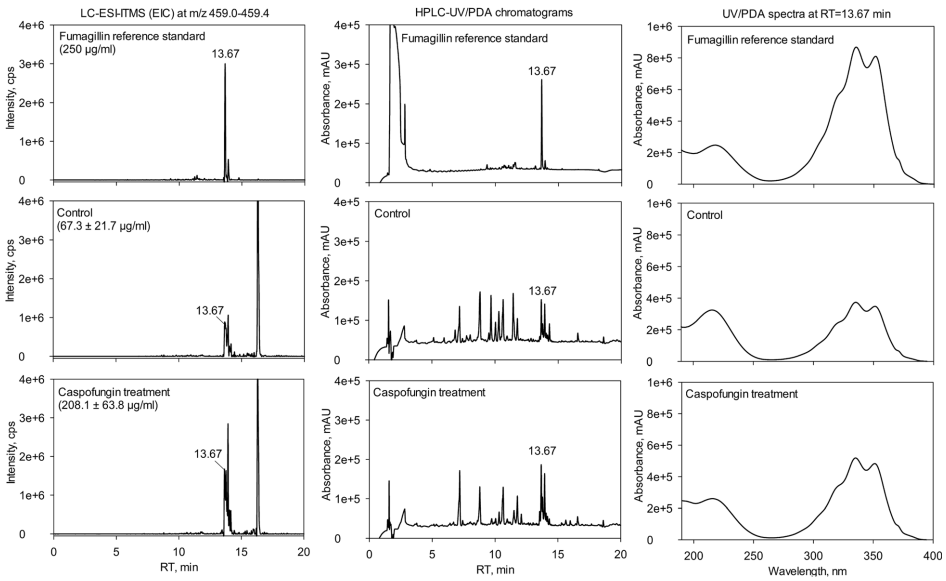


Fig. 7 Caspofungin-induced increased production of the secondary metabolite fumagillin. LC-ESI-ITMS extracted ion chromatograms (EIC) at m/z 459.0–459.4 amu (left), HPLC-UV/PDA chromatograms (center) and UV/PDA spectra at RT = 13.67 min (right) of 250 µg/ml fumagillin reference standard (top) and crude extract of *A. fumigatus* without (center) and with caspofungin treatment (bottom)

components by covering more than 60% of KPM module components. Considering the modules of the single time point datasets, e.g. secretome at 8 h, we found an up to 6.5-fold higher component number by covering up to 93% of KPM components. Hence, we focused on the results received by MD. Nevertheless, additional KPM analyses regarding the overlap of molecular levels and the estimation of the best match of transcriptomic and proteomic time points are shown in Additional file 2: Figures S1 and S2 and Additional file 4.

Discussion

In this study, we focused on the integration of omics data derived from heterogeneous sources. Therefore, we used experimental data of an *A. fumigatus* study investigating the stress response to the antifungal drug caspofungin at different molecular levels and time points. For the analyses, we applied SA considering only DEGs/DSyPs/DSePs and the regulatory module-detecting single-seed MD approach considering DEGs/DSyPs/DSePs, non-DEGs/DSyPs/DSePs as well as structural PPIN information. We focused on the single-seed approach instead of the also available multi-seed MD approach since the single-seed approach is comparable with other well-established maximal clique enumeration problem-based algorithms (e.g., Barrenäs et al. [45] or Gustafsson et al. [46]). In addition, Vlaic et al. showed that the multi-seed-identified modules can be essentially considered as an extension of the single-seed modules. However, we also applied the multi-seed approach to our experimental data set. In summary, the multi-seed MD approach allows for effectively integrating multilevel omics data. Multi-seed-generated results contain the regulatory modules received by the single-seed approach and are even more comprehensive. The overall regulatory module generated by the multi-seed approach confirms the already observed key players and significantly associated processes. Details on the analyses can be found in the Additional files 2 and 5.

Table 5 Comparison of ModuleDiscoverer- and KeyPathwayMiner-detected regulatory modules

Underlying experimental dataset	Component number of MD modules	Overlap (percentage value regarding KPM module)	Component number of KPM modules
Transcriptome 0.5 h	511	134 (75.7%)	177
Transcriptome 1 h	256	62 (63.9%)	97
Transcriptome 4 h	313	123 (74.1%)	166
Transcriptome 8 h	256	89 (65.0%)	137
Proteome 4 h	147	36 (75.0%)	48
Proteome 8 h	124	30 (63.8%)	47
Secretome 8 h	293	42 (93.3%)	45
Overall regulatory module	894	343 (59.6%)	576

Comparison of ModuleDiscoverer (MD) and KeyPathwayMiner (KPM) regarding their number of module components. The overlap is defined as fraction of the intersection of the respective datasets from the KPM datasets

Relation of transcriptomic, proteomic and secretomic data

The comparison of all three molecular levels regarding all measured, SA- or MD-considered components resulted in only small overlap values. This observation is in agreement with other integrative transcriptomic and proteomic studies reporting that there is no or only a weak correlation between different molecular levels [47–49]. Potential explanations are biological (e.g., translational regulation or differences in protein and mRNA half-lives in vivo) or methodological origins (e.g., detection limits of the techniques or the choice of measured time points) [48, 49]. Figures 2a and 3 show an apparently contradictory outcome regarding the overlap of datasets of different molecular levels: Fig. 2a shows the highest overlap percentage value for transcriptome and proteome, Fig. 3 for proteome and secretome. This can be explained by the fact that Figs. 2a and 3 are based on analyses that considered diverse datasets. For Fig. 2a, all detected genes and proteins were analyzed. In contrast, Fig. 3 comprises only a fraction of these components because of a further filtering step to only compare DEGs/DSyPs/DSePs (SA) or regulatory module components (DEGs/DSyPs/DSePs and associated background proteins, MD). Actually, in Fig. 3, both approaches MD and SA showed the highest overlap between proteome and secretome. On the one hand, this highest overlap percentage reflects the same underlying measurement technique. In this study, the transcriptome was measured by RNA-Seq, the proteome and secretome by LC-MS/MS. As the techniques themselves are very different, also differences in their respective outcome can be expected. Therefore, as the intracellular proteome and secretome are based on the same measurement technique, they are more similar to each other than, for instance, transcriptome and proteome. On the other hand, the highest overlap also demonstrates the biological similarity in terms of immediately consecutive protein-based levels. Thus, both levels consist of proteins which differ only in the secretion step via classical (i.e., N-terminal secretory signal peptide triggered) or non-classical (i.e., without involvement of N-terminal signal peptides) secretory pathways [50]. Hence, proteome and secretome can be considered as immediately consecutive levels which can both be measured by LC-MS/MS.

By a general comparison of MD- and SA-received results, we determined up to 12-fold higher overlap values provided by MD than those calculated by SA. This is reasonable as SA focuses on the comparison of lists of DEGs, DSyPs and DSePs, exclusively. Hence, non-DEGs/

DSyPs/DSePs measured in the experimental background were not considered which results in a high loss of data for the analyses. In contrast, the additional information considered by MD led to a much higher number of (overlapping) components.

Analysis of the overall fungal response and potential key factors

With the aid of the ORM, we analyzed the *A. fumigatus* response to caspofungin over all molecular levels and time points. We found that ORM clusters are significantly enriched with biological functions like (1,3)-alpha-D-glucan biosynthesis and carbohydrate metabolic processes, actin filament-based processes, activation of protein kinase activity and response to oxidative stress. These results are in agreement with a genome-wide expression profiling study of *Aspergillus niger* in response to caspofungin [51]. Here, many of the upregulated genes were predicted or confirmed to function in cell wall assembly and remodeling, cytoskeletal organization, signaling and oxidative stress response. Also, genes and proteins of the electron transport chain were specifically enriched which supports the hypothesis that caspofungin acts as an effector of mitochondrial oxidative phosphorylation [52]. This is consistent with results from Cagas et al. [47] who analyzed the proteomic response of *A. fumigatus* to caspofungin and identified the largest change in a mitochondrial protein that has a role in mitochondrial respiratory chain complex IV assembly. The significant enrichment of genes and proteins of the amino acid metabolic process is best explained by the growth inhibitory activity of caspofungin that leads to the downregulation of the primary metabolisms including amino acid biosynthesis [53].

The cluster 5 represents (gene-associated) proteins involved in the activation of protein kinase activity. Mitogen-activated kinases (MAPK) are important regulators in the fungal response to stress that is induced by environmental changes or the disruption of cell wall integrity ([54], and references therein) which are both consequences of the caspofungin treatment. Also cellular transport mechanisms were influenced by this antifungal drug leading to osmotic stress as already reported in Altwasser et al. [26]. In addition, we observed the association of ORM cluster components with the (1,3)-alpha-D-glucan biosynthesis as well as carbohydrate metabolic processes. Consistently, caspofungin inhibits the synthesis of β-(1,3)-glucan which is the principal component of the fungal cell wall [55]. As a compensatory response, the production of other cell wall polymers was stimulated. Another interesting finding was the increased production of the secondary metabolite fumagillin upon exposure of *A. fumigatus* to caspofungin. So far, only the release of the secondary metabolite gliotoxin has been reported for cultures of *A. fumigatus* in the presence of caspofungin [56]. Fumagillin has anti-angiogenic activity [57] and induces cell death in erythrocytes [58]. It is therefore possible that administration of caspofungin induces the production of secondary metabolites that have adverse effects on host cells during the infection. Another interesting aspect of our finding is that the induction of fumagillin production upon caspofungin exposure may represent a form of 'microbial communication' between fungi, in particular taking into account that echinocandins like caspofungin are produced by a diverse set of fungi [59].

As Wang et al. [13] reported, studying key factors of a drug-induced response by analyzing the underlying network structure may help to better understand the position and dynamics of drug targets and associated proteins potentially involved in drug-caused side effects. Here, in addition to the main target β-(1,3)-D-glucan synthase, we detected polyubiquitin UbiD among the top five nodes of the ORM ranked by both node degree and betweenness centrality. Polyubiquitin is known to encode multiple ubiquitin units in tandem, each of these transcribed as a single transcript. It is involved in several metabolic pathways and plays an important role in the regulation of the proteasome-based protein degradation processes [43, 60]. Some recent studies have already reported the importance of polyubiquitin in the fungal stress response. In the pathogenic yeast *Candida albicans*, Leach et al. [61] have shown that polyubiquitin is required for the adaption to sudden stress induced, e.g., by heat or caspofungin and is critical for the fungus' pathogenicity. In another study in *S. cerevisiae*, Lesage et al. [62] described ubiquitin-related protein degradation as an important process in the compensation for defects in glucan biosynthesis. We hypothesize that polyubiquitin is an important player in the compensatory response of *A. fumigatus* to caspofungin. In line, the corresponding gene *ubi4* was shown to be induced upon heat-shock in *A. nidulans* [43].

Exemplarily, CBF/NF-Y family transcription factor was detected among the list of TFs. Its *C. albicans* ortholog DPB4 represents a putative DNA polymerase epsilon subunit D and was shown to be involved in filamentous growth and maintenance of the mitochondrial DNA genome [63]. This role in mitochondrial processes in conjunction with caspofungin treatment is in agreement with the in previous studies shown importance of mitochondrial functions for drug tolerance and virulence of fungal pathogens ([47], and references therein). Also for *C. albicans*, Khamooshi et al. [64] have reported that deletion of DPB4 results in a decreased resistance to caspofungin in drop plate assays. These facts could indicate an involvement of CBF/NF-Y family transcription factor in the resistance of *A. fumigatus* to caspofungin.

Interestingly, in our study, both the polyubiquitin and the CBF/NF-Y family transcription factor were detected in all transcriptome and, in case of CBF/NF-Y family transcription factor, proteome time points but neither as DEG nor as DSyP. However, their location within the

ORM had shown that they are closely related to DEGs, DSyPs or DSePs. Consequently, by considering DEGs, DSyPs or DSePs for data analyses by SA, these proteins would not have been taken into account as factors in the fungal response despite the fact that they likely have a strong influence on DEGs, DSyPs or DSePs as shown in the ORM. To our knowledge, the role of both the polyubiquitin and the CBF/NF-Y family transcription factor has not been examined yet in the context of caspofungin-induced stress in *A. fumigatus*. Hence, our analyses offer novel hypotheses which have to be verified in future studies.

The module-detecting approach KeyPathwayMiner

In addition to MD, also other approaches identifying regulatory modules are available, for instance, KPM. Similar to MD, KPM can be used for the analyses of both, single-level and multilevel omics data. However, it does not make assumptions about community structures. KPM combines DEGs, DSyPs or DSePs with non-DEG/DSyP/DSeP exception nodes acting as 'bridges' to detect maximal connected sub-networks [15]. The comparison of MD- and KPM-generated regulatory modules showed that MD generates modules with a significant higher number of components than KPM. Additionally, these MD module components cover most of the KPM components. As these findings indicate that MD-generated modules are more comprehensive than modules derived by KPM, we focused on the results obtained by MD.

PPIN information as limiting factor

The basis of module-detecting approaches like MD or KPM is information from underlying organism-specific PPINs. Hence, the quality of results provided by these approaches also depends on the comprehensiveness of the underlying PPIN itself. Only those components of the experimental data which do also occur in the PPIN are considered for the regulatory module. For example, the PPIN of *A. fumigatus* strain A1163 downloaded from STRING consists of 4123 proteins. But according to current information provided by CADRE, the fungus itself is known to comprise 9916 protein-coding genes. Hence, more than half of the known fungal components cannot be considered for analyses based on this PPIN. Consequently, the available PPIN information can be considered as limiting factor in the data analyses. Thus, while our results highlight the benefits and potential provided by the regulatory module detection-based analysis of multilevel omics data, future studies will have to focus on the expansion of organism-specific PPINs.

Conclusion

PPINs enable the consideration of both structural and functional relationships between network proteins. Thus, they facilitate a focused view on closely related components in terms of modules. In this study, we demonstrated so far untested capacity of the module-detecting MD approach to integrate omics data coming from different molecular levels and time points. Moreover, we showed that this level of integration is not achievable using a simple approach of comparing lists of DEGs/DSyPs/DSePs. The integration of these data in one ORM can provide an overview of the overall organism's response to an external stimulus. We presented several approaches for analyzing this response and potential key factors contributing to, e.g., drug-caused side effects in more detail. With the aid of the regulatory module-detecting approach, it is possible to identify potential response key factors which cannot be detected in commonly used approaches comparing DEGs, DSyPs and DSePs, exclusively.

Additional files

Additional file 1: Lists of differentially expressed genes, differentially synthesized proteins and differentially secreted proteins. (XLSX 227 kb)

Additional file 2: Supplementary Materials. (PDF 3110 kb)

Additional file 3: Quantification of the secondary metabolites fumagillin and pseurotin A. (XLSX 21 kb)

Additional file 4: KeyPathwayMiner-generated overall regulatory module and significantly enriched biological processes. (XLSX 55 kb)

Additional file 5: Significantly enriched biological processes of the MD-multi-seed-based overall regulatory module. (XLSX 49 kb)

Acknowledgements

The authors would like to thank Silke Steinbach for excellent technical assistance. We also thank Dominik Driesch for fruitful discussions.

Funding

This work was supported by the Jena School for Microbial Communication (JSMC) [to TC], Deutsche Forschungsgemeinschaft (DFG) CRC/Transregio 124 'Pathogenic fungi and their human host: Networks of interaction' (subprojects A1 [to AAB], C5 [to IJ], INF [to JL, RG] and Z2 [to OK, TK]), Thüringer Aufbaubank (TAB) [to JL] and the German Federal Ministry of Education & Research (BMBF FKZ 0315439) [to AAB, VV].

Availability of data and materials

The datasets supporting the conclusion of this article are included within the article and its additional files. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE [65] partner repository with the dataset identifier PXD008153. The RNA-Seq data that support the findings of this study are available as mentioned in Altwasser et al. [26].

Authors' contributions

TC, SGH, TK, JL and SV performed the data analyses. OK, DJM and VV performed the experiments. TC, OK, SGH, RG, IDJ, AAB, SV and JL interpreted the results. TC, OK, TK, SV and JL wrote the paper. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Author details

[1]Systems Biology/Bioinformatics, Leibniz Institute for Natural Product Research and Infection Biology – Hans Knöll Institute, Jena, Germany. [2]Molecular and Applied Microbiology, Leibniz Institute for Natural Product Research and Infection Biology – Hans Knöll Institute, Jena, Germany. [3]BioControl Jena GmbH, Jena, Germany. [4]Biobricks of Microbial Natural Product Syntheses, Leibniz Institute for Natural Product Research and Infection Biology – Hans Knöll Institute, Jena, Germany. [5]Microbial Immunology, Leibniz Institute for Natural Product Research and Infection Biology – Hans Knöll Institute, Jena, Germany. [6]Institute for Microbiology, Friedrich Schiller University, Jena, Germany. [7]Research Group PiDOMICs, Leibniz Institute for Natural Product Research and Infection Biology – Hans Knöll Institute, Jena, Germany. [8]Institute for Bacterial Infections and Zoonoses, Federal Research Institute for Animal Health – Friedrich Loeffler Institute, Jena, Germany. [9]Present address: PerkinElmer Inc., Rodgau, Germany.

Received: 12 March 2018 Accepted: 8 October 2018
Published online: 20 October 2018

References

1. Ebrahim A, Brunk E, Tan J, O'Brien EJ, Kim D, Szubin R, et al. Multi-omic data integration enables discovery of hidden biological regularities. Nat Commun. 2016;7:13091.
2. Wu Y, Williams EG, Dubuis S, Mottis A, Jovaisaite V, Houten SM, et al. Multilayered genetic and omics dissection of mitochondrial activity in a mouse reference population. Cell. 2014;158:1415–30. https://doi.org/10.1016/j.cell.2014.07.039.
3. Chen R, Mias GI, Li-Pook-Than J, Jiang L, Lam HYK, Chen R, et al. Personal omics profiling reveals dynamic molecular and medical phenotypes. Cell. 2012;148:1293–307.
4. Aebersold R, Mann M. Mass-spectrometric exploration of proteome structure and function. Nature. 2016;537:347–55.
5. Michalski A, Cox J, Mann M. More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS. J Proteome Res. 2011;10:1785–93.
6. List M, Alcaraz N, Dissing-Hansen M, Ditzel HJ, Mollenhauer J, Baumbach J. KeyPathwayMinerWeb: online multi-omics network enrichment. Nucleic Acids Res. 2016;44:W98–104.
7. Peng C, Li A, Wang M. Discovery of bladder Cancer-related genes using integrative heterogeneous network modeling of multi-omics data. Sci Rep. 2017;7:15639.
8. Hua J, Koes D, Kou Z. Finding motifs in protein-protein interaction networks. Proj Final Rep. 2003. www.cs.cmu.edu/~dkoes/research/prot-prot.pdf.
9. Tornow S. Functional modules by relating protein interaction networks and gene expression. Nucleic Acids Res. 2003;31:6283–9. https://doi.org/10.1093/nar/gkg838.
10. Stuart JM, Segal E, Koller D, Kim SK. A gene-coexpression network for global discovery of conserved genetic modules. Science (80-). 2003;302:249–55.
11. Trevino V, Cassese A, Nagy Z, Zhuang X, Herbert J, Antzack P, et al. A network biology approach identifies molecular cross-talk between Normal prostate epithelial and prostate carcinoma cells. PLoS Comput Biol. 2016;12(4):e1004884.
12. McGee SR, Tibiche C, Trifiro M, Wang E. Network analysis reveals a signaling regulatory loop in the PIK3CA-mutated breast Cancer predicting survival outcome. Genomics Proteomics Bioinformatics. 2017;15:121–9.
13. Wang X, Thijssen B, Yu H. Target essentiality and centrality characterize drug side effects. PLoS Comput Biol. 2013;9(7):e1003119.
14. Hartwell LH, Hopfield JJ, Leibler S, Murray AW. From molecular to modular cell biology. Nature. 1999;402:C47–52. https://doi.org/10.1038/35011540.
15. Ulitsky I, Shamir R. Identification of functional modules using network topology and high-throughput data. BMC Syst Biol. 2007;1:8.
16. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, et al. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet. 2003;34:166–76.
17. Ulitsky I, Krishnamurthy A, Karp RM, Shamir R. DEGAS: De novo discovery of dysregulated pathways in human diseases. PLoS One. 2010;5(10):e13367.
18. Alcaraz N, Pauling J, Batra R, Barbosa E, Junge A, Christensen AGL, et al. KeyPathwayMiner 4.0: condition-specific pathway analysis by combining multiple omics studies and networks with Cytoscape. BMC Syst Biol. 2014;8:99.
19. Vlaic S, Conrad T, Tokarski-Schnelle C, Gustafsson M, Dahmen U, Guthke R, et al. ModuleDiscoverer: identification of regulatory modules in protein-protein interaction networks. Sci Rep. 2018;8(1):433.
20. Van De Veerdonk FL, Gresnigt MS, Romani L, Netea MG, Latgé JP. *Aspergillus fumigatus* morphology and dynamic host interactions. Nat Rev Microbiol. 2017;15:661–74.
21. Moreno-Velásquez SD, Seidel C, Juvvadi PR, Steinbach WJ, Read ND. Caspofungin-mediated growth inhibition and paradoxical growth in *Aspergillus fumigatus* involve fungicidal hyphal tip lysis coupled with regenerative intrahyphal growth and dynamic changes in β-1,3-glucan synthase localization. Antimicrob Agents Chemother. 2017;61. https://doi.org/10.1128/AAC.00710-17.
22. Spriggs KA, Bushell M, Willis AE. Translational regulation of gene expression during conditions of cell stress. Mol Cell. 2010;40:228–37.
23. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, et al. Orchestrating high-throughput genomic analysis with Bioconductor. Nat Methods. 2015;12:115–21.
24. da Silva Ferreira ME, Kress MR, Savoldi M, Goldman MH, Härtl A, Heinekamp T, et al. The akuB KU80 mutant deficient for nonhomologous end joining is a powerful tool for analyzing pathogenicity in *Aspergillus fumigatus*. Eukaryot Cell. 2006;5:207–11.
25. Brakhage AA, Van den Brulle J. Use of reporter genes to identify recessive trans-acting mutations specifically involved in the regulation of *Aspergillus nidulans* penicillin biosynthesis genes. J Bacteriol. 1995;177:2781–8.
26. Altwasser R, Baldin C, Weber J, Guthke R, Kniemeyer O, Brakhage AA, et al. Network modeling reveals cross talk of MAP kinases during adaptation to caspofungin stress in *aspergillus fumigatus*. PLoS One. 2015;10(9):e0136932.
27. Cerqueira GC, Arnaud MB, Inglis DO, Skrzypek MS, Binkley G, Simison M, et al. The *Aspergillus* genome database: multispecies curation and incorporation of RNA-Seq data to improve structural gene annotations. Nucleic Acids Res. 2014;42:D705–10.
28. Mabey J, Anderson M, Giles P, Miller C, Attwood T, Paton N, et al. CADRE: the central *Aspergillus* data REpository. Nucleic Acids Res. 2004;1:D401–5.
29. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/ Bioconductor package biomaRt. Nat Protoc. 2009;4:1184–91.
30. Baldin C, Valiante V, Krüger T, Schafferer L, Haas H, Kniemeyer O, et al. Comparative proteomics of a tor inducible *Aspergillus fumigatus* mutant reveals involvement of the Tor kinase in iron regulation. Proteomics. 2015;15:2230–43.
31. *Aspergillus fumigatus* Af293 Sequence. www.aspergillusgenome.org/download/sequence/A_fumigatus_Af293/current/A_fumigatus_Af293_current_orf_trans_all.fasta.gz. Accessed 27 Sept 2015.
32. Jöhnk B, Bayram Ö, Abelmann A, Heinekamp T, Mattern DJ, Brakhage AA, et al. SCF ubiquitin ligase F-box protein Fbx15 controls nuclear co-repressor localization, stress response and virulence of the human pathogen *Aspergillus fumigatus*. PLoS Pathog. 2016;12:e1005899–9.
33. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 2015;43:D447–52.
34. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13:2498–504.
35. Yip AM, Horvath S. Gene network interconnectedness and the generalized topological overlap measure. BMC Bioinformatics. 2007;8:22.
36. Neuwirth E. RColorBrewer: ColorBrewer palettes. R Package version 11–2. 2014. https://CRAN.R-project.org/package=RColorBrewer.
37. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9:559.
38. Pagès H, Carlson M, Falcon S, Li N. AnnotationDbi: Annotation Database Interface. R Package version 1382. 2017.
39. Morgan M, Falcon S, Gentleman R. GSEABase: Gene set enrichment data structures and methods. R Package version 1382. 2017.
40. Carlson M. GO.db: A set of annotation maps describing the entire Gene Ontology. R Package version 341. 2017.

41. Falcon S, Gentleman R. Using GOstats to test gene lists for GO term association. Bioinformatics. 2007;23:257–8.
42. Gentleman R. Category: Category Analysis. R Package version 2421. 2017.
43. Noventa-Jordão MA, do Nascimento AM, Goldman MH, Terenzi HF, Goldman GH. Molecular characterization of ubiquitin genes from *Aspergillus nidulans*: mRNA expression on different stress and growth conditions. Biochim Biophys Acta. 2000;1490:237–44 http://www.ncbi.nlm.nih.gov/pubmed/10684969.
44. Finley D, Özkaynak E, Varshavsky A. The yeast polyubiquitin gene is essential for resistance to high temperatures, starvation, and other stresses. Cell. 1987; 48:1035–46.
45. Barrenäs F, Chavali S, Alves AC, Coin L, Jarvelin MR, Jörnsten R, et al. Highly interconnected genes in disease-specific networks are enriched for disease-associated polymorphisms. Genome Biol. 2012;13(6):R46.
46. Gustafsson M, Edström M, Gawel D, Nestor CE, Wang H, Zhang H, et al. Integrated genomic and prospective clinical studies show the importance of modular pleiotropy for disease susceptibility, diagnosis and treatment. Genome Med. 2014;6(2):17.
47. Cagas SE, Jain MR, Li H, Perlin DS. Profiling the *Aspergillus fumigatus* proteome in response to caspofungin. Antimicrob Agents Chemother. 2011;55:146–54.
48. Nie L, Wu G, Culley DE, Scholten JCM, Zhang W. Integrative analysis of transcriptomic and proteomic data: challenges, solutions and applications. Crit Rev Biotechnol. 2007;27:63–75.
49. Albrecht D, Guthke R, Brakhage AA, Kniemeyer O. Integrative analysis of the heat shock response in *Aspergillus fumigatus*. BMC Genomics. 2010;11:32.
50. Bendtsen JD, Jensen LJ, Blom N, Von Heijne G, Brunak S. Feature-based prediction of non-classical and leaderless protein secretion. Protein Eng Des Sel. 2004;17:349–56.
51. Meyer V, Damveld RA, Arentshorst M, Stahl U, Van Den Hondel CAMJJ, Ram AFJ. Survival in the presence of antifungals: genome-wide expression profiling of *aspergillus niger* in response to sublethal concentrations of caspofungin and fenpropimorph. J Biol Chem. 2007;282:32935–48.
52. Shingu-Vazquez M, Traven A. Mitochondria and fungal pathogenesis: drug tolerance, virulence, and potential for antifungal therapy. Eukaryot Cell. 2011;10:1376–83.
53. Bowman JC, Hicks PS, Kurtz MB, Rosen H, Schmatz DM, Liberator PA, et al. The antifungal echinocandin caspofungin acetate kills growing cells of *Aspergillus fumigatus* in vitro. Antimicrob Agents Chemother. 2002;46:3001–12.
54. May GS, Xue T, Kontoyiannis DP, Gustin MC. Mitogen activated protein kinases of *Aspergillus fumigatus*. Med Mycol. 2005;43(Suppl 1):S83–6.
55. Mayr A, Aigner M, Lass-Flörl C. Caspofungin: when and how? The microbiologist's view. Mycoses. 2012;55:27–35.
56. Eshwika A, Kelly J, Fallon JP, Kavanagh K. Exposure of *Aspergillus fumigatus* to caspofungin results in the release, and de novo biosynthesis, of gliotoxin. Med Mycol. 2013;51:121–7.
57. Sin N, Meng L, Wang MQW, Wen JJ, Bornmann WG, Crews CM. The anti-angiogenic agent fumagillin covalently binds and inhibits the methionine aminopeptidase, MetAP-2. Proc Natl Acad Sci. 1997;94:6099–103. https://doi.org/10.1073/pnas.94.12.6099.
58. Zbidah M, Lupescu A, Jilani K, Lang F. Stimulation of suicidal erythrocyte death by fumagillin. Basic Clin Pharmacol Toxicol. 2013;112:346–51.
59. Netzker T, Fischer J, Weber J, Mattern DJ, König CC, Valiante V, et al. Microbial communication leading to the activation of silent fungal secondary metabolite gene clusters. Front Microbiol. 2015;6:299.
60. Alfano C, Faggiano S, Pastore A. The ball and chain of Polyubiquitin structures. Trends Biochem Sci. 2016;41:371–85.
61. Leach MD, Stead DA, Argo E, Maccallum DM, Brown AJP. Molecular and proteomic analyses highlight the importance of ubiquitination for the stress resistance, metabolic adaptation, morphogenetic regulation and virulence of *Candida albicans*. Mol Microbiol. 2011;79:1574–93.
62. Lesage G, Sdicu AM, Ménard P, Shapiro J, Hussein S, Bussey H. Analysis of β-1,3-glucan assembly in *Saccharomyces cerevisiae* using a synthetic interaction network and altered sensitivity to caspofungin. Genetics. 2004; 167:35–49.
63. Skrzypek MS, Binkley J, Binkley G, Miyasato SR, Simison M, Sherlock G. The *Candida* genome database (CGD): incorporation of assembly 22, systematic identifiers and visualization of high throughput sequencing data. Nucleic Acids Res. 2017;45:D592–6.
64. Khamooshi K, Sikorski P, Sun N, Calderone R, Li D. The Rbf1, Hfl1 and Dbp4 of *Candida albicans* regulate common as well as transcription factor-specific mitochondrial and other cell activities. BMC Genomics. 2014;15:56.
65. Vizcaíno JA, Csordas A, Del-Toro N, Dianes JA, Griss J, Lavidas I, et al. 2016 update of the PRIDE database and its related tools. Nucleic Acids Res. 2016; 44:D447–56.

2.4 Manuskript 4: „Strategies of pathogenic *Candida* species to survive in human blood have evolved independently“

Status

Manuskript eingereicht in PLOS Genetics

Literaturangabe

Kämmer, P., McNamara, S., Wolf, T., Conrad, T., Hünniger, K., Kurzai, O., Guthke, R., Hube, B., Linde, J., Brunke, S. Strategies of pathogenic *Candida* species to survive in human blood have evolved independently. *Eingereicht in PLOS Genetics.*

Übersicht

In der Arbeit von Kämmer *et al.* konnte mit der Untersuchung dynamischer Zusammenhänge ein weiteres Merkmal des ModuleDiscoverers aufgezeigt werden. In dieser Studie wurden die WPPI im humanen Vollblutmodell untersucht, wobei das Vollblut mit den vier verschiedenen *Candida*-Spezies *C. albicans, C. glabrata, C. parapsilosis* und *C. tropicalis* infiziert wurde. Es sollte die Frage geklärt werden, welche Infektionsstrategien in den pathogenen Pilzen konserviert sind und welche einer evolutionären Entwicklung unterliegen. Mithilfe eines dualen RNA-Seq-Ansatzes wurden Transkriptomdaten von sowohl Wirt als auch Pathogen gleichzeitig gemessen. Am Beispiel von *C. albicans* wurde ein gesamtregulatorisches Modul erzeugt, das alle zur Verfügung stehenden Zeitreihendaten des Pilzes integriert. So war es zum einen möglich, einen zeitlich globalen Überblick auf alle stattfindenden Interaktionen zu erlangen. Zum anderen konnten die zeitbedingten Änderungen in der Aktivität biologischer Prozesse nachvollzogen werden. Die Arbeit von Kämmer *et al.* hat gezeigt, dass die angeborene menschliche Immunantwort auf Transkriptionsebene größtenteils speziesunabhängig stereotyp agiert. Im Gegensatz dazu verfolgen die einzelnen pathogenen Pilze überwiegend unterschiedliche Strategien, um ihr Überleben im Wirt zu sichern. Die Ergebnisse weisen darauf hin, dass die Anpassung der *Candida*-Spezies als Antwort auf die wirtsinitiierten Abwehrmechanismen kein phylogenetisches Merkmal ist.

Beiträge

KP, MS, JL, HB und BS konzipierten die Studie. KP und HK waren für die methodische Durchführung verantwortlich. Die Datenanalyse erfolgte durch KP, MS, WT, CT und BS. KP und MS schrieben das Manuskript. Für das Layout waren KP, MS, WT und CT zuständig. Alle Autoren waren an der Überprüfung und Überarbeitung des Manuskripts beteiligt.

PLOS Genetics

Strategies of pathogenic Candida species to survive in human blood have evolved independently

--Manuscript Draft--

Manuscript Number:	
Full Title:	Strategies of pathogenic Candida species to survive in human blood have evolved independently
Short Title:	Comparative transcriptomics in fungal blood infections
Article Type:	Research Article
Section/Category:	General
Keywords:	Pathogen evolution; Candida albicans; Candida glabrata; Candida tropicalis; Candida parapsilosis; Candida species; human fungal pathogens; dual-species RNA-seq; blood infection model; host-pathogen interaction; comparative transcriptomics
Corresponding Author:	Sascha Brunke Hans-Knöll-Institute Jena, GERMANY
Corresponding Author's Institution:	Hans-Knöll-Institute
First Author:	Philipp Kämmer
Order of Authors:	Philipp Kämmer
	Sylvie McNamara
	Thomas Wolf
	Theresia Conrad
	Kerstin Hünniger
	Oliver Kurzai
	Reinhard Guthke
	Bernhard Hube
	Jörg Linde
	Sascha Brunke
Abstract:	Only four species, Candida albicans, C. glabrata, C. parapsilosis, and C. tropicalis, together account for about 90% of all Candida bloodstream infections and are among the most common causes of invasive fungal infections of humans. However, virulence potential varies among these species, and the phylogenetic tree reveals that their pathogenicity may have emerged several times independently during evolution. We therefore tested these four species in a human blood infection model to determine, via comprehensive dual-species RNA-sequencing analyses, which fungal infection strategies are conserved and which are evolutionary recent developments. The ex vivo infection progressed from initial immune cell interactions to nearly complete killing of all fungal cells. During the course of infection, we characterized important parameters of pathogen-host interactions, like fungal survival, types of interacting immune cells, and cytokine release. On the transcriptional level, we obtained a predominantly uniform and species-independent human response governed by a strong up-regulation of pro-inflammatory processes, which was down-regulated at later time points after most fungi had been killed. In stark contrast, we observed that the different fungal species pursue predominantly individual strategies and show significantly different global transcriptome patterns. Among other findings, our functional analyses revealed that the fungal species rely on different metabolic pathways and virulence factors to survive the host-imposed stress. These data show that adaptation of Candida species as response to the host is not a phylogenetic trait, but rather has likely evolved independently as a prerequisite to cause human infections.

Strategies of pathogenic *Candida* species to survive in human blood have evolved independently

Philipp Kämmer[1¶], Sylvie McNamara[2¶], Thomas Wolf[2], Theresia Conrad[2], Kerstin Hünniger[3,4], Oliver Kurzai[3,4], Reinhard Guthke[2], Bernhard Hube[1,5,6*], Jörg Linde[2,7,8] and Sascha Brunke[1*]

[1] Department of Microbial Pathogenicity Mechanisms, Leibniz Institute for Natural Product Research and Infection Biology – Hans Knöll Institute, Jena, 07745, Germany

[2] Research Group Systems Biology and Bioinformatics, Leibniz Institute for Natural Product Research and Infection Biology – Hans Knöll Institute, Jena, 07745, Germany

[3] Research Group Fungal Septomics, Leibniz Institute for Natural Product Research and Infection Biology – Hans Knöll Institute, Jena, 07745, Germany

[4] Institute for Hygiene and Microbiology, University of Würzburg, 97080, Germany

[5] Friedrich Schiller University Jena, 07743, Germany

[6] Center for Sepsis Control and Care, Jena University Hospital, 07747, Germany

[7] Research Group PiDOMICS, Leibniz Institute for Natural Product Research and Infection Biology – Hans Knöll Institute, Jena, 07745, Germany

[8] Institute for Bacterial Infections and Zoonoses, Federal Research Institute for Animal Health – Friedrich Loeffler Institute, Jena, 07743, Germany

[¶] These authors contributed equally to this work

[*] Corresponding authors: sascha.brunke@leibniz-hki.de, bernhard.hube@leibniz-hki.de

Abstract

Only four species, *Candida albicans*, *C. glabrata*, *C. parapsilosis*, and *C. tropicalis*, together account for about 90% of all *Candida* bloodstream infections and are among the most common causes of invasive fungal infections of humans. However, virulence potential varies among these species, and the phylogenetic tree reveals that their pathogenicity may have emerged several times independently during evolution. We therefore tested these four species in a human blood infection model to determine, *via* comprehensive dual-species RNA-sequencing analyses, which fungal infection strategies are conserved and which are evolutionary recent developments. The *ex vivo* infection progressed from initial immune cell interactions to nearly complete killing of all fungal cells. During the course of infection, we characterized important parameters of pathogen-host interactions, like fungal survival, types of interacting immune cells, and cytokine release. On the transcriptional level, we obtained a predominantly uniform and species-independent human response governed by a strong up-regulation of pro-inflammatory processes, which was down-regulated at later time points after most fungi had been killed. In stark contrast, we observed that the different fungal species pursue predominantly individual strategies and show significantly different global transcriptome patterns. Among other findings, our functional analyses revealed that the fungal species rely on different metabolic pathways and virulence factors to survive the host-imposed stress. These data show that adaptation of *Candida* species as response to the host is not a phylogenetic trait, but rather has likely evolved independently as a prerequisite to cause human infections.

Author summary

To ensure their survival, pathogens have to adapt immediately to new environments in their hosts, for example during the transition from the gut to the bloodstream. Here we investigate the basis of this adaptation in a group of fungi which are among the most common causes of hospital-acquired infections, the *Candida* species. In a human full blood infection model we studied which genes and processes are active over the course of an infection in both, the host and four different *Candida* pathogens. To our surprise we found that, while the human host response is predominantly uniform, the pathogens pursue largely individual strategies and each regulate genes involved in largely disparate processes in the blood. Since the host environment is very similar for each fungus, our results reveal that *C. albicans*, *C. glabrata*, *C. parapsilosis*, and *C. tropicalis* all have developed individual strategies in the host. This indicates that their pathogenicity in humans has evolved several times independently, and that genes which are central for survival in the host for one species may be irrelevant in another.

Introduction

An inadequate and dysregulated systemic inflammatory immune response to microbial bloodstream infections can lead to severe organ dysfunctions, also termed as sepsis [1], which is a major public health concern with high mortality rates [2, 3]. While mostly bacterial in origin, fungi are also known to be a major cause of sepsis, and among them *Candida* species are most common [4]. However, of more than 150 different known *Candida* species, only four account for at least 90% of all bloodstream infections, namely *C. albicans*, *C. glabrata*, *C. parapsilosis*, and *C. tropicalis* [5-7].

With the exception of *C. glabrata*, which is more closely related to the baker's yeast *Saccharomyces cerevisiae*, these species can all be found in the so-called CTG clade of fungi, sharing a unique difference in codon translation. This hints towards an at least partially shared basis of their pathogenicity, mediated by strategies which have evolved at the root of this branch of the phylogenetic tree. However, non-pathogenic species are present interspersed with the pathogens [8, 9], which rather suggests independent evolutionary origins of their pathogenicity. Comparative genomic analyses suggest that certain lineages are predisposed by previous adaptations to develop pathogenicity, but it is unclear so far whether these species then follow parallel strategies in the host [10].

Common to the pathogenic *Candida* species is that they are usually commensals which colonize human skin or intestinal and oral mucosal surfaces without causing clinical symptoms. In patients with immunodeficiencies or damaged anatomical barriers, dissemination into the bloodstream can occur from these sites or from biofilms on medical devices [11, 12]. Some clinical differences exist among the infecting species: While *C. glabrata* has a high incidence in elder individuals, *C. parapsilosis* causes high mortality rates in low-birth-weight neonates [13, 14]. *C. tropicalis* is more often associated with neutropenia or malignancy and more frequently isolated in some regions of Asia than other *Candida* species [15, 16]. *C. albicans* remains the most prevalent cause of invasive candidiasis, but the frequency of other *Candida* species has increased to about 50% [11, 17,

18]. Notably, non-*albicans Candida* species, in particular *C. glabrata*, are often more resistant against commonly used antifungals [19, 20] and thus require different antifungal therapies. However, due to the lack of differentiable clinical symptoms and slow diagnostic tools, the discrimination between infections caused by different *Candida* species remains a challenge. Detecting distinct differences in the host or fungal response to infection may therefore hold the potential to find specific biomarkers for a quick determination of the infecting species.

Upon entering the bloodstream, *Candida* cells face an entirely new environment characterized by harsh conditions: Access to nutrients is strictly limited and the innate immune system combats invading pathogens immediately. The complement system enables opsonization and promotes phagocytosis [21]. Monocytes and particularly neutrophils act as the first line of cellular defense to clear fungi from the bloodstream [22]. Consequently, neutropenia is associated with poor prognosis in candidemic patients [23] and neutrophils have been shown to be pivotal in governing the transcriptional response of *C. albicans* and eliminating this fungus in human blood [24, 25]. It is currently unclear to which extent these observations are specific for *C. albicans* or also apply to other pathogenic *Candida* species.

Models to investigate clinical events in the laboratory have various limitations. *In vitro* infection models of primary immune cells have identified important virulence factors of *Candida* species [26-31], but lack in the complex interplay among the different components of the immune response. Animal *in vivo* models, mainly mice, give a better understanding of the onset and progress of disseminated candidiasis [28, 32-39], but most *Candida* species are not normal commensals or pathogens of these model hosts [40], and the animals' immune system differs in important aspects from humans [reviewed in 41]. Using human whole blood *ex vivo* can overcome some of these limitations [42]: Our own previous studies explored the transcriptional responses of *C. albicans* or the host during blood infections [24, 43-45], or characterized the interplay of innate immune cells and molecular blood components with *C. albicans* or *C. glabrata* [25, 46-48]. Further multiple studies explored

various aspects of *Candida*-phagocyte interactions using isolated immune cells [for example, reviewed in 22, 49, 50-52].

Here, we employ a complex time-resolved *ex vivo* whole blood infection model, which mimics the early dissemination stage of candidemia [24, 25], to (i) characterize molecular and cellular events during infection and to (ii) investigate the interdependent transcriptional patterns of both, the human host and the most common *Candida* species, *C. albicans*, *C. glabrata*, *C. parapsilosis* and *C. tropicalis* by a dual-species RNA-sequencing (RNA-seq) approach. With the exception of some distinct species-specific fungal association with innate immune cells, we show that the human host responds predominantly uniformly to the challenge with *Candida* spp. A strong pro-inflammatory cytokine response is accompanied by highly similar transcriptional changes upon infection with any of the four *Candida* species. In contrast, the fungal responses are dominated by species-specific adaptations, indicating that their pathogenicity has evolved independently. Based on these data sets, we determined interspecies regulatory networks of *C. albicans* (as the predominant cause of candidemia) and its host to predict new species-specific molecular markers of host-pathogen interactions.

Results

Mimicking main features of human *Candida* bloodstream infections *ex vivo*

Dissemination *via* the bloodstream is a hallmark of invasive *Candida* infection [46, 53]. We applied an improved *ex vivo* blood infection model [described in 25] to simulate early dissemination stages of *Candida* species, using an up to fifty times less infection dose (10^6 cells/ml) than previous studies [24, 43, 44]. That way we mimicked more closely the clinical situation and were able to characterize fungal killing kinetics, interactions with blood cells and immunological parameters. As *C. parapsilosis* and *C. tropicalis* have never been characterized in this model, we compared fungal killing kinetics, interactions with blood cells and immunological parameters of these species with *C. albicans* and *C. glabrata*.

Within 30 minutes post infection (mpi), a substantial part of the fungal population of all species had been killed, demonstrating the efficiency of the antifungal activity of blood (Figure 1A). Within 60 mpi, about 80% of fungal cells were killed, with the notable exception of *C. albicans* (57.3% killing). This trend continued for up to four hours post infection with a significantly higher proportion of surviving *C. albicans* (19.1%) than *C. tropicalis*, *C. parapsilosis* or *C. glabrata* cells (5.2, 1.7 or 2.7%, respectively). Therefore, the immune response is able to largely clear the infecting pathogens in our *ex vivo* whole blood infection model.

Immediately after entering the bloodstream, invading pathogens face cells of the innate immune system that contribute to fungal killing. We characterized these interactions using FITC-labeled *Candida* yeasts and immunofluorescence-stained immune cells. *Candida* cells interacted rapidly with leukocytes in human blood: 60 mpi (Figure 1B) the vast majority of fungal cells was in contact predominantly with neutrophils (45.1 – 73.1%), and to a much lower extent with monocytes (3.1 – 9.5%). We observed some species-specific differences in these associations, both at the early and late time points: *C. albicans* was more frequently associated with neutrophils at 240 mpi (80% vs. 50 – 60% for the other species). At the same

156 time, we found fewer *C. albicans* cells which were not associated with any immune cell
157 based on the lack of concurrent FITC- and immune cell staining (14.1% vs. 29.7 – 38%).
158 *C. glabrata* interacted much more avidly with monocytes than the other species starting at
159 60 mpi (9.6% vs. 3.1 – 3.9%), which is in line with previous results [46, 48]. At 240 mpi,
160 *C. parapsilosis* was also more frequently associated with monocytes than *C. albicans* and
161 *C. tropicalis* (7.5% vs. 3.4 – 3.7%). The flow cytometry-based data were qualitatively
162 validated by blood smears. For each fungal species, contact to blood cells was observed
163 microscopically (Figure S 1A), and only *C. albicans* formed its typical filaments starting at
164 60 mpi.

165 As *Candida* cells interacted predominantly with neutrophils in blood, we determined
166 neutrophil activation *via* presence of specific surface markers. Surface levels of the general
167 early activation marker CD69 (AIM) were slightly elevated 240 mpi in presence of all four
168 *Candida* species, with *C. glabrata* inducing the largest increase (Figure S 1B). CD11b
169 (CR3A/ITGAM), mediating leukocyte adhesion, and the degranulation marker CD66b
170 (CEACAM8) were also increased, compared to mock infection, with significantly elevated
171 CD66b upon *C. albicans* infection. In contrast, CD16 (FcγRIII) decreased noticeably under all
172 four fungal infection regimes. Together these results demonstrate significant neutrophil
173 activation upon *Candida* blood infection by all four species.

174 **The human transcriptional response is mainly species-independent**

175 Having characterized the overall interaction pattern of Candida species and blood-borne
176 immune cells and having identified some differences, we next monitored the global
177 transcriptional response of host and fungal cells in a kinetic, dual-species RNA-seq
178 approach. In total, we acquired 1.3 to 2.3 billion high-quality mixed human and fungal raw
179 reads, and another 225 million from three mock-infected blood samples, to reach 14- to 25-
180 fold transcriptome coverage for fungi and host, respectively (Table S 1).

181 A first global overview of the human samples by principle component analyses (PCA)
182 showed that the time point post-infection rather than the infecting species governs

183 transcriptional variance, which clearly differed from the non-infected control samples (Figure
184 2A). This human host response is characterized initially by a small number of regulated
185 genes that rapidly increased over time from only 30 – 50 at 15 mpi for all species to a
186 maximum of 1,940 differentially expressed genes (DEGs) at 240 mpi during *C. parapsilosis*
187 infection (Figure 2B). For *C. albicans*, *C. tropicalis*, and *C. glabrata* similar maxima of 1,513,
188 1,599, and 1,609 DEGs were observed at 240 mpi, respectively (Table S 2).

189 By comparing host transcriptional changes over all time points (Figure 2C) we found a
190 common core response to fungal infections of about 670 up- and 490 down-regulated genes;
191 here called “quadruple” genes, differentially regulated at least once in all four infection
192 kinetics. There are smaller numbers of genes which are only differentially expressed for one
193 *Candida* species compared to the respective infection control for each species, from a
194 maximum of 381 DEGs down-regulated for *C. parapsilosis* to a minimum of 9 DEGs down-
195 regulated for *C. glabrata* infection. In summary, the human transcriptional response to
196 infecting *Candida* species is predominantly uniform, with few detectable unique regulations.

197

198 **Immune system processes govern the human transcriptional response**

199 We went on to characterize the human core response. Functional gene ontology (GO) term
200 analyses showed that this response is governed by innate immune system processes: In
201 particular genes involved in inflammatory responses, cytokine-mediated signaling, and
202 chemotaxis were significantly up-regulated 240 mpi (Figure 3A). In apparent contradiction,
203 genes associated with other immune processes like endocytosis and Toll-like receptor
204 signaling pathway were down-regulated 240 mpi. This likely indicates a shift from pro- to anti-
205 inflammatory processes when most fungal cells were killed, dampening the immune
206 response.

207 As we found mainly immune system processes differentially expressed in response to all four
208 *Candida* infections, we thus investigated the regulation of immunoregulatory genes of the
209 human core response in detail (Figure S 2, Table S 3). The majority of these genes were

uniformly regulated in response to different *Candida* species, including an immediate up-regulation of major pro-inflammatory cytokine- and chemokine-encoding genes like *IL1B*, *IL6*, *CXCL8*, and *TNF*, which increased up to 2,000-fold. Among the pattern recognition receptor genes (PRR), galectin-3 (*LGALS3,* recognizing β-mannan of *Candida* spp.) transcription was up-regulated in response to all species. Except for *CLEC5A*, several C-type lectin family member genes were down-regulated (*e.g. CLEC10A*, *CLEC11A*, *CLEC12A*), as were the Toll-like receptor genes *TLR1*, *TLR5*, *TRL6*, *TLR7*, *TLR8*, and *TLR10*. The gene coding for TLR2, known to be critical for immune responses during candidiasis [54], was predominantly up-regulated in response to *C. glabrata*, *C. parapsilosis*, and *C. tropicalis* (and to a lesser extend in response to *C. albicans*). Thus, the human transcriptional response was found to be mainly uniform to early *Candida* blood infections.

As we detected an up-regulation of major pro-inflammatory cytokine-encoding genes upon *Candida* blood infections (Figure S 2), we measured plasma cytokine levels at 240 mpi. Pro-inflammatory IL-1β, IL-6 and TNF-α were markedly increased upon any infection compared to mock control and higher during *C. glabrata* or *C. parapsilosis* infections (Figure 3B) as compared to *C. albicans* and *C. tropicalis* infections. Since this triad of cytokines is mainly released by monocytes [55], we propose that this may reflect their higher association rates with *C. glabrata* and *C. parapsilosis*. Of note, *C. glabrata* induced lower plasma levels of IL-8, a potent chemoattractant for neutrophils, than any of the other three species (Figure 3B).

In summary, similar immune system activation by different *Candida* species was detected on several levels in the whole blood infection model: Immune cells, mainly neutrophils and monocytes, associated rapidly with *Candida* cells, which were efficiently killed during the course of infection. Concurrently, surface activation markers and pro-inflammatory cytokines were up-regulated. This was accompanied by a predominantly uniform response on the transcriptional level, governed by processes of the innate immune system.

The few commonly regulated fungal pathways are highly conserved

Concurrently to the host approach, we went on to analyze the transcriptomes of *C. albicans*, *C. glabrata*, *C. parapsilosis*, and *C. tropicalis* during blood infection. Given the similar host response and the relationship of the fungal species, we expected that the fungi employ comparable survival strategies in blood. However, we found significant differences: In contrast to their host, *Candida* species, except *C. glabrata*, regulated a significant fraction of their transcriptome already 15 mpi (Table S 4). This response was robust during the whole course of infection: For *C. albicans*, *C. tropicalis*, and *C. parapsilosis* 38.6% (2427 of 7074), 35.7% (2236 of 6258), and 46.3% (2759 of 5984) of their genetic repertoire was differently regulated compared to the pre-culture at least at one time point, respectively. In stark contrast, only 10% of *C. glabrata* genes (541 of 5566) were differentially expressed at any time point during infection. Moreover, the direction of regulation differed significantly: While the majority of genes was down-regulated in *C. albicans*, *C. glabrata* and *C. parapsilosis*, most transcripts were up-regulated in *C. tropicalis* (Figure 4A). For example, 60 mpi 57% of *C. tropicalis* genes were up-regulated, compared to only 13.4 – 34.3% in the other species.

Although the genomes of the four *Candida* species share among them more than 3,500 orthologs, surprisingly, only 189 of these were commonly regulated (Figure 4B) at any time. Interestingly, the transcriptional variance of this fungal core response was neither determined by the time point post-infection nor the infecting species, as indicated by PCA (Figure 4C). To characterize this conserved regulation, we performed GO term analyses (Figure 4D). A key feature of this fungal core response is an extensive shutdown of protein biosynthesis and related processes like rRNA processing, translation initiation, and purine biosynthesis. The glycolytic genes *ENO1*, *HXK2*, and *PFK1* were likewise universally down-regulated, as were genes associated with fatty acid synthesis (*FAS1*, *FAS2*), indicating a metabolic rearrangement. In contrast, genes of the general stress response, e.g. coding for heat shock proteins (*HSP78*, *HSP104*), were commonly up-regulated. Likewise, all four species also increased the expression of genes encoding hydrolytic enzymes such as secreted aspartyl proteinase (*SAP*) genes and their orthologs, which have been linked to several aspects of

Candida pathogenicity [56, 57]. We consider these responses, based on similar regulations of orthologous genes, to be a common evolutionary trait which preceded and likely helped to enable the appearance of individual pathogenicity in the different *Candida* species.

Candida species pursue custom tailored strategies to survive in blood

As the fungal transcriptome adaptation was found to be surprisingly individual for each *Candida* species, we went on to characterize these individual responses in more detail. We aimed to estimate whether the overall strategies for survival in the host, based on functional annotations of the regulated genes, differed significantly between the species – which would indicate independent evolutionary adaptations. Using the well-annotated *C. albicans* genome as a model, we first obtained a regulatory module to demonstrate the kinetics in the fungal response. Regulatory modules are sub-networks of protein-protein interaction networks comprising sets of co-expressed genes sharing a common function [58]. *Via* GO term analyses of clusters within the regulatory module containing strongly connected network components (Figure 5, Table S 5), we were thus able to characterize *C. albicans*' overall response to the host. Using this as a template, we compared the responses of all species based on orthologous genes.

A first hallmark of *C. albicans* response was an immediate (within the first 15–30 minutes) and stable up-regulation of the glyoxylate cycle (*ICL1* and *MLS1*) and fermentative energy production (*ADH2* and *ALD6*) (Figure 5, cluster 3), indicating fast glucose restriction and alternative carbon source utilization upon phagocytosis by neutrophils or monocytes. Both *C. tropicalis* and *C. parapsilosis* responded similarly and furthermore strongly up-regulated genes involved in β-oxidation (*POX1-3*, *PXP2*, *FOX2*, *FOX3*, *POT1*, and *ECI1*) as an immediate reaction. However, *C. parapsilosis* significantly down-regulated the glyoxylate cycle in the later phase and strikingly, *C. glabrata* did not react with nutrient acquisition markers in the early phase, but rather down-regulated transporters for carbohydrates, amino acids, and ammonium.

290 Similar to the carbon starvation response, *C. albicans* up-regulated the expression of several
291 amino acids biosynthesis genes early during infection, including glutamate and branched-
292 chain amino acids (cluster 11), and its arginine biosynthesis genes were constantly highly
293 expressed until 240 mpi (cluster 12). Interestingly, both patterns were, at least partially, also
294 found for *C. tropicalis* and *C. parapsilosis*, but not for *C. glabrata*.

295 Like carbon, metals are strictly limited in the host, and efficient metal acquisition is
296 recognized as a key pathogenic trait [59]. We found that, to a large extent, the iron-related
297 response of *C. tropicalis* is similar to that of the well-studied [60] *C. albicans* (Figure 6A):
298 Hemoglobin uptake (*PGA7*, *RBT5*), the ferric reductase genes *FRP1* and *FRP2* and vacuolar
299 iron transport (*FTH1*) were up-regulated, while the multicopper ferroxidase and iron
300 permease genes were down-regulated. Strikingly, for *C. glabrata*, none of the recently
301 described iron uptake-related genes [61] was found regulated upon blood infection (Figure
302 6A), with the only exception of *SIT1* (60 mpi: $\log_2$FC 2.43; p=0.051), again indicating different
303 survival strategies.

304 Upon engulfment by phagocytes, fungal cells are exposed to harmful reactive oxygen
305 species (ROS) to which *Candida* species have a variety of generally conserved detoxifying
306 enzymes. However, we found unique patterns in the regulation of these genes in each
307 species (Figure 6B): While *C. albicans* strongly expressed the superoxide dismutase genes
308 *SOD1*, *SOD4*, and, in particular, *SOD5*, *C. parapsilosis* and *C. tropicalis* most of all up-
309 regulated an alkyl hydroperoxide reductase (*AHP1*) and putative glutathione S-transferase
310 (*GTT12*, *GTT13*) genes. In contrast to the other fungi, *C. glabrata* exhibits a very restrained
311 oxidative stress response with only a very slightly up-regulation of *CTA1* (240 mpi: $\log_2$FC
312 1.09). Evidently, the *Candida* species evolved to use very different strategies to detoxify ROS
313 during blood infection.

314 Finally, adhesion to endothelial cells is an essential prerequisite for escape from the
315 bloodstream. Large families of adhesin genes are found in the genomes of all investigated
316 *Candida* species, and several of them were regulated during blood infection. *C. albicans* up-
317 regulated adhesins with gene-specific kinetics: While *ALS1* and *HWP2* expression

318 decreased over time, transcriptional levels of *HWP1* and *ALS3* remained almost stable at a
319 high level. Although *HWP1* is the second most highly up-regulated *C. albicans* gene during
320 blood infection (30 mpi: $\log_2$FC 12.6), its orthologs were either not regulated or even down-
321 regulated in *C. tropicalis* and *C. parapsilosis*, respectively. Remarkably, of the 67 predicted
322 genes for adhesin-like proteins in *C. glabrata* [62] only *EPA6*, *EPA7*, and *PWP1* were
323 immediately up-regulated. Thus, each fungus expresses a different subset of adhesin genes,
324 and even closely related genes differ in their regulation.

325

326 **Gene regulatory networks depict host pathogen interactions on a**
327 **transcriptional level**

328 Overall, we thus found predominantly individual fungal host adaptation mechanisms,
329 indicating that most of the survival strategies in blood evolved independently in response to
330 the host. We assume that genes involved in these specific adaptations should have a
331 significant influence on both the host response and fungal survival. Consequently, we
332 leveraged our simultaneous acquisition of both fungal and host transcriptomes over the
333 course of an infection to generate an interdependent gene regulatory network of both
334 partners which reflect their mutual transcriptional responses. We then predicted so far
335 unknown intra- and interspecies regulatory connections [33, 63] in the computationally
336 accessible subnetwork of the infection-relevant process oxidative stress response, which is
337 highly induced in *C. albicans*, but not in *C. glabrata* (Figure 6B), and thus shows hallmarks of
338 a recent adaptation processes, possibly as a result of co-evolution during host-pathogen
339 interactions.

340 As a first result our interspecies regulatory network confirmed several previously known
341 interactions (Figure 7A), and more importantly, it predicted a multitude of novel intra- and
342 interspecies regulatory interactions. For example, the up-regulation of *C. albicans SOD5* was
343 predicted to (indirectly) promote transcription of the human proinflammatory cytokines *IL6*
344 and *IL1B*, while *SOD5* itself seemed positively linked to *CAT1* expression levels and
345 enhances *GPX2* transcription in fungal cells. On the other hand, human *IL6* expression was

predicted to promote fungal *HSP21* transcription, which in turn is repressed by *SOD5*, putting the *SOD5* gene at a central position of this host-pathogen gene regulatory network.

We therefore obtained a *sod5*ΔΔ deletion mutant and quantified transcription levels of human and fungal genes by quantitative PCR (qPCR) during infection. This largely confirmed our RNA-seq results (Figure 7B). While contrary to our mathematical model the deletion of fungal *SOD5* did not change expression of human *NFKB1*, *IL1B*, and *IL6*, it did in fact result in higher transcript levels of *C. albicans GPX2* and *HSP21*. Given the predicted importance of the oxidative stress network in interactions between host and *C. albicans*, we tested the survival of *sod5*ΔΔ, *cat1*ΔΔ, and *cap1*ΔΔ, as *CAP1* being the central regulator of *C. albicans* oxidative stress response [64], deletion mutants in human blood. Already 15 mpi, *sod5*ΔΔ survival was reduced compared to the wild type (46.8 vs. 64.3%, Figure 7C) and continued to be significantly diminished at 30 and 60 mpi. Similarly, a clear trend to a fitness defect was observed between 60 and 240 mpi upon *CAT1* deletion and for all time points except 240 mpi when *CAP1* was deleted. These results support the important role of Sod5 and possibly Cat1 in infection, as indicated by our transcriptome models. Accordingly, although catalases are well-conserved among fungal species, the lack of strong up-regulation of *CTA1* in *C. glabrata* (Figure 6B) was reflected in the survival of a *C. glabrata cta1*Δ strain in blood: In contrast to *C. albicans cat1*ΔΔ, *cta1*Δ showed no discernable fitness defect during the whole course of infection (Figure 7D). All in all, these data therefore support our transcriptome-based notion that the fungal pathogens, although equipped with a partially conserved genetic repertoire, pursue unique strategies to survive in blood that are largely divergent for each fungus.

Discussion

We aimed to investigate the strategies employed by different pathogenic *Candida* in a central step of a disseminating infection, and determine whether these are evolutionary conserved, have evolved analogously along trajectories dictated by common ancestral adaptations, or evolved completely independently. To this end, we applied an *ex vivo* whole blood infection model to simulate a key step of dissemination.

Upon entering the bloodstream, *Candida* cells face a new and hostile environment: Nutrients are restricted, pH and other physical factors change, and arguably most importantly, the host immune system starts to combat the invaders. In this study we have used a whole blood infection model developed and refined in our laboratories [24, 25] for a global comparative transcriptional analysis of the four most common pathogenic *Candida* species causing life threatening bloodstream infections. The model has previously been used, among others, to confirm the central role of neutrophils in the immune response against *C. albicans*, to describe the complement component C5a as a central player in *C. albicans* blood infections and to dissect differences in host cell association of *C. albicans* versus *C. glabrata* [24, 25, 44, 46, 47]. Data on immune cell interactions, cytokine release, fungal survival rates, and kinetics of the mutual host and *Candida* spp. transcriptional responses obtained in this study revealed an unexpected level of unique regulation on the fungal side facing a mostly uniform host response.

In detail, previous studies showed fast and efficient killing of *C. albicans* and *C. glabrata* [24, 25, 48] which we here extended systematically to include *C. parapsilosis*, and *C. tropicalis*. These showed an even lower overall survival than *C. albicans*, indicating that the latter is slightly better adapted to survive in human blood. We also detected significant differences in the types of immune cells which were associated with fungal cells in the early infection phase. Although the cellular immune response to all *Candida* spp. was dominated by neutrophils, monocytes were associated with a measurable proportion of *C. glabrata* and *C. parapsilosis* cells – much more than with *C. albicans* and *C. tropicalis*. These *ex vivo*

findings confirmed and corroborated earlier *in vitro* and *in vivo* studies where *C. glabrata* attracted monocytes more strongly than neutrophils and was more efficiently phagocytosed than *C. albicans* [46, 48]. Similarly, a higher rate of macrophage migration towards *C. parapsilosis* (compared to *C. albicans*) as well as intracellular replication of *C. parapsilosis* was shown by Tóth *et al.* [65]. Consequently, it has been suggested that phagocytosis by and survival in monocytes is a fungal-driven mechanism of *C. glabrata* [26, 56, 66, 67] and *C. parapsilosis* [65] to evade immune surveillance. This model would predict an early high association rate to blood monocytes, which is supported by our *ex vivo* data.

Infection of human blood with any *Candida* species led to the release of the chemokine IL-8 and pro-inflammatory cytokines (IL-1β, IL-6, TNF-α). This triad of cytokines is mainly produced by blood monocytes and is crucial for driving the acute phase response to pathogens [55]. *C. glabrata* and *C. parapsilosis* induced more IL-1β, IL-6, and TNF-α, suggesting stronger activation of monocytes, in agreement with their higher association rates. The higher release of the neutrophil attractant IL-8 during *C. albicans* infection confirms previous work [46] and reflects our observed frequent association of *C. albicans* with neutrophils.

On the transcriptional level, the host responded slowly and with steadily increasing transcriptional changes to all *Candida* infections. We showed that – with very few exceptions – the response is time- rather than species-dependent. This indicates the recognition of a common pattern, leading to a uniform response by the immune system in the early phase of any *Candida* blood infection. We detected mainly functional categories of innate immunity to be significantly regulated in our transcriptional data, like inflammatory response or regulation of cytokine secretion. In the short term at least, the transcriptional immune reaction is thus largely independent of the infecting *Candida* species. Pro-inflammatory cytokine genes like *IL6* and *TNF* and chemokine genes like *CCL20* were among the most up-regulated genes. A previous study found the same genes up-regulated in infections with species as diverse as *C. albicans*, *Aspergillus fumigatus*, *Escherichia coli* and *Staphylococcus aureus* [45]. Genes that were specific for infections with the two fungi in that study, like *FOSB* and *TBC1D7*, were

424 similarly regulated in our experimental setting with all *Candida* species. This supports their
425 potential role as general immune response markers for fungal infections. On the host side,
426 the *ex vivo* whole blood infection model therefore mimics vital characteristics of an early
427 *Candida* bloodstream infection: Rapid association of immune with fungal cells trigger efficient
428 *Candida* killing and pro-inflammatory cytokine release, which does not require immediate and
429 major changes in the transcriptional response [25]. Most importantly, we found that the
430 restrained transcriptional response to infection with different *Candida* species follows a
431 strikingly uniform program – despite measurable differences in physical immune cell
432 interactions and a severe divergence of the fungal transcriptome kinetics.

433 Fradin *et al*. were the first to interrogate the fungal transcriptional response to human blood
434 in a *C. albicans* infection [43]. By using a refined blood infection model [25], we looked
435 beyond *C. albicans* to determine whether *Candida* species follow evolutionary conserved
436 transcriptional patterns or pursue different strategies to survive in blood. All species, except
437 *C. glabrata*, showed an immediate (15 mpi) and highly divergent regulation of a substantial
438 subset of their genomes. We found a smaller core response of commonly regulated
439 orthologs, comprised mainly of translational shutdown, up-regulation of extracellular
440 hydrolytic enzymes and onset of a general heat shock response, which altogether we
441 consider an evolutionary older response preceding and enabling the development of
442 individual pathogenicity programs. The translational shutdown especially is likely a response
443 to the nutrient limitation in blood, and corroborates earlier studies of *C. albicans* blood [24]
444 and macrophage [68] infections, *C. glabrata* infection of macrophages [56] as well as
445 *A. fumigatus* blood infection [69]. This indicates that down-regulation of translation in the
446 early infection phase is a common principle in *Candida* or even fungal pathogenesis.

447 Interestingly, most of the fungal transcriptional regulations were diverse between or even
448 unique to one of the *Candida* species, which concerns almost all aspects of fungal
449 adaptation to the host environment, from the use of alternative energy sources to the
450 expression of pathogenicity mechanisms. For instance, while glycolytic enzymes were
451 commonly down-regulated, the alternative glyoxylate cycle was up-regulated in three of the

four tested species early upon infection with the exception of *C. glabrata*, which increased expression of respective genes only in the later phase. The glyoxylate cycle has been suggested as a potential drug target due to its ubiquitous up-regulation in microbial infections [70-74]. For some fungi, fatty acids may serve as energy and – *via* the glyoxylate cycle – carbon source during infection, as indicated by the up-regulation of genes for β-oxidation, lipases and carnitine transport most prominently by *C. parapsilosis* and, to a somewhat lesser extent, by *C. tropicalis*. The carbon metabolic response of these fungi is therefore quite distinct from *C. albicans*, where we did not observe strong induction of fatty acid catabolism, and especially from *C. glabrata*. Similar observation can be made with other central processes:

Trace metal access *via* sophisticated metal uptake systems plays a major role for the outcome of fungal infections [reviewed in 59]. We therefore expected a comprehensive metal deprivation response, and indeed observed a strong iron uptake response in *C. albicans* and *C. tropicalis*, indicated by the up-regulation of hemoglobin uptake systems [75] and reductive iron transport [60], but no such response in *C. parapsilosis* nor *C. glabrata*: While *C. parapsilosis* seems to rely on vacuolar iron storage, *C. glabrata* regulated none of the known iron homeostatic genes [60, 61, 76]. Thus, the strategies of *Candida* spp. to overcome iron limitation are highly diverse with few overlaps, and *C. parapsilosis* and *C. glabrata* do not seem to require extracellular iron uptake during early blood infection – or use mechanisms that are unknown yet.

The expression of adhesins enables attachment to the blood vessel endothelium. In agreement with previous findings, we found strong and rapid induction of adhesin gene families in *C. albicans* [43] known to be involved in endothelial cell adherence [reviewed in 77, 78]. Among the members of the gene families in *C. albicans*, *C. tropicalis* and *C. parapsilosis*, we found clearly different expression patterns, in accordance with the high genetic variability of *ALS* and *IFF*/*HYR* gene families within the CTG clade members [79]. The *C. glabrata* genome contains a repertoire of unrelated adhesion-mediating genes, comprising mainly the large *EPA* gene family [62, 78]. Indeed, *EPA6* and *EPA7*, known to

480 mediate adherence to endothelial cells [80], were early up-regulated in our model. These
481 species-specific expression patterns of adhesin genes indicate that each *Candida* species
482 follows the same strategy of adhesion, but acquired its adhesion capability to host cells
483 independently.

484 The presence of such diverse solutions to a similar host environment led us to conclude that
485 based on a common core response, the individual realization of pathogenesis has evolved
486 mostly independently in the four *Candida* species and was not necessarily following the
487 same evolutionary trajectories. We thus targeted a functional module with species-specific
488 adaptations, the oxidative stress response, where only *C. albicans* seems to rely mainly on
489 superoxide dismutases.

490 Our dual-species approach allowed us to leverage the transcriptome data and create
491 interspecies regulatory networks to reveal new intra- and interspecies interactions. In the
492 oxidative stress response of *C. albicans* during infection *SOD5* inhabited a central position in
493 the network. While its deletion caused no change in human gene expression, it still led to
494 increased expression of its predicted fungal network targets *GPX2* and *HSP21*, suggesting a
495 compensatory role of these genes upon *SOD5* deletion, in agreement with previously
496 ascribed functions [68, 81]. Importantly, *SOD5* as well as *CAT1* deletion was detrimental to
497 *C. albicans* survival in blood, while deletion of the catalase gene *CTA1* did not have an effect
498 on *C. glabrata* survival, as predicted by its lack of up-regulation in our model. The same
499 functional units therefore had different effects in *Candida* bloodstream infections, supporting
500 our notion of independently evolved strategies.

501 Taken together, we created a comprehensive dataset of *Candida* blood infections showing
502 that the human transcriptome, governed by an innate immune system response, is largely
503 species-independent and highly stereotypical. In stark contrast, strategies of different
504 *Candida* species of different evolutionary relatedness differ strongly when facing human
505 blood. As indicated by the presence of interspersed non-pathogenic species in the
506 phylogenetic tree [10], the investigated *Candida* species have evidently independently
507 evolved strategies to survive in the harsh blood environment. In addition, we found

indications of a smaller common set of reactions, including stress and metabolic responses, which may have enabled the fungi to evolve their independent strategies by allowing basic survival in the host, including as commensals.

These findings also have several important consequences. For example, while it will be difficult to identify fungal gene products as general biomarkers for fungal bloodstream infections, it is more likely that species-specific fungal markers and general host biomarkers for fungal infections can be identified. Our data further suggest that the use of *C. albicans* as the model organism for *Candida* virulence can lead to inaccurate concepts of pathogenicity. This is, for example, demonstrated by *C. glabrata* with its very limited transcriptional response. As all four pathogens are major causes of candidemia, our concept of fungal virulence in general, even within the *Candida* species, likely needs to change even more toward the concept of multiple virulence strategies.

Methods

Ethics approval and consent to participate

Human peripheral blood was collected from healthy volunteers with written informed consent. This study was conducted according to the principles expressed in the Declaration of Helsinki. The blood donation protocol and use of blood for this study were approved by the institutional ethics committee of the University Hospital Jena (permission number 2207-01/08).

Strains and culture conditions

C. albicans SC5314, *C. glabrata* ATCC2001, *C. tropicalis* DSM 4959 and *C. parapsilosis* GA1 strains (Table 1) were maintained as glycerol stocks and re-streaked on YPD agar plates. For experiments, single colonies were grown overnight in YPD at 30 °C and reinoculated in fresh YPD at 30 °C to reach mid-log phase.

Table 1. Strains used in the study.

Strain	Description	Reference
C. albicans	*C. albicans* WT strain SC5314	[105]
C. glabrata	*C. glabrata* WT strain ATCC2001	[106]
C. parapsilosis	*C. parapsilosis* WT strain GA1	[107]
C. tropicalis	*C. tropicalis* WT strain DSM 4959	[108]
*cat1*ΔΔ	SC5314, ura3::imm434/ura3::imm434 his1::hisG/his1::hisG cat1::loxP-URA3-loxP/cat1::HIS1	A. Brown (unpublished)
*cap1*ΔΔ	SC5314, Δcap1 / Δcap1	A. Brown (unpublished)
*sod5*ΔΔ	SC5314, Δsod5::hisG / Δsod5::hisG	[24]
*cta1*Δ	ATCC2001, CAGL0K10868g::NAT1	[109]

Whole blood infection model

Cells of respective strains were harvested in 1x PBS (phosphate buffered saline) and diluted in an appropriate concentration. Human whole blood was freshly drawn from healthy volunteers and anticoagulated with recombinant Hirudin (Sarstedt, Nuremberg Germany). Immediately, yeast cells were added at a concentration of 1×10^6 cells per ml blood and further incubated at 37 °C as indicated. For mock infection samples 1× PBS was used.

Flow cytometry of immune cell interaction and activation

C. albicans, *C. glabrata*, *C. tropicalis*, and *C. parapsilosis* strains were grown as previously described (strains and culture conditions). Aliquots were stained with FITC (fluorescein isothiocyanate), added at a concentration of 1×10^6 cells per ml blood and incubated at 37 °C as indicated. To distinguish different immune cell populations, whole blood was stained with mouse anti-human CD3-PerCP (clone SK7, T cells), CD19-PE (clone HIB19, B cells), CD45-PE-Cy7 (clone HI30, leukocytes), CD56-V450 (clone B159, NK cells) and CD66b-PE (clone G10F5, PMN) obtained from BioLegend®. Monocytes were stained with mouse anti-human CD14-PerCP (clone 47-3D6) from Abcam. Stained samples were treated with FACS Lysing Solution (BD), washed and acquired immediately. For raw data analysis FlowJo v10.0.8 software was used. The presence of activation markers was determined with mouse anti-human CD11b-V450 (clone ICRF44) from BD and CD16-BV510 (clone 3G8), CD69-APC (clone FN50) from BioLegend®. Stained samples were treated with FACS Lysing Solution (BD), washed and acquired immediately. For raw data analysis FlowJo v10.0.8 software was used.

Blood smears

Blood smears of *C. albicans*-, *C. glabrata*-, *C. tropicalis*- and *C. parapsilosis*-infected samples were prepared at indicated time points and stained with May-Grünwald-Giemsa staining, dried and microscopically visualized.

RNA isolation

At indicated time points infected blood samples were split into aliquots for separated fungal and human RNA isolations. For mock infections aliquots were used for human RNA isolation at 240 mpi only. To isolate human RNA aliquots were added to a PAXgene® Blood RNA Tube (PreAnalytiX) and processed with the PAXgene® Blood RNA Kit (PreAnalytiX) according to the manufacturer's protocol. For fungal RNA isolation aliquots were added to ice-cold water, centrifuged and immediately frozen in liquid nitrogen. The cell pellet was further processed with the RiboPure™-Yeast Kit (Thermo Fisher Scientific) according to the manufacturer's protocol. RNA quantity was determined with NanoDrop 1000

569 Spectrophotometer (Thermo Fisher Scientific) and RNA quality was verified with an Agilent
570 2100 Bioanalyzer (Agilent Technologies). Fungal and human RNA samples were pooled
571 subsequently in a quantitative ratio of 1:10. All samples were prepared in three biological
572 replicates with independent donors at independent time points.

573 **RNA sequencing**

574 Library preparation and RNA sequencing was carried out at GATC Biotech (Konstanz,
575 Germany). After poly(A) filtering, mRNA was fragmented and cDNA libraries were generated
576 for each sample. 50 bp single sequence reads were produced using the Illumina HiSeq 2500
577 platform.

578 **RNA-seq data preprocessing**

579 Single-end, 50 bp Illumina HiSeq 2500 raw reads were quality trimmed with Trimmomatic
580 v0.32 [82]. The *H. sapiens* genome GRCh38 and annotation were downloaded from the
581 ENSEMBL database [83]. *C. albicans* SC5314 assembly 22, *C. glabrata* CBS138,
582 *C. parapsilosis* CDC317 and *C. tropicalis* MYA-3404 genomes and corresponding genome
583 annotations were downloaded from the Candida Genome Database (CGD, [84]. For
584 *C. albicans* transcriptionally active regions identified by RNAseq [85] were added to the
585 annotation. All sequencing reads were mapped against concatenated genomes of *H. sapiens*
586 and one out of four *Candida* species using TopHat v2.1.0 [86]. Read mapping was carried
587 out and only uniquely aligned hits were kept for further analysis. Transcriptome coverage
588 was calculated as mapped reads multiplied by read length and divided by transcriptome
589 length. featureCounts v1.4.3 [87] was applied to count the number of reads within annotated
590 genes. Human and pathogen genes were tested individually for significant differential
591 expression. DESeq2 [88] was used to calculate adjusted p-values based on count values.
592 Mean RPKM and $\log_2$FC values were calculated manually. Afterwards the following cutoffs
593 were applied: adjusted p-value ≤ 0.01, abs($\log_2$FC) ≥ 1.5 and RPKM ≥ 1 on at least one time
594 point. The RNA-seq dataset generated and analyzed during the current study has been
595 deposited in NCBI's Gene Expression Omnibus [89]. The accession number will be given as
596 soon as received from NCBI.

Expression data analyses

The "prcomp" function provided by the GNU R package Stats [90] was utilized to apply PCA of log_2FC values for all host genes to any *Candida* or mock infection. The mock infection samples (240 mpi) have no dedicated counterpart at 0 mpi. We calculated four separate log_2FC values for the comparison against 0 mpi infected with *C. albicans*, *C. glabrata*, *C. parapsilosis*, and *C. tropicalis*.

Functional Gene ontology categories enriched for DEGs were identified with FungiFun2 [88] using hypergeometric distribution and Benjamini-Hochberg corrected p-values < 0.05 and REVIGO [91]. Ortholog information of *Candida* species was retrieved from the CGD. DEGs with orthologs in *C. albicans*, *C. glabrata*, *C. parapsilosis* and *C. tropicalis* were quantitatively compared. DEGs of *H. sapiens* datasets were quantitatively compared.

Quantification of cytokines

The amount of IL-1β, IL-6, IL-8, and TNF-α was determined by ELISA according to the manufacturer's protocol (eBioscience). After 240 mpi infected blood samples were centrifuged to obtain plasma and immediately frozen in liquid nitrogen. Cytokine levels were calculated from standard dilutions of the respective recombinant cytokine.

Module

ModuleDiscoverer was applied as described [58] to identify the regulatory module. For this analysis, only genes which were differentially expressed in at least one of the measured time points were considered. In addition, a high-confidence (score > 0.7) PPIN of *C. albicans* was downloaded from STRING version 9.1 [92]. Both DEGs and the *C. albicans* PPIN were taken as input for ModuleDiscoverer. Identifier annotations provided by CGD [93] were used. Sub-modules of the resulting regulatory module with a number of network components < 10 were not considered. The clustering of the regulatory module was performed in the programming language R version 3.4.1 using the generalized topological overlap measure regarding second-order connections as described in [94]. A cutoff of 0.65 was chosen to receive the clusters. Cytoscape version 3.2.1 [95] was used for visualizing the regulatory module. For

624 performing GO term enrichment analyses concerning biological processes, FungiFun2 [88]
625 including Fisher's exact test and Benjamini-Hochberg False Discovery Rate correction was
626 applied to each sub-module and cluster. GO terms composed of at least two members,
627 associated with at least two components and leading to adjusted p-values > 0.05 were
628 considered as significantly enriched.

629 **Gene selection for network inference and network prediction**

630 All GO categories corresponding to the keyword 'oxidative stress' were retrieved through the
631 advanced search of AmiGo [96] and filtered for the ontology source 'biological process'.
632 Differentially expressed members of a GO category were identified for *C. albicans* through a
633 GO term gene association file retrieved from the CGD. For human candidate genes,
634 members of a GO category 'oxidative stress' were retrieved through biomaRt [97] and filtered
635 for DEGs. In total, ten human and ten *C. albicans* DEGs were selected for the prediction of a
636 gene regulatory network (network inference).

637 We applied the extended NetGenerator tool [63, 98-100] to predict a small-scale gene
638 regulatory network. The main inputs are (i) $\log_2$FCs of candidate genes over time and (ii)
639 prior knowledge about candidate gene interactions. Prior knowledge for human genes was
640 gathered from the following sources: Transcription factor binding sites were retrieved from
641 oPOSSUM [101] and Qiagen website (http://www.sabiosciences.com/chipqpcrsearch.php).
642 Regulator-target interactions were extracted from TRANSFAC version 2015.4 [102].
643 Furthermore, known interactions of the type 'promoter binding' were extracted from Pathway
644 Studio 10 (ResNet11, Version/Update von Q3 2015) [103]. *C. albicans* prior knowledge was
645 manually extracted from gene descriptions provided by CGD [93]. Again, regulator-target
646 interactions were extracted from TRANSFAC (version 2015.4). Interactions of genes involved
647 in oxidative stress response were extracted from [104]. A score of 0.5 was assigned to prior
648 knowledge supported by ≥ 2 prior knowledge sources. Otherwise the prior knowledge score
649 was set to 0.3.

Reverse transcription quantitative real-time PCR (RT-qPCR)

500 ng of high quality DNase I-treated RNA samples were reversely transcribed into cDNA by using oligo-dT primers and SuperScript™ III Reverse Transcriptase (Invitrogen). Subsequently 1 µl of diluted cDNA was used for gene expression analyses with GoTaq® Green Master Mix (Promega) and a C1000 thermocycler (Bio-Rad CFX96™). Expression levels of biological triplicates were normalized to the reference genes *B2M* (*H. sapiens*) or ACT1 (*C. albicans*). Primers uses for qPCR analyses are listed in Table 2.

Table 2. Primers used for gene expression analyses.

Target gene	Name	Sequence (5' → 3')
B2M	rt-B2M_fw	TGGGTTTCATCCATCCGACA
	rt-B2M_rev	TTCACACGGCAGGCATACTC
NFKB1	rt-NFKB1_fw	GAGCTCCGAGACAGTGACAG
	rt-NFKB1_rev	GGTCCTTCCTGCCCATAATCA
IL1B	rt-IL1b_fw	ACATCAGCACCTCTCAAGCA
	rt-IL1b_rev	TGGGTACAGCTCTCTTTAGGAA
IL6	rt-IL6_fw	GACCCAACCACAAATGCCAG
	rt-IL6_rev	ATTTGCCGAAGAGCCCTCAG
SOD5	rt-SOD5_fw	TCCTGCTGCTCATGAAGTTG
	rt-SOD5_rev	GTGTTAGCATTGCCGTGTCC
CAT1	rt-CAT1_fw	ACTCCAGTGTTTTTCATTAGAG
	rt-CAT1_rev	GTGTGACCAGAGTAACCATTCA
GPX2	rt-GPX2_fw	TTGGTGTGACTTTCCCCGTA
	rt-GPX2_rev	TGCCACAACATTACCATCTTGA
HSP21	rt-HSP21_fw	TCTCTCGTTATGGTGCTGGTG
	rt-HSP21_rev	AACCACGTATTTGTCGGATTCT
ACT1	rt-ACT1_fw	ACGGTGAAGAAGTTGCTGCT
	rt-ACT1_rev	TGGATTGGGCTTCATCACCA

Statistical analyses (not RNA-seq data)

All experiments were done in at least three biological replicates with blood from non-identical donors and independent fungal cell cultures. Data sets are reported as mean ± standard deviation (SD). Statistical significance was calculated using 2way ANOVA (killing, immune cell association and activation, qPCR) or 1way ANOVA (cytokine release) with multiple comparison correction. Probability values are indicated as follows: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.005$, and **** $p < 0.0001$.

666 **Acknowledgements**

667 We thank Ilse D. Jacobsen, Bianca Schulze and Maria J. Niemiec (Microbial Immunology,
668 HKI, Jena, Germany), Marie von Lilienfeld-Toal (Infections in Hematology/Oncology, UKJ
669 and HKI, Jena, Germany), Kathleen Kämmer (Jena, Germany) and Annika König (Microbial
670 Pathogenicity Mechanisms, HKI, Jena, Germany) for technical support and
671 Gianni Panagiotou for critical reading of the manuscript and helpful discussion. We also
672 thank all voluntary blood donors.

673

References

1. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA. 2016;315(8):801-10. doi: 10.1001/jama.2016.0287. PubMed PMID: 26903338; PubMed Central PMCID: PMCPMC4968574.

2. Bone RC, Balk RA, Cerra FB, Dellinger RP, Fein AM, Knaus WA, et al. Definitions for Sepsis and Organ Failure and Guidelines for the Use of Innovative Therapies in Sepsis. Chest. 1992;101(6):1644-55. doi: 10.1378/chest.101.6.1644.

3. Martin GS. Sepsis, severe sepsis and septic shock: changes in incidence, pathogens and outcomes. Expert Rev Anti Infect Ther. 2012;10(6):701-6. doi: 10.1586/eri.12.50. PubMed PMID: 22734959; PubMed Central PMCID: PMCPMC3488423.

4. Delaloye J, Calandra T. Invasive candidiasis as a cause of sepsis in the critically ill patient. Virulence. 2014;5(1):161-9. doi: 10.4161/viru.26187. PubMed PMID: 24157707; PubMed Central PMCID: PMCPMC3916370.

5. Pfaller MA, Diekema DJ. Epidemiology of invasive candidiasis: a persistent public health problem. Clin Microbiol Rev. 2007;20(1):133-63. doi: 10.1128/CMR.00029-06. PubMed PMID: 17223626; PubMed Central PMCID: PMCPMC1797637.

6. Guinea J. Global trends in the distribution of *Candida* species causing candidemia. Clin Microbiol Infect. 2014;20 Suppl 6:5-10. doi: 10.1111/1469-0691.12539. PubMed PMID: 24506442.

7. Turner SA, Butler G. The *Candida* pathogenic species complex. Cold Spring Harb Perspect Med. 2014;4(9):a019778. doi: 10.1101/cshperspect.a019778. PubMed PMID: 25183855; PubMed Central PMCID: PMCPMC4143104.

8. Kurtzman C, Robnett C. Identification of clinically important ascomycetous yeasts based on nucleotide divergence in the 5'end of the large-subunit (26S) ribosomal DNA gene. Journal of clinical microbiology. 1997;35(5):1216-23.

9. Mühlhausen S, Kollmar M. Molecular phylogeny of sequenced *Saccharomycetes* reveals polyphyly of the alternative yeast codon usage. Genome Biol Evol. 2014;6(12):3222-37. doi: 10.1093/gbe/evu152. PubMed PMID: 25646540; PubMed Central PMCID: PMCPMC4986446.

10. Gabaldón T, Carreté L. The birth of a deadly yeast: tracing the evolutionary emergence of virulence traits in *Candida glabrata*. FEMS Yeast Res. 2016;16(2):fov110. doi: 10.1093/femsyr/fov110. PubMed PMID: 26684722.

11. Yapar N. Epidemiology and risk factors for invasive candidiasis. Ther Clin Risk Manag. 2014;10:95-105. doi: 10.2147/TCRM.S40160. PubMed PMID: 24611015; PubMed Central PMCID: PMCPMC3928396.

12. Spellberg B, Filler SG, Marr KA. *Candida*: What Should Clinicians and Scientists Be Talking About? 2012:1-8. doi: 10.1128/9781555817176.ch1.

13. Krcmery V, Barnes AJ. Non-*albicans Candida* spp. causing fungaemia: pathogenicity and antifungal resistance. J Hosp Infect. 2002;50(4):243-60. doi: 10.1053/jhin.2001.1151. PubMed PMID: 12014897.

14. Trofa D, Gácser A, Nosanchuk JD. *Candida parapsilosis*, an emerging fungal pathogen. Clin Microbiol Rev. 2008;21(4):606-25. doi: 10.1128/CMR.00013-08. PubMed PMID: 18854483; PubMed Central PMCID: PMCPMC2570155.

15. Colombo AL, Guimarães T, Silva LR, de Almeida Monfardini LP, Cunha AK, Rady P, et al. Prospective observational study of candidemia in São Paulo, Brazil: incidence rate, epidemiology, and predictors of mortality. Infect Control Hosp Epidemiol. 2007;28(5):570-6. doi: 10.1086/513615. PubMed PMID: 17464917.

16. Yang YL, Cheng MF, Wang CW, Wang AH, Cheng WT, Lo HJ, et al. The distribution of species and susceptibility of amphotericin B and fluconazole of yeast pathogens isolated from sterile sites in Taiwan. Med Mycol. 2010;48(2):328-34. doi: 10.3109/13693780903154070. PubMed PMID: 20141372.

17. Perlroth J, Choi B, Spellberg B. Nosocomial fungal infections: epidemiology, diagnosis, and treatment. Med Mycol. 2007;45(4):321-46. doi: 10.1080/13693780701218689. PubMed PMID: 17510856.

18. Silva S, Negri M, Henriques M, Oliveira R, Williams DW, Azeredo J. *Candida glabrata*, *Candida parapsilosis* and *Candida tropicalis*: biology, epidemiology, pathogenicity and antifungal resistance. FEMS Microbiol Rev. 2012;36(2):288-305. doi: 10.1111/j.1574-6976.2011.00278.x. PubMed PMID: 21569057.

19. Diekema DJ, Messer SA, Brueggemann AB, Coffman SL, Doern GV, Herwaldt LA, et al. Epidemiology of Candidemia: 3-Year Results from the Emerging Infections and the Epidemiology of Iowa Organisms Study. Journal of Clinical Microbiology. 2002;40(4):1298-302. doi: 10.1128/jcm.40.4.1298-1302.2002.

20. Eggimann P, Garbino J, Pittet D. Epidemiology of *Candida* species infections in critically ill non-immunosuppressed patients. The Lancet Infectious Diseases. 2003;3(11):685-702. doi: 10.1016/s1473-3099(03)00801-6.

21. Luo S, Skerka C, Kurzai O, Zipfel PF. Complement and innate immune evasion strategies of the human pathogenic fungus *Candida albicans*. Mol Immunol. 2013;56(3):161-9. doi: 10.1016/j.molimm.2013.05.218. PubMed PMID: 23809232.

22. Netea MG, Joosten LA, van der Meer JW, Kullberg BJ, van de Veerdonk FL. Immune defence against *Candida* fungal infections. Nat Rev Immunol. 2015;15(10):630-42. doi: 10.1038/nri3897. PubMed PMID: 26388329.

23. Shoham S, Levitz SM. The immune response to fungal infections. Br J Haematol. 2005;129(5):569-82. doi: 10.1111/j.1365-2141.2005.05397.x. PubMed PMID: 15916679.

24. Fradin C, De Groot P, MacCallum D, Schaller M, Klis F, Odds FC, et al. Granulocytes govern the transcriptional response, morphology and proliferation of *Candida albicans* in human blood. Mol Microbiol. 2005;56(2):397-415. doi: 10.1111/j.1365-2958.2005.04557.x. PubMed PMID: 15813733.

25. Hünniger K, Lehnert T, Bieber K, Martin R, Figge MT, Kurzai O. A virtual infection model quantifies innate effector mechanisms and *Candida albicans* immune escape in human blood. PLoS Comput Biol. 2014;10(2):e1003479. doi: 10.1371/journal.pcbi.1003479. PubMed PMID: 24586131; PubMed Central PMCID: PMCPMC3930496.

26. Seider K, Brunke S, Schild L, Jablonowski N, Wilson D, Majer O, et al. The facultative intracellular pathogen *Candida glabrata* subverts macrophage cytokine

762 production and phagolysosome maturation. J Immunol. 2011;187(6):3072-86. doi:
763 10.4049/jimmunol.1003730. PubMed PMID: 21849684.

764 27. Miramón P, Dunker C, Windecker H, Bohovych IM, Brown AJ, Kurzai O, et al.
765 Cellular responses of *Candida albicans* to phagocytosis and the extracellular
766 activities of neutrophils are critical to counteract carbohydrate starvation, oxidative
767 and nitrosative stress. PLoS One. 2012;7(12):e52850. doi:
768 10.1371/journal.pone.0052850. PubMed PMID: 23285201; PubMed Central PMCID:
769 PMCPMC3528649.

770 28. Chen YL, Yu SJ, Huang HY, Chang YL, Lehman VN, Silao FG, et al.
771 Calcineurin controls hyphal growth, virulence, and drug tolerance of *Candida*
772 *tropicalis*. Eukaryot Cell. 2014;13(7):844-54. doi: 10.1128/EC.00302-13. PubMed
773 PMID: 24442892; PubMed Central PMCID: PMCPMC4135728.

774 29. Ermert D, Niemiec MJ, Rohm M, Glenthøj A, Borregaard N, Urban CF.
775 *Candida albicans* escapes from mouse neutrophils. J Leukoc Biol. 2013;94(2):223-
776 36. doi: 10.1189/jlb.0213063. PubMed PMID: 23650619.

777 30. Priest SJ, Lorenz MC. Characterization of Virulence-Related Phenotypes in
778 *Candida* Species of the CUG Clade. Eukaryot Cell. 2015;14(9):931-40. doi:
779 10.1128/EC.00062-15. PubMed PMID: 26150417; PubMed Central PMCID:
780 PMCPMC4551586.

781 31. Tóth A, Zajta E, Csonka K, Vágvölgyi C, Netea MG, Gácser A. Specific
782 pathways mediating inflammasome activation by *Candida parapsilosis*. Sci Rep.
783 2017;7:43129. doi: 10.1038/srep43129. PubMed PMID: 28225025; PubMed Central
784 PMCID: PMCPMC5320503.

785 32. Jacobsen ID, Brunke S, Seider K, Schwarzmüller T, Firon A, d'Enfért C, et al.
786 *Candida glabrata* persistence in mice does not depend on host immunosuppression
787 and is unaffected by fungal amino acid auxotrophy. Infect Immun. 2010;78(3):1066-
788 77. doi: 10.1128/IAI.01244-09. PubMed PMID: 20008535; PubMed Central PMCID:
789 PMCPMC2825948.

790 33. Tierney L, Linde J, Müller S, Brunke S, Molina JC, Hube B, et al. An
791 Interspecies Regulatory Network Inferred from Simultaneous RNA-seq of *Candida*
792 *albicans* Invading Innate Immune Cells. Front Microbiol. 2012;3:85. doi:
793 10.3389/fmicb.2012.00085. PubMed PMID: 22416242; PubMed Central PMCID:
794 PMCPMC3299011.

795 34. Nguyen LN, Cesar GV, Le GT, Silver DL, Nimrichter L, Nosanchuk JD.
796 Inhibition of *Candida parapsilosis* fatty acid synthase (Fas2) induces mitochondrial
797 cell death in serum. PLoS Pathog. 2012;8(8):e1002879. doi:
798 10.1371/journal.ppat.1002879. PubMed PMID: 22952445; PubMed Central PMCID:
799 PMCPMC3431346.

800 35. Pérez JC, Kumamoto CA, Johnson AD. *Candida albicans* commensalism and
801 pathogenicity are intertwined traits directed by a tightly knit transcriptional regulatory
802 circuit. PLoS Biol. 2013;11(3):e1001510. doi: 10.1371/journal.pbio.1001510. PubMed
803 PMID: 23526879; PubMed Central PMCID: PMCPMC3601966.

804 36. Amorim-Vaz S, Tran Vdu T, Pradervand S, Pagni M, Coste AT, Sanglard D.
805 RNA Enrichment Method for Quantitative Transcriptional Analysis of Pathogens In
806 Vivo Applied to the Fungus *Candida albicans*. MBio. 2015;6(5):e00942-15. doi:
807 10.1128/mBio.00942-15. PubMed PMID: 26396240; PubMed Central PMCID:
808 PMCPMC4600103.

809 37. Bruno VM, Shetty AC, Yano J, Fidel PL, Jr., Noverr MC, Peters BM.
810 Transcriptomic analysis of vulvovaginal candidiasis identifies a role for the NLRP3
811 inflammasome. MBio. 2015;6(2). doi: 10.1128/mBio.00182-15. PubMed PMID:
812 25900651; PubMed Central PMCID: PMCPMC4453569.

813 38. Xu W, Solis NV, Ehrlich RL, Woolford CA, Filler SG, Mitchell AP. Activation
814 and alliance of regulatory pathways in *C. albicans* during mammalian infection. PLoS
815 Biol. 2015;13(2):e1002076. doi: 10.1371/journal.pbio.1002076. PubMed PMID:
816 25693184; PubMed Central PMCID: PMCPMC4333574.

817 39. Cheng S, Clancy CJ, Xu W, Schneider F, Hao B, Mitchell AP, et al. Profiling of
818 *Candida albicans* gene expression during intra-abdominal candidiasis identifies
819 biologic processes involved in pathogenesis. J Infect Dis. 2013;208(9):1529-37. doi:
820 10.1093/infdis/jit335. PubMed PMID: 24006479; PubMed Central PMCID:
821 PMCPMC3789567.

822 40. Savage DC, Dubos RJ. Localization of indigenous yeast in the murine
823 stomach. J Bacteriol. 1967;94(6):1811-6. PubMed PMID: 16562156; PubMed Central
824 PMCID: PMCPMC276909.

825 41. Mestas J, Hughes CCW. Of Mice and Not Men: Differences between Mouse
826 and Human Immunology. The Journal of Immunology. 2004;172(5):2731-8. doi:
827 10.4049/jimmunol.172.5.2731.

828 42. Duggan S, Leonhardt I, Hünniger K, Kurzai O. Host response to *Candida*
829 *albicans* bloodstream infection and sepsis. Virulence. 2015;6(4):316-26. doi:
830 10.4161/21505594.2014.988096. PubMed PMID: 25785541; PubMed Central
831 PMCID: PMCPMC4601378.

832 43. Fradin C, Kretschmar M, Nichterlein T, Gaillardin C, d'Enfért C, Hube B.
833 Stage-specific gene expression of *Candida albicans* in human blood. Mol Microbiol.
834 2003;47(6):1523-43. PubMed PMID: 12622810.

835 44. Fradin C, Mavor AL, Weindl G, Schaller M, Hanke K, Kaufmann SH, et al. The
836 early transcriptional response of human granulocytes to infection with *Candida*
837 *albicans* is not essential for killing but reflects cellular communications. Infect Immun.
838 2007;75(3):1493-501. doi: 10.1128/IAI.01651-06. PubMed PMID: 17145939; PubMed
839 Central PMCID: PMCPMC1828553.

840 45. Dix A, Hünniger K, Weber M, Guthke R, Kurzai O, Linde J. Biomarker-based
841 classification of bacterial and fungal whole-blood infections in a genome-wide
842 expression study. Front Microbiol. 2015;6:171. doi: 10.3389/fmicb.2015.00171.
843 PubMed PMID: 25814982; PubMed Central PMCID: PMCPMC4356159.

844 46. Duggan S, Essig F, Hünniger K, Mokhtari Z, Bauer L, Lehnert T, et al.
845 Neutrophil activation by *Candida glabrata* but not *Candida albicans* promotes fungal
846 uptake by monocytes. Cell Microbiol. 2015;17(9):1259-76. doi: 10.1111/cmi.12443.
847 PubMed PMID: 25850517.

848 47. Hünniger K, Bieber K, Martin R, Lehnert T, Figge MT, Löffler J, et al. A second
849 stimulus required for enhanced antifungal activity of human neutrophils in blood is
850 provided by anaphylatoxin C5a. J Immunol. 2015;194(3):1199-210. doi:
851 10.4049/jimmunol.1401845. PubMed PMID: 25539819.

852 48. Timme S, Lehnert T, Praußе MTE, Hünniger K, Leonhardt I, Kurzai O, et al.
853 Quantitative simulations predict treatment strategies against fungal infections in

854 virtual neutropenic patients. Frontiers in Immunology. 2018;9:667. doi:
855 10.3389/fimmu.2018.00667.

856 49. Erwig LP, Gow NA. Interactions of fungal pathogens with phagocytes. Nat Rev
857 Microbiol. 2016;14(3):163-76. doi: 10.1038/nrmicro.2015.21. PubMed PMID:
858 26853116.

859 50. Gilbert AS, Wheeler RT, May RC. Fungal Pathogens: Survival and Replication
860 within Macrophages. Cold Spring Harb Perspect Med. 2014;5(7):a019661. doi:
861 10.1101/cshperspect.a019661. PubMed PMID: 25384769.

862 51. Lionakis MS. New insights into innate immune control of systemic candidiasis.
863 Med Mycol. 2014;52(6):555-64. doi: 10.1093/mmy/myu029. PubMed PMID:
864 25023483; PubMed Central PMCID: PMCPMC4823972.

865 52. Jiménez-López C, Lorenz MC. Fungal immune evasion in a model host-
866 pathogen interaction: *Candida albicans* versus macrophages. PLoS Pathog.
867 2013;9(11):e1003741. doi: 10.1371/journal.ppat.1003741. PubMed PMID: 24278014;
868 PubMed Central PMCID: PMCPMC3836912.

869 53. Mavor A, Thewes S, Hube B. Systemic Fungal Infections Caused by *Candida*
870 Species: Epidemiology, Infection Process and Virulence Attributes. Current Drug
871 Targets. 2005;6(8):863-74. doi: 10.2174/138945005774912735.

872 54. Netea MG, Van Der Graaf CA, Vonk AG, Verschueren I, Van Der Meer JW,
873 Kullberg BJ. The role of toll-like receptor (TLR) 2 and TLR4 in the host defense
874 against disseminated candidiasis. J Infect Dis. 2002;185(10):1483-9. doi:
875 10.1086/340511. PubMed PMID: 11992285.

876 55. McInnes IB. Cytokines. Kelley and Firestein's Textbook of Rheumatology.
877 102017. p. 2288.

878 56. Kaur R, Ma B, Cormack BP. A family of glycosylphosphatidylinositol-linked
879 aspartyl proteases is required for virulence of *Candida glabrata*. Proc Natl Acad Sci U
880 S A. 2007;104(18):7628-33. doi: 10.1073/pnas.0611195104. PubMed PMID:
881 17456602; PubMed Central PMCID: PMCPMC1863504.

882 57. Naglik JR, Challacombe SJ, Hube B. *Candida albicans* Secreted Aspartyl
883 Proteinases in Virulence and Pathogenesis. Microbiology and Molecular Biology
884 Reviews. 2003;67(3):400-28. doi: 10.1128/mmbr.67.3.400-428.2003.

885 58. Vlaic S, Conrad T, Tokarski-Schnelle C, Gustafsson M, Dahmen U, Guthke R,
886 et al. ModuleDiscoverer: Identification of regulatory modules in protein-protein
887 interaction networks. Sci Rep. 2018;8(1):433. doi: 10.1038/s41598-017-18370-2.
888 PubMed PMID: 29323246; PubMed Central PMCID: PMCPMC5764996.

889 59. Gerwien F, Skrahina V, Kasper L, Hube B, Brunke S. Metals in Fungal
890 Virulence. FEMS Microbiol Rev. 2017. doi: 10.1093/femsre/fux050. PubMed PMID:
891 29069482.

892 60. Almeida RS, Wilson D, Hube B. *Candida albicans* iron acquisition within the
893 host. FEMS Yeast Res. 2009;9(7):1000-12. doi: 10.1111/j.1567-1364.2009.00570.x.
894 PubMed PMID: 19788558.

895 61. Gerwien F, Safyan A, Wisgott S, Hille F, Kaemmer P, Linde J, et al. A Novel
896 Hybrid Iron Regulation Network Combines Features from Pathogenic and
897 Nonpathogenic Yeasts. MBio. 2016;7(5). doi: 10.1128/mBio.01782-16. PubMed
898 PMID: 27795405; PubMed Central PMCID: PMCPMC5082906.

62. de Groot PW, Kraneveld EA, Yin QY, Dekker HL, Gross U, Crielaard W, et al. The cell wall of the human pathogen *Candida glabrata*: differential incorporation of novel adhesin-like wall proteins. Eukaryot Cell. 2008;7(11):1951-64. doi: 10.1128/EC.00284-08. PubMed PMID: 18806209; PubMed Central PMCID: PMCPMC2583536.

63. Schulze S, Henkel SG, Driesch D, Guthke R, Linde J. Computational prediction of molecular pathogen-host interactions based on dual transcriptome data. Front Microbiol. 2015;6:65. doi: 10.3389/fmicb.2015.00065. PubMed PMID: 25705211; PubMed Central PMCID: PMCPMC4319478.

64. Alarco AM, Raymond M. The bZip transcription factor Cap1p is involved in multidrug resistance and oxidative stress response in *Candida albicans*. J Bacteriol. 1999;181(3):700-8. PubMed PMID: 9922230; PubMed Central PMCID: PMCPMC93433.

65. Tóth R, Tóth A, Papp C, Jankovics F, Vágvölgyi C, Alonso MF, et al. Kinetic studies of *Candida parapsilosis* phagocytosis by macrophages and detection of intracellular survival mechanisms. Front Microbiol. 2014;5:633. doi: 10.3389/fmicb.2014.00633. PubMed PMID: 25477874; PubMed Central PMCID: PMCPMC4238376.

66. Otto V, Howard D. Further studies on the intracellular behavior of *Torulopsis glabrata*. Infection and immunity. 1976;14(2):433-8.

67. Roetzer A, Gratz N, Kovarik P, Schüller C. Autophagy supports *Candida glabrata* survival during phagocytosis. Cell Microbiol. 2010;12(2):199-216. doi: 10.1111/j.1462-5822.2009.01391.x. PubMed PMID: 19811500; PubMed Central PMCID: PMCPMC2816358.

68. Lorenz MC, Bender JA, Fink GR. Transcriptional response of *Candida albicans* upon internalization by macrophages. Eukaryot Cell. 2004;3(5):1076-87. doi: 10.1128/EC.3.5.1076-1087.2004. PubMed PMID: 15470236; PubMed Central PMCID: PMCPMC522606.

69. Irmer H, Tarazona S, Sasse C, Olbermann P, Löffler J, Krappmann S, et al. RNAseq analysis of *Aspergillus fumigatus* in blood reveals a just wait and see resting stage behavior. BMC Genomics. 2015;16:640. doi: 10.1186/s12864-015-1853-1. PubMed PMID: 26311470; PubMed Central PMCID: PMCPMC4551469.

70. Lorenz MC, Fink GR. Life and death in a macrophage: role of the glyoxylate cycle in virulence. Eukaryot Cell. 2002;1(5):657-62. PubMed PMID: 12455685; PubMed Central PMCID: PMCPMC126751.

71. Muñoz-Elías EJ, McKinney JD. *Mycobacterium tuberculosis* isocitrate lyases 1 and 2 are jointly required for *in vivo* growth and virulence. Nat Med. 2005;11(6):638-44. doi: 10.1038/nm1252. PubMed PMID: 15895072; PubMed Central PMCID: PMCPMC1464426.

72. Derengowski LS, Tavares AH, Silva S, Procópio LS, Felipe MS, Silva-Pereira I. Upregulation of glyoxylate cycle genes upon *Paracoccidioides brasiliensis* internalization by murine macrophages and *in vitro* nutritional stress condition. Med Mycol. 2008;46(2):125-34. doi: 10.1080/13693780701670509. PubMed PMID: 18324491.

73. Cheah HL, Lim V, Sandai D. Inhibitors of the glyoxylate cycle enzyme ICL1 in *Candida albicans* for potential use as antifungal agents. PLoS One.

2014;9(4):e95951. doi: 10.1371/journal.pone.0095951. PubMed PMID: 24781056; PubMed Central PMCID: PMCPMC4004578.

74. Lorenz MC, Fink GR. The glyoxylate cycle is required for fungal virulence. Nature. 2001;412(6842):83-6. doi: 10.1038/35083594. PubMed PMID: 11452311.

75. Weissman Z, Kornitzer D. A family of *Candida* cell surface haem-binding proteins involved in haemin and haemoglobin-iron utilization. Mol Microbiol. 2004;53(4):1209-20. doi: 10.1111/j.1365-2958.2004.04199.x. PubMed PMID: 15306022.

76. Gerwien F, Safyan A, Wisgott S, Brunke S, Kasper L, Hube B. The Fungal Pathogen *Candida glabrata* Does Not Depend on Surface Ferric Reductases for Iron Acquisition. Front Microbiol. 2017;8:1055. doi: 10.3389/fmicb.2017.01055. PubMed PMID: 28642757; PubMed Central PMCID: PMCPMC5463049.

77. Grubb SE, Murdoch C, Sudbery PE, Saville SP, Lopez-Ribot JL, Thornhill MH. *Candida albicans*-endothelial cell interactions: a key step in the pathogenesis of systemic candidiasis. Infect Immun. 2008;76(10):4370-7. doi: 10.1128/IAI.00332-08. PubMed PMID: 18573891; PubMed Central PMCID: PMCPMC2546854.

78. de Groot PW, Bader O, de Boer AD, Weig M, Chauhan N. Adhesins in human fungal pathogens: glue with plenty of stick. Eukaryot Cell. 2013;12(4):470-81. doi: 10.1128/EC.00364-12. PubMed PMID: 23397570; PubMed Central PMCID: PMCPMC3623432.

79. Butler G, Rasmussen MD, Lin MF, Santos MA, Sakthikumar S, Munro CA, et al. Evolution of pathogenicity and sexual reproduction in eight *Candida* genomes. Nature. 2009;459(7247):657-62. doi: 10.1038/nature08064. PubMed PMID: 19465905; PubMed Central PMCID: PMCPMC2834264.

80. Zupancic ML, Frieman M, Smith D, Alvarez RA, Cummings RD, Cormack BP. Glycan microarray analysis of *Candida glabrata* adhesin ligand specificity. Mol Microbiol. 2008;68(3):547-59. doi: 10.1111/j.1365-2958.2008.06184.x. PubMed PMID: 18394144.

81. Mayer FL, Wilson D, Jacobsen ID, Miramón P, Slesiona S, Bohovych IM, et al. Small but crucial: the novel small heat shock protein Hsp21 mediates stress adaptation and virulence in *Candida albicans*. PLoS One. 2012;7(6):e38584. doi: 10.1371/journal.pone.0038584. PubMed PMID: 22685587; PubMed Central PMCID: PMCPMC3369842.

82. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114-20. doi: 10.1093/bioinformatics/btu170. PubMed PMID: 24695404; PubMed Central PMCID: PMCPMC4103590.

83. International Human Genome Sequencing C. Finishing the euchromatic sequence of the human genome. Nature. 2004;431(7011):931-45. doi: 10.1038/nature03001. PubMed PMID: 15496913.

84. Inglis DO, Arnaud MB, Binkley J, Shah P, Skrzypek MS, Wymore F, et al. The *Candida* genome database incorporates multiple *Candida* species: multispecies search and analysis tools with curated gene and protein information for *Candida albicans* and *Candida glabrata*. Nucleic Acids Res. 2012;40(Database issue):D667-74. doi: 10.1093/nar/gkr945. PubMed PMID: 22064862; PubMed Central PMCID: PMCPMC3245171.

991 85. Bruno VM, Wang Z, Marjani SL, Euskirchen GM, Martin J, Sherlock G, et al.
992 Comprehensive annotation of the transcriptome of the human fungal pathogen
993 *Candida albicans* using RNA-seq. Genome Res. 2010;20(10):1451-8. doi:
994 10.1101/gr.109553.110. PubMed PMID: 20810668; PubMed Central PMCID:
995 PMCPMC2945194.

996 86. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2:
997 accurate alignment of transcriptomes in the presence of insertions, deletions and
998 gene fusions. Genome Biol. 2013;14(4):R36. doi: 10.1186/gb-2013-14-4-r36.
999 PubMed PMID: 23618408; PubMed Central PMCID: PMCPMC4053844.

1000 87. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose
1001 program for assigning sequence reads to genomic features. Bioinformatics.
1002 2014;30(7):923-30. doi: 10.1093/bioinformatics/btt656. PubMed PMID: 24227677.

1003 88. Priebe S, Kreisel C, Horn F, Guthke R, Linde J. FungiFun2: a comprehensive
1004 online resource for systematic analysis of gene lists from fungal species.
1005 Bioinformatics. 2015;31(3):445-6. doi: 10.1093/bioinformatics/btu627. PubMed PMID:
1006 25294921; PubMed Central PMCID: PMCPMC4308660.

1007 89. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene
1008 expression and hybridization array data repository. Nucleic Acids Res.
1009 2002;30(1):207-10. PubMed PMID: 11752295; PubMed Central PMCID:
1010 PMCPMC99122.

1011 90. Love MI, Huber W, Anders S. Moderated estimation of fold change and
1012 dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550. doi:
1013 10.1186/s13059-014-0550-8. PubMed PMID: 25516281; PubMed Central PMCID:
1014 PMCPMC4302049.

1015 91. Supek F, Bošnjak M, Škunca N, Šmuc T. REVIGO summarizes and visualizes
1016 long lists of gene ontology terms. PLoS One. 2011;6(7):e21800. doi:
1017 10.1371/journal.pone.0021800. PubMed PMID: 21789182; PubMed Central PMCID:
1018 PMCPMC3138752.

1019 92. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, et al.
1020 STRING v9.1: protein-protein interaction networks, with increased coverage and
1021 integration. Nucleic Acids Res. 2013;41(Database issue):D808-15. doi:
1022 10.1093/nar/gks1094. PubMed PMID: 23203871; PubMed Central PMCID:
1023 PMCPMC3531103.

1024 93. Skrzypek MS, Binkley J, Binkley G, Miyasato SR, Simison M, Sherlock G. The
1025 *Candida* Genome Database (CGD): incorporation of Assembly 22, systematic
1026 identifiers and visualization of high throughput sequencing data. Nucleic Acids Res.
1027 2017;45(D1):D592-D6. doi: 10.1093/nar/gkw924. PubMed PMID: 27738138; PubMed
1028 Central PMCID: PMCPMC5210628.

1029 94. Yip AM, Horvath S. Gene network interconnectedness and the generalized
1030 topological overlap measure. BMC Bioinformatics. 2007;8:22. doi: 10.1186/1471-
1031 2105-8-22. PubMed PMID: 17250769; PubMed Central PMCID: PMCPMC1797055.

1032 95. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al.
1033 Cytoscape: a software environment for integrated models of biomolecular interaction
1034 networks. Genome Res. 2003;13(11):2498-504. doi: 10.1101/gr.1239303. PubMed
1035 PMID: 14597658; PubMed Central PMCID: PMCPMC403769.

96. Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S, et al. AmiGO: online access to ontology and annotation data. Bioinformatics. 2009;25(2):288-9. doi: 10.1093/bioinformatics/btn615. PubMed PMID: 19033274; PubMed Central PMCID: PMCPMC2639003.

97. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat Protoc. 2009;4(8):1184-91. doi: 10.1038/nprot.2009.97. PubMed PMID: 19617889; PubMed Central PMCID: PMCPMC3159387.

98. Guthke R, Möller U, Hoffmann M, Thies F, Töpfer S. Dynamic network reconstruction from gene expression data applied to immune response during bacterial infection. Bioinformatics. 2005;21(8):1626-34. doi: 10.1093/bioinformatics/bti226. PubMed PMID: 15613398.

99. Töpfer S, Guthke R, Driesch D, Woetzel D, Pfaff M. The NetGenerator Algorithm: Reconstruction of Gene Regulatory Networks. In: Tuyls K, Westra R, Saeys Y, Nowé A, editors. Knowledge Discovery and Emergent Complexity in Bioinformatics: First International Workshop, KDECB 2006, Ghent, Belgium, May 10, 2006 Revised Selected Papers. Berlin, Heidelberg: Springer Berlin Heidelberg; 2007. p. 119-30.

100. Weber M, Henkel SG, Vlaic S, Guthke R, van Zoelen EJ, Driesch D. Inference of dynamical gene-regulatory networks based on time-resolved multi-stimuli multi-experiment data applying NetGenerator V2.0. BMC Syst Biol. 2013;7:1. doi: 10.1186/1752-0509-7-1. PubMed PMID: 23280066; PubMed Central PMCID: PMCPMC3605253.

101. Kwon AT, Arenillas DJ, Worsley Hunt R, Wasserman WW. oPOSSUM-3: advanced analysis of regulatory motif over-representation across genes or ChIP-Seq datasets. G3 (Bethesda). 2012;2(9):987-1002. doi: 10.1534/g3.112.003202. PubMed PMID: 22973536; PubMed Central PMCID: PMCPMC3429929.

102. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, et al. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res. 2006;34(Database issue):D108-10. doi: 10.1093/nar/gkj143. PubMed PMID: 16381825; PubMed Central PMCID: PMCPMC1347505.

103. Nikitin A, Egorov S, Daraselia N, Mazo I. Pathway studio - the analysis and navigation of molecular networks. Bioinformatics. 2003;19(16):2155-7. doi: 10.1093/bioinformatics/btg290.

104. Wang Y, Cao YY, Jia XM, Cao YB, Gao PH, Fu XP, et al. Cap1p is involved in multiple pathways of oxidative stress response in *Candida albicans*. Free Radic Biol Med. 2006;40(7):1201-9. doi: 10.1016/j.freeradbiomed.2005.11.019. PubMed PMID: 16545688.

105. Gillum AM, Tsay EY, Kirsch DR. Isolation of the *Candida albicans* gene for orotidine-5'-phosphate decarboxylase by complementation of *S. cerevisiae URA3* and *E. coli* pyrF mutations. Mol Gen Genet. 1984;198(2):179-82. PubMed PMID: 6394964.

106. Koszul R, Malpertuy A, Frangeul L, Bouchier C, Wincker P, Thierry A, et al. The complete mitochondrial genome sequence of the pathogenic yeast *Candida* (*Torulopsis*) *glabrata*. FEBS Letters. 2003;534(1-3):39-48. doi: 10.1016/s0014-5793(02)03749-3.

1083 107. Gácser A, Salomon S, Schäfer W. Direct transformation of a clinical isolate of
1084 *Candida parapsilosis* using a dominant selection marker. FEMS Microbiol Lett.
1085 2005;245(1):117-21. doi: 10.1016/j.femsle.2005.02.035. PubMed PMID: 15796988.

1086 108. Rüchel R. A variety of *Candida* proteinases and their possible targets of
1087 proteolytic attack in the host. Zentralblatt für Bakteriologie, Mikrobiologie und Hygiene
1088 1 Abt Originale A, Medizinische Mikrobiologie, Infektionskrankheiten und
1089 Parasitologie. 1984;257(2):266-74. PubMed PMID: 6385564.

1090 109. Schwarzmüller T, Ma B, Hiller E, Istel F, Tscherner M, Brunke S, et al.
1091 Systematic phenotyping of a large-scale *Candida glabrata* deletion collection reveals
1092 novel antifungal tolerance genes. PLoS Pathog. 2014;10(6):e1004211. doi:
1093 10.1371/journal.ppat.1004211. PubMed PMID: 24945925; PubMed Central PMCID:
1094 PMCPMC4063973.

1095

Figure legends

Figure 1. *Candida* species interact with human immune cells and are killed immediately upon blood exposure.

(A) Within one hour of blood infection the majority of fungal cells are killed. Although the killing kinetics between the fungal species are similar, *C. albicans* survives to an overall larger extent. (B) Already 60 mpi the majority of fungal cells are associated with immune cells of human blood, predominantly with neutrophils. *C. glabrata* and *C. parapsilosis* associate to monocytes in a higher amount than *C. tropicalis* and *C. albicans* 240 mpi. Results show means of four (A) or three (B) independent experiments from different donors ± SD. **P*<0.05, **P<0.01, ***P<0.001, ****P<0.0001 (2way ANOVA test).

Figure 2. *Candida* species induce a mainly species-independent human transcriptional core response.

(A) Principle component analyses (PCA) reveal a higher similarity between samples of one time point (same color) than one species (same icon). Mock infected control samples (▽) are clustered together and clearly separated from all infection samples. (B) The transcriptional host response starts restrained with a few regulated genes 15 mpi but distinctly increases during the time course of infection with similar kinetics between the four *Candida* infections. (C) Venn diagrams illustrate that in response to *Candida* blood infections about 670 and 490 human genes are commonly up- (left) and down- (right) regulated, respectively.

Figure 3. Immune system processes govern the human core response of up- and down-regulated genes and pro-inflammatory cytokines and chemokines are released upon *Candida* blood infection.

(A) Functional gene ontology (GO) analyses were performed to identify enriched biological processes of common up- or down-regulated human genes. Immune system processes like

inflammatory response, cytokine-mediated signaling (both up-regulated) or toll-like receptor signaling (down-regulated) govern the human core response indicating a strong but balanced response to *Candida* blood infections. (B) Plasma levels of the pro-inflammatory cytokines IL-1β, IL-6 and TNF-α and the chemokine IL-8 were increased 240 mpi compared to mock infection. *C. parapsilosis* and *C. glabrata* cause higher levels of the pro-inflammatory cytokines compared to *C. albicans* and *C. tropicalis* infections. IL-8 plasma levels were lower upon *C. glabrata*-infection than infection with the other three *Candida* species. Results show means of at least three (B) independent experiments from different donors ± SD (1way ANOVA test).

Figure 4. Species-specific responses govern fungal transcriptomes upon blood infection.

(A) *Candida* genes are regulated with different kinetics in response to blood infection. With exception of *C. glabrata*, substantial subsets of the fungal genomes are immediately regulated. Furthermore, the proportions of up- (left graph) and down-regulated (right graph) genes vary between the four species. (B) A Venn diagram shows that only 189 orthologs are commonly regulated in response to *Candida* blood infection. (C) PCA analyses of the fungal core response reveal no clear similarity between samples of one time point (same color) or one species (same icon). (D) Enriched categories of the common fungal core response comprise, among others, the up-regulation of the unfolded protein response and the down-regulation of several translational processes.

Figure 5. Transcriptional regulation of genes within clusters of a regulatory module is a highly dynamic process.

C. albicans DEGs were used to generate a protein-protein interaction network-based regulatory module containing sets of co-expressed genes sharing a common function. Color-coded clusters within the regulatory module contain strongly connected network components

1149 which are significantly associated with distinct biological processes (Table S 5), e.g. cluster
1150 12 arginine biosynthesis. The regulation of each single gene is indicated by intense
1151 (differentially expressed) or transparent (not differentially expressed) colouring, according to
1152 the respective time point. The regulation of most of the clusters is highly dynamic during the
1153 infection process. **Figure 5** is also provided as animated **Figure 5.gif**.

1154

1155 **Figure 6. *Candida* species regulate species-specific subsets of genes involved in iron**
1156 **acquisition and oxidative stress response.**

1157 (A) *Candida* species exploit miscellaneous iron sources during blood infection. Note, genes
1158 without a respective ortholog are shown in white. (B) *C. albicans*, *C. tropicalis* and
1159 *C. parapsilosis* up-regulate several genes of oxidative stress response, while *C. glabrata*
1160 does not. Note, genes without a respective ortholog are shown in white.

1161

1162 **Figure 7. *C. albicans* oxidative stress response genes contribute to *C. albicans***
1163 **survival in blood.**

1164 (A) An interspecies network of human (red circles) and *C. albicans* (blue circles) genes
1165 involved in oxidative stress response predicts a multitude of new interspecies interactions on
1166 the transcriptional level (green lines). Activation of gene expression is indicated by arrow
1167 heads and repression by bars. Prior knowledge is shown in black lines (dashed: not re-
1168 inferred, solid: re-inferred). (B) Gene expression (qPCR) of *NFKB1*, *IL1B*, and *IL6* are not
1169 changed due to *SOD5* deletion compared to *C. albicans* wild type-infection 240 mpi.
1170 Expression of *C. albicans GPX2* and *HSP21* are increased upon *sod5*ΔΔ blood infection.
1171 (C) Whole blood infection with mutants of the oxidative stress response but change the
1172 survival of respective mutants in whole blood. (D) In contrast to its *C. albicans* ortholog
1173 *CAT1*, the catalase encoding gene *CTA1* is dispensable for *C. glabrata* survival in blood.
1174 Indeed, *CTA1* lacking mutant (*cta1*Δ, light grey) showed better survival than the respective
1175 wild type (ATCC2001) 15 mpi. Results show means of results from at least three

1176 independent experiments from different donors ± SD. *P<0.05, **P<0.01, ***P<0.001 (B, D:
1177 2way ANOVA test, C: 1way ANOVA test).
1178

1179 ## Supporting information

1180 **Figure S 1 *Candida* cells associate with immune cells and activate neutrophils during**
1181 **blood infection.**

1182 (A) Blood smears of *Candida* spp. blood infections display the interaction of host and fungal
1183 cells. *C. albicans* changes the morphology by forming germ tubes 60 mpi and filaments
1184 120 mpi, respectively. Morphological alterations were not observed for *C. tropicalis*,
1185 *C. parapsilosis*, and *C. glabrata*. Scale bar: 10 µm. (B) The PMN activation markers CD69,
1186 CD66b, CD11b are increased and CD16 decreased upon *Candida* blood infection compared
1187 to mock infection.
1188

1189 **Figure S 2. *Candida* blood infections cause comprehensive regulations of the immune**
1190 **system.**

1191 A multitude of immunomodulatory genes of the human common core response is similarly
1192 up- (red) or down-regulated (blue) in response to *Candida* blood infections 240 mpi with *IL1A*
1193 is the most strongly up-regulated human gene (almost 100,000 times upon *C. parapsilosis*
1194 infection).
1195

1196 **Table S 1. Statistics of dual species RNA-sequencing.**

1197 This table provides an overview of the RNA-seq statistics for all four *Candida* infections.
1198

1199 **Table S 2. Differentially expressed human genes.**

1200 Human DEGs in response to *C. albicans* blood infection are listed in an xlsx-file.
1201

Table S 3. Regulation of immunoregulatory genes of the human core response.

Immunoregulatory genes of the human core response 240 mpi are listed in an xlsx-file.

Table S 4. Differentially expressed fungal genes.

DEGs from *C. albicans* in response to *Candida* blood infection are listed in an xlsx-file.

Table S 5. Clusters and cluster-associated enriched biological processes of *C. albicans* DEG-based regulatory module.

Cluster-associated enriched biological processes of all clusters and the respective cluster-associated enriched biological processes are provided as xlsx-file.

Figure 1

Click here to download Figure Figure 1.png

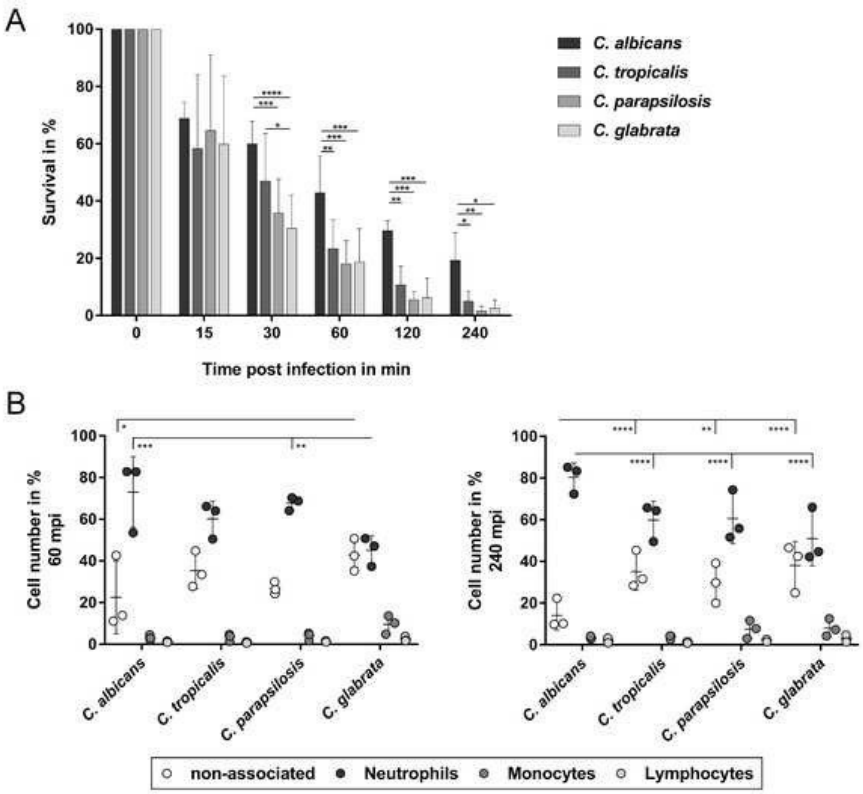


Figure 2

Click here to download Figure Figure 2.png

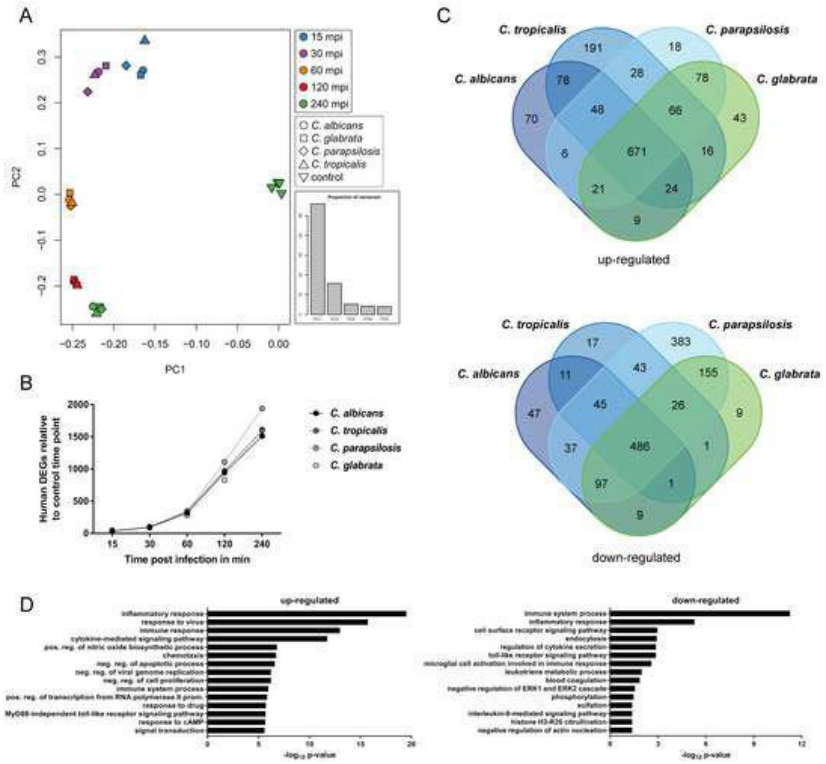


Figure 3

Click here to download Figure Figure 3.png

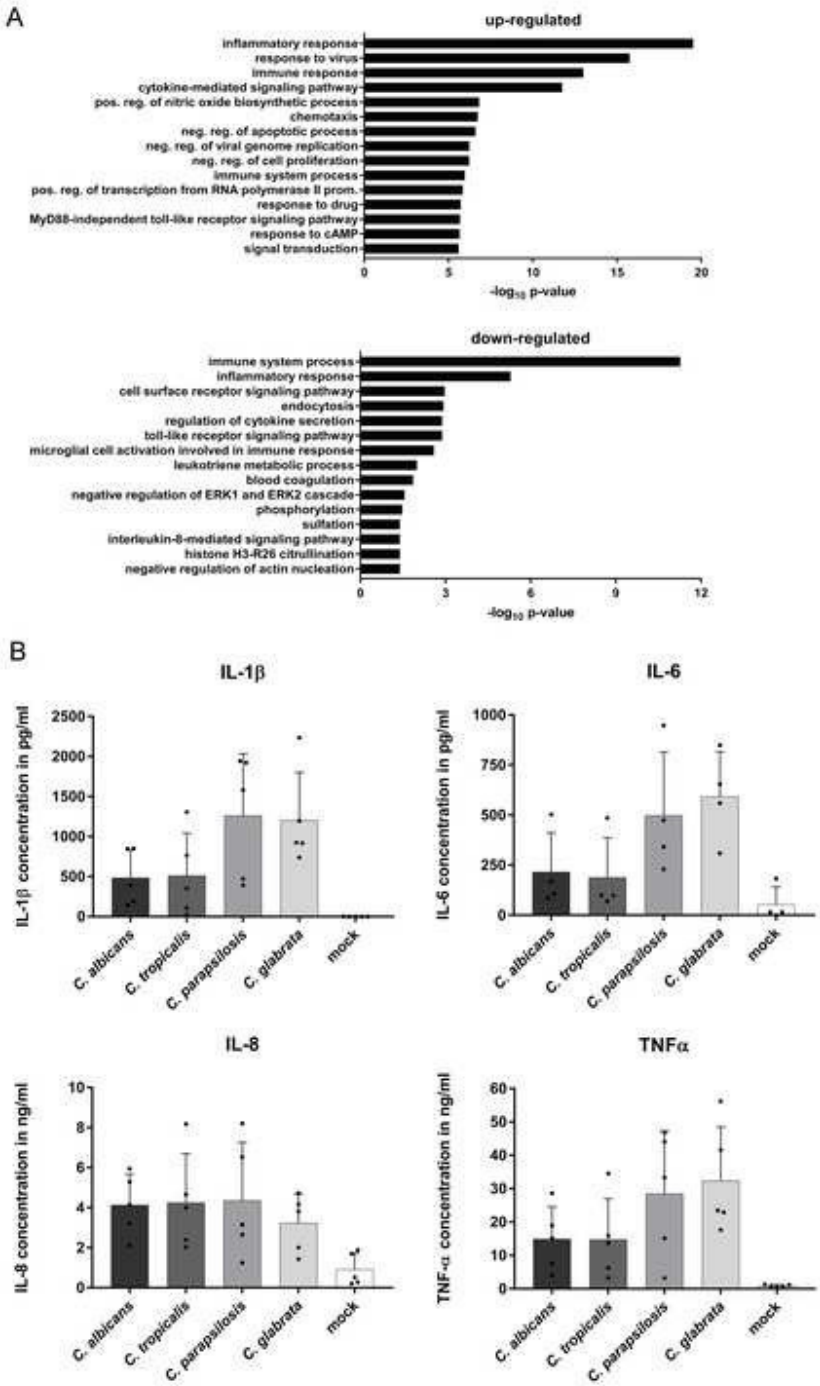


Figure 4

Click here to download Figure Figure 4.png

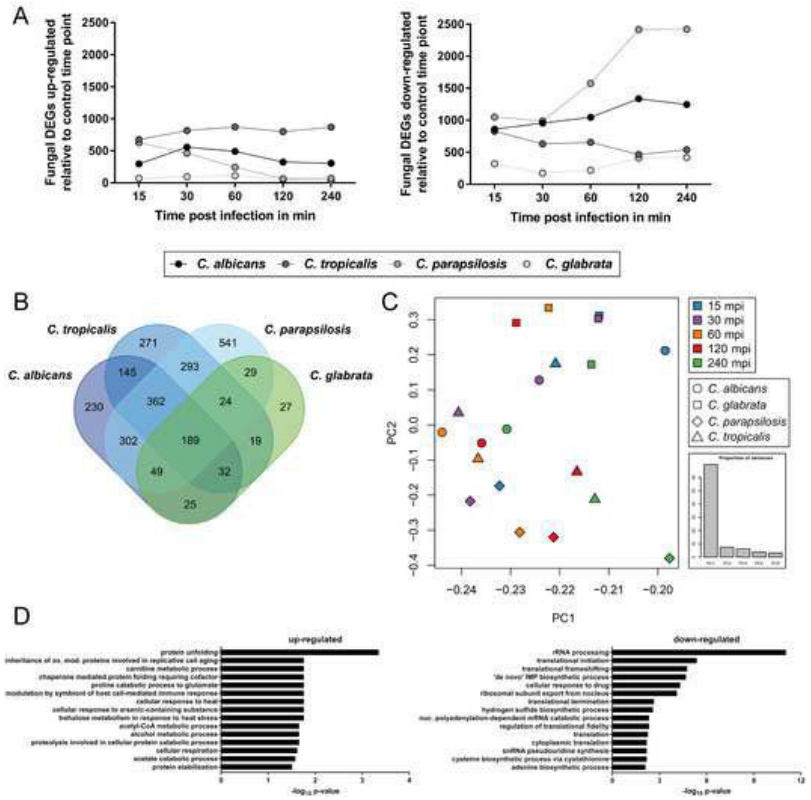


Figure 5

Click here to download Figure Figure 5.png

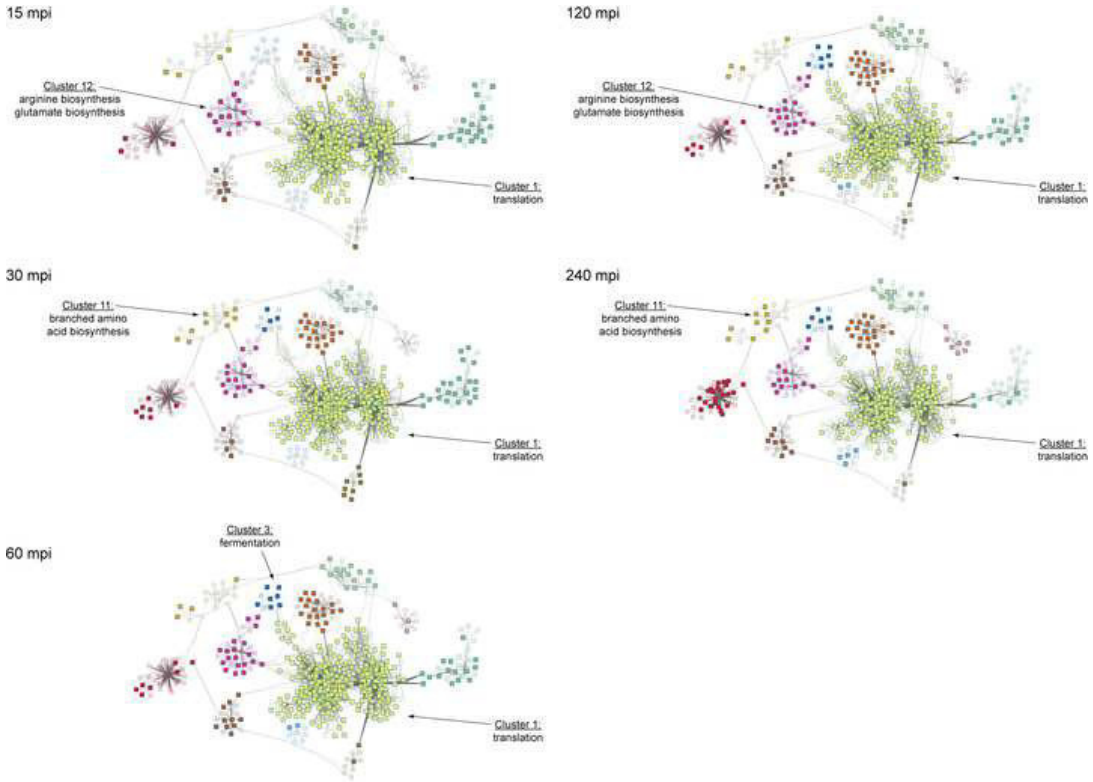


Figure 5_animation

Click here to download Figure Figure 5_animation_web.gif

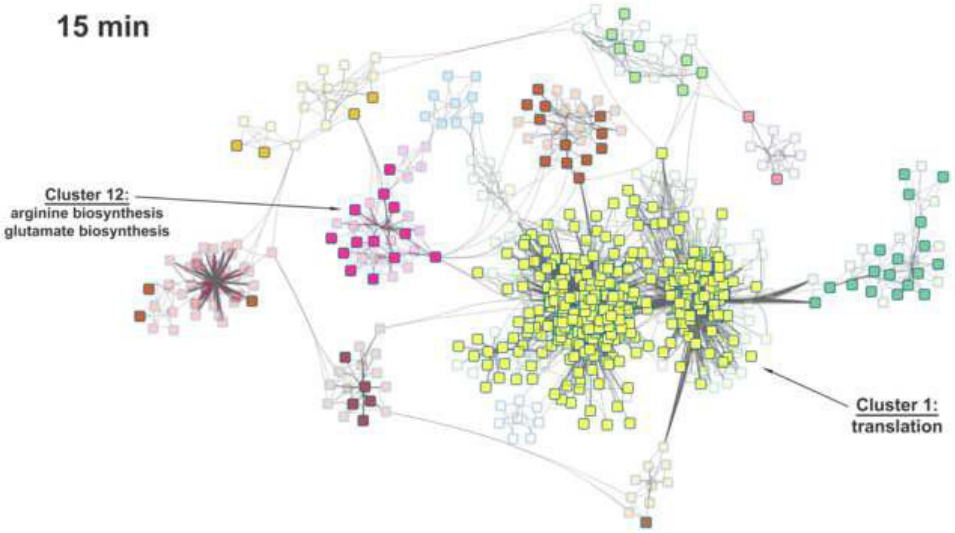


Figure 6

Click here to download Figure Figure 6.png

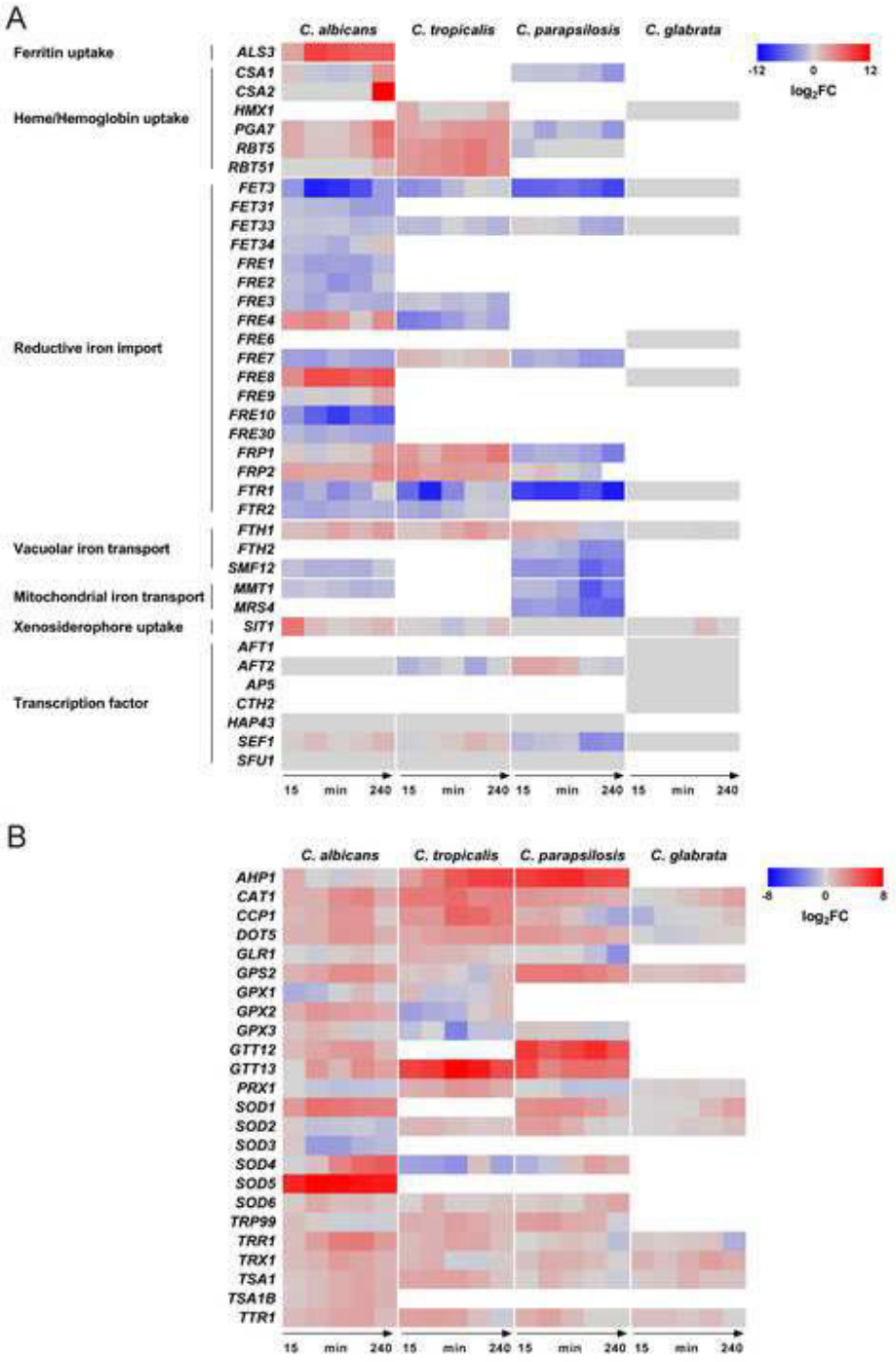
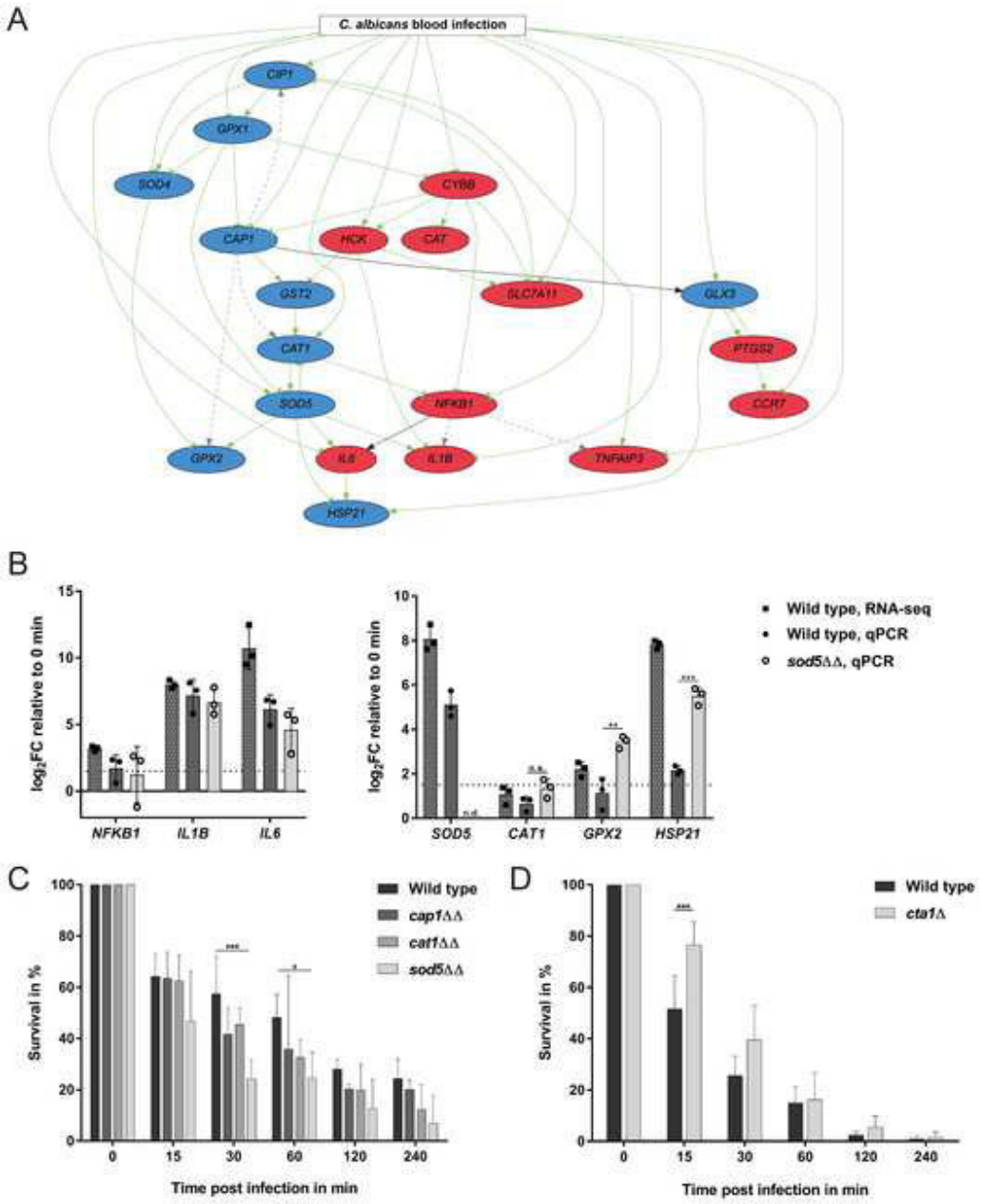


Figure 7

Click here to download Figure Figure 7.png



2.5 Manuskript 5: „Facing the challenges of multiscale modelling of bacterial and fungal pathogen-host interactions“

Status

Veröffentlicht im Februar 2016

Literaturangabe

Schleicher, J.*, Conrad, T.*, Gustafsson, M., Cedersund, G., Guthke, R., Linde, J. (2016). Facing the challenges of multiscale modelling of bacterial and fungal pathogen–host interactions. *Briefings in Functional Genomics,* 16(2). https://doi.org/10.1093/bfgp/elv064

Übersicht

Das Review von Schleicher und Conrad *et al.* konzentriert sich auf einen weiteren Ansatz zur Integration von Multiskalen- und Multi-Omik-Daten: Die Modellierung. In dieser Arbeit wird ein Überblick über die derzeit verfügbaren, computergestützten Modellierungsansätze für WPI-basierte Daten und die dabei zu bewältigenden Herausforderungen gegeben. Es werden sowohl zeitkontinuierliche, zeitdiskrete als auch zeitunabhängige Modellierungsansätze betrachtet. Daneben werden verschiedene Beispiele für kombinierte Ansätze vorgestellt. Eine Recherche in der aktuellen Literatur hat gezeigt, dass vor allem die Forschung bezüglich bakterieller Infektionen schon weit vorangeschritten ist. Im Gegensatz dazu befindet sich die Modellierung von WPPI derzeit noch am Anfang. Somit liefert dieses Review einen Ausblick darauf, inwiefern die Anwendung und die Kombination verschiedener Modellierungsansätze die Datenintegration unterstützen und zu einem tieferen Verständnis der WPPI beitragen können. Die im Review vorgestellten Ansätze zur Multiskalen- und Multi-Omik-Modellierung stellen einen guten Ausgangspunkt für weitere Forschungsarbeiten dar, die im Anschluss an diese Dissertation erfolgen könnten.

* geteilte Erstautorenschaft

Beiträge

SJ und CT waren für die Konzipierung der Studie zuständig. SJ, CT, GR und LJ übernahmen die Literaturrecherche, die Interpretation der Rechercheergebnisse und das Schreiben des Manuskripts. GM und CG unterstützten den Interpretations- und Schreibprozess. Alle Autoren waren an der Überprüfung und Überarbeitung des Manuskripts beteiligt.

OXFORD

Briefings in Functional Genomics, 16(2), 2017, 57–69

doi: 10.1093/bfgp/elv064
Advance Access Publication Date: 8 February 2016
Review paper

Facing the challenges of multiscale modelling of bacterial and fungal pathogen–host interactions

Jana Schleicher*, Theresia Conrad*, Mika Gustafsson, Gunnar Cedersund, Reinhard Guthke, and Jörg Linde

Corresponding author: Jörg Linde, Leibniz Institute for Natural Product Research and Infection Biology—Hans Knöll Institute, Jena, Germany. Tel.: +49-3641-532-1290; E-mail: Joerg.Linde@leibniz-hki.de
*These two authors contributed to this work equally.

Abstract

Recent and rapidly evolving progress on high-throughput measurement techniques and computational performance has led to the emergence of new disciplines, such as systems medicine and translational systems biology. At the core of these disciplines lies the desire to produce multiscale models: mathematical models that integrate multiple scales of biological organization, ranging from molecular, cellular and tissue models to organ, whole-organism and population scale models. Using such models, hypotheses can systematically be tested. In this review, we present state-of-the-art multiscale modelling of bacterial and fungal infections, considering both the pathogen and host as well as their interaction. Multiscale modelling of the interactions of bacteria, especially *Mycobacterium tuberculosis*, with the human host is quite advanced. In contrast, models for fungal infections are still in their infancy, in particular regarding infections with the most important human pathogenic fungi, *Candida albicans* and *Aspergillus fumigatus*. We reflect on the current availability of computational approaches for multiscale modelling of host–pathogen interactions and point out current challenges. Finally, we provide an outlook for future requirements of multiscale modelling.

Key words: infection; host–pathogen interaction; mathematical modelling; multiscale modelling

Introduction

A computational model is a simplified representation of a more complex, real system. Using models and data from the real system, one can deduct and infer properties about that system. Modelling has successfully been applied in various areas ranging from physics and economics to biology. There are numerous examples where

Jana Schleicher is a postdoctoral scientist in the Research Group Systems Biology and Bioinformatics (SBI) of the Leibniz Institute for Natural Product Research and Infection Biology—Hans Knöll Institute (HKI) in Jena, Germany. Her expertise is in both biology and mathematics. Her research focuses on mathematical modelling of molecular, cellular and organ scale processes to identify pathogenicity mechanisms.
Theresia Conrad is a PhD student at the SBI of the HKI within the Jena School for Microbial Communications. She is working on multiscale modelling of fungus–host interactions.
Mika Gustafsson is a senior lecturer at the Linköping University, Sweden, and a member of the Department of Physics, Chemistry, and Biology/Bioinformatics with expertise in bioinformatics and systems biology, especially constructing and analysing large-scale gene regulatory networks.
Gunnar Cedersund is a senior lecturer at the Linköping University, Sweden, a member of the Department of Clinical and Experimental Medicine and the Department of Biomedical Engineering. He is head of the Integrative Systems Biology Group, developed a new modelling approach—core-box modelling—and also developed a multiscale model for insulin resistance in type 2 diabetes.
Reinhard Guthke is Professor for Systems Biology at Jena University. He leads the Research Group SBI of the HKI. He is experienced in knowledge- and data-based modelling in medicine and biotechnology.
Jörg Linde is a postdoctoral scientist at the Research Group SBI of the HKI. He is experienced in genome-scale data analysis and network model inference with a focus on fungal infections and using dual RNA-Seq data.

even simple models are sufficient to draw conclusions that could not have been drawn without the models, or where previous conclusions drawn without the models have led to erroneous results [1, 2]. However, sometimes the structure of the underlying problem requires the use of more complex models: multiscale models.

Multiscale modelling is applied to systems that have important features across many orders of magnitude in time and space. For instance, computational weather forecasts became more realistic in the early 1980s by including the interactions of soil and vegetation with the atmosphere. The development of multiscale modelling started in the 1970s in various disciplines such as physics, meteorology and chemistry. This was driven by the advent of powerful computing platforms and the availability of a huge amount of measured data. In 2013, the Nobel Prize in Chemistry was awarded for the development of multiscale models of large complex chemical systems and biochemical reactions such as protein folding [3]. After 2000, with the development of more holistic approaches in biology and medicine (so-called 'systems biology' [4] and 'systems medicine'), the practice of multiscale modelling became more common in the life sciences. Its aim is to describe and support the understanding of human (patho)physiological functions. In the past few years, multiscale modelling has been applied successfully to the dynamics of the heart [5], liver [6–8], human metabolism [9–11] and immune system [12–15], which all are systems regulated at multiple scales of time and space and involve multiple compartments (e.g. cells, tissues and organs). This progress in the life sciences has been driven by the availability of a vast quantity of high-throughput measurements, so-called *omics* data, at the genome, transcriptome (i.e. microarray and RNA-Seq data), proteome and metabolome scales as well as progress in imaging technologies [16]. Walpole *et al.* [17] and Castiglione *et al.* [14] reviewed best practices in multiscale modelling of complex biological systems, coupling continuous and discrete modelling techniques. The 'Coordinating Action for the Implementation of Systems Medicine' across Europe published recommendations for multiscale modelling in systems medicine, including the establishment of ontologies, suitable information technology infrastructure and the development of standard operating procedures for data management and modelling [18].

Recently, we reviewed computational methods for modelling host–pathogen interactions (HPIs) [19]. It was highlighted that the systems biology of immune defence and pathogen activities needs to model HPI by including multiple scales. For example, models of the interplay between pathogens and immune cells have to include cellular interactions elucidated by the emerging image-based systems biology of infection [20, 21].

Current research in infection biology focuses on the involvement of multiple spatial and temporal scales in HPI as well as in the diagnosis and treatment of infections. Multiscale modelling in biology is the computational requisite for functional genomics studies with clinical applications; it is based on genome-wide approaches involving high-throughput methods rather than the more traditional 'gene-by-gene' approach. Here, systems medicine aims to develop multiscale computational models that integrate data and knowledge from the clinical and basic sciences. In other words, knowledge and data derived from *in vitro* experiments and animal models will be translated to the situation of individual patient's [18]. To cope with this task, modelling of HPI has to be carried out at different scales (Figure 1):

i. Molecular scale, including the genome, transcriptome, proteome and metabolome. This scale encompasses the interactome and complex molecular processes such as gene expression, gene regulatory networks, signalling and metabolic pathways involved in immunity and inflammation.
ii. Cellular scale, including the activities and behaviour of the different immune cells (e.g. T-cells and neutrophils) and different pathogen processes (e.g. bacteria or fungal conidia and hyphae formation).
iii. Inter-cellular and tissue scale, including inflammation processes and biofilm formation (e.g. quorum sensing mechanisms).
iv. Organ scale, including specific environmental conditions in each organ relevant for the infection process and the connection between organs (e.g. transfer of signals, toxins).
v. Body system scale, including multi-organ failure in sepsis and the population dynamics of the pathogen.

Our review provides a summary of state-of-the-art multiscale modelling of the interactions of microbial pathogens with the human host. While previous reviews mainly focus on bacterial infection, we additionally include results from the evolving modelling approach for fungal infections. Epidemiological studies and multiscale modelling of viral infections [22–24] are out of the scope of the present review. Vodovotz *et al.* ([25] and references therein) mainly focus on inflammation in the body, including multiscale models of sepsis. However, there is a lack of models considering both sides of HPI, i.e. both the pathogen and the host side. In future, research has to be focused more on these interaction, but bearing in mind that the interaction between pathogens themselves (see e.g. [26, 27]) is also important. Since the 2000s, papers have been published on multiscale modelling of bacterial HPIs (e.g. [28]), in particular for tuberculosis [29–38], whereas for fungal infections, the integration of multiple scales is currently in its early stages [39]. The low number of multiscale models simulating the interaction between a fungal pathogen and its host can be attributed to the more complex fungal genome and cell structure in comparison with bacteria, putting challenges on the development of suitable technical as well as computational approaches. But also important, research on fungal pathogens has attracted attention just in the past few decades, whereas bacterial pathogens had a longer research history. The increased attention may be attributed to the increasing infection rate of fungal pathogens [40]. We present—to our knowledge—the first overview of these early fungal HPI models. Our aim is to discover core areas for further research efforts and to identify the main challenges in the field of multiscale modelling.

The benefit of state-of-the-art multiscale models

Systems biology of microbial infections intents to describe and analyse the confrontation of a host with bacterial and fungal pathogens [41]. Therefore, the interactions of the host s immune system with components of the pathogen should be elucidated by iteratively using computational approaches and experimental studies that provide spatiotemporal data. The ultimate aim of systems biology is to unravel the key mechanisms of pathogenicity and then apply this knowledge to identify diagnostic biomarkers and potential drug targets, thereby improving the treatment of infectious diseases. For instance, multiscale and multicompartment models of tuberculosis were used for integration of data from multiple model systems over multiple length and time scales of the *in vivo* immune response to *Mycobacterium tuberculosis* [29–38]. Modelling development of decades has reached a state that allows the application of

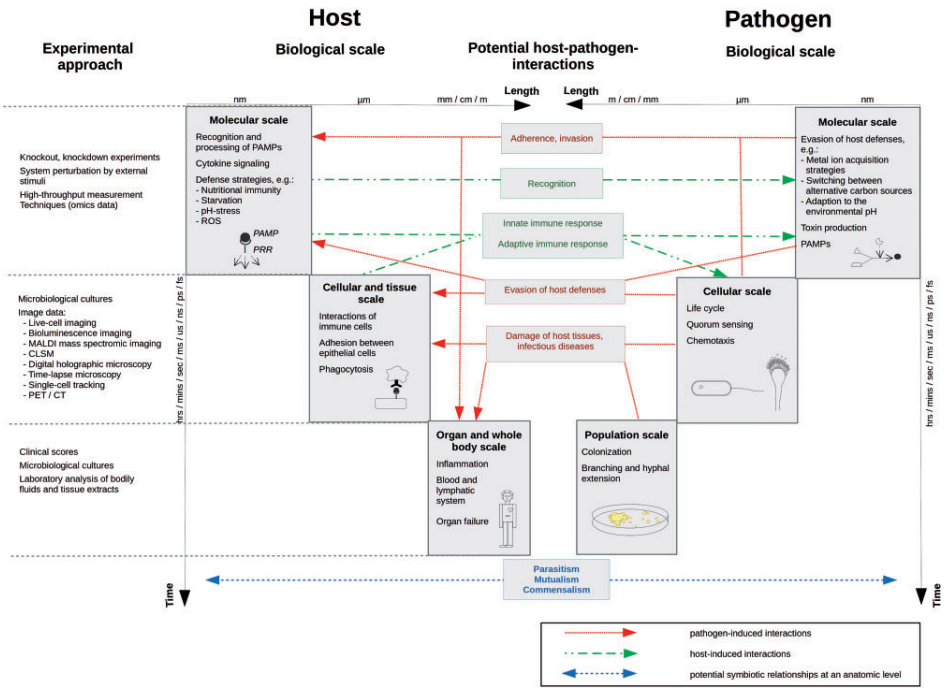


Figure 1. Schematic diagram of the complex spatiotemporal nature of HPIs, including a summary of experimental methods, which can be used at each scale. PAMP = pathogen-associated molecular pattern; PRR = pattern recognition receptor; PET = positron emission tomography; CT = computer tomography; CLSM = confocal laser scanning microscopy; MALDI = matrix-assisted laser desorption/ionization.

model predicted hypothesis in clinical settings. With help of model-based simulations, Linderman *et al.* [37] designed therapeutic interventions by immunomodulation with tumor necrosis factor alpha and interleukin 10, i.e. pro-inflammatory and anti-inflammatory cytokines, by antibiotic administration and, finally, the effect of vaccination. Cilfone *et al.* [36] applied such models to compare different therapeutic regimes for the treatment of tuberculosis. They found that inhaled formulation of the antibiotic isoniazid given at a significantly reduced dose frequency has better sterilizing efficacy and reduced toxicity than the conventional oral regimen. For modelling, they combined dynamics of lung granuloma, carrier release kinetics, pharmacokinetics and pharmacodynamics.

In general, multiscale models of HPI share the same benefits as other models in systems biology. They provide a deeper insight into the complex interplay of hosts and pathogens by providing a mechanistic understanding of the interaction network. In particular, agent-based models (ABMs) are used to model the interaction between hosts and pathogens and to improve our understanding of the underlying mechanisms (see below 'Multiscale modelling approaches of HPI'). This is because the investigated system can be modelled in a natural way by interacting individuals, and the model output allows us to capture emergent phenomena (i.e. complex patterns emerge on a higher scale through the interaction of individuals on a lower scale; [42]). HPI models can also be used to guide the setting up of experiments by *in silico* generation of hypotheses, which can be experimentally validated, frequently in an iterative cycle [43, 44]. In the field of translational systems biology, multiscale models are used to improve diagnosis of infectious diseases by biomarker discovery or to predict the clinical outcome of infections. Furthermore, such models can be used to make predictions about how a patient reacts under defined conditions or how a therapy can be optimized (i.e. therapy decision support and therapy optimization [25]).

For decades, multiscale models have been applied in pharmaceutical research and industry for drug development to predict the absorption, distribution, metabolism and excretion of synthetic or natural substances in the host [45]. Historically, such models have only been multiscale in the sense that they include both descriptions of the internal drug dynamics within an individual and the variation of key parameters across the population. In other words, in such models, which often are formulated using so-called non-linear mixed-effects models, the pharmacokinetics and pharmacodynamics are captured using simple quasi-phenomenological descriptions (PKPD models). In the past one to two decades, there has been an increasing push to also develop more realistic models, based on the physiological understanding of the involved processes. Such so-called physiologically based pharmacokinetic (PBPK) models are

compartmental and regression models, which include human or animal anatomy, physicochemical and biochemical mechanisms or toxicological effects. This push has gained further momentum through the rise of the field Systems Pharmacology, which attempts to combine intracellular systems biology models with whole-body scale PBPK models. In general, these kinds of pharmacometric models, independently of the degree of detail, have been used to successfully optimize the drug administration regimes and to extrapolate from animal models to the human host.

A typical application of multiscale models in infection biology is antibiotic administration. Frequently, different drugs with different molecular features are compared. Predictive chemistry models, namely the so-called quantitative structure–activity relationship (QSAR) models, may be integrated in multiscale models. QSAR models have been used for risk management. They are recommended by regulatory authorities for registration, evaluation, authorization and restriction of chemicals [46].

An important potential and benefit of multiscale modelling of HPI is the replacement, refinement and reduction of animal trials in research, the so-called '3Rs', by *in silico* experiments during the transition from *in vitro* experiments to clinical trials. Regarding this, a major breakthrough was recently achieved in type 1 diabetes: now the Food and Drug Administration allows for the usage of a multi-PBPK model for glucose homeostasis instead of test animals when certifying certain insulin treatments [47].

Experimental methods relevant for computational modelling

A central requirement for multiscale modelling in HPIs is the availability of suitable measurement data (Figure 1). These data are necessary to estimate model parameter values and to refine model structure, as well as to validate the models by testing the model-derived predictions.

At the molecular scale, various high-throughput measurement techniques have been developed over the past decades. Next-generation sequencing [48] allows us to assemble complete high-quality genomes of microbes, to structurally and functionally annotate genomes [49] and to identify genomic changes as risk factors on the host side [50]. Expression data can be used for diagnosis. For instance, in a genome-wide expression study, a supervised machine learning approach was applied for classification of bacterial and fungal whole-blood infections [51]. The latest advances in hybrid tandem mass spectrometry [e.g. triple quadrupole, quadrupole time-of-flight, Orbitrap hybrid mass spectrometer (tandem-in-space instruments) and ion-trapping mass spectrometers (tandem-in-time instruments)] make it possible to analyse complex proteoms with a high resolution, sensitivity and mass accuracy. In addition, various mass spectrometry imaging [e.g. matrix-assisted laser desorption/ionization (MALDI imaging)] and Raman spectroscopic imaging techniques can be used to measure the abundances and spatial distributions of proteins and metabolites in a tissue.

As eukaryotes, fungi have larger and more complex genomes than bacteria. Therefore, complete sequenced genomes of fungi were available at a later time point than bacterial genomes. Availability of the genome sequence allows identification of specific infection and interaction pathways, the discovery of drug targets, as well as species-specific microarrays. Moreover, genetic manipulations (knock-out, knock-down, overexpression) of fungi are more challenging.

A challenge in connecting the molecular scale to the cellular scale is the heterogeneous nature of biological samples, i.e. samples are composed of cell types with different gene expression profiles. In infection biology, this issue is most pronounced for organ samples (e.g. lung, liver and brain) and blood assays. To deal with mixed samples in gene expression analyses, in the past decade, several groups developed expression deconvolution algorithms, e.g. [52–57]. These algorithms allow the extraction of information on a cell-based scale from heterogeneous biological samples (for an introduction see [58, 59]). A variety of these algorithms were combined in the R package CellMix [60], which allows for an efficient estimation of cell type proportions and cell type-specific expression profiles in mixed samples. Similarly, the R package DeconRNASeq also enables deconvolution of mRNA-Seq data from mixed samples [61].

For storage and access of *omics* data, several data repositories are available (e.g. GenBank, Gene Expression Omnibus, ArrayExpress, PRIDE). Other repositories provide knowledge on functional genomics, i.e. genome annotation of both hosts and pathogens [19, 39, 62]. The database PHISTO, a web-based HPI search tool, stores known molecular relations between pathogens and the human host, extracted by text mining from scientific papers [63]. Such molecular biological databases have been used to infer interolog-based networks for the molecular interaction of the pathogen *Candida albicans* with its animal and human hosts [64, 65].

While advances in *omics* techniques drive the progress of multiscale modelling on the molecular scale, there has also been significant progress on the cellular scale based on imaging data from positron emission tomography/computer tomography, bioluminescence imaging, confocal laser scanning microscopy, live cell imaging, time-lapse microscopy, single-cell tracking, digital holographic microscopy and MALDI mass spectrometric imaging. Although the automated analysis of image and video data from HPI remains a challenging task [66–68], it holds great potential because it automatically extracts important parameters such as velocity or turning angles for individual cells. Moreover, automatic analysis identifies interactions between individual host and pathogen cells, such as touching events, adherence or phagocytosis. Such data drive the emerging image-based systems biology of infection [20, 21]. The integration of both *omics* and image-based sub-models in multiscale models is challenging owing to the requirement of combining different modelling techniques. Here, an outstanding task is to combine the non-spatial *omics* data with the image-based sub-models that generally have an inherent spatial scale.

For modelling at the cellular, tissue and organ scales, biomechanical, rheological and physicochemical parameters become important. For example, cytometric data and data quantifying the deformability of erythrocytes (e.g. *Plasmodium falciparum*-parasitized red blood cells) were analysed and modelled using a particle-based simulation technique (i.e. dissipative particle dynamics) for different stages of malaria [69].

In general, the analysis tools to investigate HPI at the cellular, tissue, organ and whole body scales stem from various medical disciplines such as radiology, clinical/medical microbiology, clinical immunology, cytopathology, clinical chemistry/medical biochemistry, haematology and clinical pathology. Diagnostics of infectious diseases affecting the whole body are based on the laboratory analysis of body fluids, such as blood, urine, sputum and tissue extracts by macroscopic or microscopic analysis. Clinical scores summarize the status of an infection by

Table 1: A selection of modelling approaches used to examine HPIs

Reference	Host	Pathogen	Scales*	Time-independent modelling approach		Continuous time modelling approach		Discrete time modelling approach				
				Constraint-based	Game theory	ODEs	PDEs	Agent-based	State-based	Cellular automata-based	Boolean	Probabilistic
Single application of the main modelling approach												
Thakar *et al.* [74]	Mammalian host	*Bordetella bronchiseptica*	1: Cytokines 2: Immune cells 3: Lung, lymph nodes, bacterial growth	x								
Hummert *et al.* [75]	Human host	*Candida albicans*	2: Macrophages, ingested yeast cells, fungal survival strategies 3: Fungal growth (fitness)		x							
Eswarappa [76]	Mammalian host	Pathogenic bacteria (persistent infections)	1, 2: Extra-, intra-cellular compartments and defence mechanisms 3: Bacterial growth		x							
Tierney *et al.* [77]	Murine host	*Candida albicans*	1: Gene expression, cytokines 2: Innate immune cells			x						
Boswell *et al.* [78]	Plant root system	*Rhizoctonia solani*	1: Fungal uptake of external substrate 2: Growth of fungal mycelia				x					
Tokarski *et al.* [79]	Human host	*Aspergillus fumigatus*	1: Chemical communication 2: Phagocytes, chemotaxis, clearing efficiency, conidia, lung 3: Fitness					x				
Hünniger *et al.* [80]	Human host	*Candida albicans*	1: Antifungal effector molecules, cytokines 2: Immune cells, whole-blood 3: Distribution of fungal cells						x			
Wcislo *et al.* [81]	Wheat	*Fusarium graminearum (invasion, colonization)*	1: Nutrient concentrations, secreted substances 2: Plant cells, fungal cells 3: Fungal growth							x		
Thakar and Albert [82]	Mammalian host	*Helicobacter pylori, Bordetella bronchiseptica, Bordetella pertussis*	1: Cytokines, antibodies 2: Immune cells, bacterial cells								x	
Grant *et al.* [83]	Murine host	*Salmonella enterica*	1: Bacterial genetic diversity 2: Infected cells 3: Liver, spleen, blood; bacterial growth and death									x

(continued)

Table 1. Continued

Reference	Host	Pathogen	Scales*	Time-independent modelling approach		Continuous time modelling approach		Discrete time modelling approach				
				Constraint-based	Game theory	ODEs	PDEs	Agent-based	State-based	Cellular automata-based	Boolean	Pro-babilistic
Combined application of different modelling approaches												
Cilfone *et al.* [36]	Non-human primates	*Mycobacterium tuberculosis*	1: Cytokines, granuloma function, antibiotics (carrier) 2: Immune cells, granulomas formation, receptor-ligand dynamics, lung			x	x	x				
Linderman *et al.* [37]	Non-human Primates	*Mycobacterium tuberculosis*	1: Cytokines, antibodies, granuloma function, antibiotics 2: Immune cells, granuloma formation, receptor-ligand dynamics, lung 3: Bacterial growth			x	x	x				
Tyc [84]	Human host	*Candida albicans*	1: Drug treatment, environmental conditions, virulence factors 2: Immune cells, virulence factors 3: Fungal growth and phenotypes		x	x		x				
Carbo *et al.* [85]	Murine host	*Helicobacter pylori*	1: Virulence factors, cytokines 2: Immune cells, gastric lumen, epithelium, lamina propria, lymph nodes 3: Gastric mucosa, bacterial colonization			x		x				
Pollmächer *et al.* [86]	Human host	*Aspergillus fumigatus*	1: Chemokines 2: Leucocytes, alveoli, conidia					x**				

Besides introducing models that only use one modelling approach to simulate various scales, we also provide references of models in which multiple approaches were combined. These models provide valuable ideas how the problem of combining different models of various scales can be sorted out.

ODE = ordinary differential equation; PDE = partial differential equation.

*1: Molecular scale; 2: Cellular and tissue scale; 3: Organ and whole body scale, population scale.

**ABM combination of migration and interaction in continuous space with spatio-temporal modelling on a discrete grid.

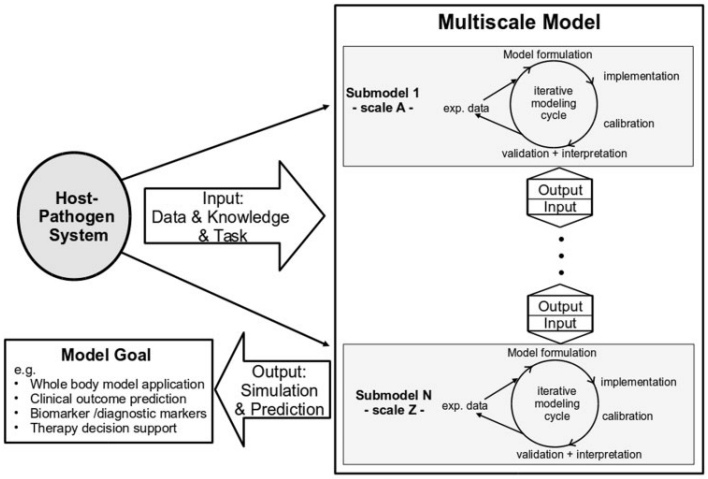


Figure 2. Schematic overview of a multiscale model structure. Sub-models on various scales are used to examine multiscale HPIs. In each sub-model the iterative cycle of modelling and experimental calibration and validation has to be passed through.

combining clinical parameters with observation of the infected individual. For example, the 'Clinical Pulmonary Infection Score' (CPIS) and the 'Sepsis-related Organ Failure Assessment Score' (SOFA) [70], used to classify patients with severe sepsis, are sometimes recorded as easily attainable data on the host side. Clinical scores are also used in animal trials of infection studies and usually include body temperature, weight, activity and feeding patterns. In contrast, pathogens are identified in the laboratory using microbiological cultures. The multiscale model-based personalized treatment of infectious diseases will be based on the stratification of patients by analysing both observed clinical phenomena of physiologic variability and molecular patterns that characterize the immunological state.

In general, for experimental validation of multiscale models, adequate experimental systems that focus on individual modules of interest are needed. To make model results reliable and useful, for example for clinics, verification of a multiscale model has to be conducted on each implemented scale. Therefore, it is necessary to obtain experimental data from the different spatio-temporal scales (Figure 1). In the future, so-called microphysiological systems, e.g. organs-on-chips or tissue-engineered 3D organ constructs that use human cells, provide an alternative to animal testing [71]. On the bioinformatics side of validation, Pârvu and Gilbert [72] developed a methodology for automatic validation of multiscale computational models.

Multiscale modelling approaches of HPI

Dada and Mendes [73] and Walpole *et al.* [17] have characterized the main modelling approaches; here, we provide an overview of their application in multiscale HPI modelling (Table 1). Simple modelling approaches, which can include multiple scales of HPI but neglect time, are (evolutionary) game theoretical concepts and constraint-based models. Often, HPI modelling requires the behaviour of the simulated system over time to be considered (e.g. with regard to infection time or time for immune response). In dynamic modelling, a system can be simulated in a continuous or discrete-time context, depending on the model aim and the chosen computational approach. In this section, we start by reviewing time-independent models of HPI, followed by models in continuous time and models using discrete time. Finally, combining the advantages of different modelling approaches on different scales offers an opportunity for multiscale modelling. Thus, we introduce mixed models linking different modelling approaches and exemplify their application to HPI.

In general, game theory concepts [87] are used to examine the possible outcomes of interactions, in which real world entities are represented as 'players' who take part in a 'game' with the aim of optimizing some sort of pay-off. Players can choose between different strategies. To find an optimal solution for the game, the approach takes into account the costs and benefits of each strategy in relation to the strategy chosen by the other player. The application of this concept to evolving organisms or populations is termed evolutionary game theory. With this approach the evolutionary dynamics of strategy changes of interacting species can be examined depending on the frequencies of strategies and the fitness gain for each strategy. As a recently published example, Li *et al.* [88] studied the *in vitro* population dynamics of two commensal bacteria that synergistically protect the metazoan host *Hydra vulgaris* from fungal infection. Another example is the modelling of interplay of drug-resistant and drug-sensitive pathogens under antibiotic treatment [89]. In HPI, evolutionary game theory may be used in future to elucidate the adaptation of evasion strategies of pathogens or defence strategies of the host over time. With a more phenotypic and generalized view, game theory can also be applied to model the interaction between pathogens and humans or the interaction between different pathogens. For example, this approach was used to understand what advantages the human fungal pathogen *C. albicans* experiences by changing its morphological form [75, 90] in the context of interacting with the

host's immune cells. Additionally, persistent bacterial infection was described by developing a game theoretical model; predictions regarding persistent bacterial infections were drawn by considering the ability of a pathogen to survive extracellularly and intracellularly, within an immune cell [76].

Another knowledge-based approach for large-scale modelling is so-called constraint-based modelling. The idea of constraint-based modelling is to describe a biological system by a set of knowledge-based constraints, which characterize its possible behaviours but in general do not allow a precise prediction to be made. This modelling approach has mainly been applied to the modelling of metabolic networks [91–93]. Jamshidi and Raghunathan [94] outlined a systematic procedure to produce constraint-based HPI models. Interestingly, a constraint-based network model of HPI was also presented to describe the dynamic outcome of the interplay between host immune components and *Bordetella bronchiseptica* virulence factors [74].

Typical modelling approaches using a continuous time context consist of ordinary differential equations (ODEs) or, if space is included in the model, partial differential equations. ODE-based modelling is widely used at the molecular scale, such as for gene regulatory network models [95]. Here, gene expression analysis by RNA-Seq offers the opportunity to monitor and model the transcriptome of both the pathogen and host, as shown for the interaction of *C. albicans* with murine dendritic cells using an ODE-based approach [77]. Generalizing this work, methods for exploiting dual RNA-Seq data for the inference of gene regulatory networks of HPI has been presented [96].

In addition to their utility at the molecular scale, ODE modelling is also applicable for whole cell simulations and at the body scale (e.g. pathogen population scale; PBPK models [45]). Palsson *et al.* [13] published a fully integrated immune response model (FIRM) consisting of multiple sub-models, a multi-organ structure, circulating blood, lymphoid tissue, different immune cell types and cytokines and immune cell recruitment. FIRM was tested by simulating the response to a blood-borne pathogen (i.e. tuberculosis infection). An ODE-based simulator has the flexibility to be expanded. It is suitable for step-by-step interactive integration of further sub-models, describing the processes within the pathogen and their interaction with the host. FIRM may be a starting point for multiscale modelling of HPI.

In addition, the Lotka-Volterra model, well known for simulations of predator–prey interactions, can be used for multiscale modelling of HPI. The system consists primarily of a pair of first-order, non-linear ODEs, but the equations can be generalized to include, for example, trophic interactions, spatial structures and more than two species (e.g. [97]). Stein *et al.* [98] studied the dynamic stability of intestinal microbiota by use of a generalized Lotka-Volterra model for focal species to account for external perturbations representing antibiotics or diet.

Some aspects of HPI require the application of discrete time intervals, therefore permitting the use of agent-based [99], state-based [80] and cellular automata-based [81] Boolean [82] or probabilistic models [83, 100]. These approaches were used to model the HPI taking into account individual genes or cells (e.g. immune and pathogen cells) in time and, partly, space [20, 21].

In an ABM, the behaviour and interaction of autonomous agents are simulated over time to examine the emergence of complex phenomena on a higher scale. Each agent gets a set of rules determining its method of interaction and behaviour, thus making ABMs a promising tool for studying HPI and, more generally, infectious diseases and inflammatory processes [101]. The advantage of the agent-based modelling approach is the possibility of relatively easily integrating space (e.g. as a discrete grid) and, additionally, accounting for variability (e.g. in behaviour or movement) among individual cells and/or cell types.

HPI in anastomotic leaks was examined by using the agent-based modelling approach [102]. An ABM of epithelial restitution was augmented by individual *Pseudomonas aeruginosa* agents interacting with the epithelium. The simulation of different killing mechanisms leads to a mechanistic understanding of tissue destruction.

An agent-based approach was also used by Tokarski *et al.* [79] to investigate the clearance efficiency of *Aspergillus fumigatus* conidia by neutrophil granulocytes. A combination of live imaging and grid-based modelling of individual cells allows *in silico* testing of different hypotheses for hunting strategies of immune cells. This modelling approach demonstrated that chemokine sensing by immune cells is the most efficient strategy. The ABM was implemented in the free software tool NetLogo [103, 104]. This well-established tool facilitates a user-friendly and efficient programming of ABMs. SPARK (Simple Platform for Agent-based Representation of Knowledge) is an alternative tool for multiscale ABMs that runs faster [105].

Besides the ABM approach, theoretical modelling in discrete time can also be realized by the use of Boolean networks [106]. In the past, Boolean models were developed to describe and simulate within-host immune interactions (reviewed by [82]). This heuristic modelling approach allows prediction of new interaction pathways or drug targets within the host–pathogen infection system. For example, a Boolean modelling technique was applied to model the signal transduction of the hepatocyte growth factor pathway of the human host in response to infection by *Helicobacter pylori* [107]. This model predicts new molecular targets against *H. pylori* infection, which were experimentally verified.

The combined application of different modelling approaches on multiple scales may facilitate multiscale modelling of HPI. As a prominent example, for multiscale modelling of *M. tuberculosis* infection, a system of ODEs to capture intracellular signalling pathways was combined with a discrete probabilistic ABM that describes cellular behaviour at the tissue scale [29–38] and references therein). Also, for the interaction of *C. albicans* with the human host, ODE-based, agent-based and game theory-based modelling methods were compared and partially combined [84].

A combination of ODE-based modelling with ABM was used to model the mucosal responses during *H. pylori* infection [85]. This hybrid model considers immune effector cells (i.e. macrophages, T-helper cells and pro-inflammatory epithelial cells) that secrete cytokines and chemokines, which recruit immune cells and promote their activation and differentiation to inflammatory phenotypes, and, finally, secrete effector molecules that destroy bacteria and may cause tissue damage.

A multiscale model simulating the distribution of chemokine concentrations in *A. fumigatus*-infected human alveoli was developed by combining an ABM of migration and interaction in continuous space with spatiotemporal modelling on a discrete grid [86].

In general, multiscale modelling has to reuse and link different sub-models (Figure 2). This requires a multiscale computational infrastructure and (sub-)model repositories. The systems biology markup language is the most developed standard concept at the moment and is increasingly used to support the exchange of models in the modelling community. In future, this concept may be expanded to support also multiscale models.

Challenges and outlook

Systems biology of microbial infection encompasses all scales of the pathogen and the host's immune system, leading to a complex interaction network on multiple scales. A common challenge in systems biology is the successful combination of both experimental and theoretical approaches. In this context, an iterative cycle should be applied in which model development and refinement, parameter calibration and *in silico* experiments alternate with experimental data collection and hypothesis/prediction validation (Figure 2). The application of an iterative cycle of experiments and model refinement for fungal pathogens is at the moment connected to more effort than for bacterial pathogens. In bacteria, genes of the same pathway and also virulence genes are often clustered together in an operon allowing a shared regulation. This clustering facilitates the study of regulatory mechanisms and enables a relatively simple mathematical replication of the regulatory processes in bacteria. Such structured, regulatory units are not present in fungi which makes it more difficult to find virulence genes, to understand their regulation and finally to develop a representative mathematical model of the regulatory network. In addition, in fungi there are secondary metabolite gene clusters characterized by complex structures of co-regulation [108, 109].

Moreover, fungi have complex life cycles with multiple morphological forms. They may occur as unicellular yeast or in a filamentous form (dimorphism). This indicates the need of implementing a broader spatial scale in fungal models than in bacterial interaction models. Furthermore, fungi have developed multiple sophisticated, specific and unique pathogenicity mechanisms including immune evasion strategies. Only a few of them are modelled by game-theoretical methods and ABMs [110] and many of the pathogenicity mechanisms are not well understood, e.g. the production of hydrophobins on the spore surface as an immune evasion strategy of the environmental fungus *A. fumigatus*.

At the start of developing a multiscale model of a complex biological system, researchers have to bear in mind that the aim of a model is not to completely mirror the real system. Essentially, the most important aspect of modelling relates to the wise reduction of the complexity of the investigated system to identify key properties. The parts to be implemented in a model and the parts to be left out are dictated by the biological question(s) that will be addressed with the model. Kirschner *et al.* proposed a 'tunable resolution' for multiscale models [111], in which sub-models at different scales are defined and connected. In case a specific question requires additional parameters, these can be added to one of the sub-models. *Vice versa*, more coarse-grained sub-models can be applied if the details on lower scales are not needed for the question in focus.

A further challenge in multiscale modelling of HPI is the combination of different time scales. While regulatory interactions on the transcriptomic scale take place in minutes, it may cause effects on the cell-, tissue- or organ-scale hours or days later. Approaches allowing a transition from one time scale to another need to be developed. For example, Chaves *et al.* [112] presented three asynchronous algorithms to meet this challenge for genetic regulatory networks using the example of Boolean models, which could also be applicable to multiscale problems.

Moreover, the number of features per scale is variable. The human body consists of several organs; each organ itself consists of millions of cells, and each cell has several thousands of proteins and transcripts. Thus, to develop an efficient multiscale model, stringent feature selection with a strong focus on the simulated phenomenon and model aim is essential.

A large number of parameters is an integral part of multiscale models, but the many parameters are also associated with some problems that must be overcome. The allowance of many parameters has positive implications, because it allows for a more realistic description of the system. In contrast, few parameters and minimal models may often imply that overly simplified and lumped descriptions of states and processes have to be used, which may be hard to interpret physiologically. One of the reasons why a high degree of parameters is negative is that it is hard to ensure that all of them have realistic values. Granted, some of the parameters may have values that can be determined in independent experiments, but in biology, and especially for large multiscale models, there are often many parameters that have to be inferred simultaneously from systems-wide dynamic data. This determination is often not unique: a problem known as parameter unidentifiability [113]. If untreated, such unidentifiability implies that all model predictions come with an arbitrarily large uncertainty range, i.e. the predictions are suggestions and not unique consequences of the model and data. Fortunately, recent progress in model analysis has allowed for the identification of such uniquely inferred predictions. Such predictions, sometimes called core predictions, are predictions that are uniquely determined from the data, even though the parameter values are non-unique. In practice, a three-step approach has been proposed by Cedersund, which allows for the accurate identification of the outer boundaries of such predictions [113]. Nevertheless, these methods are still only applicable to small- and medium-sized models, ranging between 1 and 50 parameters. For truly large multiscale models, further method developments are needed.

For ethical reasons, experiments designed to calibrate parameter values in a human model might not be realizable in the human body or in animals. In these cases, the parameter values have to be identified using sub-models and exploiting data measured under *in vitro* conditions despite the fact that the value may be different *in vivo*. The same may be true for model validation, especially for HPI models. Suitable *in vitro* systems that mimic the pathophysiology of infection in humans have to be used.

In summary, HPIs should be described by a combination of spatiotemporal models with interacting molecular networks of both the host and the pathogen. Considerable advances in multiscale modelling of microbial HPI have been made in the study of tuberculosis. In this case, and for other human infections including fungal infection by *C. albicans* and *A. fumigatus*, ODE-based, state-based and ABMs are the main techniques for successful modelling at the molecular and cellular scales. In the future, high-throughput *omics* and image data should be simultaneously considered and modelled in an integrated manner.

A promising approach to multiscale modelling is hierarchical modelling (see also Figure 2), in which the sub-models at each scale appear in a well-defined place in a super-model, in a so-called tree-structure. Such models have a natural modular structure, where one version of sub-models can be replaced for one another, to better suit the particular data and question that is studied. This approach has been relatively well-developed in technical systems, and in biological systems an important application involves glucose homeostasis and diabetes [9]. By

combining input–output data with the data for a sub-module, one can consider the modelling of a sub-module as an isolated modelling problem. Any model for the sub-module can be expected to fit into the dynamics of the super-model as long as the sub-model reproduces the measured output when exposed to the measured input. Thus, these input–output data also allow for a meaningful way of combining both *in vitro* and *in vivo* data. This approach was presented in [9] and [10], and in [11] the approach was used to unravel both where insulin resistance appears inside individual fat cells, and how this resistance spreads to the rest of the body.

Regarding hierarchical modelling of HPI, the software tool SPARKS [106] might be useful, but the development of easy-to-use software applications for multiscale modelling will be an important task in the coming decades. Currently, there are further initiatives to establish a computational framework for multiscale modelling [114]. Andasari *et al.* [115] presented a multiscale, individual-based simulation environment that integrates a lattice-based Cellular Potts Model on the cellular scale (CompuCell3D) and an ODE-based Bionetsolver for intracellular modelling of reaction-kinetic network dynamics. This hybrid system has been applied to cancer research. The system may also be suitable for HPI modelling. Furthermore, the WholeCellKB is an open-source web-based software program for multiscale *omics* modelling and, in particular, WholeCellKB-MG enables whole-cell modelling of the human pathogen *Mycoplasma genitalium* by integrating diverse data sources into a single database [116, 117].

From a systems medicine perspective, the multi-layered HPI models should ideally also make use of all the available clinical information at hand, such as information regarding a patient's disease history and life-style factors, which probably will be extended to genotype information in the future. These factors may affect the system's response to stimuli via differences in the initial conditions of the state variables or by altering the effective regulatory interactions.

Finally, we want to highlight the fact that interactions also take place between different populations of bacteria and fungi (e.g. quorum sensing), which may positively or negatively influence the infection process in the host and determine the outcome of HPI (e.g. [118]). It is a future task to develop models that account for the interactions among three or more species (i.e. parasites, fungi, bacteria, viruses and host) by integrating the species and their unique characteristics at various temporal and spatial scales. Additionally, multiscale models need to combine models for intra-species communication (e.g. [26, 27]) with models for HPI.

Key Points

- Multiscale modelling of host–pathogen interactions has to consider processes at various temporal and spatial resolutions.
- The scales range from minutes to days and include the molecular, cellular, tissue and organism scales of the host and the molecular, cellular and population scales of the pathogen.
- Integrating across these scales requires multiple modelling approaches, such as ordinary and partial differential equations, state-based models and agent-based models.

Funding

This work was supported by the Jena School for Microbial Communication [to T.C.] and the Deutsche Forschungsgemeinschaft CRC/Transregio 124 'Pathogenic fungi and their human host: Networks of interaction' subproject INF [to J.L., R.G.].

References

1. Jullesson D, Johansson R, Rajan MR, *et al.* Dominant negative inhibition data should be analyzed using mathematical modeling–re-interpreting data from insulin signaling. *FEBS J* 2015;**282**(4):788–802.
2. Nyman E, Lindgren I, Lövfors W, *et al.* Mathematical modeling improves EC50 estimations from classical dose-response curves. *FEBS J* 2015;**282**(5):951–62.
3. Fersht AR. Profile of Martin Karplus, Michael Levitt, and Arieh Warshel, 2013 nobel laureates in chemistry. *Proc Natl Acad Sci USA* 2013;**110**(49):19656–7.
4. Kitano H. Systems biology: a brief overview. *Science* 2002;**295**(5560):1662–4.
5. Qu Z, Garfinkel A, Weiss JN, Nivala M. Multi-scale modeling in biology: how to bridge the gaps between scales? *Prog Biophys Mol Biol* 2011;**107**(1):21–31.
6. Schliess F, Hoehme S, Henkel SG, *et al.* Integrated metabolic spatial-temporal model for the prediction of ammonia detoxification during liver damage and regeneration. *Hepatology* 2014;**60**(6):2040–51.
7. Kuepfer L, Kerb R, Henney AM. Clinical translation in the virtual liver network. *CPT Pharmacometrics Syst Pharmacol* 2014;**3**:e127.
8. Schwen LO, Schenk A, Kreutz C, *et al.* Representative sinusoids for hepatic four-scale pharmacokinetics simulations. *PLoS One* 2015;**10**(7):e0133653.
9. Cedersund G, Strålfors P. Putting the pieces together in diabetes research: towards a hierarchical model of whole-body glucose homeostasis. *Eur J Pharm Sci* 2009;**36**(1):91–104.
10. Nyman E, Brännmark C, Palmér R, *et al.* A hierarchical whole-body modeling approach elucidates the link between *in vitro* insulin signaling and *in vivo* glucose homeostasis. *J Biol Chem* 2011;**286**(29):26028–41.
11. Brännmark C, Nyman E, Fagerholm S, *et al.* Insulin signaling in type 2 diabetes: experimental and modeling analyses reveal mechanisms of insulin resistance in human adipocytes. *J Biol Chem* 2013;**288**(14):9867–80.
12. Santoni D, Pedicini M, Castiglione F. Implementation of a regulatory gene network to simulate the TH1/2 differentiation in an agent-based model of hypersensitivity reactions. *Bioinformatics* 2008;**24**(11):1374–80.
13. Palsson S, Hickling TP, Bradshaw-Pierce EL, *et al.* The development of a fully-integrated immune response model (FIRM) simulator of the immune response through integration of multiple subset models. *BMC Syst Biol* 2013;**7**:95.
14. Castiglione F, Pappalardo F, Bianca C, *et al.* Modeling biology spanning different scales: an open challenge. *Biomed Res Int* 2014;**2014**:902545.
15. Azhar N, Vodovotz Y. Innate immunity in disease: insights from mathematical modeling and analysis. *Adv Exp Med Biol* 2014;**844**:227–43.
16. Watabe M, Arjunan SN V, Fukushima S, *et al.* A computational framework for bioimaging simulation. *PLoS One* 2015;**10**(7):e0130089.

17. Walpole J, Papin JA, Peirce SM. Multiscale computational models of complex biological systems. *Annu Rev Biomed Eng* 2013;**15**:137–54.
18. Wolkenhauer O, Auffray C, Brass O, *et al*. Enabling multiscale modeling in systems medicine. *Genome Med* 2014;**6**(3):21.
19. Durmuş S, Çakir T, Özgür A, Guthke R. A review on computational systems biology of pathogen-host interactions. *Front Microbiol* 2015;**6**:235.
20. Figge MT, Murphy RF. Image-based systems biology. *Cytometry A* 2015;**87**(6):459–61.
21. Medyukhina A, Timme S, Mokhtari Z, Figge MT. Image-based systems biology of infection. *Cytometry A* 2015;**87**(6):462–70.
22. Heldt FS, Frensing T, Pflugmacher A, *et al*. Multiscale modeling of influenza A virus infection supports the development of direct-acting antivirals. *PLoS Comput Biol* 2013;**9**(11):e1003372.
23. Rong L, Perelson AS. Mathematical analysis of multiscale models for hepatitis C virus dynamics under therapy with direct-acting antiviral agents. *Math Biosci* 2013;**245**(1):22–30.
24. Murillo LN, Murillo MS, Perelson AS. Towards multiscale modeling of influenza infection. *J Theor Biol* 2013;**332**:267–90.
25. Vodovotz Y, An G. *Translational Systems Biology. Concepts and Practice for the Future of Biomedical Research*. Elsevier Science Publishing Co Inc, London, UK, 2014.
26. Ben-Jacob E, Becker I, Shapira Y, *et al*. Bacterial linguistic communication and social intelligence. *Trends Microbiol* 2004;**12**:366–72.
27. Ben-Jacob E. Social behavior of bacteria: from physics to complex organization. *Eur Phys J* 2008;**3**:315–22.
28. Foteinou PT, Calvano SE, Lowry SF, Androulakis IP. Multiscale model for the assessment of autonomic dysfunction in human endotoxemia. *Physiol Genomics* 2010;**42**(1):5–19.
29. Wigginton JE, Kirschner D. A model to predict cell-mediated immune regulatory mechanisms during human infection with *Mycobacterium tuberculosis*. *J Immunol* 2001;**166**(3):1951–67.
30. Fallahi-Sichani M, El-Kebir M, Marino S, *et al*. Multiscale computational modeling reveals a critical role for TNF-α receptor 1 dynamics in tuberculosis granuloma formation. *J Immunol* 2011;**186**(6):3472–83.
31. Marino S, Linderman JJ, Kirschner DE. A multifaceted approach to modeling the immune response in tuberculosis. *Wiley Interdiscip Rev Syst Biol Med* 2011;**3**(4):479–89.
32. Fallahi-Sichani M, Kirschner DE, Linderman JJ. NF-κB signaling dynamics play a key role in infection control in tuberculosis. *Front Physiol* 2012;**3**:170.
33. Fallahi-Sichani M, Flynn JL, Linderman JJ, Kirschner DE. Differential risk of tuberculosis reactivation among anti-TNF therapies is due to drug binding kinetics and permeability. *J Immunol* 2012;**188**(7):3169–78.
34. Ghosh S, Baloni P, Mukherjee S, *et al*. A multi-level multiscale approach to study essential genes in *Mycobacterium tuberculosis*. *BMC Syst Biol* 2013;**7**(1):132.
35. Cilfone NA, Perry CR, Kirschner DE, Linderman JJ. Multi-scale modeling predicts a balance of tumor necrosis factor-α and interleukin-10 controls the granuloma environment during *Mycobacterium tuberculosis* infection. *PLoS One* 2013;**8**(7):e68680.
36. Cilfone NA, Pienaar E, Thurber G, *et al*. Systems pharmacology approach toward the design of inhaled formulations of rifampicin and isoniazid for treatment of tuberculosis. *CPT Pharmacometrics Syst Pharmacol* 2015;**4**(3):193–203.
37. Linderman JJ, Cilfone NA, Pienaar E, *et al*. A multi-scale approach to designing therapeutics for tuberculosis. *Integr Biol* 2015;**7**(5):591–609.
38. Pienaar E, Cilfone NA, Lin PL, *et al*. A computational tool integrating host immunity with antibiotic dynamics to study tuberculosis treatment. *J Theor Biol* 2015;**367**:166–79.
39. Horn F, Heinekamp T, Kniemeyer O, *et al*. Systems biology of fungal infection. *Front Microbiol* 2012;**3**:1–20.
40. Rolston K. Overview of systemic fungal infections. *Oncology (Williston Park)* 2001;**15**(11 Suppl 9):11–4.
41. Guthke R, Linde J, Mech F, Figge MT. Systems biology of microbial infection. *Front Microbiol* 2012;**3**:328.
42. Bonabeau E. Agent-based modeling: methods and techniques for simulating human systems. *Proc Natl Acad Sci USA* 2002;**99**(Suppl 3):7280–7.
43. Somvanshi PR, Venkatesh KV. A conceptual review on systems biology in health and diseases: from biological networks to modern therapeutics. *Syst Synth Biol* 2014;**8**(1):99–116.
44. Cedersund G, Roll J. Systems biology: model based evaluation and comparison of potential explanations for given biological data. *FEBS J* 2009;**276**(4):903–22.
45. Jones HM, Chen Y, Gibson C, *et al*. Physiologically based pharmacokinetic modeling in drug discovery and development: a pharmaceutical industry perspective. *Clin Pharmacol Ther* 2015;**97**(3):247–62.
46. Dearden JC, Rowe PH. Use of artificial neural networks in the QSAR prediction of physicochemical properties and toxicities for REACH legislation. *Methods Mol Biol* 2015;**1260**:65–88.
47. Kovatchev B, Breton M, DallaMan C, Cobelli C. *In silico Model and Computer Simulation Environment Approximating the Human Glucose/Insulin Utilization*. Food Drug Administration Master File MAF-1521, Silver Spring, USA, 2008.
48. Mardis ER. Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet* 2008;**9**:387–402.
49. Linde J, Duggan S, Weber M, *et al*. Defining the transcriptomic landscape of *Candida glabrata* by RNA-Seq. *Nucleic Acids Res* 2015;**43**(3):1392–406.
50. Liu JZ, Hov JR, Folseraas T, *et al*. Dense genotyping of immune-related disease regions identifies nine new risk loci for primary sclerosing cholangitis. *Nat Genet* 2013;**45**(6):670–5.
51. Dix A, Hünniger K, Weber M, *et al*. Biomarker-based classification of bacterial and fungal whole-blood infections in a genome-wide expression study. *Front Microbiol* 2015;**6**:171.
52. Bar-Joseph Z, Farkash S, Gifford DK, *et al*. Deconvolving cell cycle expression data with complementary information. *Bioinformatics* 2004;**20**(Suppl 1):i23–30.
53. Hoffmann M, Pohlers D, Koczan D, *et al*. Robust computational reconstitution - a new method for the comparative analysis of gene expression in tissues and isolated cell fractions. *BMC Bioinformatics* 2006;**7**:369.
54. Clarke J, Seo P, Clarke B. Statistical expression deconvolution from mixed tissue samples. *Bioinformatics* 2010;**26**(8):1043–9.
55. Shen-Orr SS, Tibshirani R, Khatri P, *et al*. Cell type-specific gene expression differences in complex tissues. *Nat Methods* 2010;**7**(4):287–9.
56. Gong T, Hartmann N, Kohane IS, *et al*. Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complex clinical blood samples. *PLoS One* 2011;**6**(11):e27156.
57. Li Y, Xie X. A mixture model for expression deconvolution from RNA-seq in heterogeneous tissues. *BMC Bioinformatics* 2013;**14**(Suppl 5):S11.

58. Zhao Y, Simon R. Gene expression deconvolution in clinical samples. *Genome Med* 2010;**2**(12):93.
59. Shen-Orr SS, Gaujoux R. Computational deconvolution: extracting cell type-specific information from heterogeneous samples. *Curr Opin Immunol* 2013;**25**(5):571–8.
60. Gaujoux R, Seoighe C. CellMix: a comprehensive toolbox for gene expression deconvolution. *Bioinformatics* 2013;**29**(17):2211–2.
61. Gong T, Szustakowski JD. DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. *Bioinformatics* 2013;**29**(8):1083–5.
62. Galperin MY, Rigden DJ, Fernández-Suárez XM. The 2015 Nucleic Acids Research Database Issue and molecular biology database collection. *Nucleic Acids Res* 2015;**43**(Database issue):D1–5.
63. Durmuş Tekir SD, Ülgen KÖ. Systems biology of pathogen-host interaction: networks of protein-protein interaction within pathogens and pathogen-human interactions in the post-genomic era. *Biotechnol J* 2013;**8**(1):85–96.
64. Wang Y-C, Lin C, Chuang M-T, *et al.* Interspecies protein-protein interaction network construction for characterization of host-pathogen interactions: a *Candida albicans*-zebrafish interaction study. *BMC Syst Biol* 2013;**7**:79.
65. Remmele CW, Luther CH, Balkenhol J, *et al.* Integrated inference and evaluation of host-fungi interaction networks. *Front Microbiol* 2015;**6**:764.
66. Mech F, Thywissen A, Guthke R, *et al.* Automated image analysis of the host-pathogen interaction between phagocytes and *Aspergillus fumigatus*. *PLoS One* 2011;**6**(5):e19591.
67. Brandes S, Mokhtari Z, Essig F, *et al.* Automated segmentation and tracking of non-rigid objects in time-lapse microscopy videos of polymorphonuclear neutrophils. *Med Image Anal* 2015;**20**(1):34–51.
68. Kraibooj K, Schoeler H, Svensson C-M, *et al.* Automated quantification of the phagocytosis of *Aspergillus fumigatus* conidia by a novel image analysis algorithm. *Front Microbiol* 2015;**6**:549.
69. Fedosov DA, Lei H, Caswell B, *et al.* Multiscale modeling of red blood cell mechanics and blood flow in malaria. *PLoS Comput Biol* 2011;**7**(12):e1002270.
70. Macdonald SPJ, Arendts G, Fatovich DM, Brown SGA. Comparison of PIRO, SOFA, and MEDS scores for predicting mortality in emergency department patients with severe sepsis and septic shock. *Acad Emerg Med* 2014;**21**(11):1257–63.
71. Wikswo JP. The relevance and potential roles of microphysiological systems in biology and medicine. *Exp Biol Med (Maywood)* 2014;**239**(9):1061–72.
72. Pârvu O, Gilbert D. Automatic validation of computational models using pseudo-3D spatio-temporal model checking. *BMC Syst Biol* 2014;**8**:124.
73. Dada JO, Mendes P. Multi-scale modelling and simulation in systems biology. *Integr Biol (Camb)* 2011;**3**(2):86–96.
74. Thakar J, Saadatpour-Moghaddam A, Harvill ET, Albert R. Constraint-based network model of pathogen-immune system interactions. *J R Soc Interface* 2009;**6**(36):599–612.
75. Hummert S, Hummert C, Schröter A, *et al.* Game theoretical modelling of survival strategies of *Candida albicans* inside macrophages. *J Theor Biol* 2010;**264**(2):312–8.
76. Eswarappa SM. Location of pathogenic bacteria during persistent infections: insights from an analysis using game theory. *PLoS One* 2009;**4**(4):e5383.
77. Tierney L, Linde J, Müller S, *et al.* An interspecies regulatory network inferred from simultaneous RNA-seq of *Candida albicans* invading innate immune cells. *Front Microbiol* 2012;**3**:85.
78. Boswell GP, Jacobs H, Davidson FA, *et al.* A positive numerical scheme for a mixed-type partial differential equation model for fungal growth. *Appl Math Comput* 2003;**138**(2-3):321–40.
79. Tokarski C, Hummert S, Mech F, *et al.* Agent-based modeling approach of immune defense against spores of opportunistic human pathogenic fungi. *Front Microbiol* 2012;**3**:129.
80. Hünniger K, Lehnert T, Bieber K, *et al.* A virtual infection model quantifies innate effector mechanisms and *Candida albicans* immune escape in human blood. *PLoS Comput Biol* 2014;**10**(2):e1003479.
81. Wcislo R, Miller S, Dzwinel W. PAM: Particle automata model in simulation of *Fusarium graminearum* pathogen expansion. *J Theor Biol* 2016;**389**:110–22.
82. Thakar J, Albert R. Boolean models of within-host immune interactions. *Curr Opin Microbiol* 2010;**13**(3):377–81.
83. Grant AJ, Restif O, McKinley TJ, *et al.* Modelling within-host spatiotemporal dynamics of invasive bacterial disease. *PLoS Biol* 2008;**6**(4):757–70.
84. Tyc KM. A modeling perspective on *Candida albicans* interactions with its human host. Dissertation, Berlin, 2012.
85. Carbo A, Bassaganya-Riera J, Pedragosa M, *et al.* Predictive computational modeling of the mucosal immune responses during *Helicobacter pylori* infection. *PLoS One* 2013;**8**(9):e73365.
86. Pollmächer J, Figge MT. Deciphering chemokine properties by a hybrid agent-based model of *Aspergillus fumigatus* infection in human alveoli. *Front Microbiol* 2015;**6**:503.
87. Colman A. *Game Theory and its Application in the Social and Biological Sciences*. New York, NY: Routledge, 1999.
88. Li XY, Pietschke C, Fraune S, *et al.* Which games are growing bacterial populations playing? *J R Soc Interface* 2015;**12**(108):20150121.
89. Gao D, Lietman TM, Porco TC. Antibiotic resistance as collateral damage: the tragedy of the commons in a two-disease setting. *Math Biosci* 2015;**263**:121–32.
90. Tyc KM, Kühn C, Wilson D, Klipp E. Assessing the advantage of morphological changes in *Candida albicans*: a game theoretical study. *Front Microbiol* 2014;**5**:41.
91. Beste DJ, Hooper T, Stewart G, *et al.* GSMN-TB: a web-based genome-scale network model of *Mycobacterium tuberculosis* metabolism. *Genome Biol* 2007;**8**(5):R89.
92. Raghunathan A, Reed J, Shin S, *et al.* Constraint-based analysis of metabolic capacity of *Salmonella typhimurium* during host-pathogen interaction. *BMC Syst Biol* 2009;**3**:38.
93. Rienksma RA, Suarez-Diez M, Spina L, *et al.* Systems-level modeling of mycobacterial metabolism for the identification of new (multi-)drug targets. *Semin Immunol* 2014;**26**(6):610–22.
94. Jamshidi N, Raghunathan A. Cell scale host-pathogen modeling: another branch in the evolution of constraint-based methods. *Front Microbiol* 2015;**6**:1032.
95. Linde J, Schulze S, Henkel S, Guthke R. Data- and knowledge-based modeling of gene regulatory networks: an update. *EXCLI J - Exp Clin Sci* 2015;**14**:346–78.
96. Schulze S, Henkel SG, Driesch D, *et al.* Computational prediction of molecular pathogen-host interactions based on dual transcriptome data. *Front Microbiol* 2015;**6**:65.
97. Sprott JC, Wildenberg JC, Azizi Y. A simple spatiotemporal chaotic Lotka–Volterra model. *Chaos, Solitons and Fractals* 2005;**26**:1035.
98. Stein RR, Bucci V, Toussaint NC, *et al.* Ecological modeling from time-series inference: insight into dynamics and

stability of intestinal microbiota. *PLoS Comput Biol* 2013;**9**(12):e1003388.
99. Bauer AL, Beauchemin CA, Perelson AS. Agent-based modeling of host-pathogen systems: the successes and challenges. *Inf Sci (Ny)* 2009;**179**(10):1379–89.
100. Wylie DC, Hori Y, Dinner AR, Chakraborty AK. A hybrid deterministic-stochastic algorithm for modeling cell signaling dynamics in spatially inhomogeneous environments and under the influence of external fields. *J Phys Chem B* 2006;**110**(25):12749–65.
101. An G. Introduction of an agent-based multi-scale modular architecture for dynamic knowledge representation of acute inflammation. *Theor Biol Med Model* 2008;**5**(1):11.
102. Stern JR, Olivas AD, Valuckaite V, *et al.* Agent-based model of epithelial host-pathogen interactions in anastomotic leak. *J Surg Res* 2013;**184**(2):730–8.
103. Wilensky U. NetLogo. http://ccl.northwestern.edu/netlogo/ Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL, 1999.
104. Chiacchio F, Pennisi M, Russo G, *et al.* Agent-based modeling of the immune system: NetLogo, a promising framework. *Biomed Res Int* 2014;**2014**:907171.
105. Solovyev A, Mikheev M, Zhou L, *et al.* SPARK. *Int J Agent Technol Syst* 2010;**2**(3):18–30.
106. Albert I, Thakar J, Li S, *et al.* Boolean network simulations for life scientists. *Source Code Biol Med* 2008;**3**:16.
107. Franke R, Müller M, Wundrack N, *et al.* Host-pathogen systems biology: logical modelling of hepatocyte growth factor and *Helicobacter pylori* induced c-Met signal transduction. *BMC Syst Biol* 2008;**2**:4.
108. Brakhage AA, Schroeckh V. Fungal secondary metabolites - strategies to activate silent gene clusters. *Fungal Genet Biol* 2011;**48**:15–22.
109. Brakhage AA. Regulation of fungal secondary metabolism. *Nat Rev Microbiol* 2013;**11**:21–32.
110. Dühring S, Germerodt S, Skerka C, *et al.* Host-pathogen interactions between the human innate immune system and *Candida albicans* - understanding and modeling defense and evasion strategies. *Front Microbiol* 2015;**6**:625.
111. Kirschner DE, Hunt CA, Marino S, *et al.* Tuneable resolution as a systems biology approach for multi-scale, multi-compartment computational models. *Wiley Interdiscip Rev Syst Biol Med* 2014;**6**(4):225–45.
112. Chaves M, Albert R, Sontag ED. Robustness and fragility of Boolean models for genetic regulatory networks. *J Theor Biol* 2005;**235**(3):431–49.
113. Cedersund G. Conclusions via unique predictions obtained despite unidentifiability–new definitions and a general method. *FEBS J* 2012;**279**(18):3513–27.
114. McKeever S, Johnson D. The role of markup for enabling interoperability in health informatics. *Front Physiol* 2015;**6**:152.
115. Andasari V, Roper RT, Swat MH, Chaplain MAJ. Integrating intracellular dynamics using CompuCell3D and Bionetsolver: applications to multiscale modelling of cancer cell growth and invasion. *PLoS One* 2012;**7**(3):e33726.
116. Karr JR, Sanghvi JC, Macklin DN, *et al.* WholeCellKB: model organism databases for comprehensive whole-cell models. *Nucleic Acids Res* 2013;**41**(Database issue):D787–92.
117. Karr JR, Sanghvi JC, Macklin DN, *et al.* A whole-cell computational model predicts phenotype from genotype. *Cell* 2012;**150**(2):389–401.
118. Morales DK, Hogan DA. *Candida albicans* interactions with bacteria in the context of human health and disease. *PLoS Pathog* 2010;**6**(4):e1000886.

2.6 Arbeitsanteile der Autoren

Tabelle 2.1: Prozentualer Arbeitsanteil der Autoren an den Manuskripten

Titel	**Autor**	**Anteil**
Manuskript 1	Betty Hebecker*	25 %
„Dual-species transcriptional profiling	Sebastian Vlaic*	25 %
during systemic candidiasis reveals	**Theresia Conrad**	**10 %**
organ-specific host-pathogen	Michael Bauer	3 %
interactions“	Sascha Brunke	5 %
	Mario Kapitan	5 %
	Jörg Linde	4 %
	Bernhard Hube	3 %
	Ilse D. Jacobsen	20 %
Manuskript 2	Sebastian Vlaic	45 %
„ModuleDiscoverer: Identification of	**Theresia Conrad**	**25 %**
regulatory modules in protein-protein	Christian Tokarski-Schnelle	10 %
interaction networks“	Mika Gustafsson	5 %
	Uta Dahmen	5 %
	Reinhard Guthke	5 %
	Stefan Schuster	5 %
Manuskript 3	**Theresia Conrad**	**50 %**
„Module-detection approaches for the	Olaf Kniemeyer	10 %
integration of multilevel omics data	Sebastian G. Henkel	7 %
highlight the comprehensive response of	Thomas Krüger	6 %
Aspergillus fumigatus to caspofungin“	Derek J. Mattern	1 %
	Vito Valiante	1 %
	Reinhard Guthke	3 %
	Ilse D. Jacobsen	1 %
	Axel A. Brakhage	1 %
	Sebastian Vlaic	10 %
	Jörg Linde	10 %

Titel	**Autor**	**Anteil**
Manuskript 4	Philipp Kämmer	50 %
„Strategies of pathogenic *Candida*	Sylvie McNamara	20 %
species to survive in human blood	Thomas Wolf	9 %
have evolved independently“	**Theresia Conrad**	**2 %**
	Kerstin Hünniger	1 %
	Oliver Kurzai	1 %
	Reinhard Guthke	1 %
	Bernhard Hube	3 %
	Jörg Linde	4 %
	Sascha Brunke	9 %
Manuskript 5	Jana Schleicher*	30 %
„Facing the challenges of multiscale	**Theresia Conrad***	**30 %**
modelling of bacterial and fungal	Mika Gustafsson	5 %
pathogen-host interactions“	Gunnar Cedersund	5 %
	Reinhard Guthke	15 %
	Jörg Linde	15 %

* geteilte Erstautorenschaft

3 Diskussion

Die Zielstellung dieser Dissertation ist es, anhand verschiedener Integrationsansätze für Multiskalen- und Multi-Omik-Daten neue Erkenntnisse im Bereich der WPPI zu erlangen. Um diese Zielstellung erfüllen zu können, wurden im Hauptteil einige der Publikationen vorgestellt, an denen ich im Rahmen der Dissertation mitgewirkt habe. In diesem Kapitel soll näher auf die Frage eingegangen werden, welche biologischen und methodischen Ursachen der Heterogenität von Daten zugrunde liegen. Im Anschluss daran erfolgt eine kritische Betrachtung der in den vorgestellten Arbeiten angewandten Ansätze zur Integration von Multiskalen- und Multi-Omik-Daten.

3.1 Gründe für die Heterogenität von Multiskalen- und Multi-Omik-Daten

3.1.1 Biologische Aspekte

Die zeitlich und strukturell bedingte zelluläre Antwort auf externe Stimuli

Ein wichtiger Aspekt für alle in dieser Dissertation vorgestellten Manuskripte zur Integration von Multiskalen- und Multi-Omik-Daten ist deren Heterogenität. Als biologische Ursachen spielen vor allem strukturelle und zeitliche Faktoren innerhalb des biologischen Systems eine bedeutende Rolle. Am Beispiel einer Zelle sollen diese Faktoren in vereinfachter Form erläutert werden (Abbildung 3.1). Aktiviert ein Stimulus einen zellulären Rezeptor, wird ein intrazelluläres Signal erzeugt. Diese Information wird mittels Signalkaskaden zu ihrem jeweiligen Bestimmungsort transportiert und löst dort eine entsprechende zelluläre Reaktion aus. Signalkaskaden können entweder „direkt“ oder „indirekt“ zu einer zellulären Reaktion führen. Im

ersten Fall kommt es „direkt“ zu einer Änderung des Proteoms und der daran geknüpften Antwort/Reaktion auf den Stimulus. „Indirekte“ Antworten beeinflussen zunächst die Genexpression im Zellkern oder in den Mitochondrien, führen damit zu Änderungen im Transkriptom und nehmen erst anschließend Einfluss auf das Proteom und weitergehende Reaktionen. Die direkte zelluläre Antwort kann innerhalb von Sekunden bzw. weniger Minuten erfolgen. Die indirekte dagegen dauert länger und kann Zeitspannen von einigen Minuten bis zu mehreren Stunden umfassen [Castiglione *et al.,* 2014; Murphy *et al.,* 2014].

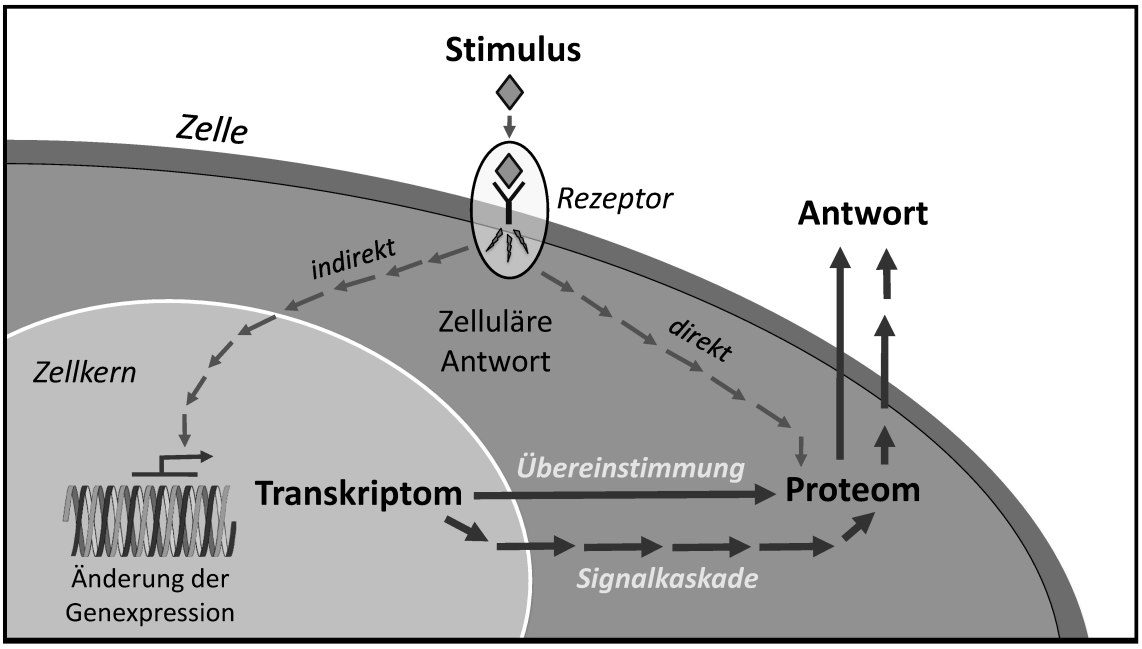


Abbildung 3.1: Schematische Darstellung der direkten und indirekten (hier über den Zellkern) zellulären Antwort auf einen Stimulus

In Abhängigkeit von den involvierten Signalwegen, deren gegenseitiger Beeinflussung, Rückkopplungen und dem jeweils betrachteten Zeitpunkt liegt eine mehr oder auch weniger starke zelluläre Antwort vor. Dementsprechend kann zu einem bestimmten Zeitpunkt immer auch nur ein kleiner Ausschnitt der zellulären Gesamtantwort betrachtet werden. Bezüglich Transkriptom- und Proteomstudien bedeutet das Folgendes: Die Expression eines bestimmten Gens und die Synthese des zugehörigen Proteins können als zwei aufeinander folgende Prozesse angesehen werden, die nicht zum selben Zeitpunkt stattfinden. Damit spiegeln sich zwar Änderungen in der Genexpression auch auf Proteinebene wider, allerdings zu jeweils unterschiedlichen Zeitpunkten

[Liu *et al.,* 2016]. Für experimentelle Studien spielt die Wahl der Messzeitpunkte also eine entscheidende Rolle dafür, ob bzw. in welchem Ausmaß Zusammenhänge zwischen den Omik-Ebenen beobachtet werden können. Dabei sind diese Zusammenhänge jedoch nicht nur von übereinstimmenden Omik-Ebenenkomponenten geprägt (z. B. Genen und deren zugehörigen synthetisierten Proteinen), sondern werden auch durch Signalwege charakterisiert, die auf die Transkription mehrerer Gene und, in der Folge, mehrerer Proteine wirken. Ein Beispiel hierfür wird im Kapitel 3.2.4, Abbildung 3.3 gezeigt. Die Komponenten dieser Signalwege, insbesondere Proteine, müssen nicht zwingend miteinander übereinstimmen, können sich aber gegenseitig beeinflussen und im Zusammenspiel eine konzertierte zelluläre Antwort erzeugen [Conrad *et al.,* 2018].

Proteinumsatz

Einer der hauptverantwortlichen posttranslationalen Faktoren, welche die Korrelation zwischen Transkriptom- und Proteomebene beeinflussen können, ist der Abbau der Transkripte und Proteine, charakterisiert durch deren Halbwertszeit. Damit im Zusammenhang steht der Proteinumsatz. Der Proteinumsatz beschreibt die Bilanz von Proteinsynthese und Proteinabbau. Er stellt einen effektiven Weg für die Erhaltung eines funktionellen Proteoms dar, wobei alte und/oder möglicherweise geschädigte und toxische Proteine abgebaut und durch neu synthetisierte ersetzt werden. Jedes Protein verfügt über eine individuelle Umsatzrate, die Zeitspannen von einigen Sekunden bis hin zu mehreren Tagen umfassen kann. Sie hängt von Faktoren ab wie der intrinsischen Proteinstabilität, der N-terminalen Aminosäure mit ihrem Einfluss auf die Proteinabbaugeschwindigkeit, der posttranslationalen Verarbeitung, der Ubiquitinierung oder der Lokalisation des Proteins [Liu *et al.,* 2016; Maier *et al.,* 2009].

Analog ist die Halbwertszeit der Transkripte zu betrachten. Auch deren Abbaugeschwindigkeit ist von Transkript zu Transkript unterschiedlich. Zudem kann sie durch verschiedene Einflussfaktoren geregelt werden, wie beispielsweise durch die Reduktion des mRNA-Poly(A)-Schwanzes oder durch verschiedene Enzyme, die den Transkriptabbau katalysieren [Lugowski *et al.,* 2018].

Biologische Variabilität

In den vorgestellten Manuskripten wurden Organismen wie z. B. *A. fumigatus* und *C. albicans* auf der Pathogenseite sowie Maus und Mensch auf der Wirtsseite untersucht. Sowohl zwischen den Organismen als auch innerhalb eines einzelnen Organismus muss die interindividuelle und die intraindividuelle biologische Variabilität berücksichtigt werden. Faktoren wie Spezies, Alter, Größe, Geschlecht, genetische Variabilität oder auch Umweltbedingungen spielen hier eine Rolle. Die genetische Variabilität wird unter anderem durch Polymorphismen in der Gensequenz beeinflusst und trägt maßgeblich zur Stabilität und Funktionalität von einzelnen Proteinen, Proteinkomplexen etc. bei. Bekannte Beispiele sind Einzelnukleotidpolymorphismen, Insertions-/Deletionspolymorphismen, Inversionen oder Kopienzahlvarianten [Fraser, 2001; Fraser, 2017].

3.1.2 Methodische Aspekte

Neben den biologischen Aspekten tragen auch methodische Faktoren zur Quantität, Qualität und der letztendlichen Vergleichbarkeit der Daten bei. Dazu zählen etwa das Experimentdesign inklusive der Wahl der Kontrollen, Replikate und Messzeitpunkte sowie die Wahl der Messmethoden [Hasin *et al.,* 2017]. Zusätzliche Fehler- oder Heterogenitätsquellen können bei der Probenaufbereitung entstehen, unter anderem durch den proteolytischen Verdau und das Aufkonzentrieren oder Derivatisieren von Komponenten einer Probe. Ein Beispiel hierfür liefert die Studie von Müller *et al.* aus dem Jahr 2017. Sie beschäftigt sich mit Glykoproteinen und deren Beitrag zur Endotoxintoleranz in Monozyten. Glykoproteine spielen eine wichtige Rolle in der molekularen und zellulären Erkennung sowie der Modulation von intra- und interzellulärem *Crosstalk.* Dadurch repräsentieren sie wichtige Angriffspunkte für medikamentöse Behandlungen. Die Schwierigkeit bei der Untersuchung von Glykoproteinen und auch anderen membranständigen Proteinen besteht darin, dass sie aufgrund ihrer geringen Häufigkeit und ungünstigen biochemischen Eigenschaften nur schwer über Methoden wie der Massenspektrometrie zu untersuchen sind. Müller *et al.* konnten in der Arbeit jedoch zeigen, dass die Verwendung einer hydrazidbasierten Methode die Anreicherung von Glykoproteinen ermöglicht und somit den Grundstein für umfassende Analysen von Glykoproteinen und deren Regulation legt [Müller *et al.,* 2017; Wells *et al.,* 2011].

Limitationen der Microarray-Analyse

Seit ihrer Einführung in den 1990er Jahren haben sich Microarrays als Hochdurchsatztechnologie für Genexpressionsuntersuchungen etabliert. In der Studie von Hebecker und Vlaic *et al.* aus dem Jahr 2016 (Manuskript 1) bilden sie die Grundlage für die Erstellung eines umfassenden Genexpressionsprofils von Maus und *C. albicans* in verschiedenen Organen und zu unterschiedlichen Zeitpunkten. Microarrays als hybridisierungsbasierte Ansätze sind im Vergleich zu anderen Ansätzen wie RNA-Seq verhältnismäßig kostengünstig, weisen allerdings auch einige Nachteile auf. So führt beispielsweise die Benutzung von speziell designten, vorwissensbasierten Sonden dazu, dass für bisher im Genom nicht-annotierte Gene in der Regel keine Sonden eingesetzt werden und diese Gene somit auch nicht detektiert werden. Die Affinität und Intensität einer Sonde hängen stark von den gegebenen Hybridisierungsbedingungen ab. Außerdem liefert die Signalstärke einer Sonde nur relative Werte der Transkriptkonzentration, sofern sich Signalstärke und Konzentration (in bestimmten Wertebereichen) proportional verhalten. Allerdings ist eine lineare Proportionalität aufgrund von Hybridisierungskinetiken nicht immer gegeben, sodass eine direkte Vergleichbarkeit der Intensitäten nicht möglich ist. Liegen hohe Konzentrationen der zu bindenden Sequenzen vor, kommt es zur Signalsättigung und damit zu einem nichtlinearen Zusammenhang zwischen Signalstärke und Konzentration. Bei niedrigen Konzentrationen besteht das Problem, dass entweder die Hybridisierung gar nicht erst stattfindet oder das Signal durch Rauschen überdeckt wird. Somit verfügen die Sonden nur über einen beschränkten Detektionsbereich. Neben der Konzentration tragen unter anderem auch Kreuzhybridisierung, fehlerhafte Sondenannotationen, alternatives Spleißen oder Polymorphismen dazu bei, dass die Signalstärke der Sonden und die tatsächliche Genexpression, d. h. die Konzentration der Transkripte, nicht miteinander korrelieren [Bumgarner, 2013; Zhao *et al.,* 2014].

Limitationen der RNA-Seq-Analyse

Seit einigen Jahren löst die RNA-Seq-Technologie die Microarrays zunehmend ab. Sie wurde beispielsweise in der Arbeit von Conrad *et al.* (Manuskript 3) angewandt, um die caspofungininduzierte transkriptionelle Stressantwort von *A. fumigatus* über mehrere Zeitpunkte hinweg zu untersuchen. Kämmer *et al.* (Manuskript 4) haben mithilfe eines dualen RNA-Seq-Ansatzes sowohl wirts- als auch pilzinduzierte WPPI im

humanen Vollblutmodell untersucht, wobei das Vollblut mit verschiedenen *Candida*-Spezies infiziert wurde. RNA-Seq ermöglicht es, alle in einer Probe enthaltenen RNAs gleichzeitig zu identifizieren, ihre Sequenzen zu charakterisieren und die Häufigkeit zu quantifizieren. Im Gegensatz zu Microarrays benötigt diese Technologie keine Sonden, wodurch die damit zusammenhängenden Bias vermieden werden. Es können mit RNA-Seq auch neue Transkripte, allelspezifische Expressionen, Sequenzvarianzen oder Isoformen detektiert sowie Exon-Intron-Grenzen ermittelt werden. Diese Technologie ist empfindlicher gegenüber Genen mit geringer Expression und genauer bei der Detektion von Genen, deren Transkripte in großer Menge vorliegen. Nichtsdestotrotz weist auch die RNA-Seq-Technologie einige Limitationen bzw. Bias auf. So beschreibt beispielsweise ein sequenzspezifisches Bias die aufgrund unterschiedlich präferierter Genregionen ungleiche Verteilung von *Reads* (d. h. RNA-Fragmente). Damit kann es die Sequenziertiefe beeinflussen. Genetische Variationen, Sequenzwiederholungen, sehr kurze und/oder untereinander sehr ähnliche *Read*-Sequenzen sowie Sequenzierfehler führen zu Unsicherheiten und damit potenziellen Schwankungen beim Ermitteln der Genabundanzen. Auch die Softwareeinstellungen der verschiedenen Schritte einer RNA-Seq-Analyse sowie methodische und biologische Variabilität haben Einfluss auf die Ergebnisse [Finotello *et al.,* 2015; Zhao *et al.,* 2014].

Limitationen der Massenspektrometrie

Massenspektrometrie ermöglicht die Hochdurchsatzidentifizierung und -quantifizierung von Proteinen und hat sich damit für die Untersuchung von Proteomproben bewährt. In Conrad *et al.* (Manuskript 3) z. B. ist die Massenspektrometrie ein wichtiger Bestandteil der Multi-Omik-Datenerhebung bezüglich der caspofungininduzierten Stressantwort von *A. fumigatus*. Die Methode wurde sowohl für die Untersuchung des Proteoms als auch des Sekretoms (d. h. von potenziell sekretierten Proteinen, die sich in der Zellumgebung befinden) verwendet. Aber auch in anderen Studien findet die Massenspektrometrie Anwendung [Aebersold *et al.,* 2016], so beispielsweise in der zu Beginn des Kapitels erwähnten Arbeit zur Glykoproteinanalyse von Müller *et al.* aus dem Jahr 2017. Für die Bewertung der durch die Massenspektrometrie erzeugten Ergebnisse müssen allerdings einige Limitationen der Methode berücksichtigt werden. Das Detektionslimit des Instruments spielt dabei eine wichtige Rolle. Es beschreibt die benötigte Konzentration einer Probe, um sich von dem vom Instrument selbst erzeugten Rauschen abzuheben. Dem Rauschen liegen Fluktuationen des Instru-

menthintergrundniveaus zugrunde, die auch auftreten, wenn keine Probe vorliegt. Darüber hinaus ist das sogenannte *Undersampling* ein wohlbekanntes Problem in der Massenspektrometrie. Es beruht zum einen auf der nicht ausreichenden Aufnahmerate für hochkomplexe Proben, zum anderen auf den technisch bedingten Schwankungen am Detektionslimit. So führen wiederholte Messungen von ein und derselben Probe trotz gleicher Bedingungen zu voneinander abweichenden Ergebnissen. *Undersampling* tritt umso stärker auf, je komplexer und dynamischer die zu untersuchenden Proben sind. Die Dynamik bezieht sich hierbei auf das Intervall zwischen dem kleinsten und dem größten Konzentrationswert für die einzelnen Komponenten der Probe. Auch das Zeitfenster für den Messprozess ist entscheidend. Dieses wird vor allem durch die angewandten (im Messgerät implementierten) Algorithmen und deren zugrunde liegenden Ionenstatistik limitiert [Wells *et al.,* 2011].

3.2 Bioinformatisch-methodische Aspekte

3.2.1 Klassische Ansätze: Komponentenvergleich und Signalweganalyse

Ein intuitiver und weit verbreiteter Ansatz für die Analyse multipler Datensätze basiert auf dem Vergleich von Listen differenziell regulierter Komponenten. Dieser klassische Ansatz kommt in vielen Studien zur Anwendung, wie beispielsweise in der von Hebecker und Vlaic *et al.* (Manuskript 1), Conrad *et al.* (Manuskript 3) oder Kämmer *et al.* (Manuskript 4). Er erlaubt es, Aussagen über die Gemeinsamkeiten der untersuchten Datensätze zu treffen. Aussagen über potenzielle Unterschiede sind dagegen kritisch zu betrachten, da es diverse Gründe für das Vorhandensein oder auch Nicht-Vorhandensein von differenziell regulierten Komponenten gibt. Einige biologische (z. B. der Proteinumsatz und die biologische Variabilität) und methodische (z. B. das Experimentdesign und die Limitationen der jeweiligen Analysemethode) Ursachen wurden bereits im vorangegangenen Kapitel erläutert. Indem bei einer Analyse ausschließlich differenziell regulierte Komponenten betrachtet werden, wird vor allem auf solche Komponenten fokussiert, die einen direkten biologischen Zusammenhang mit der zugrunde liegenden Fragestellung aufweisen. Diese biologischen

Zusammenhänge können als Signalwege beschrieben werden, mit denen die Komponenten signifikant assoziiert sind. Dabei tragen allerdings nicht nur die unter den gegebenen Bedingungen (z. B. das Vorhandensein bzw. Nicht-Vorhandensein eines externen Stimulus) als differenziell reguliert identifizierten Komponenten zur Aktivität eines Signalwegs bei. Auch andere Komponenten könnten einen entscheidenden Einfluss haben, wurden aber aufgrund der genannten Ursachen fälschlicherweise als nicht-differenziell reguliert identifiziert und somit nicht weiter berücksichtigt. Je nach Verhältnis der differenziell/nicht-differenziell regulierten Signalwegkomponenten und in Abhängigkeit der gewählten Grenzwerte bestimmt die angewandte Statistik maßgeblich, ob ein Signalweg als signifikant mit der Fragestellung assoziiert wird oder nicht. Daher kann der Fokus auf ausschließlich differenziell regulierte Komponenten dazu führen, dass ein Großteil der potenziell mit der Fragestellung assoziierten Signalwege nicht erkannt wird [García-Campos *et al.,* 2015; Pavlidis *et al.,* 2004].

Ein großer Vorteil von Signalweganalysen liegt darin, dass Daten verschiedener Omik-Technologien miteinander verknüpft und somit Zusammenhänge zwischen diesen Daten im Hinblick auf den biologischen Kontext aufgedeckt werden können. Dafür wird als Eingabe neben den experimentell gemessenen Daten auch Vorwissen aus Wissensdatenbanken benötigt. Damit haben Umfang bzw. Qualität des Vorwissens einen maßgeblichen Einfluss auf die Ergebnisse [García-Campos *et al.,* 2015; Hasin *et al.,* 2017]. Ein Beispiel hierfür liefert die Arbeit von Conrad *et al.* (Manuskript 3). Um regulatorische Module detektieren zu können, werden sowohl experimentelle Daten als auch ein von der Datenbank STRING [Szklarczyk *et al.,* 2019] bereitgestelltes *A. fumigatus*-PPIN genutzt. Dieses PPIN besteht aus 4123 Proteinen. Laut dem *Central Aspergillus Data Repository* (CADRE) [Gilsenan *et al.,* 2012] umfasst das Genom von *A. fumigatus* allerdings 9916 proteinkodierende Gene. Somit werden mehr als die Hälfte der eigentlich für den Pilz bekannten Komponenten nicht durch das PPIN betrachtet. Sie fließen daher nicht mit in das regulatorische Modul ein, obwohl sie eventuell sogar experimentell gemessen wurden.

Eine weitere Schwierigkeit besteht darin, eine gemeinsame Terminologie zwischen den verschiedenen Datensätzen und auch der Wissensdatenbanken zu schaffen. Oftmals geht durch die Verwendung unterschiedlicher Annotationen (Zuordnungen von biologischen Informationen) eine Vielzahl an Daten verloren. Auch die Überset-

zung einer Annotation in eine andere stellt durch fehlende Zuordnungen oder durch Mehrfachzuordnungen eine potenzielle Fehlerquelle dar. Darüber hinaus können verschiedene Autoren voneinander abweichende Bezeichnungen für den gleichen Sachverhalt verwenden. Das erhöht die Falsch-Negativ-Rate in den Ergebnissen. Das Gleiche gilt im umgekehrten Fall, wenn mehrere Autoren für verschiedene Sachverhalte dieselbe Bezeichnung nutzen. Das erhöht die Falsch-Positiv-Rate in den Ergebnissen [Furnham *et al.,* 2012; Gillis *et al.,* 2013; Heyer *et al.,* 2017;]. Einen Nachteil stellen auch die fehlenden einheitlichen Vorschriften dar, nach denen neue Forschungserkenntnisse zeitnah bzw. überhaupt in die vorhandenen Wissensdatenbanken eingepflegt werden müssten. Dadurch entsteht eine Diskrepanz zwischen dem tatsächlich vorhandenen und dem in Datenbanken zur Verfügung stehenden Wissen [García-Campos *et al.,* 2015; Hasin *et al.,* 2017].

3.2.2 Clustering-basierte Ansätze

Ein typisches Beispiel für die Anwendung von Clustering-Ansätzen liefert die Studie von Hebecker und Vlaic *et al.* (Manuskript 1). In dieser wurden verschiedene Clustering-Verfahren angewendet, um zeitlich bedingte Genexpressionsänderungen in den Zellen von Wirt und Pathogen in verschiedenen Organen nachvollziehen zu können. Aufgrund der Vielfältigkeit der verfügbaren Clustering-Ansätze (siehe Kapitel 1.5.1) und der damit einhergehenden Vor- und Nachteile stellt die Auswahl eines passenden Ansatzes für die jeweilige Fragestellung oftmals eine Herausforderung dar. Eine Limitation von hierarchischen Clustering-Verfahren wie DIANA und AGNES ist beispielsweise, dass einmal getroffene ungünstige Entscheidungen bezüglich der Zusammenlegung oder dem Aufsplitten von Clustern nicht mehr rückgängig gemacht und so algorithmisch (automatisch) optimiert werden können. Folglich tendieren die Ansätze dazu, in lokalen Optima gefangen zu sein und das globale Optimum nicht zu finden. Darüber hinaus sind hierarchische Verfahren und auch Partitionierungsansätze empfindlich gegenüber Rauschen und Ausreißern in den Eingabedaten. Die k-Means-Methode hat zusätzlich den Nachteil, dass die Auswahl der Startpunkte für die Clustergenerierung vom Nutzer festgelegt werden muss oder zufällig erfolgt. Das trägt zu einer starken Abhängigkeit der Ergebnisse von der Wahl dieser Punkte bei. Dem kann durch wiederholte Clustering-Prozesse mit zufälligen Startpunkten entgegengewirkt werden, wobei allerdings die Reproduzierbarkeit der

Ergebnisse nicht zwingend gegeben ist. Außerdem besteht auch hier das Problem der lokalen Optima. Eine weitere Herausforderung bei der Anwendung von Clustering-Ansätzen ist die Anzahl der zulässigen Cluster. Diese muss durch den Nutzer oder computergestützt abgeschätzt werden und sollte so gewählt sein, dass sie die Struktur der zu untersuchenden Daten optimal widerspiegelt. Auch die Wahl der zugrunde liegenden Metrik (Abstandsfunktion) zur Berechnung von Distanzen zwischen den Clustern obliegt meist dem Nutzer [Embrechts *et al.,* 2013; Oyelade *et al.,* 2016; R Core Team, 2018].

Die vorgenannten Parameter (Startpunkte, Clusteranzahl etc.) können algorithmisch optimiert werden, verlangen dafür aber ein Optimalitätskriterium. Dessen Auswahl muss der Nutzer treffen, da es viele verschiedene und keine allgemein gültigen Kriterien gibt. Für die Wahl des optimalen Clustering-Ansatzes wird eine Validierungsmethode angewandt. Ziel dabei ist die Generierung von Clustern, die neben der Bereitstellung biologisch relevanter Ergebnisse auch gute statistische Eigenschaften bezüglich Dichte, Kompaktheit, Separierung oder Stabilität aufweisen. Um eine Validierung von solchen intrinsischen bzw. Stabilitätsmaßen zu ermöglichen, wurden verschiedene Indizes (Optimalitätskriterien, Validierungsmaße) definiert, die beispielsweise über das R-Paket [R Core Team, 2018] *clValid* von Brock *et al.* aus dem Jahr 2008 berechnet werden können. Das Maß für die Dichte bezieht sich darauf, dass benachbarte Datenpunkte auch die gleiche Clusterzugehörigkeit erhalten. Kompaktheit bewertet Intra-Cluster-Distanzen, die Aussagen darüber treffen, wie eng die einzelnen Datenpunkte zusammenliegen. Separierung hingegen beschreibt die Inter-Cluster-Distanzen, welche die Cluster voneinander trennen. Indizes, die Kompaktheit und Separierung miteinander kombinieren, sind der Dunn-Index oder der Silhouettenkoeffizient. Validierungsmaße für die Stabilität vergleichen die Clustering-Ergebnisse basierend auf allen zu clusternden Daten mit denen, die durch die Löschung einzelner Proben aus den zugrunde liegenden Daten erzeugt werden. Beispiele hierfür sind *average proportion of non-overlap* (APN), *average distance* (AD), *average distance between means* (ADM) oder *figure of merit* (FOM) [Brock *et al.,* 2008; Oyelade *et al.,* 2016].

Die genannten Validierungsmaße zur Bewertung der Clustering-Ergebnisse und der gewählten Anzahl an Clustern wurden von Hebecker und Vlaic *et al.* (Manuskript 1)

für die Untersuchung der WPPI von Maus und *C. albicans* eingesetzt. Abbildung 3.2 zeigt die einzelnen Validierungsindizes der Clustering-Ergebnisse für die Wirtsseite. Hier wurden fünf Clustering-Algorithmen aus dem *clValid*-Paket (hierarchisches Clustering, k-Means, SOTA, DIANA und CLARA) entsprechend ihrer Ergebnisse für die Anzahl von 3 bis 15 Cluster verglichen. Die resultierenden Werte von jeder Validierungsmessung wurden skaliert und so transformiert, dass ein Wert von 1 das beste Ergebnis repräsentiert. Alle Indizes wurden schließlich miteinander kombiniert und als Durchschnitt der Mittelwerte der internen und Stabilitätsvalidierungsmaße dargestellt. Der maximale Wert dieser Kombination verweist auf den jeweils geeignetsten Clustering-Ansatz und die Anzahl an benötigten Clustern. Im konkreten Fall wurden drei Cluster als optimal ermittelt.

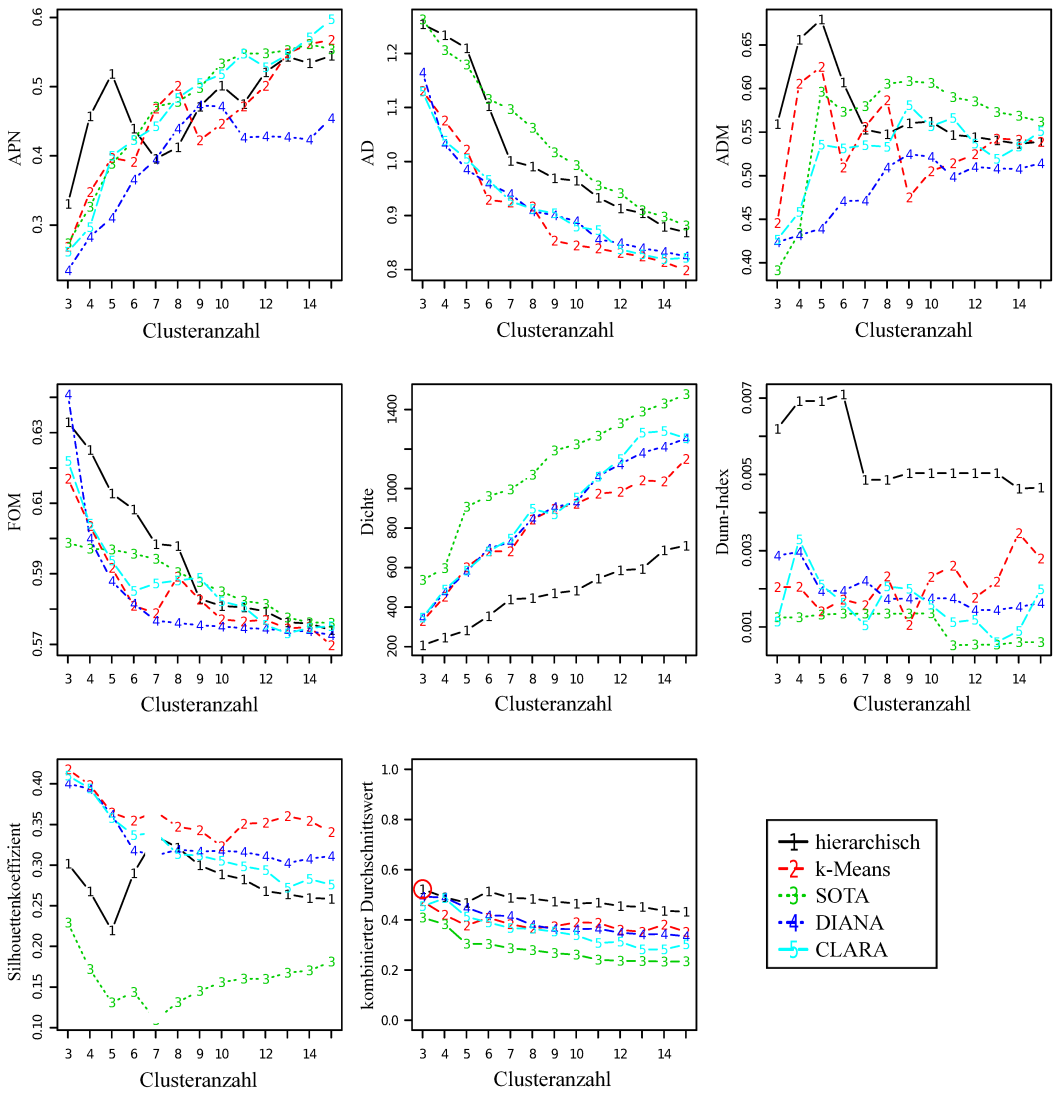


Abbildung 3.2: Clustervalidierung für die im Jahr 2016 von Hebecker und Vlaic *et al.* (Manuskript 1) publizierten Ergebnisse bezüglich der Wirtsseite. Beispiel: Clusterstabilität gemessen mit 7 verschiedenen Validierungskriterien sowie schließlich einer Kombination („kombinierter Durchschnittswert“, rechts unten) aus diesen

3.2.3 Interaktionsnetzwerkbasierte Ansätze

Wie die Arbeiten von Conrad *et al.* (Manuskript 3) und Kämmer *et al.* (Manuskript 4) zeigen, sind netzwerkbasierte Ansätze zum einen in der Lage, Multiskalen- und Multi-

Omik-Daten mithilfe von Experimental- und Wissensdatenbanken zu integrieren, zum anderen ermöglichen sie die Vorhersage bisher unbekannter Interaktionen zwischen Komponenten mittels Netzwerkinferenz. Hierzu wird im Review von Guthke *et al.* aus dem Jahr 2016 ein umfassender Überblick zum aktuellen Forschungsstand datenbasierter Rekonstruktionen von GRN gegeben. Der Vorteil bei der Integration von heterogenen Datensätzen liegt in der Fähigkeit, ein gewisses Maß an verrauschten oder fehlenden Daten zu kompensieren. Wird beispielsweise ein bestimmtes Protein nicht in den Proben detektiert, so kann es aufgrund seiner durch Vorwissen bekannten Konnektivitätsmuster trotzdem im resultierenden Netzwerk integriert werden. Allerdings wird durch die Benutzung von Vorwissen auch eine starke Abhängigkeit von dessen Umfang und Qualität geschaffen. So beruhen beispielsweise die vorgestellten ModuleDiscoverer-Anwendungen in den Manuskripten 2, 3 und 4 auf PPIN, die von der Datenbank STRING bereitgestellt wurden. Werden nun in experimentellen Datensätzen Komponenten detektiert, die nicht Teil des PPIN sind, so werden diese Komponenten nicht in das regulatorische Modul aufgenommen. Vor allem, wenn Schlüsselkomponenten im zugrunde liegenden PPIN fehlen, ist es schwer, Zusammenhänge zwischen den Datensätzen zu identifizieren. Auch andere bereits in Kapitel 3.2.1 beschriebene Faktoren bezüglich des Vorwissens oder der Verwendung verschiedener Terminologien oder Annotationen rufen bei den netzwerkbasierten Ansätzen Probleme hervor. An dieser Stelle soll noch einmal auf die Schwierigkeit der Übersetzung verschiedener Annotationen ineinander eingegangen werden. Ein Beispiel dafür ist die Integration von Genexpressionsdaten in ein PPIN. Dafür müssen zunächst die Gennamen der experimentellen Daten in die Proteinnamen des Netzwerks übersetzt werden. Durch z. B. alternatives Spleißen kommt es aber vor, dass ein Gen mehrere Proteine kodiert. Eine eineindeutige Gen-Protein-Zuordnung ist somit nicht möglich. Mehrfachzuordnungen können auch auftreten, wenn das in den PPIN abgebildete Wissen auf verschiedenen Spezies basiert. Das hat zur Folge, dass orthologe Relationen betrachtet werden müssen, die nicht selten mehrdeutig sind [Guthke *et al.,* 2016; Hasin *et al.,* 2017; Priebe *et al.,* 2013].

Die resultierenden Mehrfachzuordnungen sowie die Einbeziehung von Daten, die nicht zwingend mit der zugrunde liegenden Fragestellung in Verbindung stehen, können zur Generierung uneindeutiger Netzwerke führen, die mehrere gleichwahrscheinliche und möglicherweise unspezifische Erklärungen für eine Fragestellung bieten. Durch die

Integration verschiedener Datensätze lässt sich diese Problematik allerdings steuern und erlaubt eine höhere Fokussierung auf die tatsächlich für die experimentelle Fragestellung relevanten Komponenten und deren Zusammenhänge. So bietet die Identifikation von regulatorischen Modulen den Vorteil, dass sie Vorwissen und differenziell regulierte Komponenten aus den experimentellen Daten eng miteinander verknüpft. Damit können sie maßgeblich dazu beitragen, dass nur eine reduzierte Anzahl potenziell nicht mit der Fragestellung assoziierter Komponenten in das regulatorische Modul aufgenommen wird.

3.2.4 Der Umgang mit fehlenden Omik-Ebenen am Beispiel von ModuleDiscoverer

Ein kritischer Punkt bezüglich der Integration von Multiskalen- und Multi-Omik-Daten ist der Umgang mit einzelnen fehlenden Daten oder kompletten Ebenen. In der Statistik wurden dafür Methoden der sogenannten Imputation entwickelt, die auch in der Systembiologie angewendet werden [Albrecht *et al.,* 2010]. Ein Beispiel für eine Studie mit einer fehlenden Omik-Ebene ist die Arbeit von Lehmann *et al.* aus dem Jahr 2018. In dieser haben sich die Autoren mit der Frage beschäftigt, inwieweit *Toll-like*-Rezeptoren (TLR) humaner Epithelzellen ähnliche oder auch spezifische Signalkaskaden als Antwort auf TLR-spezifische Stimuli aufweisen. Dabei wurden sowohl das Transkriptom als auch das Sekretom berücksichtigt und unter anderem mittels Komponentenvergleich und Signalweganalysen in Zusammenhang gebracht [Lehmann *et al.,* 2018]. Proteine innerhalb der Zelle, die einen potenziellen Zwischenschritt zwischen Transkriptom und Sekretom darstellen können (Abbildung 3.3), wurden nicht betrachtet. Es stellt sich daher die Frage, inwiefern sich die fehlende Proteinebene über netzwerkbasierte Ansätze wie ModuleDiscoverer nachträglich rekonstruieren lässt.

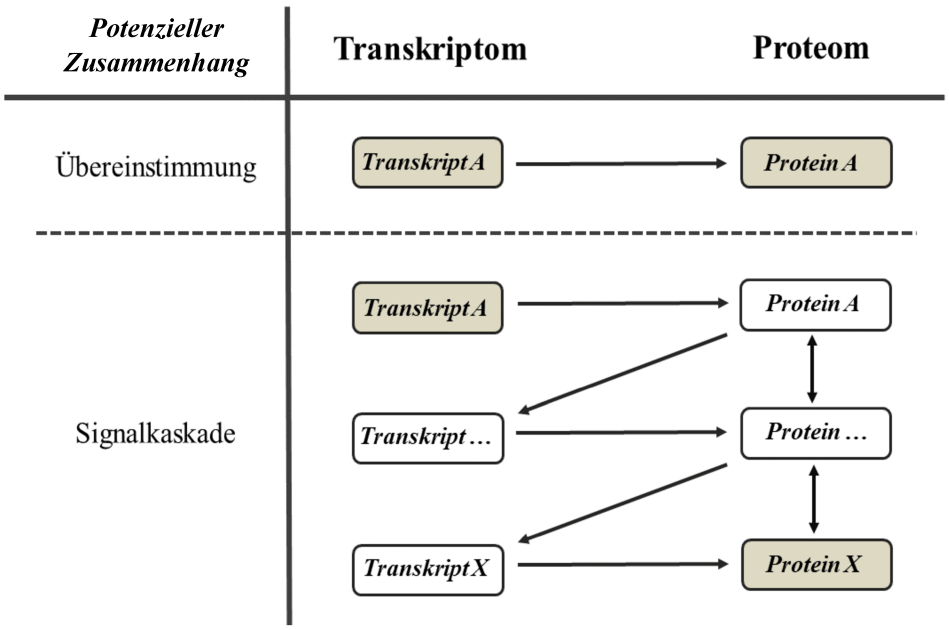


Abbildung 3.3: Potenzielle Zusammenhänge zwischen Transkriptom und Proteom; grau hinterlegte Komponenten wurden durch experimentelle Messungen detektiert

In Kapitel 3.1.1 wurden bereits die potenziellen Zusammenhänge zwischen Omik-Ebenen in Form von übereinstimmenden Komponenten oder ebenenübergreifenden Signalwegen erwähnt. Abbildung 3.3 zeigt hierzu noch einmal beispielhafte Szenarien für Transkriptom- und Proteomdaten. Dabei repräsentieren grau hinterlegte Komponenten die durch experimentelle Messungen detektierten Daten. Weiß hinterlegte Komponenten stellen dagegen nicht detektierte Komponenten dar, die allerdings zusammen mit den detektierten an einem gemeinsamen Signalweg beteiligt sind. Auf diese Weisen können auch vermeintlich voneinander abweichende experimentelle Ergebnisse auf einer gemeinsamen Basis beruhen. Die dargestellten Szenarien lassen sich leicht auf die Sekretomebene erweitern. Wird das Beispiel aus Lehmann *et al.* auf den Zusammenhang der übereinstimmenden Komponenten bezogen, dann heißt das, dass ein exprimiertes Gen (Transkriptomebene) auch in seiner Form als synthetisiertes Protein (fehlende Proteinebene) vorhanden ist, um letztendlich über Sekretionssignalwege aus der Zelle ausgeschleust zu werden (Sekretomebene) [Lehmann *et al.,* 2018]. Folgende Bedingungen müssen für die betroffenen Komponenten erfüllt sein, damit potenzielle Zusammenhänge über ModuleDiscoverer abgedeckt werden können:

(1) Sie sind als Komponenten im vorwissensbasierten PPIN enthalten; (2) sie sind differenziell reguliert und/oder weisen eine hohe Konnektivität mit anderen differenziell regulierten Netzwerkkomponenten auf; (3) sie sind Teil einer Clique im vorwissensbasierten PPIN. Sind die Bedingungen (1) bis (3) erfüllt und handelt es sich um den Zusammenhang der übereinstimmenden Komponenten, lässt sich die fehlende Proteinebene über das jeweils exprimierte Gen und/oder das sekretierte Protein im regulatorischen Modul abbilden. Wird der Ebenenzusammenhang allerdings durch Signalwege charakterisiert, müssen komplexere Szenarien berücksichtigt werden. Einzelne fehlende Komponenten können gegebenenfalls mithilfe ihrer durch Vorwissen bekannten Konnektivitätsmuster im Netzwerk bzw. Modul ergänzt werden. Allerdings können die bereits zuvor diskutierten Faktoren wie Proteinumsatz, Rückkopplung, die Wahl der Zeitpunkte etc. die Szenarien zusätzlich verkomplizieren. Grundsätzlich ist die Abbildung einer fehlenden Ebene im regulatorischen Modul davon abhängig, ob die Komponenten der fehlenden Ebene eine hohe Konnektivität zu den gemessenen Komponenten der untersuchten Ebenen aufweisen. Dementsprechend lassen sich fehlende Ebenen nur bedingt und in Abhängigkeit vom Vorwissen sowie der experimentellen Daten rekonstruieren.

3.3 Schlusswort

Die Integration von Daten aus heterogenen Quellen, wie den Multiskalen- und Multi-Omik-Daten, stellt eine wohlbekannte Herausforderung in der heutigen Biowissenschaft dar. Durch die Integration ist es möglich, ein tieferes Verständnis für die Organisation und das Zusammenspiel biologischer Systeme zu entwickeln. Die in dieser Arbeit vorgestellten Ansätze stellen zum einen bereits etablierte Herangehensweisen zur Analyse und Interpretation von Daten heterogener Quellen dar. Zum anderen wurde mit dem moduldetektierenden ModuleDiscoverer ein erst kürzlich entwickeltes, bisher nur an einem Beispiel der Leberpathologie eingesetztes, Werkzeug vorgestellt, das nun erstmals für eine infektionsbiologische Fragestellung angewendet wurde. Mit diesem können sowohl einzelne als auch multiple Datensätze unterschiedlicher Omik-Ebenen berücksichtigt werden. Durch die Einbeziehung zeitlicher Skalen lässt sich außerdem die zeitliche Dynamik des betrachteten Systems nachahmen. Es

hat sich gezeigt, dass ModuleDiscoverer, bzw. moduldetektierende Ansätze im Allgemeinen, die Integration von Multi-Omik-Daten effektiv unterstützen und dadurch tiefere Einblicke in die komplexen biologischen Zusammenhänge der einzelnen Ebenen gewähren. Dabei können auch potenzielle Schlüsselfaktoren der WPPI identifiziert werden, deren Detektion über andere klassische Ansätze nicht möglich ist. Mit der vorliegenden Dissertation zur Integration von Multiskalen- und Multi-Omik-Daten konnte ein wichtiger Beitrag zur Untersuchung von WPI am Beispiel von pathogenen Pilzen geleistet werden. Allerdings sind aufgrund ihrer Komplexität und der Grenzen der derzeit zur Verfügung stehenden Experimental- und Wissensdatenbanken sowie bioinformatischen Werkzeuge weiterführende Forschungsarbeiten nötig, um ein umfassenderes Verständnis erlangen zu können.

Literaturverzeichnis

Aebersold *et al.*, 2016

Aebersold, R., Mann, M. (2016). Mass-spectrometric exploration of proteome structure and function. *Nature,* 537(7620).
https://doi.org/10.1038/nature19949

Albrecht *et al.*, 2010

Albrecht, D., Kniemeyer, O., Brakhage, A. A., Guthke, R. (2010). Missing values in gel-based proteomics. *Proteomics,* 10(6).
https://doi.org/10.1002/pmic.200800576

Alcaraz *et al.*, 2011

Alcaraz, N., Kücük, H., Weile, J., Wipat, A., Baumbach, J. (2011). Keypathwayminer: Detecting case-specific biological pathways using expression data. *Internet Mathematics,* 7(4).
https://doi.org/10.1080/15427951.2011.604548

Alcaraz *et al.*, 2014

Alcaraz, N., Pauling, J., Batra, R., Barbosa, E., Junge, A., Christensen, A. G. L., ... Baumbach, J. (2014). KeyPathwayMiner 4.0: Condition-specific pathway analysis by combining multiple omics studies and networks with Cytoscape. *BMC Systems Biology,* 8(1).
https://doi.org/10.1186/s12918-014-0099-x

Altwasser *et al.*, 2015

Altwasser, R., Baldin, C., Weber, J., Guthke, R., Kniemeyer, O., Brakhage, A. A., ... Valiante, V. (2015). Network modeling reveals cross talk of MAP kinases during adaptation to caspofungin stress in *Aspergillus fumigatus. PLoS ONE,* 10(9).
https://doi.org/10.1371/journal.pone.0136932

Ashburner *et al.*, 2000

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., ... Sherlock, G. (2000). Gene ontology: Tool for the unification of biology. *Nature Genetics,* 25(1).
https://doi.org/10.1038/75556

Blickensdorf *et al.*, 2019

Blickensdorf, M., Timme, S., Figge, M. T. (2019). Comparative Assessment of Aspergillosis by Virtual Infection Modeling in Murine and Human Lung. *Frontiers in Immunology,* 10.
https://doi.org/10.3389/fimmu.2019.00142

Brock *et al.*, 2008

Brock, G., Pihur, V., Datta, S., Datta, S. (2008). clValid: An R Package for Cluster Validation. *Journal of Statistical Software,* 25(4).
https://doi.org/citeulike-article-id:2574494

Brown *et al.*, 2012

Brown, G.D., Denning, D.W., Gow, N.A., Levitz, S.M., Netea, M.G., White, T.C. (2012). Hidden killers: human fungal infections. *Science Translational Medicine,* 4(165).
https://doi.org/10.1126/scitranslmed.3004404

Bumgarner *et al.*, 2013

Bumgarner, R. (2013). DNA microarrays: Types, Applications and their future. *Current Protocols in Molecular Biology,* 0(22).
https://doi.org/10.1002/0471142727.mb2201s101

Butcher *et al.*, 2004

Butcher, E. C., Berg, E. L., Kunkel, E. J. (2004). Systems biology in drug discovery. *Nature Biotechnology,* 22.
https://doi.org/10.1038/nbt1017

Cassini *et al.*, 2016

Cassini, A., Plachouras, D., Eckmanns, T., Abu Sin, M., Blank, H. P., Ducomble, T., ... Suetens, C. (2016). Burden of Six Healthcare-Associated Infections on European Population Health: Estimating Incidence-Based Disability-Adjusted Life Years through a Population Prevalence-Based Modelling Study. *PLoS Medicine,* 13(10).
https://doi.org/10.1371/journal.pmed.1002150

Castiglione *et al.*, 2014

Castiglione, F., Pappalardo, F., Bianca, C., Russo, G., Motta, S. (2014). Modeling biology spanning different scales: An open challenge. *BioMed Research International,* 2014.
https://doi.org/10.1155/2014/902545

Conrad *et al.*, 2018

Conrad, T., Kniemeyer, O., Henkel, S., Krüger, T., Mattern, D. J., Valiante, V., ... Linde, J. (2018). Module-detection approaches for the integration of multilevel omics data highlight the comprehensive response of *Aspergillus fumigatus* to caspofungin. *BMC Systems Biology,* 12(1).
https://doi.org/10.1186/s12918-018-0620-8

De Keersmaecker *et al.*, 2006

De Keersmaecker, S. C. J., Thijs, I. M. V, Vanderleyden, J., Marchal, K. (2006). Integration of omics data: How well does it work for bacteria? *Molecular Microbiology,* 62(5).
https://doi.org/10.1111/j.1365-2958.2006.05453.x

Dühring *et al.*, 2017

Dühring, S., Ewald, J., Germerodt, S., Kaleta, C., Dandekar, T., Schuster, S. (2017). Modelling the host-pathogen interactions of macrophages and *Candida albicans* using Game Theory and dynamic optimization. *Journal of the Royal Society Interface,* 14(132).
https://doi.org/10.1098/rsif.2017.0095

Embrechts *et al.*, 2013

Embrechts, M. J., Gatti, C. J., Linton, J., Roysam, B. (2013). Hierarchical clustering for large data sets. *Studies in Computational Intelligence. Springer Verlag,* 410.
https://doi.org/10.1007/978-3-642-28696-4_8

Finotello *et al.*, 2015

Finotello, F., Di Camillo, B. (2015). Measuring differential gene expression with RNA-seq: Challenges and strategies for data analysis. *Briefings in Functional Genomics,* 14(2).
https://doi.org/10.1093/bfgp/elu035

Fraser, 2001

Fraser, C. G. (2001). Biological Variation: From Principles to Practice. *American Association for Clinical Chemistry,* 1. Auflage.
ISBN: 9781890883492

Fraser, 2017

Fraser, C. G. (2017). Biological variation: a rapidly evolving aspect of laboratory Medicine. *Journal of Laboratory and Precision Medicine,* 2(35).
http://dx.doi.org/10.21037/jlpm.2017.06.09

Furnham *et al.*, 2012

Furnham, N., de Beer, T. A. P., Thornton, J. M. (2012). Current challenges in genome annotation through structural biology and bioinformatics. *Current Opinion in Structural Biology,* 22(5).
https://doi.org/10.1016/j.sbi.2012.07.005

García-Campos *et al.*, 2015

García-Campos, M. A., Espinal-Enríquez, J., Hernández-Lemus, E. (2015). Pathway analysis: State of the art. *Frontiers in Physiology,* 6.
https://doi.org/10.3389/fphys.2015.00383

Gillis *et al.*, 2013

Gillis, J., Pavlidis, P. (2013). Assessing identity, redundancy and confounds in Gene Ontology annotations over time. *Bioinformatics,* 29(4).
https://doi.org/10.1093/bioinformatics/bts727

Gilsenan *et al.*, 2012

Gilsenan, J. M., Cooley, J., Bowyer, P. (2012). CADRE: The Central Aspergillus Data REpository 2012. *Nucleic Acids Research,* 40(D1).
https://doi.org/10.1093/nar/gkr971

Guthke *et al.*, 2016

Guthke, R., Gerber, S., Conrad, T., Vlaic, S., Durmus, S., Çakir, T., ... Linde, J. (2016). Data-based reconstruction of gene regulatory networks of fungal pathogens. *Frontiers in Microbiology,* 7.
https://doi.org/10.3389/fmicb.2016.00570

Hasin *et al.*, 2017

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

Hebecker *et al.*, 2016

Hebecker, B., Vlaic, S., Conrad, T., Bauer, M., Brunke, S., Kapitan, M., ... Jacobsen, I. D. (2016). Dual-species transcriptional profiling during systemic candidiasis reveals organ-specific host-pathogen interactions. *Scientific Reports,* 6.
https://doi.org/10.1038/srep36055

Heinekamp *et al.*, 2015

Heinekamp, T., Schmidt, H., Lapp, K., Pähtz, V., Shopova, I., Köster-Eiserfunke, N., ... Brakhage, A. A. (2015). Interference of *Aspergillus fumigatus* with the immune response. *Seminars in Immunopathology,* 37(2).
https://doi.org/10.1007/s00281-014-0465-1

Heyer *et al.*, 2017

Heyer, R., Schallert, K., Zoun, R., Becher, B., Saake, G., Benndorf, D. (2017). Challenges and perspectives of metaproteomic data analysis. *Journal of Biotechnology,* 261.
https://doi.org/10.1016/j.jbiotec.2017.06.1201

Jacobsen *et al.*, 2017

Jacobsen, I. D., Hube, B. (2017). *Candida albicans* morphology: still in focus. *Expert Review of Anti-Infective Therapy,* 15(4).
https://doi.org/10.1080/14787210.2017.1290524

Joyce *et al.*, 2006

Joyce, A. R., Palsson, B. Ø. (2006). The model organism as a system: Integrating „omics“ data sets. *Nature Reviews Molecular Cell Biology,* 7(3).
https://doi.org/10.1038/nrm1857

Kämmer *et al.*, eingereicht in PLOS Genetics

Kämmer, P., McNamara S., Wolf T., Conrad T., Hünniger K., Kurzai O., ... Brunke S. Strategies of pathogenic *Candida* species to survive in human blood have evolved independently. *Eingereicht in PLOS Genetics.*

Kim *et al.*, 2011

Kim, J., Sudbery, P. (2011). *Candida albicans,* a major human fungal pathogen. *Journal of Microbiology (Seoul, Korea),* 49(2).
https://doi.org/10.1007/s12275-011-1064-7

Krüger *et al.*, 2015

Krüger, T., Luo, T., Schmidt, H., Shopova, I., Kniemeyer, O. (2015). Challenges and Strategies for Proteome Analysis of the Interaction of Human Pathogenic Fungi with Host Immune Cells. *Proteomes,* 3(4). https://doi.org/10.3390/proteomes3040467

Lehmann *et al.*, 2018

Lehmann, R., Müller, M. M., Klassert, T. E., Driesch, D., Stock, M., Heinrich, A., ... Slevogt, H. (2018). Differential regulation of the transcriptomic and secretomic landscape of sensor and effector functions of human airway epithelial cells. *Mucosal Immunology,* 11(3). https://doi.org/10.1038/mi.2017.100

Lin *et al.*, 2015

Lin, C. Y., Lee, T. L., Chiu, Y. Y., Lin, Y. W., Lo, Y. S., Lin, C. T., Yang, J. M. (2015). Module organization and variance in protein-protein interaction networks. *Scientific Reports,* 5. https://doi.org/10.1038/srep09386

Linde *et al.*, 2010

Linde, J., Wilson, D., Hube, B., Guthke, R. (2010). Regulatory network modelling of iron acquisition by a fungal pathogen in contact with epithelial cells. *BMC Systems Biology,* 4. https://doi.org/10.1186/1752-0509-4-148

Linde *et al.*, 2012

Linde, J., Hortschansky, P., Fazius, E., Brakhage, A. A., Guthke, R., Haas, H. (2012). Regulatory interactions for iron homeostasis in *Aspergillus fumigatus* inferred by a Systems Biology approach. *BMC Systems Biology,* 6. https://doi.org/10.1186/1752-0509-6-6

Linde *et al.*, 2015

Linde, J., Schulze, S., Henkel, S. G., Guthke, R. (2015). Data- and knowledge-based modeling of gene regulatory networks: An update. *EXCLI Journal,* 14. https://doi.org/10.17179/excli2015-168

Liu *et al.*, 2016

Liu, Y., Beyer, A., Aebersold, R. (2016). On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell,* 165(3). https://doi.org/10.1016/j.cell.2016.03.014

Lugowski *et al.*, 2018

Lugowski, A., Nicholson, B., Rissland, O. S. (2018). Determining mRNA half-lives on a transcriptome-wide scale. *Methods,* 137.
https://doi.org/10.1016/j.ymeth.2017.12.006

Maier *et al.*, 2009

Maier, T., Güell, M., Serrano, L. (2009). Correlation of mRNA and protein in complex biological samples. *FEBS Letters,* 583(24).
https://doi.org/10.1016/j.febslet.2009.10.036

***„Microbiology by numbers“*, 2011**

Microbiology by numbers. (2011). *Nature Reviews Microbiology,* 9(9).
https://doi.org/10.1038/nrmicro2644

Müller *et al.*, 2017

Müller, M. M., Lehmann, R., Klassert, T. E., Reifenstein, S., Conrad, T., Moore, C., ... Slevogt, H. (2017). Global analysis of glycoproteins identifies markers of endotoxin tolerant monocytes and GPR84 as a modulator of TNFα expression. *Scientific Reports,* 7(1).
https://doi.org/10.1038/s41598-017-00828-y

Murphy *et al.*, 2014

Murphy, K. M., Travers, P., Walport, M. (2014). Janeway Immunologie. *Springer-Verlag Berlin Heidelberg,* 7. Auflage.
ISBN: 9783662442272

Oyelade *et al.*, 2016

Oyelade, J., Isewon, I., Oladipupo, F., Aromolaran, O., Uwoghiren, E., Ameh, F., ... Adebiyi, E. (2016). Clustering Algorithms: Their Application to Gene Expression Data. *Bioinformatics and Biology Insights,* 10.
https://doi.org/10.4137/BBI.S38316

Pavlidis *et al.*, 2004

Pavlidis, P., Qin, J., Arango, V., Mann, J. J., Sibille, E. (2004). Using the gene ontology for microarray data mining: A comparison of methods and application to age effects in human prefrontal cortex. *Neurochemical Research,* 6.
https://doi.org/10.1023/B:NERE.0000023608.29741.45

Pirofski *et al.*, 2012

Pirofski, L., Casadevall, A. (2012). Q and A What is a pathogen? A question that begs the point. *BMC Biology,* 10.
https://doi.org/10.1186/1741-7007-10-6

Priebe *et al.*, 2013

Priebe, S., Menzel, U. (2013). Assignment of orthologous genes by utilization of multiple databases the orthology package in R. *In BIOINFORMATICS 2013 - Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms.*
https://doi.org/10.1016/j.jhazmat.2004.12.030

R Core Team, 2018

R Core Team. (2018). R: A language and environment for statistical computing. *Software.*
https://doi.org/ISBN 3-900051-07-0,

Roy *et al.*, 2018

Roy, S., Schmeier, S., Kaczkowski, B., Arner, E., Alam, T., Ozturk, M., ... Suzuki, H. (2018). Transcriptional landscape of *Mycobacterium tuberculosis* infection in macrophages. *Scientific Reports,* 8(1).
https://doi.org/10.1038/s41598-018-24509-6

Schleicher *et al.*, 2016

Schleicher, J., Conrad, T., Gustafsson, M., Cedersund, G., Guthke, R., Linde, J. (2016). Facing the challenges of multiscale modelling of bacterial and fungal pathogen–host interactions. *Briefings in Functional Genomics,* 16(2).
https://doi.org/10.1093/bfgp/elv064

Szklarczyk *et al.*, 2019

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., ... Von Mering, C. (2019). STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research,* 47(D1).
https://doi.org/10.1093/nar/gky1131

Vlaic *et al.*, 2012

Vlaic, S., Schmidt-Heck, W., Matz-Soja, M., Marbach, E., Linde, J., Meyer-Baese, A., ... Gebhardt, R. (2012). The extended TILAR approach: A novel tool for dynamic modeling of the transcription factor network regulating the adaption to in vitro cultivation of murine hepatocytes. *BMC Systems Biology,* 6.
https://doi.org/10.1186/1752-0509-6-147

Vlaic *et al.*, 2018

Vlaic, S., Conrad, T., Tokarski-Schnelle, C., Gustafsson, M., Dahmen, U., Guthke, R., Schuster, S. (2018). ModuleDiscoverer: Identification of regulatory modules in protein-protein interaction networks. *Scientific Reports,* 8(1).
https://doi.org/10.1038/s41598-017-18370-2

Wells *et al.*, 2011

Wells, G., Prest, H., Russ IV, C. W. (2011). Signal, Noise, and Detection Limits in Mass Spectrometry.*Agilent Technologies, Inc., Technical Note, Chemical Analysis Group.*
https://www.agilent.com/cs/library/technicaloverviews/public/5990-7651EN.pdf

Zhao *et al.*, 2014

Zhao, H., Li, Y., Wang, S., Yang, Y., Wang, J., Ruan, X., ... Fang, X. (2014). Whole transcriptome RNA-seq analysis: Tumorigenesis and metastasis of melanoma. *Gene,* 548(2).
https://doi.org/10.1016/j.gene.2014.07.038

Abbildungsverzeichnis

Tabellenverzeichnis

Ehrenwörtliche Erklärung

Die geltende Promotionsordnung der Fakultät für Biowissenschaften ist mir bekannt. Die vorliegende Dissertation habe ich selbst angefertigt. Ich habe keine Textabschnitte eines Dritten oder eigener Prüfungsarbeiten ohne Kennzeichnung übernommen. Ich habe alle von mir benutzten Hilfsmittel, persönliche Mitteilungen und Quellen angegeben. Unterstützung bei der Auswahl und Auswertung des Materials sowie bei der Herstellung der Manuskripte, habe ich nur von den genannten Autoren und den in der Danksagung genannten Personen erhalten. Die Hilfe eines Promotionsberaters habe ich nicht in Anspruch genommen. Dritte haben weder unmittelbar noch mittelbar geldwerte Leistungen von mir für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen. Ich habe die Dissertation nicht bereits zuvor als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung eingereicht. Ich habe die gleiche, eine in wesentlichen Teilen ähnliche oder eine andere Abhandlung bei keiner anderen Hochschule als Dissertation eingereicht.

Berlin, 11. Juni 2019