COMPARATIVE GENOMICS AND TRANSCRIPTOMICS ELUCIDATE VIRULENCE MECHANISMS AND HOST RESPONSES IN INFECTIOUS DISEASES

DISSERTATION

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF

“DOCTOR RERUM NATURALIUM” (DR. RER. NAT.)

SUBMITTED TO THE COUNCIL OF

THE FACULTY OF BIOLOGICAL SCIENCES

OF FRIEDRICH SCHILLER UNIVERSITY JENA

by M.Sc. Tongta Sae-Ong

born on 15 November 1990 in Phuket (Thailand)

Reviewers:

1. Assoc. Prof. Dr. Gianni Panagiotou (Hans-Knöll-Institut Jena)
2. Prof. Dr. Michael Bauer (Universitätsklinikum Jena)
3. Prof. Dr. Thomas Dandeker (Universität Würzburg)

Date of public defense: 04.10.2022 in Jena

SUMMARY

The main thematic area of the present thesis is the development and application of bioinformatics pipelines, namely whole-genome sequence (WGS) analysis and transcriptome profile analysis. These pipelines were applied to study the fungal pathogen *Aspergillus fumigatus* (**Manuscripts I, III, and IV**) and the early human immune mechanisms activated in response to different types of pathogens (bacteria, fungi, and co-infections) in sepsis patients (**Manuscript II**).

A. fumigatus is currently the deadliest airborne fungal pathogen causing different degrees of severity depending on the immune status of the patient. Invasive aspergillosis (IA), commonly caused by *A. fumigatus,* poses a high mortality rate of up to 50% for patients with immune impairment and up to 100% for patients infected by triazole resistant *A. fumigatus*. I initially studied 300 global *A. fumigatus* genomes (including 252 genomes newly sequenced in my study) to understand better the genomic differences between environmental and clinical strains (**Manuscript I**, *Nature Microbiology, 2021, co-first author*). The results showed that using pan-/core-genome analysis, only 69% of total identified genes were shared among all 300 *A. fumigatus* genomes. Furthermore, phylogenetic tree analysis demonstrated a cluster of clinical *A. fumigatus* isolates that possess accessory genes related to transmembrane transporters and proteins with iron-binding activity. Genome-wide association studies (GWAS) also revealed the most relevant nucleotide variants associated with clinical *A. fumigatus* strains and virulence factors.

Recent studies suggested that the triazole antifungal resistance in *A. fumigatus* could be derived from either fungicide used in agriculture or developed in-host during treatment. To understand the effects of agricultural fungicide on *A. fumigatus* genomes, we sequenced 64 *A. fumigatus* isolates from conventional and organic farms in Germany (**Manuscript III**, *mBIO, 2020, co-author*). The population structure of *A. fumigatus* from agricultural and organic farms was further analyzed based on nucleotide variants. The results showed that even though fungicides significantly reduced the azole susceptibility of *A. fumigatus* strains in the farms, they did not affect the genetic structure of the investigated 64 *A. fumigatus* isolates.

Moreover, genome-scale metabolic models (GEMs) of 252 *A. fumigatus* strains were reconstructed based on the analyzed whole-genome sequences to address the influence of genomic variations on the metabolic capacities of clinical and environmental strains (**Manuscript IV**, *in preparation, co-author*). The results revealed that 77% of the metabolic reactions were shared among all strain-specific GEMs, while 25 reactions were able to differentiate environmental and clinical strains. The potential impact of the metabolic characteristics of the clinical strains was further evaluated in connection with the lung microbiome in cystic fibrosis patients. Our analyses proposed that clinical strain-

related models were growing at a higher rate on media related to cystic fibrosis patients with a confirmed *A. fumigatus* infection.

Sepsis is a life-threatening organ dysfunction due to inappropriate host response to infections, with a mortality rate of up to 56%. Early recognition of sepsis and immediate management are necessary to reduce the mortality rate. However, early pathogen detection is still a challenging problem. The transcriptome profiles of elective surgery patients before the diagnosis of sepsis were analyzed in my thesis to study the early-activated human immune mechanisms related to pathogens' invasion (**Manuscript II**, *in preparation, first author*). Compared to bacterial infection, a fungal infection is more difficult to diagnose at an early stage. Therefore, the transcriptome profiles of 100 surgery patients, 51 patients who developed sepsis (35 bacteria sepsis, 3 fungal, 5 co-infected by bacteria and fungi, and 8 unidentified pathogens), and 49 patients that remained infection-free after surgery (controls), were comparatively analyzed. Gene co-expression patterns and network analysis among pathogen-induced sepsis revealed part of the signature immunity pathways and genes that respond specifically to different pathogens. These genes are promising to be predictive biomarkers for identifying causative pathogens in sepsis patients and should be further studied in the future.

The comparative genomic and transcriptomic analyses applied in my thesis have significantly improved our understanding of fungal pathogenicity as well as the pathogen-specific immune response mechanisms of the human host. Next to a number of novel insights, my work included in this thesis has generated a large number of new hypotheses based on big-data analysis, offering to the scientific community the possibility to design exciting new research to confirm them in future experimental studies and bring us closer to actual precision medicine for infectious diseases.

ZUSAMMENFASSUNG

Das Hauptthema der vorliegenden Arbeit ist die Entwicklung und Anwendung von Bioinformatik-Pipelines, insbesondere die Ganzgenomsequenzanalyse (WGS) und die Transkriptom-Profilanalyse. Diese Pipelines wurden zur Untersuchung des Pilzerregers *Aspergillus fumigatus* (**Manuskripte I, III und IV**) und von frühzeitigen menschlichen Immunmechanismen, die als Reaktion auf verschiedene Arten von Krankheitserregern (Bakterien, Pilze und Co-Infektionen) bei Sepsis-Patienten aktiviert werden (**Manuskript II**), eingesetzt.

A. fumigatus ist derzeit der tödlichste über die Luft übertragene Pilzerreger, der je nach Immunstatus der Patienten unterschiedliche Schweregrade aufweist. Die invasive Aspergillose (IA), die häufig durch *A. fumigatus* verursacht wird, hat eine hohe Sterblichkeitsrate von bis zu 50 % bei Patienten mit Immunschwäche und bis zu 100 % bei Patienten, die mit triazolresistenten *A. fumigatus* infiziert sind. Ich habe zunächst 300 globale *A. fumigatus*-Genome untersucht (darunter 252 Genome, die in meiner Studie neu sequenziert wurden), um die genomischen Komponenten zwischen Umwelt- und klinischen Stämmen besser zu verstehen (**Manuskript I**, *Nature Microbiology, 2021, Co-Erstautor*). Die Ergebnisse zeigten, dass unter Verwendung der Pan-/Kerngenomanalyse nur 69 % der insgesamt identifizierten Gene von allen 300 *A. fumigatus*-Genomen gemeinsam genutzt wurden. Darüber hinaus zeigte die phylogenetische Baumanalyse einen Cluster klinischer *A. fumigatus*-Isolate, welcher akzessorische Gene besitzt, die mit Transmembrantransportern und Proteinen mit eisenbindender Aktivität zusammenhängen. Genomweite Assoziationsstudien (GWAS) zeigten auch die wichtigsten Nukleotidvarianten, die mit klinischen *A. fumigatus*-Stämmen und Virulenzfaktoren assoziiert sind.

Jüngste Studien deuten darauf hin, dass die Resistenz von *A. fumigatus* gegen Triazole entweder auf den Einsatz von Fungiziden in der Landwirtschaft oder auf die Entwicklung im Wirt während der Behandlung zurückzuführen sein könnte. Um die Auswirkungen landwirtschaftlicher Fungizide auf die Genome von *A. fumigatus* zu verstehen, haben wir 64 *A. fumigatus*-Isolate aus konventionellen und ökologischen Landwirtschaftsbetrieben in Deutschland entnommen und ihre Genome sequenziert (**Manuskript III**, *mBIO, 2020, Co-Autor*). Die Populationsstruktur von *A. fumigatus* aus landwirtschaftlichen und ökologischen Betrieben wurde anhand von Nukleotidvarianten weiter analysiert. Die Ergebnisse zeigten, dass Fungizide zwar die Häufigkeit von Azol-empfindlichen *A. fumigatus*-Stämmen in den Betrieben deutlich verringerten, die genetische Struktur der untersuchten 64 *A. fumigatus*-Isolate jedoch nicht beeinflussten.

Darüber hinaus wurden genomweite Stoffwechselmodelle (GEMs) von 252 *A. fumigatus*-Stämmen auf der Grundlage der analysierten Ganzgenomsequenzen rekonstruiert, um den Einfluss genomischer Variationen auf die Stoffwechselkapazitäten

klinischer und umweltbedingter Stämme zu untersuchen (**Manuskript IV**, *in Vorbereitung, Co-Autor*). Die Ergebnisse zeigten, dass 77 % der Stoffwechselreaktionen bei allen stammspezifischen GEMs vorhanden waren, während 25 Reaktionen in der Lage waren, Umwelt- und klinische Stämme zu unterscheiden. Die potenziellen Auswirkungen der metabolischen Eigenschaften der klinischen Stämme wurden im Zusammenhang mit dem Lungenmikrobiom von Mukoviszidose-Patienten weiter untersucht. Unsere Analysen ergaben, dass Modelle, die sich auf klinische Stämme beziehen, auf Medien von Mukoviszidose-Patienten mit einer bestätigten *A. fumigatus*-Infektion eine höhere Wachstumsrate aufweisen.

Sepsis ist eine lebensbedrohliche Organfunktionsstörung, die auf eine unangemessene Reaktion des Wirts auf Infektionen zurückzuführen ist und eine Sterblichkeitsrate von bis zu 56 % aufweist. Die frühzeitige Erkennung einer Sepsis und eine sofortige Behandlung sind notwendig, um die Sterblichkeitsrate zu senken. Der frühzeitige Nachweis von Krankheitserregern ist jedoch nach wie vor ein schwieriges Problem. In meiner Dissertation wurden die Transkriptom-Profile von Patienten mit elektiven Eingriffen vor der Diagnose einer Sepsis analysiert, um die früh aktivierten menschlichen Immunmechanismen im Zusammenhang mit der Invasion von Krankheitserregern zu untersuchen (**Manuskript II**, *in Vorbereitung, Erstautor*). Im Vergleich zu einer bakteriellen Infektion ist eine Pilzinfektion in einem frühen Stadium schwieriger zu diagnostizieren. Daher wurden die Transkriptom-Profile von 100 chirurgischen Patienten, 51 Patienten, die eine Sepsis entwickelten (35 bakterielle Sepsis, 3 Pilzinfektionen, 5 Koinfektionen mit Bakterien und Pilzen, und 8 nicht identifizierte Erreger), und 49 Patienten, die nach der Operation infektionsfrei blieben (Kontrollen), vergleichend analysiert. Die Koexpressionsmuster der Gene und die Netzwerkanalyse bei der durch Krankheitserreger ausgelösten Sepsis zeigten einen Teil der charakteristischen Immunitätswege und Gene, die spezifisch auf verschiedene Krankheitserreger reagieren. Diese Gene sind vielversprechende prädiktive Biomarker für die Identifizierung der verursachenden Erreger bei Sepsispatienten und sollten in Zukunft weiter untersucht werden.

Die vergleichenden Genom- und Transkriptomanalysen, die in meiner Dissertation angewandt wurden, haben unser Verständnis der Pathogenität von Pilzen sowie der erregerspezifischen Immunantwortmechanismen des menschlichen Wirts erheblich verbessert. Neben einer Reihe neuer Erkenntnisse hat meine Arbeit im Rahmen dieser Dissertation eine große Anzahl neuer Hypothesen auf der Grundlage von Big-Data-Analysen hervorgebracht, die der wissenschaftlichen Gemeinschaft die Möglichkeit bieten, aufregende neue Forschungsarbeiten zu konzipieren, um sie in künftigen experimentellen Studien zu bestätigen und uns der eigentlichen Präzisionsmedizin für Infektionskrankheiten näher zu bringen.

TABLE OF CONTENTS

SUMMARY .... **I**

ZUSAMMENFASSUNG .... **III**

TABLE OF CONTENTS .... **V**

ABBREVIATIONS .... **VI**

CHAPTER I INTRODUCTION **1**

1. GENOMIC DATA - NEW CHALLENGES IN BIOLOGICAL RESEARCH 1
   1.1 Sequencing Technologies – The Powerful Technology for Genomic Data 1
   1.2 Applications of Sequencing Technologies - WGS becomes more feasible in microbiology 3
   1.3 Comparative Genomics - The Comprehensive Analyses of WGS 4
2. TRANSCRIPTOMIC DATA - FROM GENOME TO GENE FUNCTIONS 12
   2.1 Transcriptome data Generation 12
   2.2 Normalization and Differentially Expressed Genes (DEGs) Identification 13
   2.3 Gene co-expression Analysis and Functional Annotation 14
3. APPLICATION OF GENOMICS AND TRANSCRIPTOMICS IN HUMAN INFECTIOUS DISEASES 17
4. COMPUTATIONAL WORKFLOW FOR GENOME AND TRANSCRIPTOME DATA ANALYSES IN THIS THESIS 19
   4.1 Whole-genome sequence (WGS) analysis workflow 19
   4.2 Transcriptome profile analysis workflow 20

CHAPTER II OBJECTIVES **22**

CHAPTER III RESEARCH PUBLICATIONS **23**

List of Research Publications 23
MANUSCRIPT I 24
Overview 24
Manuscript 26
MANUSCRIPT II 42
Overview 42
MANUSCRIPT III 73
Overview 73
Manuscript 75
MANUSCRIPT IV 93
Overview 93
Manuscript 95

CHAPTER IV DISCUSSION **126**

REFERENCES **133**

APPENDICES **144**

DECLARATION **148**

CURRICULUM VITAE **149**

ABBREVIATIONS

	Full Name		*Full Name*
BP	biological process	**MF**	molecular function
BUSCO	Benchmarking Universal Single-Copy Ortholog	**ML**	maximum likelihood
BWT	Burrows-Wheeler Transform	**MRSA**	methicillin-resistant *Staphylococcus aureus*
CAD	coronary artery disease	**MSA**	multiple sequence alignment
CC	cellular compartment	**NGS**	next-generation sequencing
cDNA	complementary DNA	**OG**	orthologous group
CNV	copy number variant	**ORA**	overrepresentation analysis
CSP	putative cell surface protein	**PFAM**	protein families
DEG	differentially expressed gene	**PPI**	protein-protein interaction
ENA	European Nucleotide Archive	**QC**	quality control
EVM	EVidenceModeler	**QTL**	Quantitative Trait Locus
FC	fold change	**QUAST**	quality assessment tool for genome assemblies
FDR	false discovery rate	**RNA-seq**	RNA-sequencing
GATK	Genome Analysis Tool Kit	**RV**	Rhinovirus
GEO	Gene Expression Omnibus	**SIRS**	systemic inflammatory response syndrome
GO	gene ontology	**SMRT**	single-molecule real-time
GSEA	gene set enrichment analysis	**SNV**	single nucleotide variant
GWAS	genome-wide association study	**SRA**	Short Read Archive
HC	hierarchical clustering	**STRING**	Search Tool for the Retrieval of Interacting Genes/Proteins
HMM	Hidden Markov Model	**SV**	structural variant
HTS	high-throughput sequencing	**TGS**	targeted gene sequencing
IA	invasive aspergillosis	**TOM**	topological overlap matrix
IAV/	influenza A virus/	**UFBoot**	ultrafast bootstrap approximation approach
IBV	influenza B virus	**UniProt**	Universal Protein Knowledgebase
ICU	intensive care unit	**VEP**	ENSEMBL's Variant Effect Predictor
InDel	Insertion/deletion	**WES**	whole-exome sequencing
KEGG	Kyoto Encyclopedia of Genes and Genomes	**WGCNA**	Weighted Gene Co-expression Network Analysis
LD	linkage disequilibrium	**WGS**	Whole-genome sequence
LMM	linear mixed model	**WKS**	weighted Kolmogorov Smirnov
MCL	Markov Clustering Algorithm		

CHAPTER I INTRODUCTION

1. Genomic Data - New Challenges in Biological Research

With a growing number of sequencing data according to the advancement of sequencing technologies, bioinformatics study has become more important in biological research. The amount of releasing genomes has triggered new opportunities in biological studies that were not previously possible such as discovering new genes [1], building new genomes of unknown organisms [2], and identifying disease-associated gene markers [3].

Whole-genome analysis was initiated even before releasing the first draft of the human genome sequence. There were complete genome sequences from a group of model organisms that possess less complicated and smaller genomes than humans, such as bacterial and fungal genomes [4]. The genomic data were openly stored and shared through online databases. For example, Blackwell *et al*. recently gathered bacterial sequences from the European Nucleotide Archive (ENA) and assembled them into 661,405 bacterial genomes [5]. Similarly, more than thousands of the fungal genomes were sequenced and stored in several resources, such as MycoCosm, which also hosts the 1,000 Fungal Genomes Project [6], FungiDB [7] and Ensembl Fungi [8]. In this chapter, the sequencing data generated by new sequencing technologies, data processing, data analysis, and interpreting in the biological contexts are introduced in the following sections.

1.1 Sequencing Technologies – The Powerful Technology for Genomic Data

Over Over the last decades, sequencing technologies have been developed to increase efficiency and reduce time and cost. The first-generation sequencing technology, or “Sanger Sequencing”, was developed by Sanger and his team in the 1970s [9]. The complete human genome was first sequenced by this sanger sequencing and published in the same year with the Human Genome Project in 2001 [10,11]. This method synthesized DNA sequences by using the “chain-terminating dideoxynucleotides” technique. It has the advantages of high accuracy sequencing and can achieve long read lengths up to ~1,000 bp [12]. However, due to its relatively high cost, this technology was replaced by second-generation sequencing or more commonly called next-generation sequencing (NGS).

NGS technology uses a “clonal amplification” technique which supports massively parallel sequencing and real-time signal detection [13,14]. Therefore, NGS technology has a significant advantage of rapidly producing massive sequencing data. Sequencing platforms of this generation were developed by several companies, including 454 sequencing from Roche [15], Illumina [16] and SOLiD or small oligonucleotide ligation and detection system from Applied Biosystems [17]. Owing to the exponential reduction of NGS cost and sequencing time, the larger human genome project called the “1,000

Genomes Project" was started in 2007 [18]. Using NGS platforms, over 2,500 individual healthy people worldwide were sequenced to investigate genetic variations in the human population [18,19]. The genetic variants from this project have been used for genotype imputation in Genome-wide Association Studies (GWAS) in variant-associated diseases such as coronary artery disease (CAD) [20], type 2 diabetes [21,22] and recent pandemic disease COVID-19 [23]. However, the read lengths generated by these platforms were considerably short (~25-700 bp), which posed significant challenges in the subsequent genome assembly [24–26].

To overcome the limitation of short-read platforms, 'The third-generation sequencing' platforms were developed to sequence single DNA molecules in real-time [13,24]. This technology was first invented by HeliCos BioSciences [27] and followed by 'single-molecule real-time (SMRT)' from PacBio [28]. SMRT PacBio is capable of sequencing reads up with a length of more than 20 Kb [26]. This platform has excellent potential for *de novo* genome assembly as well as the identification of structural variants (SVs) [29]. Albeit SMRT PacBio can generate long-reads, the accuracy per base is still lower than short-reads [26]. Therefore, this platform has an explicit limitation in detecting small variants such as single-nucleotide variants (SNV) [24].

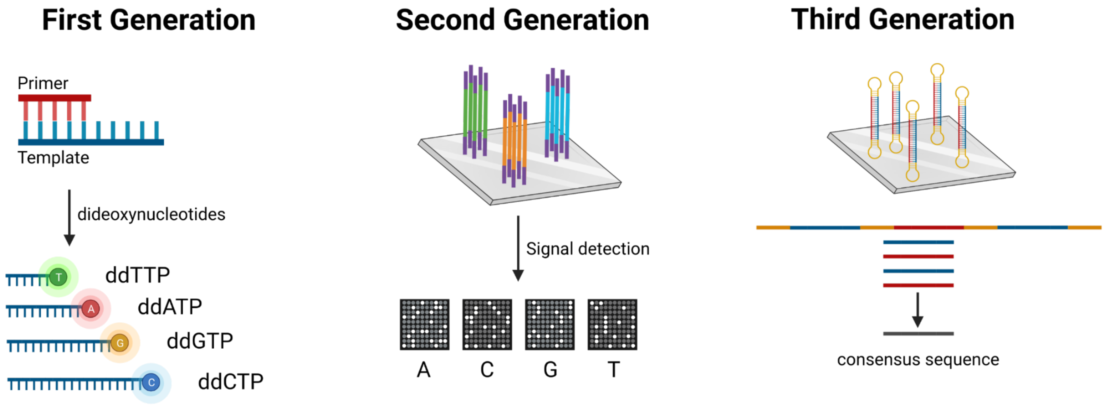


Figure 1: The first-, second-, and third-generation sequencing technologies. The first-generation sequencers generate sequences by synthesis. "Dideoxynucleotides" are tagged by fluorescence. Then, they are sequentially added, and the signals of the presented nucleotides are detected. Second-generation or next-generation sequencing (NGS) uses massively parallel sequencing to achieve high throughput and high accuracy of base detection. While the third-generation sequencing generates long reads at high throughput by using single-molecule-based method. This figure is based on the figure from Ronholm *et al.* (2016) [24].

In 2014 the "MinION" sequencer was released by the Oxford Nanopore Technologies [30,31]. In contrast to other platforms, the MinION comes with new concepts of sequencing technologies. The nanopore technology of MinION can distinguish a single nucleotide on ssDNA from its electronic and chemical characteristics, while other platforms require secondary labeling signals [25,31]. The MinION has a USB device size that can perfectly connect to a USB port and process on a personal computer [14]. Besides, the MinION can generate "ultra-long" sequences up to 200 Kb DNA [14,26]. Nonetheless, while Illumina's sequences have an error rate of less than 1%, long sequences show about 15-30% error rates [14,24].

1.2 Applications of Sequencing Technologies - WGS becomes more feasible in microbiology

Due to the great advantages of NGS, a large amount of genomic data has been generated through three main processes, including targeted gene sequencing (TGS), whole-exome sequencing (WES), and whole-genome sequencing (WGS) [32,33]. While WGS is a DNA sequencing of the entire genome of the organisms, WES and TGS are more specific by definition.

TGS and WES were designed to capture regions/genes of interest. As the sequencing regions are focused, TGS can generate ultra-high depth sequencing up to more than 1,000x [34,35], beneficial for rare variant detection [36]. WES generates a coverage depth of 100x by average, showing better quality of SNV detection than WGS, with a mean coverage depth of 30x [37]. While TGS data are more manageable, accurate, and easier to interpret than WES and WGS data, the sequencing data from WES and WGS are more uniform and widespread [38–41]. More completed sequences from WES and WGS allow the detection of complex variations such as Structural Variants (SVs) or Copy Number Variants (CNVs) [41]. While TGS data are limited to identifying variants by the genes covered in panels, WES and WGS allow discovering the new variants in novel candidate genes [33]. Moreover, the almost complete genome sequences from the WGS technique also allow for building the new genome from unknown organisms, known as *de novo* assembly [33,41,42]. Keepers *et al.* showed that the WGS method could identify 2-4 times more fungal species than the TGS method from the same environmental fields [43]. Table 1 shows the summary of the advantages and disadvantages of TGS, WES, and WGS.

Table 1: Advantages and disadvantages of TGS, WES, and WGS [33]

Sequencing Techniques	Advantages	Disadvantages
Targeted gene sequencing (TGS)	• Sequences several genes • The most cost-effective • The least time processing • Detect the most significant variants • The most accurate technique	• Only sequences regions/genes in the targeted panel • Lacking the ability to discover new genes • Low coverage of intronic regions

Whole-exome sequencing (WES)	• Sequences >90% coding exons • Lower cost than WGS • Can discover new genes	• Low coverage of pseudogenes and GC–rich regions • Poor detection of SVs • Intronic variants can be missed • Higher cost than TGS • Less accuracy than TGS
Whole-genome sequencing (WGS)	• Sequences all coding and non-coding regions • Can detect mutations in intronic or regulatory regions • Most effective for SV and CNV detection • More consistent coverage of sequences • Can discover new genes • Less false-positive rate than WES	• Identifies the most non-significant variants • Highest cost • Highest error rate • The most time consuming

As each sequencing technology has different advantages and limitations, the sequencer and sequencing techniques are mainly chosen depending on the applications of the sequences. For example, we performed second-generation sequencing with Illumina platforms to generate whole-genome sequences of *A. fumigatus* in **manuscripts I, III, and IV** to achieve high accuracy and unbiased genome sequences. Likewise, whole transcriptome sequences of the human host in **manuscript IV** were generated using the Illumina microarray, further introduced in part 2 – Transcriptomic Data.

1.3 Comparative Genomics - The Comprehensive Analyses of WGS

Various bioinformatics tools and algorithms have been developed to analyze WGS from NGS technology, from raw sequencing reads to biological interpretation. The bioinformatics workflow for NGS data analysis comprises three main parts, including (i) raw data detection and analysis, (ii) whole-genome assembly and variant detection, and (iii) variants annotation (Figure 2) [44]. In addition, whole-genome sequences can be assembled either depending on a reference genome sequence or reference-free, *de novo* assembly methods [44,45].

The first part of the NGS study is generating sequencing data, commonly done by the sequencing companies. First, the NGS machines generate sequencing reads. Then the built-in base-calling software will obtain bases by detecting the intensities of bases in the machines [46]. To access base-calling quality, a quality score per base is measured using a phred-like algorithm and reported an error probability as a logarithmic based score [47]. Phred scores ≥ 20 represent error rates ≤ 1% or base accuracy ≥ 99%, which are considered high-quality scores [48]. The most popular quality control (QC) tool, such as `FastQC`, will report the quality and statistics of sequence data [49]. Trimming tools such as `Trimmomatic` [50] and `Trim Galore` [51] will remove adapters and low-quality bases from sequence reads.

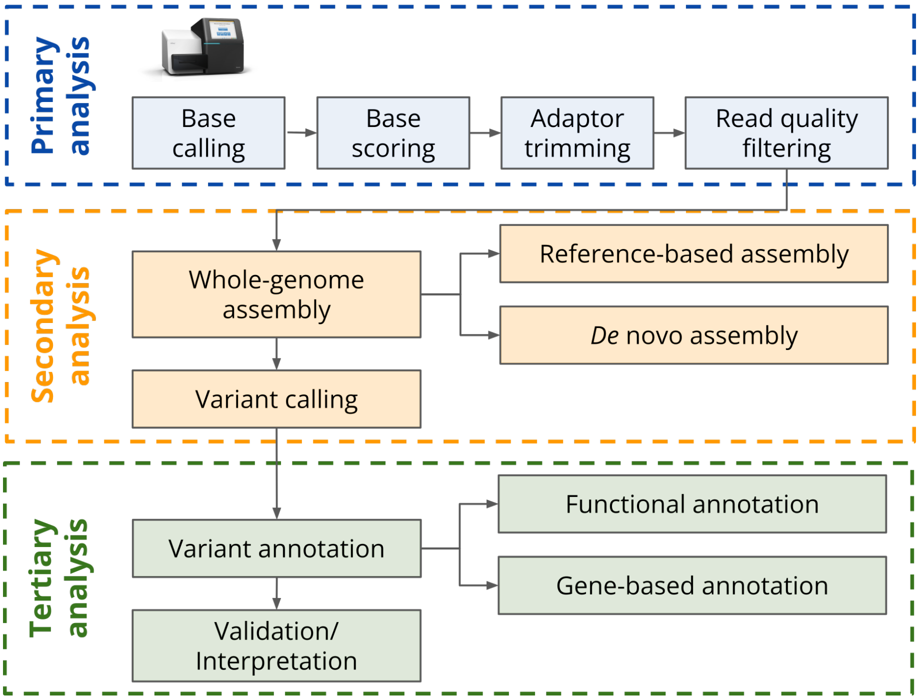


Figure 2: An overview of the NGS bioinformatics workflow. The workflow contains three main parts: the primary, secondary, and tertiary analysis. The primary data analysis is generating high-quality sequencing data. The secondary analysis includes the reconstruction of whole-genome sequences and calling variants. The tertiary includes the variant annotation and downstream analysis. This figure is based on the figure from Pereira *et al*. (2020) [44].

1.3.1 Reference-based Genome Assembly and Variant Identification

The The primary purpose of read mapping is to find the actual location of short-reads on a large reference genome and quantify the sequence similarity between short-reads and the reference [44,52]. Once individual short-reads are aligned on the reference, the sequence differences, such as single nucleotide variants (SNVs), Insertions/Deletions (InDel), and structural variants (SVs), could be identified by variant calling. In **manuscripts I and III**, I assembled *A. fumigatus* genomes using the reference-based method and *A. fumigatus* Af293 as a reference genome. The Af293 was a clinical strain and was the first complete genome sequence of *A. fumigatus*. The genomic variants between our genomes and the Af293 genome were detected and analyzed (**manuscripts I**

and III). Compared to the *de novo* assembly method, this alignment requires less computational resources and processing times. However, the error rates from base calling or mapping mismatches are considerable problems for variant identification. Furthermore, the limitation of this method is that it can fail to identify (i) multi-region mapped reads or (ii) mismapped reads [53].

The most common genetic variant is the single nucleotide variant (SNV), a single nucleotide substitution in individual genomes compared to the reference genome. SNVs that are presented in at least 1% of the populations are called single nucleotide polymorphisms (SNPs) [54]. The variants that occur in the protein-coding regions can result in (i) synonymous or no changing in amino acid translations, (ii) nonsynonymous or amino acid sequence changes, or (iii) nonsense or the amino acid changing to stop codon leading to truncation of the protein sequences [55]. Insertion/deletion (InDel) is the insertion or deletion of a short DNA sequence (<1Kb) in sequencing read compared to the reference [56]. These variants will change the length of the DNA sequence in a reading frame or frameshift, resulting in the mistranslation of protein sequences [57,58]. The large DNA sequences (>1Kb) that cause genome sequence rearrangements are known as structural variation (SV), including balanced variants, such as inversion and translocation, and unbalanced variants or copy number variation (CNV) [56,59].

Variant annotation is a further step to annotate and predict variant effects on protein-coding genes. Furthermore, a statistical method such as a genome-wide association study (GWAS) can be applied to identify phenotype-associated genomic variants.

Sequence alignment/mapping

Sequence alignment is a ubiquitous and fundamental procedure that compares two or more biological sequences, including DNA, RNA, or protein sequences. Whole-genome sequence alignment is end-to-end (global) alignments of DNA short-reads on a closely related reference genome. Since 2000, more than 60 whole-genome sequence aligners/mappers have been developed for high-throughput sequencing (HTS) data [60]. However, the two most robust and common mappers, `BWA` [61] and `Bowtie2` [62], were developed using the "Burrows-Wheeler Transform (BWT)" algorithm, which has advantages of speed and memory-efficiency [63–65]. Briefly, BWT transforms the reference sequence by fragmenting the whole genome into subsequences. The subsequences are then sorted and stored in the indexing table. Finally, the entire short-reads are aligned against the reference subsequences. This study used `BWA` as a high-accuracy mapping tool to map *A. fumigatus* short-reads against the Af293 reference genome.

Variant calling

Variant calling is a step of identifying nucleotide changing position compared to the reference genome [66]. Among all available variant calling tools, the most widely used is `Genome Analysis Tool Kit (GATK)`. `GATK` yields the best overall performance of SNV detection by using a Bayesian model to identify the true variants from base calling or mapping errors [67]. Furthermore, when applying mapped read pre-processing and low-quality variant filtering from the `GATK` framework, the variant quality can reach >99% accuracy [68,69]. `GATK` also outperformed other variant callers to detect more complicated variants such as InDels [70–72]. Moreover, this tool kit also showed the best performance in detecting variants in metagenome data [73]. In addition, `GATK` has an integrated variant annotation tool, `snpEff` [74]. Therefore, `GATK` was chosen to call variants of *A. fumigatus* sequences in this study.

Structural variants (SVs) are considerably more difficult to detect accurately and less common than small ones. However, larger variants have greatly impacted genetic functions [75,76]. Therefore, several tools have been developed to detect CNV, the most common SV impacting mRNA expression level [77]. Most CNV detection algorithms are based on imbalanced mapped reads on the reference genome [78]. These methods showed different advantages and limitations among them. In this study, the most flexible software `Control-FREEC`, which can detect CNV with/without control samples [79], was used to detect CNVs of *A. fumigatus* genomes. This tool also showed good performance in detecting CNV from WGS and WES data [80]. Moreover, it is available to analyze CNV in various organisms with different ploidy [78,79].

Variant annotation

Variant annotation is a key step in linking the detected variants to biological context. To predict the functional effects of called variants, several variant annotations have been developed, such as the most widely used tools `ANNOVAR` [81], `ENSEMBL's Variant Effect Predictor (VEP)` [82] and `snpEff` [74]. However, `snpEff` has shown a higher accuracy of variant prediction (>94%) compared to `ANNOVAR` (~80%) [83], which was used in this thesis. `SnpEff` can also annotate SNP and InDel variants on coding and non-coding region [44,84].

1.3.2 *De novo* assembly and Gene Prediction

An alternative genome assembly method, *de novo* assembly, has been primarily used when the reference genome is unavailable [44]. However, scientists recently noticed that a single reference genome is insufficient to represent the whole population [85]. In 2018, a study by Garcia-Rubio *et al*. showed the variable of called variants in *A. fumigatus* when using different reference genomes, Af293 and A1163 [86]. Their results also showed that these two reference strains belong to different lineages of *A. fumigatus* in a phylogenetic tree [86]. In this thesis, I also assembled 300 global *A. fumigatus* genomes using the *de novo* assembly method to overcome the limited materials of a single reference genome. In **manuscript I**, 300 *A. fumigatus de novo* assembled genomes were used to study pan-/core-genomes. In **manuscript IV**, 252 *A. fumigatus de novo* assembled genomes from Germany were used to build strain-specific genome-scale metabolic models

(GEMs) by Mohammad and Chen. These studies extended our understanding of genomic variations of *A. fumigatus* at genomic and metabolic levels. Compared to the read mapping method, this method provides more advantages for discovering comprehensive structural variants and functional components. However, the small variant detection is more challenging than reference-based assembly methods, as the variants could be indicated as sequencing errors [87].

De novo assembly

The main purpose of *de novo* genome assembly is to create a consensus genome sequence from the random short-reads [88]. The sequencing reads are merged based on the overlapped nucleotides to form the longer contiguous DNA sequences called "contigs" and merged "contigs" to "scaffolds" [89]. The most effective de novo assembly algorithm is the *de Bruijn* graph-based algorithm that can build the genome sequence from millions of short-reads [90]. This algorithm will construct the DNA sequence graph based on sequence similarity or overlapped based on the exact length of subsequences (*K*-mer) [91]. Several assemblers have been developed based on the *de Bruijn* graph algorithm, such as `VELVET`, `SOAPdenovo`, and `IDBA`. Among all assemblers, `IDBA` showed the overall best genome assembling for NGS data [92]. Furthermore, the *de novo* assembly method is extended by using a closely related reference genome as a guidance genome, outperforming the assembly method without using the reference genome [92].

To assess the quality of assembled genomes, the most commonly used metric is N50 [93]. N50 is calculated by sorting the contigs based on their lengths. The shortest length that contains 50% of genome length is the N50 value. Large N50 values reflect better assemblies as the genome is better combined. However, N50 is unable to provide more genome information. Therefore, other statistical metrics and genome information such as contig lengths, genome coverage, and predicted genes should be estimated to assess the assembled genome quality [94,95]. The `quality assessment tool for genome assemblies (QUAST)` is one of the most comprehensive quality assessment tools for *de novo* genome assembly [95]. `QUAST` provides complete metrics of contig sizes, misassemblies, genome elements, and functional prediction [95]. In addition, the `Benchmarking Universal Single-Copy Ortholog (BUSCO)` is another tool that has been used to measure the completeness of assembled genomes from expected gene content estimation [96].

Gene prediction and annotation

Gene prediction is the following step to identify which regions of the assembled genome contain coding genes. Two main strategies are used to predict and annotate genes in the newly assembled genomes, including (i) empirical or sequence similarity-based gene finding and (ii) the *ab initio* or *de novo* gene finding methods [97,98]. Several gene predictors have been successfully developed using a probabilistic Hidden Markov Model (HMM) for training gene sequence prediction [97,98]. For example, a transcript-based tool

`AUGUSTUS` uses the HMM to train RNA-seq data from the close species [99], while *ab initio* gene predictor `GENEMARK-ES` uses the HMM to train the query sequences themselves as known as the "self-training" method [100] for more accurate gene prediction. However, there is no "gold standard" for gene prediction software. Therefore, another strategy for gene prediction is combining results from different gene predictors to increase the total power and accuracy of gene prediction and annotation [98]. `EVidenceModeler (EVM)` is a predicted gene result combining software that weights and combines the results from several gene predictors to provide high-quality and reliable gene prediction results [101]. In this study, the `EVM` was used to combine the results from `AUGUSTUS` and `GENEMARK-ES`, following `funannotate`, a eukaryotic genome annotation pipeline [102]. Then, the predicted genes are translated to protein sequences. Next, the translated protein sequences are searched for similar protein sequences from available protein databases such as the protein families (PFAM) [103], the Universal Protein Knowledgebase (UniProt) [104] and the Kyoto Encyclopedia of Genes and Genomes (KEGG) [105,106] databases. Finally, the protein functions are assigned based on high similarity scores.

1.3.3 Genetic Diversity and Genome Evolution

Genome-wide variant detection from WGS allows us to understand evolutionary genetics. Many strategies have been used to study genome evolution and population structure, including analyses based on nucleotide variations (**manuscripts I and III**) and pan-genome analyses based on gene presence-absence variation (**manuscript I**).

Population structure

Genetic structure among populations, including intra-species and inter-species, is one of the important methods for evolutionary biology analysis. The most common and fundamental method is using statistical methods to estimate the levels of nucleotide changes within (intra-) and between (inter-) populations of organisms. Tajima's D [107], linkage disequilibrium (LD) [108] and nucleotide diversity (π) [109] are the most commonly used for intra-population genetic variations, and Wright's fixation index (*Fst*) [110], which they calculate the genetic differences based on nucleotide variations, for inter-population.

Pan-/Core-genome analysis

The pan-genome analysis is the comprehensive method of genomic diversity estimation. Pan-genome represents a complete set of orthologous and unique genes in a specific population. Orthologous genes or orthologs are groups of genes that share the same function in different species, and they evolved from the same ancestor by speciation [111]. Therefore, they are essential keys to understanding the evolutionary history of genes and/or genomes of interest. In pan-genome analysis, orthologous genes include "core" genes or genes found in all genomes and "accessory" or "dispensable" genes or genes found in some genomes [112]. The unique genes are genome-specific genes [112].

In order to identify orthologous genes, the pairwise relationship based on sequence similarity between genes is calculated. Then, the genes with high similarity scores are clustered as an orthologous group (OG). Several tools have been developed based on the "Markov Clustering Algorithm (MCL)" model, the most widely used gold standard algorithm of orthologous gene clustering [113]. `OrthoFinder` was also developed using MCL with phylogenetic trees to cluster the orthologous groups [114]. Compared to other tools that have the bias from protein lengths, `OrthoFinder` can better cluster orthologous proteins which different lengths [114]. Once the genes are clustered, core genes, accessory genes, and unique genes are identified due to the presence of genes in the population.

Genome-wide association study (GWAS)

To evaluate the association between variants and phenotype, a genome-wide association study (GWAS) is the most effective statistical method to identify the genetic variants that are responsible for phenotypic differences. GWAS employs a linear model to calculate the correlation between variants and phenotypes and tests the significance of variants based on the assumption that the variant (or SNP) does not affect the phenotype [115]. As a comprehensive screening tool for whole-genome variants in a large population, GWAS is more powerful than conventional mapping approaches such as Quantitative Trait Locus (QTL) mapping [116,117]. However, the limitation of GWAS is that it requires a very large sample size in order to reduce false-positive results [118]. Therefore, some software, such as `EMMAX` [119] and `GEMMA` [120], use a linear mixed model (LMM) instead of a simple linear model to correct the population structure effects.

Recently, larger-scale GWAS has been developed, called "panGWAS". Pan-GWAS employs the GWAS concept but screens for absent/present genes associated with the phenotypes of interest [121].

Phylogenetic tree construction

The most powerful and popular process to study evolutionary relationships among different genomes is phylogenetic tree reconstruction. The phylogenetic tree is constructed based on the closely or highly diverged genomes. The tree consists of branches representing the distance between genomes and nodes representing the genomes of interest. There are two types of phylogenetic trees; rooted and unrooted trees [122]. A rooted tree is more beneficial than an unrooted tree because it can reveal the most common ancestor genomes in the tree. To construct the rooted tree, another genome from a close species needs to be included. However, the rooted genome should be different from the genomes in the study. For example, *A. fischeri*, a closely related species of *A. fumigatus*, was used as a rooted genome in the trees in this thesis (**manuscripts I and III**).

The reconstruction of the phylogenetic tree comprises two main steps (i) multiple sequence alignment and (ii) phylogenetic tree construction. First, multiple sequence alignment (MSA) will compare and equalize the length of genomes. To construct the phylogenetic trees, there are two main approaches, including (i) distance-based and (ii) character-based (tree-searching) methods [122,123]. The main algorithm that uses distance-based is the neighbor-joining method [124]. This method will calculate the distance matrix of pairs of sequences. It is fast and computationally efficient compared to character-based methods [123,124]. It is also applicable for close genome studying [124]. However, if the genomes present high genetic variations, it could affect the accuracy of genetic distance estimation [123]. The maximum likelihood (ML) is the most popular algorithm for the character-based method [125]. ML evaluates the relative probability between genomes, while those with higher probability (likelihood) are likely to be closer than the lower probability genomes [125].

To control the accuracy of phylogenetic trees, bootstrapping is applied to estimate the confidence interval of each pair of genomes in the trees. The bootstrapping scores range from 0 to 100% and are assigned to the final tree [123,126]. The bootstrap value close to 100 represents high confidence that the genomes are closely related. In this thesis, an ultrafast bootstrap approximation approach (UFBoot) [127] that is more robust and time-efficient than the standard bootstrapping method was used to calculate the clade support of the trees.

2. Transcriptomic Data - From Genome to Gene Functions

Genomic data alone is insufficient for understanding the molecular functions of genetic elements. Transcriptomics or genome-wide expression profiling studies the totality of RNAs, which is a transcriptional level of genome linking genotypes with phenotypes in specific tissues or cell types at particular conditions or time points [128]. Furthermore, transcriptome analysis can explain how genetic variants alter gene functions [129].

To study the transcriptome data, there are three important steps, including (i) transcriptome data generation, (ii) estimating gene expression levels, and (iii) normalization of gene expression data and identifying the differentially expressed genes (DEGs). In addition, as genes do not function independently, downstream analyses and interpretations such as gene network analyses and functional annotation are performed based on sets of genes or DEGs [130].analyses and functional annotation are performed based on sets of genes or DEGs [130].

2.1 Transcriptome data Generation

The two leading technologies for whole-transcriptome profiling are microarray technology and RNA-sequencing (RNA-seq) technology. In both technologies, RNA is extracted from samples and converted to complementary DNA (cDNA). cDNA is analyzed through NGS technology (RNA-seq) or microarray technology for gene expression level detection (Figure 3). However, there are different advantages and disadvantages between these technologies.

Microarray technology is the well-established high-throughput transcriptome profiling technology that can provide massive transcriptome data from isolated RNA. In this technology, RNA is reversed into cDNA and labeled with sequence tags (targets) that correspond to genes on the array (probes) [131]. The signals from probe-target hybridization are detected for measuring the abundance of cDNA. As microarray technology has been used for decades, transcriptome data of over 5,000 organisms generated by microarray technology has been published and stored in the public database "Gene Expression Omnibus (GEO)" [132,133]. However, as this technology is an array-based method, the limitation of this technology is detecting RNA from only known gene sequences. However, this technology is flexible to customize the array for narrow specific RNA, saving cost and time. Moreover, it can generate reliable, reproducible, and high-quality transcriptome data [134]. Therefore, it is still a choice for the transcriptome profiling of protein-coding genes in model organisms and direct comparisons of the transcriptome profiling from the same array platform.

On the other hand, RNA-seq has become more standard for transcriptome profiling. This technology can provide more comprehensive genome-wide expression profiles by sequencing cDNA converted from the whole transcripts, including transcripts of novel genes. The RNA-seq reads are mapped on the reference genome. Then, gene expression levels are calculated by counting mapped reads on gene regions. More than 95% of the published RNA-seq data were generated by the Illumina short-read sequencing technology and stored on the Short Read Archive (SRA) database [135,136].

According to the GEO database, around 200 studies of sepsis microarray data were available, but less than 100 studies in the SRA database with RNA-seq generated transcriptome profiles. Since more data and studies were available, the transcriptome profiles of human hosts were generated using microarray technology in **manuscript II**.

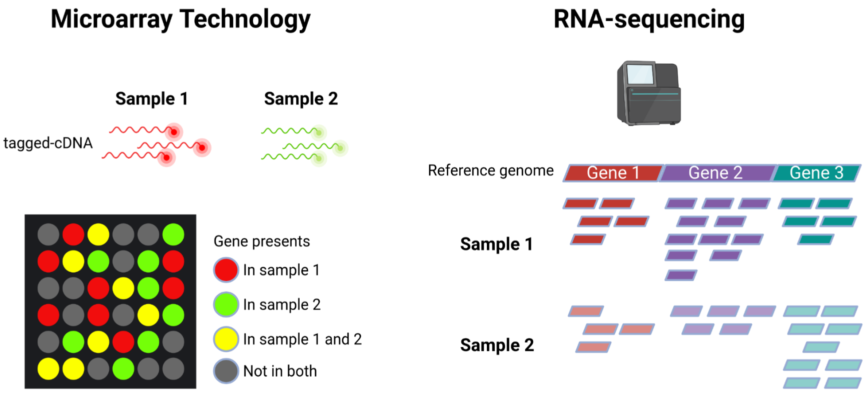


Figure 3: Microarray and RNA-sequencing (RNA-seq) technologies. The complementary DNA (cDNA) converted from RNA will be detected for the differences in expression levels between two samples/conditions. Microarray will label cDNA with fluorescent probes. The relative amount of gene expression levels will be detected by measuring fluorescence. RNA-seq will fragment cDNA and generate short-read sequences by using NGS technology. The sequences then are aligned against the reference genomes. The expression levels are calculated from the number of reads presented on genes. This figure is based on the figure from [137].

2.2 Normalization and Differentially Expressed Genes (DEGs) Identification

Normalization is the most important pre-processing step for transcriptome data analysis. Normalization aims to eliminate the experimental and technology biases for accessing the actual gene expression levels, resulting in comparable and reliable expression levels [138,139]. Microarray data normalization is done by balancing the detected color signals between genes and samples, which allows a comparison of gene expression between different datasets [140]. Then the normalized expression data is further studied for “differentially expressed genes (DEGs)”.

One of the primary purposes of transcriptomics is “differentially expressed gene (DEG)” analysis [141]. DEG analysis is the first step toward understanding the biological functions in different phenotypic conditions. DEG analysis is statistical testing that identifies the most responsible genes in different biological conditions, such as health and diseases, based on gene expression differences [130,141]. Firstly, the fold-change (FC) of each gene or a gene expression level change between conditions is estimated. Then, statistical methods have been used to estimate gene expression changes' significance in response to different conditions. In this study, the `Linear Models for Microarray Data (limma)` [142], a flexible and robust DEG analysis package in `R` software [143] that can analyze DEGs from microarray and RNA-seq data, was used to normalize the expression levels and analyze DEGs in **manuscript II**.

2.3 Gene co-expression Analysis and Functional Annotation

In biology, genes do not function independently. Genes that are expressed similarly in different conditions tend to be involved or co-regulated in the same biological pathways [144]. Therefore, to understand the biological meanings of significant DEGs, one is to extract the genes with similar expression patterns by using clustering algorithms such as the hierarchical or *K*-means clustering algorithms.

Besides, the network analysis, including gene co-expression and protein-protein interaction (PPI) networks, is increasingly used to study the system level of gene/protein functions and their interactions [145,146]. The networks comprise nodes representing genes/proteins and edges representing relationships between genes/proteins [147]. The most standard model that is widely used to reconstruct biological networks is the "scale-free" network, which can introduce the highly connected nodes or "hub" nodes [146,148]. Hubs or highly connected nodes, in this case -genes or proteins- are considered the most critical keys in biological functions as they interact with most genes/proteins in the network [149].

Gene expression matrix analysis

Gene expression matrix analysis is the most fundamental approach to discovering similarly expressed genes, as it compares the similarities or differences of expression patterns between genes and samples [150]. A gene distance matrix is firstly built by calculating the similarities/differences of pair genes between different conditions or sample groups using a distance matrix, such as the Euclidean distance method [150,151]. Then, the clustering methods, such as hierarchical clustering and *k*-means clustering, group the small distance genes into the same clusters [144,150]. In vice versa, the clusters of samples could be identified based on the similarity of gene expression patterns between samples.

Gene co-expression network

A gene co-expression network is considered the most powerful tool to estimate the gene correlations within and between modules in the network based on the expression patterns. The gene co-expression networks are categorized based on edge representative values, including "signed" or "unsigned" networks and "weighted" or "unweighted" networks [149]. Genes can have either positive or negative correlations. In "unsigned" networks, correlation values are represented as absolute values [149,152]. Thus, negatively and positively correlated genes are considered to have interactions with each other. "Signed" networks assign the negative correlation as low correlation values (<0.5) and positive correlation as high values (>0.5) [149,152].

Weighted networks represent the strength of gene correlation as continuous values from 0 to 1, while un-weighted networks represent the correlation as 0 or 1, representing no connection or genes are connected [149,153]. The most commonly used tool for gene co-expression analysis is the `Weighted Gene Co-expression Network Analysis (WGCNA)` [153]. This tool has been used widely for finding the trait-associated modules and hub genes in the identified network. The networks are built based on hierarchical clustering (HC) and tree-cutting thresholds. Then, Pearson's correlation is applied to compute the correlation between genes within and between modules. Furthermore, this method constructs the topological overlap matrix (TOM) to weigh edge scores based on the common correlated genes [153].

By using absolute values in the "unsigned" network, genes that are positively and negatively correlated could be mixed, resulting in misinterpretation of biological meaning. In contrast, the "signed" network can better separate positive and negative correlated genes, resulting in more specific biological meaning interpretation. Moreover, for a more robust result, a signed weighted network was used to construct a gene co-expression network in **manuscript II**.

Protein-protein interaction network

Protein-protein interaction (PPI) network represents the big picture of gene function interactions at the protein level [154]. The network is built by using known interactions from experimental or computational analyses in PPI databases [154]. For example, Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) is one of the most popular and vast databases of PPI that integrates protein-protein association information from high-throughput experiments, databases, literature, and computational predictions [155]. In **manuscript II**, the PPI network was also constructed based on the derived interactions from the STRING database.

Functional enrichment analysis

Functional enrichment analysis is performed to understand the changes in biological functions of the sets of genes, such as DEGs. In order to perform functional enrichment analysis, the set of genes is searched in biological annotation/pathway databases, such as Gene Ontology (GO) [156] and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways [157], which are the large and common databases. GO terms comprise three main categories of biological annotation, including biological process (BP), molecular function (MF), and cellular compartment (CC) [156]. BP represents the biological objectives or the extensive processes of gene groups involved with several genes and their function [156]. MF represents the biochemical activities of genes, and CC represents the intracellular parts that express the genes [156]. In contrast, KEGG represents the gene interaction in the biological pathways [106].

Furthermore, enrichment tools perform statistical testing to indicate the significantly enriched terms/pathways. There are several enrichment tools developed based on two main strategies, which are (i) “overrepresentation analysis (ORA)” and (ii) “gene set scoring” or “gene set enrichment analysis (GSEA)” [158]. ORA uses hypergeometric or one-sided Fisher’s exact test to evaluate the significantly overrepresented gene set in biological terms [159]. The *p*-values are corrected using the false discovery rate (FDR) [158,160]. On the other hand, GSEA will start by ranking the expression levels of a whole gene set. Genes are grouped based on their expression pattern. Then, the different enriched terms between each group are estimated using a weighted Kolmogorov Smirnov (WKS) [159].

In this thesis, the functional enrichment of DEGs was performed using the ORA strategy to enrich the specific biological meaning for pathogen-specific genes in sepsis patients from GO and KEGG databases (**manuscript II**).

3. Application of genomics and transcriptomics in human infectious diseases

Over the past decades, bioinformatics has been successfully used to study the molecular mechanisms of several human disorders, including infectious diseases, which are diseases caused by infectious agents such as viruses, bacteria, or fungi [161]. The sequencing of human and pathogen genomes drives our understanding of human susceptibility to diseases, microbial identification, microbial pathogenesis, as well as the development of new treatments [162].

The first GWAS was applied to a fungal infection study by Kumar *et al.* in candidemia patients [163]. They identified three genes associated with antifungal defense from over 200 candidemia patients [163]. In another study by Méric *et al.*, they performed the pan-genome analysis of 415 *Staphylococcus epidermidis*. They built a phylogenetic tree of *S. epidermidis* based on core genes, corresponding to 72% of the average genome size [164]. The tree suggested that each *S. epidermidis* had an equal ability to cause disease, as there was no prominent cluster of clinical isolates [164]. However, they also performed a panGWAS analysis, revealing more than 600 infection-associated genes [164]. The functions of those genes are involved in pathogenicity mechanisms such as cell toxicity, biofilm formation, and methicillin resistance [164].

More studies employed the phylogenetic tree to investigate the evolutionary relationship of human pathogens, for example, a study by Reid *et al.* [165]. They constructed a phylogenetic tree of *Escherichia coli* based on 7 housekeeping genes [165]. Their tree showed that the virulence factors in *E. coli* lineages evolve in parallel, supporting the evolution of virulence by natural selection [165]. Another example from Kiss *et al.* studied the evolution of fungal hyphae and multicellular [166]. As a result, they could identify more than 400 novel gene families with evolutionary relationships to the fungal hyphae, an important property for the invasion of several fungi [166]. Furthermore, Kim *et al.* applied a network-based approach to study methicillin-resistant *Staphylococcus aureus* (MRSA) [167]. Their results provided virulence-associated genes and suggested novel drug target genes for MRSA [167].

In addition, transcriptome profiling of blood samples from humans has been used to get insights into the complex mechanisms of the host's response to pathogens. For example, a recent study by Dissanayake *et al.* studied the transcriptomic response to different respiratory viruses, which were Rhinovirus (RV), influenza A virus (IAV), and influenza B virus (IBV) [168]. They found that chemokine- and interferon-related genes responded to all viruses [168]. Moreover, the virus-specific gene such as *ICAM5* that strongly responded to RV was also observed [168]. Another study by Parnell *et al.* compared gene expression profiles of H1N1 influenza A pneumonia, bacterial pneumonia, noninfective systemic inflammatory response syndrome (SIRS) patients, and healthy controls [169]. Using DEG analysis, clustering method, and immune cell deconvolution, their results revealed that T-cell-related immune pathways are dominantly responding to the influenza virus [169]. At the same time, neutrophil-related genes were dominantly expressed in bacterial infection [169]. As well as a study by Sweeney *et al.* used unsupervised clustering to study bacterial sepsis transcriptome profiles [170]. They found three subtypes of bacterial sepsis [170]. They also performed GO enrichment analysis to understand the differences among the three sepsis subtypes [170]. One subtype showed a significant correlation with the inflammatory signaling pathway [170]. One was significantly related to adaptive immunity and interferon signaling [170]. The third cluster was significantly related to blood coagulation pathways.

Gene and protein networks have been widely used to identify the potential biomarker genes for sepsis. For example, Zeng *et al*. identified diagnostic biomarker genes that progressively dysregulated across controls, sepsis, and septic shock patients [171]. Furthermore, they also found a decreasing in several immune cells in sepsis patients [171]. Using PPI network analysis, Zhai *et al.* can also identify signature genes for sepsis and septic shock patients [172]. Another study by Tong *et al.* identified diagnostic biomarker genes to distinguish systemic inflammatory response syndrome (SIRS) with no infection, sepsis, and septic shock patients [173].

4. Computational workflow for Genome and Transcriptome Data Analyses in this thesis

In this present thesis, I developed two bioinformatics workflows, which are "whole-genome sequence (WGS) analysis" and "transcriptome profile analysis". The WGS analysis workflow was applied in **manuscripts I, III, and IV**, to study *A. fumigatus* genomes (Figure 4). At the same time, the transcriptome profile analysis was used to characterize the human immune mechanisms responding to different types of pathogens (**manuscript II**, Figure 5).

4.1 Whole-genome sequence (WGS) analysis workflow

Aspergillus fumigatus is the most important airborne fungal pathogen. *A. fumigatus* causes different kinds of disease depending on the host's immune status [174]. In immunocompromised patients, *A. fumigatus* can cause invasive aspergillosis (IA), which poses a high mortality rate of up to 50% for immunocompromised patients [175,176]. The mortality rate can rise to 100% for patients infected with triazole-resistant *A. fumigatus* [177,178]. Recent studies suggested that the triazole antifungal resistance in *A. fumigatus* could be derived from either fungicide used in agriculture or in-host development during treatment [179]. Therefore, it is essential to understand the effects of agricultural fungicide on *A. fumigatus* genomes and the genetic variations between environmental and clinical *A. fumigatus* strains.

To understand the virulence mechanisms of *A. fumigatus*, a collection of bioinformatics tools that have been used to study fungal genomes were designed and applied to study 300 global *Aspergillus fumigatus* genomes. This WGS analysis workflow was designed for fungal genomes generated by the NGS technology. The workflow consists of quality assessment and whole-genome assembly using reference-based and *de novo* methods. The variants, including SNP/InDels and CNV, were detected and annotated the effects on the reference genes. The workflow also includes pan-/core-genome analysis and phylogenetic tree reconstruction. The meaningful variations that respond to phenotypic differences are detected using GWAS and panGWAS analysis. This workflow was applied to study 300 global *Aspergillus fumigatus* genomes.

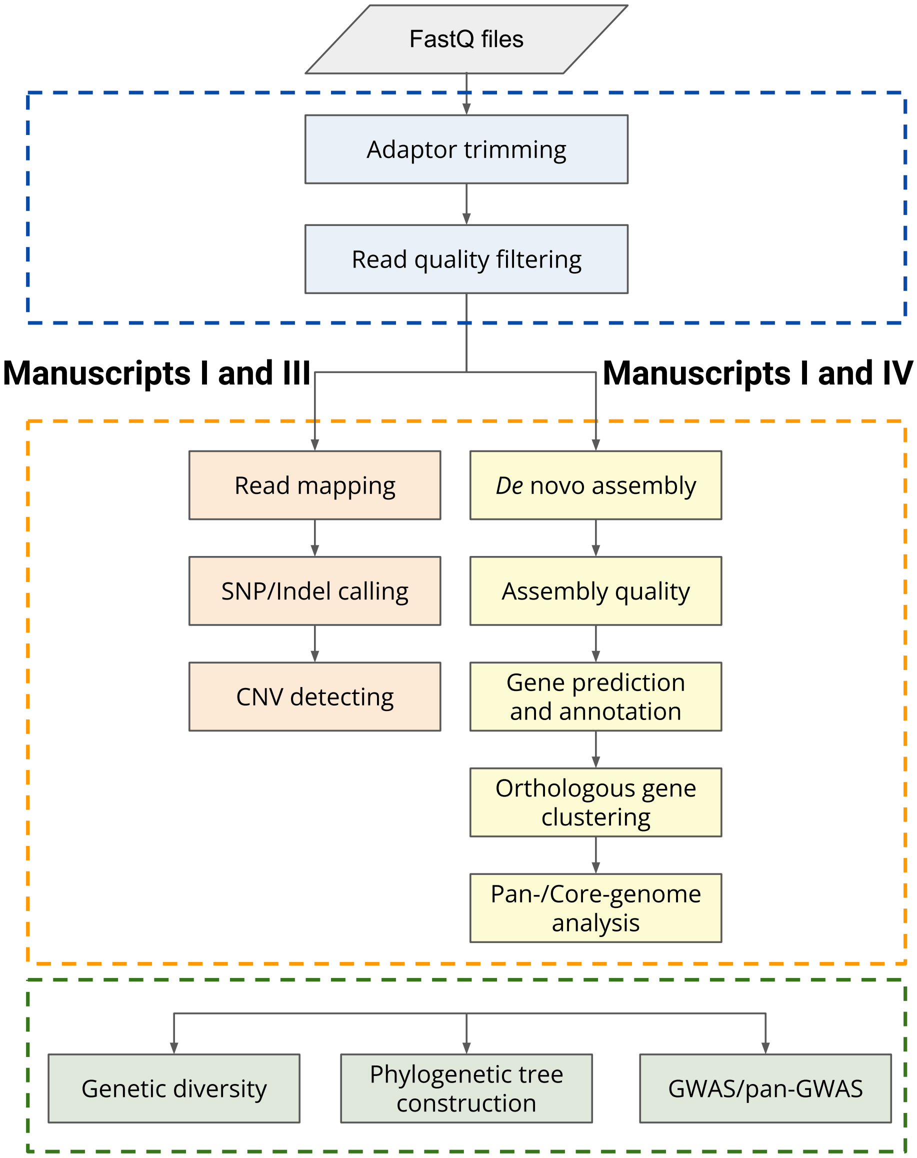


Figure 4: The whole-genome sequence analysis workflow. The workflow will use the FastQ files, which are raw sequencing data generated by NGS platforms as input. Blue boxes represent the quality control steps. Whole-genome assembly is performed using reference-based (orange, **manuscript I and III**) and *de novo* assembly (yellow, **manuscript I and IV**). The variant calling and pan-/core-genes analysis are performed. The genetic variants and pan-/core- analysis are further analyzed in downstream analysis (green).

4.2 Transcriptome profile analysis workflow

Sepsis is a life-threatening organ dysfunction due to inappropriate host response to infections, the most common cause of death among hospitalized patients in the intensive care unit (ICU) [180,181]. The mortality rates of sepsis are higher than 15% and can rise to 56% when patients present septic shock, while males have a higher mortality rate than females [182,183]. Recently, early recognition of sepsis and immediate management can reduce the mortality rate. However, early pathogen detection is still a challenging problem due to the lack of effective biomarkers [184,185]. Furthermore, compared to bacterial infection, a fungal infection is more difficult to diagnose at an early stage [186]. Therefore, gene expression data analysis is a practical approach to studying the differences in immunity mechanisms and identifying potential gene markers to detect bacterial and fungal pathogens in an early stage of infection.

In this thesis, another workflow of transcriptome or gene-expression data analysis was designed to study the differences in immune responses to bacteria, fungi, and co-infection of bacteria and fungi in sepsis patients. The workflow was divided into pre-processing and downstream analysis (Figure 5). Pre-processing analysis includes quality assessment and DEG analysis. Once DEGs are determined, the differences in biological functions are interpreted by gene clustering, network constructions, and gene set enrichment analysis. This workflow also can estimate the hub genes from gene co-expression and PPI networks to identify potential biomarkers for pathogen-specific sepsis patients.

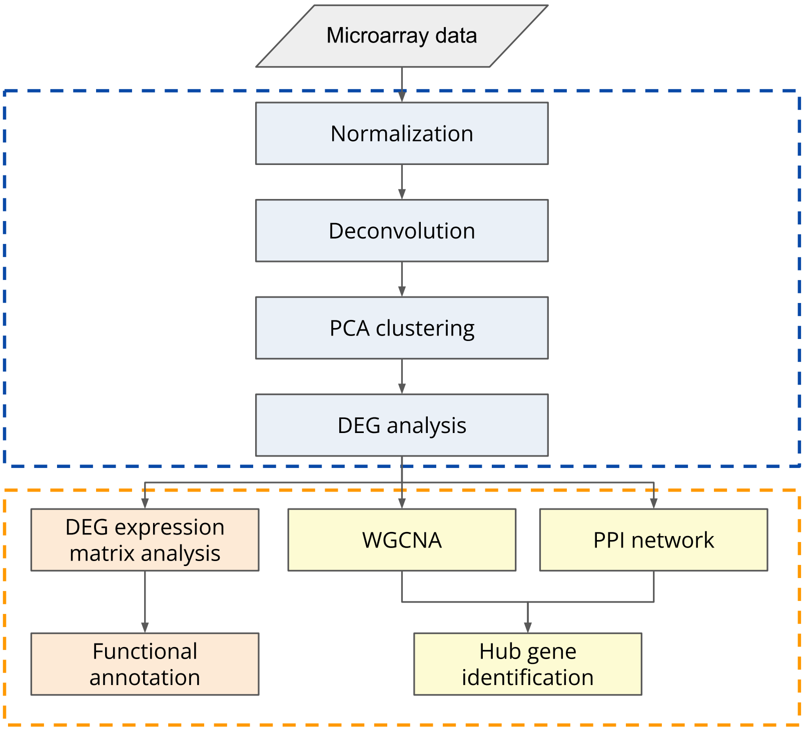


Figure 5: The microarray data analysis workflow. The relative gene expression levels generated by microarray technology are used as the input for this workflow. In quality steps, the expression levels are normalized between arrays. The whole-transcriptome profiles are observed by deconvolution and PCA analysis. DEG analysis is performed to identify the most differential gene expressions between different conditions. DEG list is further analysed to study the co-expression pattern using matrix and networks. The functional interpretation is performed by gene set enrichment analysis. Hub genes are also identified from network analysis.

CHAPTER II OBJECTIVES

This thesis aimed to implement bioinformatics approaches, including whole-genome and transcriptome analyses, to extend the knowledge of fungal pathogenicity and human immune responses to infectious microbes. During my research, state-of-art genomics methods were established and employed to generate new hypotheses and reveal genes and/or mechanisms that are important factors for *A. fumigatus* colonization and proliferation in the human host. In parallel, I successfully applied an array of computational methods (transcriptomics, network biology, and machine learning) to find pathogen-specific biomarkers for the early diagnosis of sepsis caused by bacteria, fungi, or concomitant presence of both.

The 3 research papers (2 published and one in preparation) that form part of this thesis aim to study the genetic diversity of *A. fumigatus* among different environments and overcome the limitations of using the *A. fumigatus* reference genome by generating a pan-genome using 300 globally distributed *A. fumigatus* strains. These studies served as the foundation to answer the following questions:

1. Which are the genomic similarities and differences among *A. fumigatus* strains of varied origins?
2. Which are the evolutionary relationships among *A. fumigatus* isolated within Germany and global strains?
3. Are there different evolutionary paths among *A. fumigatus* environmental and clinical strains, and how do they relate to drug resistance?
4. How does fungicide treatment affect the structure of *A. fumigatus* genomes?
5. Do the genetic differences between environmental and clinical *A. fumigatus* change their metabolic capacities and their ability to colonize the human host?

Another paper (in preparation) of my thesis put the focus on the early transcriptomic responses of the human host to a pathogen aiming to identify important immunity mechanisms and gene biomarkers specific to bacterial and/or fungal infections in sepsis patients. The study aimed to answer the following questions using blood samples from patients up to 7 days before the clinical diagnosis of sepsis:

1. Which are the common and different immune mechanisms activated in response to bacterial and fungal infections?
2. Which are the key genes that regulate those immune mechanisms?
3. Are those key genes sufficient to detect the type of pathogen (bacteria, fungal, or co-infection) at an early stage of infection?

CHAPTER III RESEARCH PUBLICATIONS

List of Research Publications

This present thesis comprises four research publications: two first-author manuscripts, including one published (**manuscript I**) and one in preparation (**manuscript II**), and two co-author manuscripts, including one published (**manuscript III**) and one in preparation (**manuscript IV**).

Manuscript I:

Barber, A.E., **Sae-Ong, T**., Kang, K. Seelbinder B., Li J., Walther G., Panagiotou G., Kurzai O. *Aspergillus fumigatus* pan-genome analysis identifies genetic variants associated with human infection. *Nat Microbiol* 6, 1526–1536 (2021). https://doi.org/10.1038/s41564-021-00993-x

Manuscript II:

Sae-Ong, T., Schäuble S., Garcia Lopez, A., Lukaszewski R.A., Singer A., Panagiotou G., Transcriptomic analysis of presymptomatic sepsis patients reveals pathogen-specific host immune responses [Manuscript in preparation]

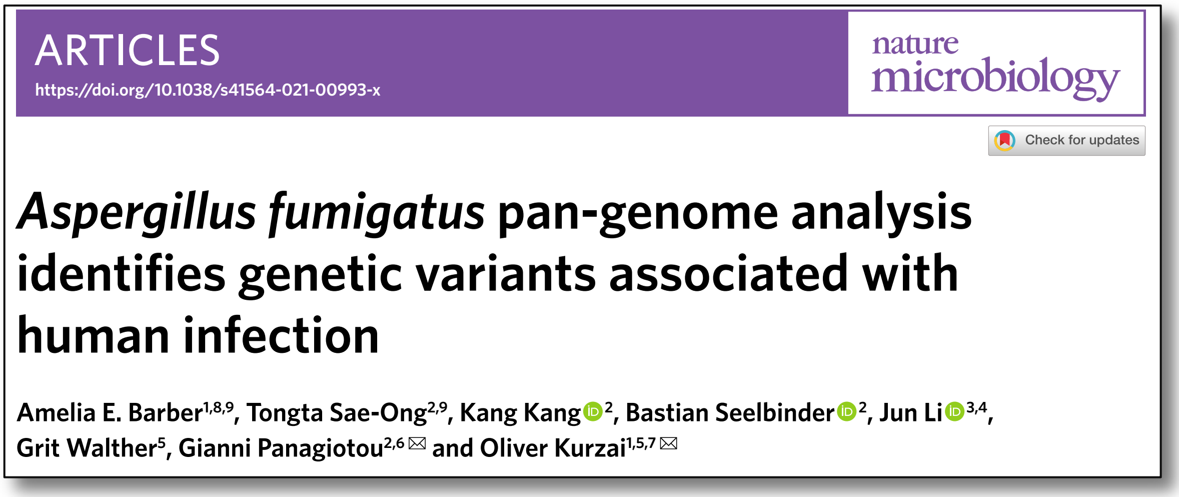
Manuscript III:

Barber A.E., Riedel J., **Sae-Ong T.**, Kang K., Brabetz W., Panagiotou G., Deising H.B., Kurzai O. 2020. Effects of agricultural fungicide use on *Aspergillus fumigatus* abundance, antifungal susceptibility, and population structure. mBio 11:e02213-20. https://doi.org/10.1128/mBio.02213-20

Manuscript IV:

Mirhakkak, M.H., Chen, X., **Sae-Ong, T.**, Xu, L., Heinekamp, T., Kurzai, O., Barber, A., Brakhage, A., Boutin, S., Schäuble, S., Panagiotou, G., A pan-genome resembling genome-scale metabolic model platform of 252 *Aspergillus fumigatus* strains reveals growth dependencies from the lung microbiome [Manuscript in preparation]

Manuscript I


ARTICLES
https://doi.org/10.1038/s41564-021-00993-x
nature microbiology

Check for updates

Aspergillus fumigatus pan-genome analysis identifies genetic variants associated with human infection

Amelia E. Barber[1,8,9], Tongta Sae-Ong[2,9], Kang Kang[2], Bastian Seelbinder[2], Jun Li[3,4], Grit Walther[5], Gianni Panagiotou[2,6] ✉ and Oliver Kurzai[1,5,7] ✉

Overview

In the first publication, we aimed to answer the questions related to genomic similarities and differences among global *A. fumigatus* strains and the evolutionary differences between environmental and clinical strains. Therefore, 300 *A. fumigatus* genomes were re-assembled using reference-based and *de novo* genome assembly methods. The results suggested the high diversity of *A. fumigatus* genomes, in which only 69% of discovered genes were shared among all strains. Furthermore, we can observe the genomic differences between environmental and clinical strains.

FORM 1

Manuscript No. 1

Manuscript title: *Aspergillus fumigatus* pan-genome analysis identifies genetic variants associated with human infection

Authors: Amelia E. Barber*, **Tongta Sae-Ong***, Kang Kang, Bastian Seelbinder, Jun Li, Grit Walther, Gianni Panagiotou and Oliver Kurzai

Bibliographic information:

Barber, A.E.*, **Sae-Ong, T**.*, Kang, K. Seelbinder B., Li J., Walther G., Panagiotou G., Kurzai O. *Aspergillus fumigatus* pan-genome analysis identifies genetic variants associated with human infection. *Nat Microbiol* 6, 1526–1536 (2021). https://doi.org/10.1038/s41564-021-00993-x

The candidate is (Please tick the appropriate box.)

☐ First author, ☒ Co-first author, ☐ Corresponding author, ☐ Co-author.

Status: published

Authors' contributions (in %) to the given categories of the publication

Author	Conceptual	Data analysis	Experimental	Writing the manuscript	Provision of material
Barber, A.E.*	30%	15%	80%	40%	
Sae-Ong, T.*	30%	80%		40%	
Seelbinder, B.		5%		5%	
Panagiotou, G.	20%			5%	50%
Kurzai, O.	20%			5%	50%
Others			20%	5%	
Total:	100%	100%	100%	100%	100%

*Authors contributed equally

Signature candidate

Signature supervisor (member of the Faculty)

ARTICLES
https://doi.org/10.1038/s41564-021-00993-x

nature microbiology

Check for updates

Aspergillus fumigatus pan-genome analysis identifies genetic variants associated with human infection

Amelia E. Barber[1,8,9], Tongta Sae-Ong[2,9], Kang Kang[2], Bastian Seelbinder[2], Jun Li[3,4], Grit Walther[5], Gianni Panagiotou[2,6] ✉ and Oliver Kurzai[1,5,7] ✉

***Aspergillus fumigatus* is an environmental saprobe and opportunistic human fungal pathogen. Despite an estimated annual occurrence of more than 300,000 cases of invasive disease worldwide, a comprehensive survey of the genomic diversity present in *A. fumigatus*—including the relationship between clinical and environmental isolates and how this genetic diversity contributes to virulence and antifungal drug resistance—has been lacking. In this study we define the pan-genome of *A. fumigatus* using a collection of 300 globally sampled genomes (83 clinical and 217 environmental isolates). We found that 7,563 of the 10,907 unique orthogroups (69%) are core and present in all isolates and the remaining 3,344 show presence/absence of variation, representing 16–22% of the genome of each isolate. Using this large genomic dataset of environmental and clinical samples, we found an enrichment for clinical isolates in a genetic cluster whose genomes also contain more accessory genes, including genes coding for transmembrane transporters and proteins with iron-binding activity, and genes involved in both carbohydrate and amino-acid metabolism. Finally, we leverage the power of genome-wide association studies to identify genomic variation associated with clinical isolates and triazole resistance as well as characterize genetic variation in known virulence factors. This characterization of the genomic diversity of *A. fumigatus* allows us to move away from a single reference genome that does not necessarily represent the species as a whole and better understand its pathogenic versatility, ultimately leading to better management of these infections.**

Diseases caused by the mould *Aspergillus fumigatus* are a major cause of human morbidity and mortality[1,2]. Invasive aspergillosis is particularly problematic in immunocompromised patients, resulting in a mortality rate of up to 50%[3,4]. Treatment of infections caused by *A. fumigatus* relies on triazole antifungal drugs. However, resistance to these frontline therapies is increasing, and the mortality rate for resistant infections is 25% higher than susceptible infections[5,6]. Although the most frequently identified resistance mutations occur in the cellular target of the triazoles—that is, *cyp51a*—up to 30% of the resistant isolates have no identifiable resistance mechanisms[7], complicating the recognition and treatment of these problematic infections.

While the host immune status is an important determinant in the development of aspergillosis, the substantial phenotypic variability observed among *A. fumigatus* isolates indicates that intraspecies diversity also plays a role in the disease[8–14]. This includes marked differences in virulence in animal models[9,10,14], fitness under hypoxia[9], growth under chemical stress(es)[11], nutritional heterogeneity[12] and induction of host inflammatory mediators[8]. As an indicator of the genomic diversity underlying the phenotypic variability observed in *A. fumigatus*, genomic comparisons between the reference strains Af293 and A1163 reveal tracts of variable gene content between the two[15], and 7% of Af293 genes are not present in A1163 (FungiDB). Despite this variation, previous studies of *A. fumigatus* have largely only analysed genomic information in the context of the reference genome and were limited to the genetic material present in Af293 due to the technical challenges of de novo eukaryotic genome analysis[16–19]. In addition, most of the isolates that have been sequenced to date are of clinical origin, thereby obscuring the genomic relationship between environmental isolates and those causing human disease.

In this study we constructed de novo genome assemblies of 300 *A. fumigatus* genomes ($n=217$ environmental isolates and $n=83$ clinical isolates) and used them to define the pan-genome of this important human fungal pathogen as well as the relationship between environmental and clinical isolates. We also leveraged the power of genome-wide association studies (GWAS) to identify genomic variation associated with human infection and triazole resistance, revealing a new range of therapeutic targets to combat these life-threatening infections.

Results

De novo assembly of 300 *A. fumigatus* genomes. In this study we used reference-guided and de novo assembly methods to analyse

---

[1]Research Group Fungal Septomics, Leibniz Institute of Natural Product Research and Infection Biology–Hans Knöll Institute, Jena, Germany. [2]Research Group Systems Biology and Bioinformatics, Leibniz Institute of Natural Product Research and Infection Biology–Hans Knöll Institute, Jena, Germany. [3]Department of Infectious Diseases and Public Health, Jockey Club College of Veterinary Medicine and Life Sciences, City University of Hong Kong, Hong Kong, China. [4]School of Data Science, City University of Hong Kong, Hong Kong, China. [5]National Reference Center for Invasive Fungal Infections (NRZMyk), Leibniz Institute of Natural Product Research and Infection Biology–Hans Knöll Institute, Jena, Germany. [6]Department of Medicine and State Key Laboratory of Pharmaceutical Biotechnology, University of Hong Kong, Hong Kong, China. [7]Institute for Hygiene and Microbiology, University of Würzburg, Würzburg, Germany. [8]Present address: Junior Research Group Fungal Informatics, Leibniz Institute of Natural Product Research and Infection Biology–Hans Knöll Institute, Jena, Germany. [9]These authors contributed equally: Amelia E. Barber, Tongta Sae-Ong. ✉e-mail: gianni.panagiotou@leibniz-hki.de; okurzai@hygiene.uni-wuerzburg.de

the genomes of 300 *A. fumigatus* isolates, representing environmental and clinical isolates from different locations across the globe. Among these, 188 samples were novel environmental and clinical isolates from Germany that were sequenced as part of this study. The remaining 112 isolates, including 64 isolates that were sequenced by us in a previous study[20], were pulled from public data repositories as raw sequence data. Our overall dataset was comprised of 217 environmental isolates and 83 clinical isolates from Europe, Asia, North America, South America and the International Space Station (Supplementary Data 1). Forty-three of 294 isolates were resistant to one or more medical triazoles, as determined by European Committee on Antimicrobial Susceptibility Testing (EUCAST) broth microdilution[21]. Azole susceptibility data were not available for six isolates. We generated de novo genome assemblies of these 300 isolates using paired-end Illumina sequencing to facilitate the unrestricted analysis of genomic diversity in *A. fumigatus*. The mean number of contigs in our assemblies was 948 and the mean N50, a marker of genome contiguity representing the weighted median contig length, was 145,494 base pairs (bp; Supplementary Table 1 and Supplementary Data 1). The mean genome size of our assemblies was 28.6 Mb (range, 26.9–30.8 Mb), with an average of 9,408 open reading frames (ORFs) per isolate and a range of 9,169 to 11,231. Using BUSCO as a measure of genome completeness, we found that an average of 97% of the expected single-copy orthologues were found and present as single copies in our genome assemblies.

To perform population genomic analyses, we aligned reads against the Af293 reference genome. We observed an average of 78,692 single nucleotide variants (SNVs) per isolate (range, 23,029–149,537) or approximately three SNVs per kilobase (Supplementary Data 1). We also detected an average of 7,383 short insertions or deletions (indels) per isolate (range, 2,528–16,134). Of the 329,405 non-redundant SNVs identified among our isolates, 33% (107,779) were not described in FungiDB, release 39. Together, this reveals a pronounced level of genetic diversity in *A. fumigatus* at the nucleotide level and considerably extends the previously recognised diversity.

The *A. fumigatus* pan-genome contains 7,563 core and 3,344 accessory genes. To examine the full genomic diversity of *A. fumigatus*, we used our de novo genome assemblies to define and characterize its pan-genome. The pan-genome is the collective gene set of a species and is composed of core genes found in all individuals and accessory genes that are not shared between all members of the species. We identified a total of 12,798 gene clusters that condensed into 10,907 non-redundant orthogroups. The *A. fumigatus* pan-genome was composed of a core genome of 7,563 orthogroups in all 300 isolates (69% of the pan-genome), 935 softcore orthogroups in >95% of the isolates (9% of the pan-genome), 1,367 shell genes in 5–95% of the isolates (13% of the pan-genome) and a cloud genome of 1,043 genes present in less than 5% of the isolates (10% of the pan-genome; Fig. 1a). Each isolate contained an average of 9,199 orthogroups (range, 8,987–9,629) and an average of 1,636 orthologous accessory-gene clusters (range, 1,424–2,066), corresponding to 16–22% of the total genome of the isolate. The pan-genome was closed—that is, the number of pan-genes did not substantially increase after the addition of approximately 250 genomes (Fig. 1b). Gene association analysis identified 53 co-occurring gene modules containing 2–251 genes (Fig. 1c).

The protein sequences of the core genes were significantly longer than the softcore or accessory genomes. The geometric mean of the length of the core genes was 436 amino acids compared with 310 amino acids for the softcore genes and 191 for the shell/cloud genes (Fig. 1d). To examine the evolutionary forces working on the core and accessory genomes, we calculated the rate of non-synonymous-to-synonymous substitutions (d_N/d_S). The geometric mean of the d_N/d_S ratio among all 10,907 pan-genes was 0.53, with significant differences between the genome compartments. The core genome showed the strongest evidence of negative or purifying selection ($d_N/d_S = 0.49$), whereas the softcore and accessory genomes had d_N/d_S ratios of 0.68 and 0.69, respectively (Fig. 1e). The lower d_N/d_S values for the core genes relative to the accessory genes indicate that they are under a higher degree of purifying selection—although neither genome compartment is evolving neutrally, as indicated by ratios of less than one.

The core genome contained a higher proportion of proteins with annotated domains, as 85% of the core genes contained at least one annotated InterPro domain compared with 71% of the softcore genes and 51% of the accessory genes (Fig. 1f). The core genome was enriched for 3,140 Pfam domains—including protein kinase domains, transcription factor domains and ABC transporters—whereas the accessory genome was enriched for 546 Pfam domains—including short-chain dehydrogenases and cytochrome P450 enzymes (Extended Data Fig. 1a). For Gene Ontology (GO) annotations, the core genome was enriched for protein binding, ATP binding, carbohydrate metabolic functions, signal transduction and 1,497 total annotations (Supplementary Data 2). The accessory genome was enriched for haem binding, response to oxidative stress and 244 total GO annotations (Supplementary Data 2).

Many of the shell and cloud genes were located on the subtelomeric ends of chromosomes 1 and 7, as measured by their position in Af293 (Extended Data Fig. 1b). Of the 10,907 orthologous gene clusters (homologous genes identified in different isolates) identified in the *A. fumigatus* pan-genome, 87% were present in Af293 (Supplementary Data 3). Overall, we identified an average of 494 genes per isolate that were absent in Af293 and a cumulative 1,934 unique ORFs were not present in Af293. In summary, the core genome of *A. fumigatus* represented 69% of the total identified orthogroups and was distinct from the accessory genome in length, function and the strength of purifying selection.

Chronic disease isolates are more genetically diverse than isolates from invasive disease and the environment. We examined the population genomics of isolates from the environment, invasive disease and chronic aspergillosis. Due to the lower number of isolates in the chronic disease group, the environmental and clinical samples were downsampled to match the number of chronic disease isolates ($n = 19$). Interestingly, the isolates from chronic disease group were significantly more diverse at the nucleotide level than isolates from invasive disease or the environment, as measured by the nucleotide diversity (π) calculated across overlapping 5 kb windows (Extended Data Fig. 2). In contrast, isolates from the invasive disease group showed less nucleotide diversity than isolates from the environment and chronic disease. The geometric mean of the genome-wide nucleotide diversity was 1.3×10^{-5} for the isolates from the chronic disease group, 8.3×10^{-6} for the environmental isolates and 6.9×10^{-6} for the isolates from the invasive disease group.

The Af293-containing genetic cluster is enriched for clinical isolates. In a phylogeny built from the coding nucleotide sequences of 5,380 single-copy orthologues, all 300 isolates formed a monophyletic group that was clearly distinct from the related outgroups of *Aspergillus oerlinghausenensis* and *Aspergillus fischeri* (Fig. 2 and Extended Data Fig. 3a). Isolates from Germany, collected and sequenced by us, intermixed with the globally sampled isolates from publicly available repositories, with no strong geographic clustering observed. We also found a high degree of congruence between the phylogeny built from the core genome sequence from de novo genome assemblies and phylogenies built using reference-guided SNV data from whole-genome SNVs and neutral loci (Extended Data Fig. 3b–d). Based on genome coverage at the MAT locus, we found an equal split of isolates of both mating types ($n = 148$ MAT1-1 isolates and $n = 149$ MAT1-2 isolates; Fig. 2).

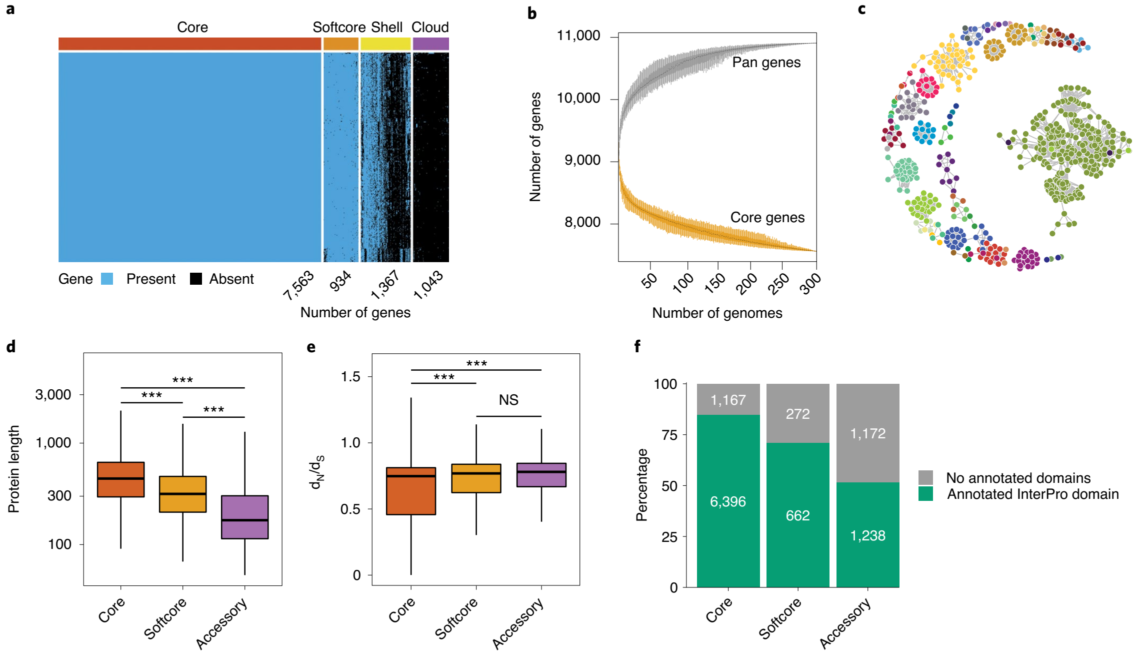


Fig. 1 | The pan-genome of *A. fumigatus*. **a**, Presence/absence matrix of 10,907 orthologous gene clusters identified from 300 *A. fumigatus* genomes. The pan-genome is subdivided into core (orthogroups present in all isolates), softcore (orthogroups present in >95% of the isolates), shell (orthogroups present in 5–95% of the isolates) and cloud (orthogroups present in less than 5% of the isolates) genomes. **b**, Pan and accessory (softcore, cloud and shell) genome size as the number of genomes included increases. Darker hues represent the 25th and 75 percentiles, while the lighter hues represent the range. **c**, Co-occurring gene modules ($n=53$) of the *A. fumigatus* accessory genome. Each circle indicates a gene and its association with other genes indicated by edges (lines). The module significance was identified using two-sided binomial exact tests with Bonferroni's correction ($P<0.05$). Only positive associations are illustrated. The colour indicates module membership. **d**, Amino-acid-sequence lengths of core, softcore and accessory (cloud and shell) genes. The exact P values were $<2\times10^{-16}$ for all of the indicated comparisons. **e**, Ratio of the non-synonymous substitutions to synonymous substitutions in core, softcore and accessory genes. Genes with ratios greater than one are under positive selection, whereas genes with ratios less than one are under purifying selection. The exact P values were: core versus softcore, $P=1.2\times10^{-7}$; core versus accessory, $P<2\times10^{-16}$; and softcore versus accessory, $P=0.15$. **c–e**, $n=7{,}563$ core, 935 softcore and 2,410 accessory orthogroups. **d**,**e**, In the box-and-whisker plots, the horizontal line in the box indicates the 50th percentile and the box extends from the 25th to the 75th percentile. The whiskers encompass the lowest and highest values within 1.5× the interquartile range. Statistical significance was determined using a two-sided Mann–Whitney *U*-test with Bonferroni's correction; $^{***}P<0.001$ and NS, not significant. **f**, Number (indicated in the bars) and fraction of core, softcore and accessory genes containing an annotated InterPro domain.

To look for evidence of genomic recombination in *A. fumigatus*, we performed a neighbour-net analysis, a phylogenetic method that allows for the representation of conflicting genetic signals that result from sexual recombination or gene conversion. The neighbour-net tree built from core genes had a highly reticulated centre, which indicates a marked degree of conflicting genetic information in the phylogenetic network and is suggestive of abundant genetic recombination in the species (Fig. 3a).

Discriminant analysis of principle components[22] was used to identify seven as the best supported number of genetic clusters in our dataset based on our de novo, reference and pan-gene count-based approaches (Extended Data Fig. 4a–c). Cluster 6 had the largest number of isolates ($n=80$), followed by cluster 2 ($n=53$), cluster 5 ($n=48$), cluster 7 ($n=43$), cluster 3 ($n=35$), cluster 4 ($n=22$) and cluster 1 ($n=19$; Fig. 4a). Interestingly, cluster 5 was enriched for clinical isolates (Fisher's exact test with Benjamini–Hochberg correction, $P=0.02$). This cluster also contained the reference strain Af293, which is a clinical isolate from a patient who died of invasive aspergillosis[23]. Together, we observed an enrichment for clinical isolates in one cluster as well as evidence of abundant genetic recombination in *A. fumigatus*.

Genetic cluster 5 contains more accessory genes and a distinct genomic profile. As genetic cluster 5 was statistically enriched for clinical isolates, we examined the genomes of each cluster to identify differences that might predispose the genetic background of cluster 5 towards human infection as well as characterize potential functional differences between the genetic clusters. Interestingly, clusters 5 and 2 contained significantly more accessory genes than the other clusters (Fig. 4a). The median number of accessory genes for cluster 5 was 1,965 compared with 1,895, 1,842, 1,882, 1,814 and 1,790 for clusters 2, 3, 1, 7 and 6, respectively (Fig. 4a). Cluster 4 had the smallest number of accessory genes, with a median of 1,749.

To predict the functional differences between the clusters, we calculated the abundance of Pfam domains and the frequency of GO annotations in the different clusters and compared the variance between clusters. A total of 170 GO annotations showed significant variation in their relative frequency between clusters (Fig. 4b and

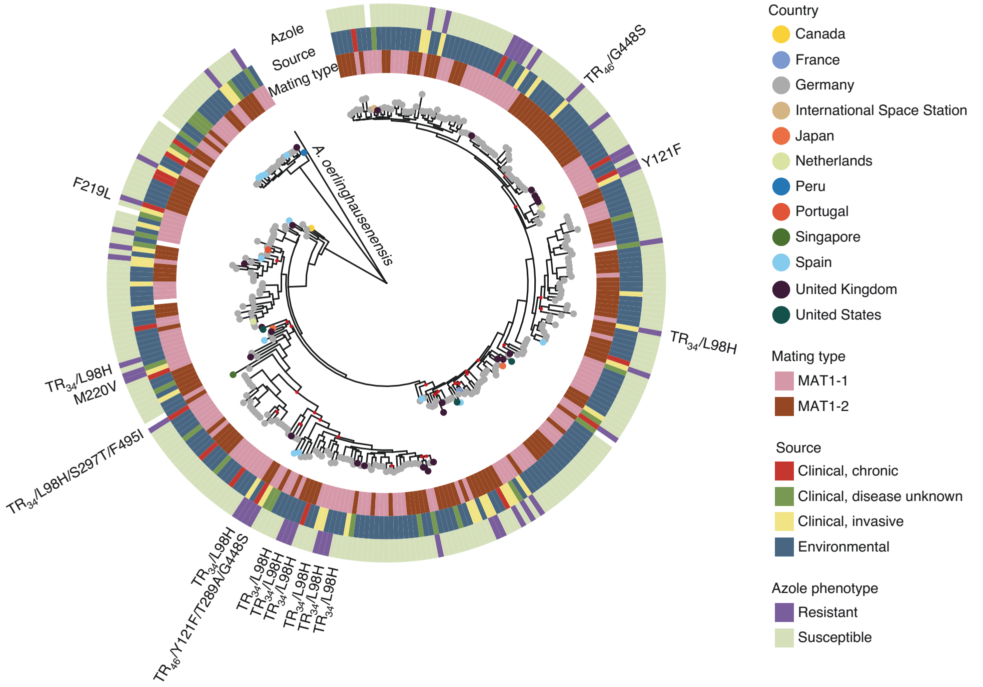


Fig. 2 | Whole-genome phylogeny of environmental and clinical *A. fumigatus*. Phylogenetic tree constructed from coding nucleotide sequences of 5,380 single-copy orthologues shared by *A. fumigatus*, *A. fischeri* and *A. oerlinghausenensis*. The phylogeny is rooted with *A. oerlinghausenensis* and the branch length was shortened for illustration. The coloured symbols at the end of branches represent the country where the sample was isolated. The red dots in the tree structure indicate nodes with ultrafast bootstrap values of less than 0.96. The metadata rings on the outside of the tree indicate the Azole phenotype (where resistance is defined as a minimum inhibitory concentration above the EUCAST breakpoint for one or more triazoles), source of the isolate and the mating type. Mutations in the *cyp51a* gene relative to the Af293 reference genome are also indicated on the outside of the tree.

Supplementary Data 4). Among these were an increased frequency of genes involved in oxidation–reduction processes, iron-ion binding, carbohydrate metabolic processes and proteolysis in cluster 5 (Fig. 4c and Supplementary Data 4). Significant variation between clusters was observed for the abundance of 269 Pfam domains (Supplementary Data 4). Cluster 5 had an increased abundance of major facilitator superfamily transporters, amino-acid permeases and chitin-recognition proteins (Fig. 4d and Supplementary Data 4).

For the GO categories and Pfam domains that did not show a significant difference in copy number between the genetic clusters, we reasoned that there could still be functional differences due to the presence of high-impact variants such as frameshifts or the gain/loss of stop codons. To examine this, we calculated the fraction of genes containing a high-impact variant(s) for each functional annotation and compared the incidence across the clusters. A total of 945 GO annotations contained significant differences in the incidence of high-impact variants between the clusters (Supplementary Data 4). Among these were a reduced number of high-impact variants in chromatin organization and mismatch repair-annotated genes in clusters 5 and 2 (Fig. 4e). We also quantified the incidence of high-impact variants in Pfam domain-containing genes and identified 482 domains with significant differences between the clusters (Supplementary Data 4). These included a reduced number of high-impact variants in cytochrome P450 enzymes and bZIP transcription factors in clusters 2 and 5. In summary, we observed distinct genomic profiles between the genetic clusters of *A. fumigatus*, including a larger number of accessory genes in clusters 2 and 5 in addition to copy-number variation and incidence of high-impact variants in functional annotations such as Pfam domains and GO categories.

A. fumigatus exhibits variation in virulence-associated genes. Using a database of 360 virulence- or fitness-associated genes for *A. fumigatus*[19,24–28] (Supplementary Data 5), we examined our 300 genomes for the presence/absence of these genes, changes in copy number relative to Af293 and the incidence of high-impact variants (for example, frameshifts and nonsense mutations). This list includes genes involved in metabolism, signalling, cell-wall biology, secondary metabolism, stress responses and antifungal drug resistance. Overall, these virulence-associated factors were well conserved. No variation in copy number, presence/absence of genes or genetic alterations anticipated to have a high functional impact was detected in 57% (205/360) of the genes. The remaining 155 virulence-associated genes had some degree of genetic variation expected to affect gene function among our 300 genomes, which can be visualized in Fig. 5 (full summary in Supplementary Data 5). Underscoring the fundamental role of these genes for the fitness of *A. fumigatus* in the environment and the human host, most cases of gene loss or high-impact genetic variation were uncommon and observed in less than 5% of the isolates ($n = 121$ genes). However, the remaining 34 genes displayed more pervasive genetic variation, including 76% of the isolates (229/300) showing frameshifts in the

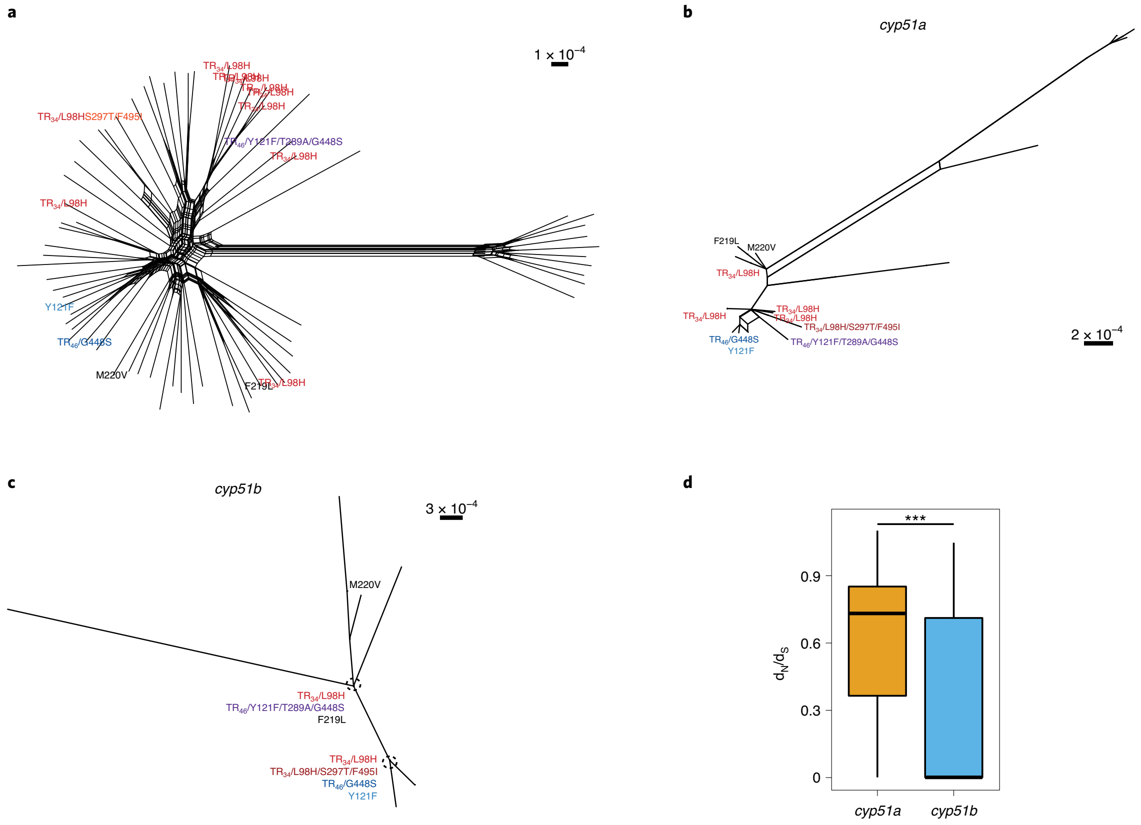


Fig. 3 | Neighbour-net trees from 70 *A. fumigatus* isolates. a, Phylogenetic network of the coding nucleotide sequence of 5,380 single-copy orthologous genes (whole-genome sequencing). The reticulated core indicates abundant conflicting genetic information among isolates, suggestive of recombination in the species. **b**,**c**, Phylogenetic networks of the coding nucleotide sequences of *cyp51a* (**b**) and *cyp51b* (**c**) plus 1,000 bp up- and downstream of the genes. Splits (parallel bands) indicate conflicting nucleotide patterns and their length is proportional to the number of bases supporting the split. The *cyp51a* genotypes of the isolates are indicated at the branch tips. Red-hued labels indicate TR_{34}-containing alleles. Blue-hued labels denote TR_{46} lineage-associated alleles. All additional *cyp51a* polymorphisms are indicated with black labels. Dashed circles in **c** denote the point on the phylogenetic network where all the indicated alleles localize. **a**–**c**, $n = 10$ isolates from each of the genetic clusters identified in the dataset. Scale bar indicates nucleotide substitutions per site. **d**, Ratio of non-synonymous-to-synonymous substitutions in *cyp51a* and *cyp51b*. The ratios were calculated from the genomes of 300 *A. fumigatus* isolates. Both genes have ratios of less than one and are under purifying selection. Statistical significance was determined using a two-sided Mann–Whitney *U*-test; $^{***}P < 0.001$; the exact *P* value was $P < 2.2 \times 10^{-16}$. The bold horizontal line indicates the 50th percentile and the filled box extends from the 25th to the 75th percentile. The whiskers encompass the lowest and highest values within 1.5× the interquartile range.

serine protease *pr1* (*Afu7g04930*) and 71% of the isolates (213/300) showing high-impact variants in the putative sensor histidine kinase *tcsB* (*Afu2g00660*).

Overall, secondary metabolism genes showed the highest variability among the virulence-associated genes, with 59 genes either being absent or showing a predicted loss of function among the 300 isolates. Interestingly, 97% of the isolates (292/300) possessed extensively degraded copies of the non-ribosomal peptide synthetase *nrps8* (also known as *pes3* or *Afu5g12730*), a gene whose deletion showed increased virulence in a murine model of invasive aspergillosis[29]. We also observed variability in the biosynthetic gene cluster encoding fumagillin, including absence of the fumagillin tailoring enzyme *fmaG* in 89% of the isolates (267/300), the absence of *fumR* in 49% of the isolates (146/300) and a complete loss of the cluster in three isolates (1%). Finally, we observed variants predicted to impact the biosynthesis of the immunosuppressive virulence factor gliotoxin in 6% of the isolates (17/300). These included high-impact variants in *gliZ* ($n = 11$ isolates); *gliA* ($n = 5$ isolates); *gliP* and *gliF* ($n = 3$ isolates each); *gliI*, *gliT* and *gliJ* ($n = 2$ isolates each); and *glicC* and *gliG* ($n = 1$ isolate each).

In addition to cases of gene loss, we observed cases of gene amplification in virulence-associated genes relative to Af293 and A1163. A total of 53 genes showed gene amplification, including 5% of the isolates ($n = 16$) with increased copy number of the putative catalase-peroxidase *cat2* (*Afu8g01670*), which is upregulated in response to neutrophils[30]. In addition, 5% of the isolates ($n = 14$) had increased copy numbers of the zinc transporter *zrfC* (*Afu4g09560*) and 3% ($n = 10$) had increased copy numbers of the putative ABC multidrug transporter *Afu5g12720*. In summary, although roughly half of the virulence-associated genes described to date were well conserved among the 300 genomes examined, we observed high-impact genetic variation in many virulence-associated genes, which could perhaps explain the wide range in virulence observed among *A. fumigatus* isolates.

GWAS-identified fungal genetic variation associated with clinical isolates. To better understand how the environmental saprobe *A. fumigatus* can cause disease in the non-native niche of the human lung, we performed a GWAS study to identify fungal variants associated with clinical isolates relative to environmental isolates as well as fungal variants associated with the specific disease states of invasive and chronic disease (Extended Data Fig. 5a). Using a linear mixed model and a minor allele frequency (MAF) > 0.05, we identified 68 genomic positions with genetic variants associated with clinical isolates relative to environmental isolates (Supplementary Data 6). These variants included hits in 27 protein-coding genes, comprising both genes with established roles in virulence as well as uncharacterized ORFs (Supplementary Table 2). Among the genes previously implicated in the virulence of *A. fumigatus* were the sterol regulatory element binding protein *srbA*, which is involved in both growth in hypoxia and iron homeostasis[31,32], the global transcriptional regulator *pacC* required for fungal invasion during pulmonary infection[33,34] and the transcription factor *acuK* that regulates gluconeogenesis and iron acquisition[35]. The analysis also identified variants in genes whose role in virulence is less established, including a microtubule spindle protein (*Afu2g16260*), a heat shock-responsive protein (*Afu4g04680*), a putative polyketide synthase (*Afu6g13930*) and histone H1 (*Afu3g06070*), which is upregulated in conidia exposed to neutrophils (AspDB).

The virulence potential of *A. fumigatus* is influenced by the host and its underlying disease status. The factors critical for the establishment of invasive infection in a neutropenic lung are probably not the same as those required for long-term survival in the human lung, as in the case of chronic diseases such as cystic fibrosis and allergic bronchopulmonary aspergillosis. We thus performed association analysis for genetic variants associated with isolates from both invasive (acute) and chronic aspergillosis. There was a high degree of overlap between the genetic variants identified in this analysis and those from the analysis of all clinical isolates, regardless of the disease status of the host, but fewer variants and genes were identified for each underlying clinical disease (Extended Data Fig. 5b). We identified 21 genomic positions with SNVs and short indels significantly associated with invasive aspergillosis (Supplementary Table 2 and Supplementary Data 6). Nine of the ten variants located in coding genes that were associated with invasive disease were shared with isolates from all clinical origins and included the transcription factors *acuK* and *pacC* as well as the tubulin beta-2 subunit *tub2*. Chronic disease had variants at five genomic positions, two of which were within coding genes: *Afu2g03540*, an orthologue of GPI-anchored cell protein *cspA* (*Afu3g08990*) and a L-cytosine transmembrane transporter (*Afu6g14530*; Supplementary Table 2 and Supplementary Data 6).

Triazole target genes display distinct phylogenetic networks and imbalanced levels of stabilizing selection. The paralogous genes *cyp51a* (*Afu4g06890*) and *cyp51b* (*Afu7g03740*) encode the molecular targets of the triazoles. Despite this, most resistance mutations and mechanisms have been described in *cyp51a*. Triazole-resistant isolates were distributed throughout the phylogeny (Fig. 2). However, most isolates carrying the TR$_{34}$/L98H allele of *cyp51a* were clustered near each other. The close genetic relationship between isolates carrying TR$_{34}$/L98H is in agreement with previous work suggesting a single origin of this allele[36].

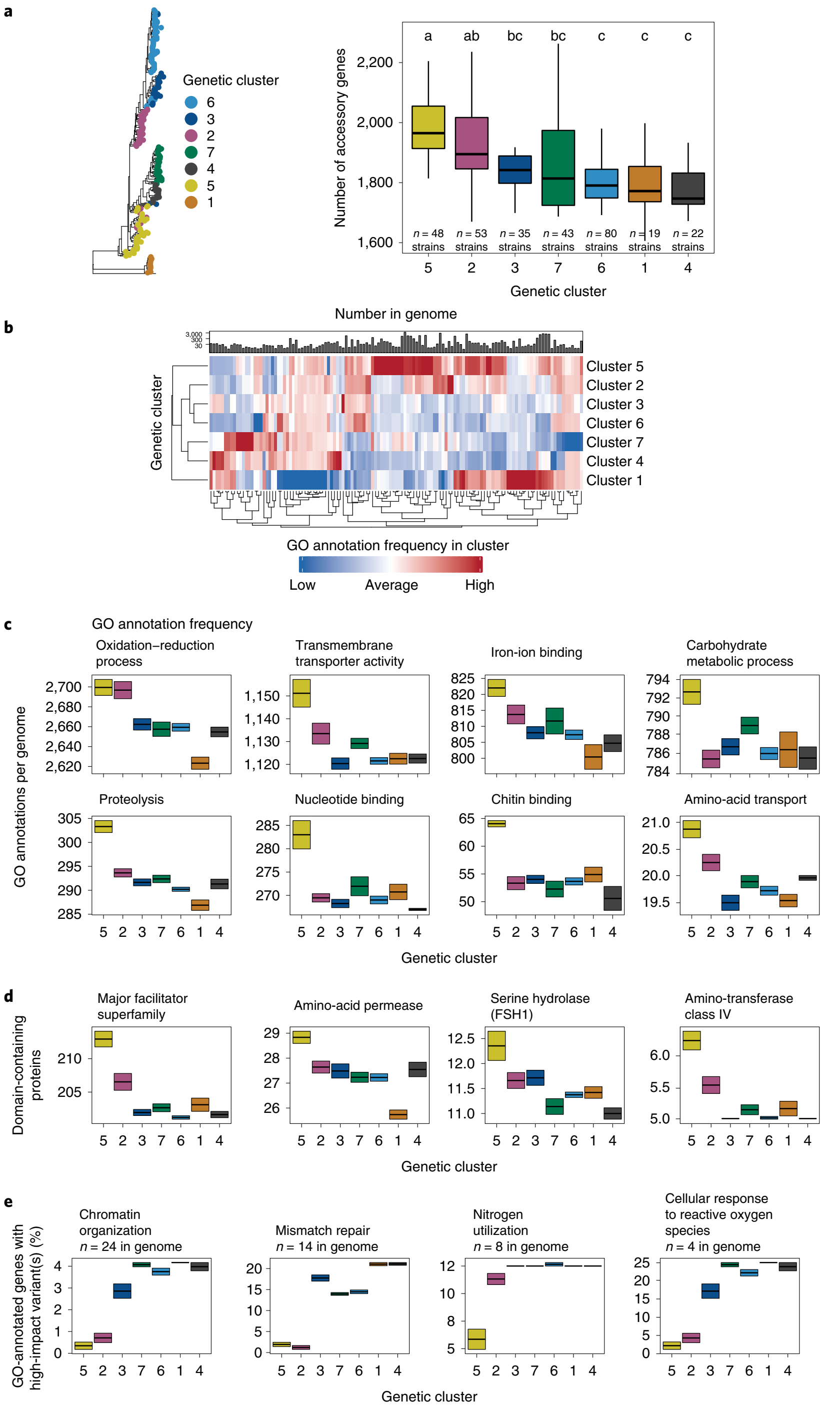
To investigate the evolutionary features of the triazole targets, we built neighbour-net trees from the coding sequence of *cyp51a* and *cyp51b* plus 1,000 bp of the up- and downstream flanking sequences. A phylogenetic network built from *cyp51a* sequences showed multiple splits (parallel bands), indicating conflicting genetic information among our isolates that could arise from recombination (Fig. 3b). Genetic recombination by isolates carrying the TR$_{34}$/L98H allele is supported by its presence in isolates of both mating types (Fig. 2). By comparison, a neighbour-net tree of *cyp51b*, which is located on a different chromosome, did not show any conflicting genetic information, as demonstrated by the lack of reticulation in the phylogenetic network (Fig. 3c). We also observed a reassortment of *cyp51a* genotypes in the tree constructed from *cyp51b* sequences relative to that constructed from *cyp51a* (Fig. 3b,c). In the tree constructed from *cyp51a* sequences, isolates carrying the TR$_{34}$/L98H allele of *cyp51a* were located at five distinct points on the phylogenetic network at positions that did not overlap with the positions of other *cyp51a* mutant alleles. In the network of *cyp51b* sequences, strains carrying the TR$_{34}$/L98H allele were found only at two positions in the network that also contained other *cyp51a* mutant alleles.

To assess the selective forces working on *cyp51a* and *cyp51b*, we examined the d_N/d_S ratios of each gene. The d_N/d_S ratios of *cyp51b* were significantly lower than *cyp51a* (mean value of 0.01 and 0.27 for *cyp51b* and *cyp51*, respectively), indicating that *cyp51b* is under a stronger degree of stabilization selection than *cyp51a* (Fig. 3d). Together, our results demonstrate higher levels of genetic disagreement in the isolate sequences of *cyp51a* compared with *cyp51b* and that *cyp51a* is under less stabilizing selection than *cyp51b*.

GWAS-identified genetic changes associated with triazole resistance. We subsequently performed variant-based GWAS to identify genomic changes associated with triazole resistance. Among the 294 samples with available susceptibility data, 44 were resistant to one or more triazole. Of these, 15 contained mutations in *cyp51a* that have been previously shown to confer triazole resistance (for example, TR$_{34}$/L98H and TR$_{46}$/Y121F/T289A/G448S) and 29 were resistant by unknown mechanisms. When we performed a linear mixed-model GWAS using a MAF > 0.01, we identified 16 genomic positions associated with triazole resistance (Supplementary Table 3 and Supplementary Data 6). These included the known TR$_{34}$ and L98H variants in the triazole target enzyme *cyp51a*. However, we repeated our analysis using a MAF > 0.05 to give a more robust variant list with fewer false positives given that association studies with smaller datasets such as ours are underpowered for the detection of true associations with rare variants (Extended Data Fig. 5c). Using this more stringent criterion, we condensed our variant list

Fig. 4 | Pan-genomic differences between the clusters of *A. fumigatus*. **a**, Number of accessory genes (right) present in the genomes of isolates belonging to each genetic cluster (left). Statistical significance was determined using a one-way analysis of variance and Tukey's honest significance test (one-sided). The letters denote significances as a compact letter display where groups that are not significantly different from each other are indicated with the same alphabet letter; $P < 0.05$. The bold line in the box-and-whisker plot indicates the 50th percentile, and the box extends from the 25th to the 75th percentiles. The whiskers denote the lowest and highest values within 1.5× the interquartile range. **b**, Heatmap showing the normalized abundance of GO annotations exhibiting significant variance in frequency between the clusters (bottom; $n = 127$ GO annotations). Statistical significance was determined using one-way analysis of variance with Bonferroni's correction ($P < 0.05$). The mean number of genes containing each GO annotation across the 300 genomes is shown (top). Note the graph is on a $\log_{10}$ scale but the *y*-axis shows actual values. **c**, Genome copy number for select GO annotations from **b** across the clusters. **d**, Genome copy number for select Pfam annotations across the clusters. **e**, The incidence of high-impact variants (for example, frameshift and loss of start) relative to Af293 was analysed for GO annotations that did not contain significant copy-number variation between the clusters. A selected subset of GO categories with significant variation in the incidence of high-impact variants between the genetic clusters is shown. **c**–**e**, The boxes denote the mean (crossbar) ± s.e.m. for the isolates of each cluster.

to variants in three protein-coding genes (Extended Data Fig. 5d and Supplementary Table 3). These included a microtubule bundle protein (*Afu2g16260*), a FGGY-family kinase induced by heat shock (*Afu4g04680*) and its adjacent, uncharacterized ORF *Afu4g04690*. The role of these genes in triazole resistance is an exciting area to follow up on.



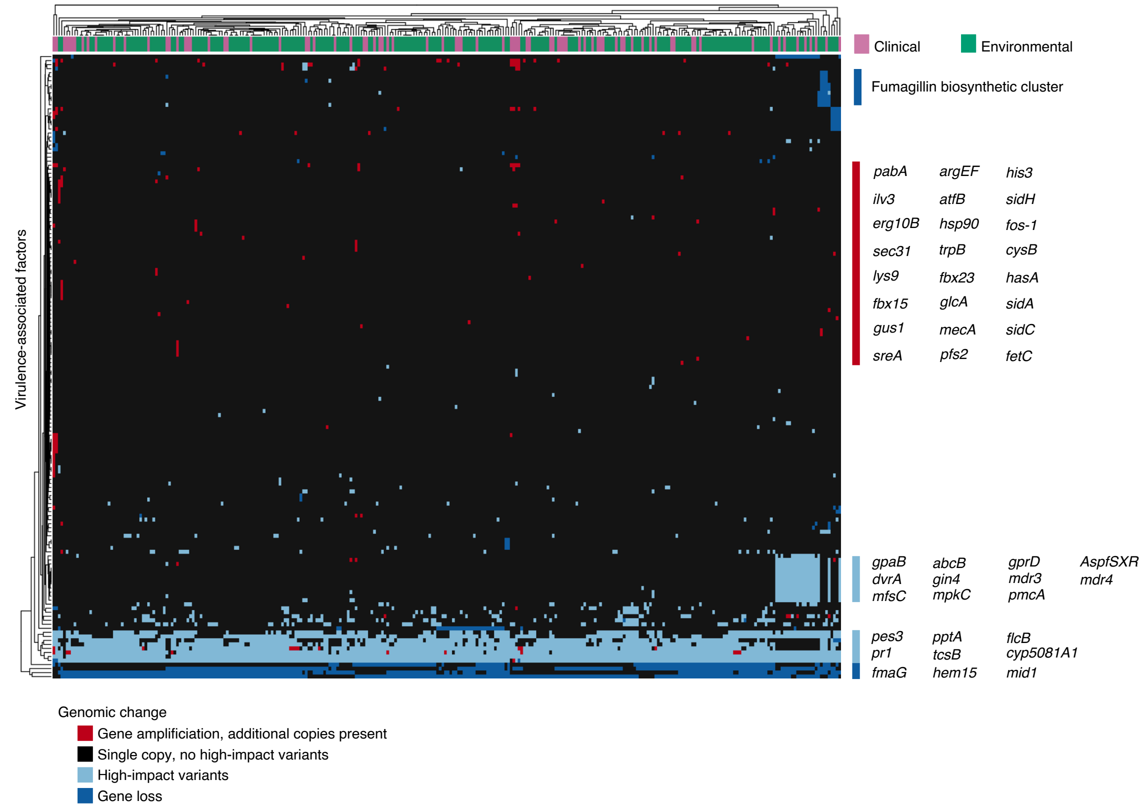


Fig. 5 | Genomic variation among known *A. fumigatus* virulence-associated factors. Heatmap of 155 virulence-associated genes where variation in copy number (gene loss or amplification) or the presence of high-impact variants (for example frameshift, loss or gain of stop codon) was observed relative to Af293 (bottom). The source of the isolate is indicated (top). Gene names for select virulence-associated factors are annotated (right).

Discussion

In this study we defined the pan-genome of *A. fumigatus* using 300 genomes, including a large number of environmental isolates largely absent from previous analyses[15–19]. Compared with *A. fumigatus*, the human commensal and opportunistic pathogen *Candida albicans* was shown to have a lower level of pan-genomic diversity, with 91% of pan-genes present in all isolates[37]. In the same study, a proof-of-concept pan-genome for *A. fumigatus* was also built using genomic data from 12 isolates and 83% of the pan-genome was found to be conserved in all isolates[37]. By contrast, our findings indicate that *A. fumigatus* has a much larger pan-genome and only 69% of the genes identified are present in all isolates; this discrepancy in results is likely to be due to the limited number of isolates included in the former study. In addition, the average BUSCO genome completeness of the assemblies used for their analysis was below 85%, suggesting that notable genetic content was unaccounted for[37]. Future work utilizing chromosome-level assemblies of *A. fumigatus* isolates will allow for a finalized pan-genome of the species with additional information on the evolutionary dynamics of chromosomal organization.

Through pan-genomic analyses we discovered notable genetic variation in virulence factors that have largely only been studied in one or two reference strains. Although most of these cases were infrequent and observed in fewer than 5% of the isolates, some, such as pseudogenization of the non-ribosomal peptide synthetase *nrps8* (or *pes3*), was observed in 97% of the isolates. The largest virulence-associated genetic variation was in secondary metabolism genes, an observation in line with a previous study of 66 isolates[38]. Both studies observed low-incidence variation in the gliotoxin and fumagillin/pseurotin biosynthetic gene clusters as well as high-incidence variation in the fumigermin biosynthetic gene cluster. In addition, although our analysis quantified high-impact genetic changes in virulence determinants, there is almost certainly additional genetic variation that impacts fungal virulence that is difficult to predict on the global scale. The genomes generated here provide a valuable resource for addressing how intraspecies variability in virulence determinants affects infection.

We observed an enrichment for clinical isolates in genetic cluster 5, suggesting that this genetic background might be more fit in the human environment. However, clinical isolates were distributed throughout the phylogeny, highlighting the overall fitness of *A. fumigatus*. In addition, this organism can take advantage of numerous, diverging clinical diseases to establish an infection. We thus performed a genome-wide association study (GWAS) to identify fungal variants associated with clinical disease in general as well as acute and chronic disease, and identified largely overlapping gene sets. However, information on the underlying clinical disease was not available for all samples and the isolates from chronic disease represented a small fraction of the dataset. Future genomic analyses including additional samples from specific underlying diseases will further illuminate the complex interplay between *A. fumigatus* and specific host disease environments.

The rising incidence of resistance to first-choice antifungals, the triazoles, is a major challenge for the management of *A. fumigatus* infections. This problem is further complicated by up to 30% of the isolates having no identifiable resistance mechanism. We performed GWAS and identified 12 genes associated with triazole resistance. These hits included previously identified variants in the triazole target gene *cyp51a* as well as genes that had not been previously linked to triazole resistance. As a caveat, association studies are underpowered at detecting associations with rare variants. Accordingly, we only screened for association of genetic variants present in at least 1% and 5% of samples. Thus, there are potentially additional resistance-associated variants that were not identified by our analysis. This is perhaps the case for the HMG-CoA demethylase *hmg1*. Clinically observed mutations in this gene conferred triazole resistance to *A. fumigatus* following reconstruction in an isogenic background[39]. Although our analysis did not identify any variants of this gene associated with triazole resistance, manual examination uncovered three triazole-resistant isolates containing single non-synonymous substitutions in *hmg1* (E306D, P309L and C369R). These variants were not considered in the GWAS due to their low prevalence in the dataset. However, the exact role these substitutions play in the resistance of this isolate is unclear, particularly for one isolate that also contained *cyp51a* alterations associated with resistance (TR_{46}/G448S). No variants were observed in *hapE* and *cyp51b*, two additional genes linked to triazole resistance.

In summary, this study provides a comprehensive view of the genetic diversity in this important human fungal pathogen. Characterization of the intraspecies diversity and moving away from a single reference genome that does not necessarily represent *A. fumigatus* as a whole will ultimately help us understand its metabolic and pathogenic versatility.

Methods

***A. fumigatus* isolates analysed in this study.** Of the 300 isolates analysed, 188 (49 clinical and 139 environmental isolates) were newly sequenced as part of this study. The 49 clinical isolates sequenced were collected by the German National Reference Center for Invasive Fungal Infections between 2014 and 2018. Bronchial alveolar lavage was the most frequent form of sample collection, representing 31% (15/49) of clinical isolates. The remaining clinical samples were isolated from other pulmonary sources, such as sputum or bronchial secretions, and the exact site of isolation was unavailable for 10% (5/49) of the samples. The 139 environmental isolates sequenced as part of this study were obtained from soil sampling of 11 farms in Germany between 2016 and 2018. Sixty-four of the remaining 112 isolates had been previously sequenced by us as part of a previous study[20] (BioProject PRJNA595552), while 48 had been previously sequenced by other groups and data downloaded from the NCBI Sequence Read Archive. In total, the dataset was comprised of 213 environmental isolates and 87 clinical isolates from Europe, Asia, North America, South America and the International Space Station. A detailed list of the isolates and their metadata characteristics can be found in Extended Data Fig. 1.

Antifungal susceptibility testing. The 188 novel isolates included in this study were screened for azole resistance using the agar-based VIPcheck Assay (Mediaproducts BV) based on EUCAST E.DEF 10.1 following the manufacturer's protocol. Isolates that showed distinguishable germination and hyphal growth on any of the azole-containing wells were subjected to EUCAST broth microdilution (protocol E.DEF 9.3.2; ref. [21]) to define the minimum inhibitory concentrations. Antifungal susceptibility in the clinical isolates was also assessed following EUCAST protocol E.DEF 9.3.2 and resistance was defined in both isolate sets using the EUCAST-established clinical breakpoints. Antifungal susceptibility data from published isolates were obtained from the source publications detailed in Supplementary Data 1.

Genome sequencing and quality assessment. Genomic DNA was extracted from isolates (cultured in Sabouraud Glucose broth at 37 °C with shaking) using a Quick-DNA fungal/bacterial miniprep kit (Zymo Research) following the manufacturer's suggested protocol. Library preparation and Illumina 2 × 150 bp paired-end sequencing were performed on a NextSeq 500 v.2 by LGC Genomics (environmental isolates) and GeneWiz (clinical isolates). Raw FASTQ files were filtered for quality using the following steps: adaptor sequences were removed, bases with an overall quality score of <20 were trimmed and reads shorter than 30 bp were removed. The remaining sequences were verified for quality using FastQC v.0.11.5 (Babraham Institute).

Reference-guided genome analysis. High-quality sequencing reads were aligned to the *A. fumigatus* Af293 reference genome v.2015-09-27 using BWA-MEM v.0.7.8-r779-dirty[40]. PCR duplicates were marked using MarkDuplicate from Picard v.2.18.25. Variant calling to detect SNVs and short indels was performed using the GATK Toolkit (v.4.1.0.0)[41]. Briefly, before variant calling, BAM files were recalibrated using GATK BaseRecalibrator, ApplyBSQ and an in-house dataset of known SNVs generated from the Af293 reference genome and SNVs present in FungiDB, release 39, with ≥80% read frequency and base call ≥ 20. Variant detection was performed using HaplotypeCaller and high-quality variants were identified using GATK best practices (SNP: QD < 2.0 || MQ < 40.0 || FS > 60.0 || MQ RankSum < −12.5 || ReadPosRankSum < −8.0; indel: QD < 2.0 || FS > 200.0 || ReadPosRankSum < −20.0). For downstream analyses, individual VCF files were combined into a single file using bcftools v.01.1.1. Variant function was predicted using SnpEff v.4.3t[42] and 1,000 bp as the cutoff for upstream and downstream flanking of the ORFs. To balance the analysis of high-quality variants with the potential bias introduced by true variants being discarded due to insufficient support, individual variants that failed the quality filter in a sample were included in the variant dataset if at least 95% of the total samples with a variant at that position passed the quality control. This hybrid-filtered variant dataset was used as the input for GWAS and genetic diversity analyses.

A. fumigatus has two mating types (MAT1-1 and MAT1-2), which are encoded within idiomorphic loci on chromosome 3 (refs. [43,44]). The mating type was assigned by calculating the genomic coverage at *Afu3g06160* and *Afu3g06170* using the knowledge that MAT1-2 isolates, including the reference strain Af293, contain a truncated copy of the HMG box mating-type transcription factor (*Afu3g06170*) and an additional gene (*MAT1-2-4*; also known as *Afu3g06160*) that are absent in MAT1-1 isolates. Isolates showing zero coverage in the genomic region of *Afu3g06160* following alignment to Af293 were assigned the mating type MAT1-1. Samples that were not assigned the mating type MAT1-1 were confirmed to be the mating type MAT1-2 through calculation of the genomic coverage at *MAT1-2-1* and *MAT1-2-4*. The ratio of coverage for *MAT1-2-1* and *MAT1-2-4* relative to the genome-wide depth of coverage was between 0.75 and 1.25 for all samples that were assigned the mating type MAT1-2.

Analysis of genomic diversity. Genomic diversity statistics were calculated based on SNV data generated as described earlier. The nucleotide diversity (π) was also calculated using VCFtools[45] with a window size of 5,000 bp and a 500 bp step size. To ensure that differences in sample sizes between the isolate populations did not skew the results, environmental and clinical samples from the acute disease group were downsampled to match the number of isolates from chronic disease in the dataset.

De novo genome assembly and annotation. Genomes were assembled de novo using IDBA-hybrid v.1.1.3 with the Af293 reference genome as a guide[46]. The quality of the genome assembly was assessed using QUAST v.5.0.2 (ref. [47]). Contigs that were shorter than 500 bp or possessing >95% identity and coverage overlap with other contigs were removed. Gene prediction and functional annotation were performed using Funannotate pipeline v.1.5.2-4cfc7f8 (ref. [48]), integrating the following steps. Assemblies were masked for repetitive elements using RepeatMasker (v.4.0.8)[49] using Dfam and RepBase repeat libraries[50]. Gene prediction was performed using EvidenceModeler v.1.1.1 (ref. [51]), incorporating evidence data generated using GeneMark-ES[52] (minimum gene length, 120 bp; and maximum intron length, 3,000 bp) and Augustus[53] (training set, *A. fumigatus*). Gene models predicted to encode peptides shorter than 50 amino acids or transposable elements, or to include span gaps were removed. Transfer-RNA prediction was performed using tRNAscan-SE v.2.0 (ref. [54]). Functional annotation was predicted using PFAM v.43 (ref. [55]), MEROPS v.12 (ref. [56]), dbCAN2 release 7.0 (ref. [57]) and BUSCO v.4.1.4 (ref. [58]). KofamScan v.1.2.0-0 (ref. [59]) was used to assign Kyoto Encyclopedia of Genes and Genomes orthologues to predicted protein sequences and InterProScan v.5.19 (ref. [60]) was used to identify the protein families.

Pan-genome analysis. OrthoFinder was used to identify and cluster orthologous genes[61]. Clustering was performed on the protein sequences of the 300 *A. fumigatus* genomes analysed in this study. In addition, protein sequences from the reference strains Af293 and A1163 were added to improve the identification of the cluster functions. Orthologous gene clusters were assigned a gene identifier from Af293 if they grouped with a single sequence of Af293. If a cluster was not assigned a Af293 gene identifier, but a single A1163 sequence was present, the cluster was assigned the gene identifier from A1163. Orthologous clusters that could not be grouped with a single Af293 or A1163 gene were queried against the NCBI RefSeq non-redundant protein database using DIAMOND using the following criteria: *E*-value cutoff of 1×10^5, percent identity > 70%, minimum query coverage > 50% and minimum subject coverage > 50%. If at least 70% of the protein sequences in the cluster were assigned to any protein in the NCBI non-redundant protein database, the cluster name was assigned to the name of the RefSeq with the highest contribution. If only 50–70% of the protein sequences in a cluster were assigned to the same protein, the matching sequences were assigned the name of the RefSeq match and the remaining sequences were left unassigned. The remaining clusters without a match in Af293, A1163 or the non-redundant database were considered novel clusters and had putative functions assigned based on their Funannotate (KofamScan and InterProScan) prediction. For these clusters that were not present in Af293 or the non-redundant database, only clusters present in at least 5% of samples were included to limit false gene predictions. The pan-genome was defined based on gene presence/absence variation in the approved cluster meeting the above criteria. Enrichment analysis was performed using a Fisher's exact test with Bonferroni's correction.

Whole-genome phylogeny. The core genome phylogeny (Fig. 2) was inferred from 5,380 single-copy orthologous genes shared by the two reference strains Af293 and A1163, the 300 *A. fumigatus* genomes analysed in this study, the related species *A. oerlinghausenensis* and *A. fischeri*, which was used to root the tree. Orthologues were identified and clustered using OrthoFinder[61]. Cluster peptide sequences were aligned using MUSCLE v.3.8.1551 (ref. [62]). The resulting peptide alignment was back-translated to a nucleotide sequence using PAL2NAL[63] and concatenated. The phylogeny was inferred from this core nucleotide alignment using IQ-TREE 2 (ref. [64]). The ModelFinder Plus module of IQ-TREE 2 was used to identify GTR + F + R8 as the best fitting substitution and site heterogeneity models for phylogeny construction. Branch support was computed using UFBoot2 ultrafast bootstraps[65]. ClonalFrameML[66] was then used to account for recombination in the phylogeny and rescale branch lengths accordingly.

The SNV-based phylogenies (Extended Data Fig. 4c,d) were constructed by first filtering out loci that showed zero coverage in any sample. For the phylogeny constructed from neutral loci, fourfold degenerate sites were used. For both the non-zero coverage and neutral loci phylogenies, SNVs were concatenated and used as the input for IQ-TREE 2. As with the core nucleotide phylogeny, ModelFinder was employed and identified GTR + F + ASC + R8 as the best fitting model for the non-zero coverage phylogeny and TVM + F + ASC + R8 as the best fitting model for the neutral loci phylogeny. Branch supports were calculated using UFBoot2.

Genetic clusters were identified using discriminant analysis of principle components[22]. To create phylogenetic network trees with clearly visible branches and network structure, the genomes were downsampled by randomly selecting ten genomes per cluster, resulting in a total of 70 samples. Neighbour-net phylogenies were inferred and visualized using the R package phangorn (v.2.5.5)[67] based on a similarity matrix of core nucleotide sequences for the whole-genome network phylogeny and nucleotide sequence alignment for the *cyp51a* and *cyp51b* genes with network phylogenies. The phylogenies were visualized using the R package Ggtree[68].

Estimation of d_N/d_S. The protein-coding sequences of each gene cluster were aligned using MUSCLE v.3.8.1551 (ref. [62]). PAL2NAL[63] was then used to convert the resulting amino-acid alignment to a nucleotide alignment that records whether a base-pair substitution resulted in a synonymous or non-synonymous change. Finally, the CODEML package of PAML[69] was used to calculate the d_N/d_S value of each orthogroup. Median values were used for comparison.

Gene co-occurrence in the pan-genome. Gene co-occurrence networks were computed using Coinfinder[70] using a presence/absence matrix of the pan-genome and a significance cutoff of 0.05 by binomial exact test with Bonferroni's correction. Networks were visualized using the R package igraph.

SNV-based GWAS and pan-GWAS. Before analysis, variant classes were assigned as follows: C, SNVs; G, insertions; D, deletions; and A, reference base. VCF files were converted to plink format using VCFtools[45] and filtered using a MAF of 0.05, which resulted in 352,306 SNVs and 24,726 indels for analysis. Positions with a missingness, or the number of individuals where there was SNV information was available, of >1% were removed from the analysis. The GWAS was performed using the EMMA eXpedited (EMMAX) software package[71], applying a linear mixed model with azole resistance (susceptible/resistant), source (environmental/clinical) or clinical disease (chronic/acute infection) as the phenotypic traits. The GEMMA, treeWAS and ECAT software packages were also tested in the framework of this project. EMMAX was ultimately selected over these tools because it accounted for sample structure the best, providing the least-inflated Q–Q plots (Extended Data Fig. 5a,c). Significant variants were determined using a cutoff of $P < 0.01$ with false-discovery-rate correction. The pan-GWAS was performed using a presence/absence matrix of the orthologous gene clusters, where zero denoted absent gene clusters and one represented gene clusters that were present in the genome. Associations between pan-gene presence/absence, isolate source and azole resistance were calculated using Scoary v.1.6.16 (ref. [72]).

Availability of isolates. The isolates that were sequenced in this study were submitted to and are publicly available in the Jena Microbial Resource Collection.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Raw FASTQ files for the isolates sequenced in this study were uploaded to the NCBI Sequence Read Archive and are publicly available under BioProject PRJNA697844. The accession numbers for the publicly available sequence data are listed in Extended Data Fig. 1. Annotated genome assemblies for sequence data generated in this study and for 64 isolates sequenced by us in a previous study[20] were submitted to NCBI GenBank and are available under the NCBI BioSample numbers listed in Extended Data Fig. 1. Datasets from FungiDB, release 39, are available at https://fungidb.org/fungidb/app/downloads/release-39/. The NCBI RefSeq non-redundant protein database v.22.01.08 is accessible at https://ftp.ncbi.nlm.nih.gov/blast/db/cloud/2018-01-22/. Source data are provided with this paper.

Received: 5 March 2021; Accepted: 8 October 2021;
Published online: 24 November 2021

References

1. Latgé, J. P. and Chamilos, G. *Aspergillus fumigatus* and Aspergillosis in 2019. *Clin. Microbiol. Rev.* https://doi.org/10.1128/CMR.00140-18 (2019).
2. *Invasive Aspergillosis. LIFE* http://www.life-worldwide.org/fungal-diseases/invasive-aspergillosis (2020).
3. Harrison, N. et al. Incidence and characteristics of invasive fungal diseases in allogeneic hematopoietic stem cell transplant recipients: a retrospective cohort study. *BMC Infect. Dis.* **15**, 584 (2015).
4. Kuster, S. et al. Incidence and outcome of invasive fungal diseases after allogeneic hematopoietic stem cell transplantation: a Swiss transplant cohort study. *Transpl. Infect. Dis.* **20**, e12981 (2018).
5. Heo, S. T. et al. Changes in in vitro susceptibility patterns of *Aspergillus* to triazoles and correlation with aspergillosis outcome in a tertiary care cancer center, 1999–2015. *Clin. Infect. Dis.* **65**, 216–225 (2017).
6. Lestrade, P. P. et al. Voriconazole resistance and mortality in invasive aspergillosis: a multicenter retrospective cohort study. *Clin. Infect. Dis.* **68**, 1463–1471 (2019).
7. Snelders, E. et al. Emergence of azole resistance in *Aspergillus fumigatus* and spread of a single resistance mechanism. *PLoS Med.* **5**, 1629–1637 (2008).
8. Rizzetto, L. et al. Strain dependent variation of immune responses to *A. fumigatus*: definition of pathogenic species. *PLoS ONE* **8**, 2–14 (2013).
9. Kowalski, C. H. et al. Heterogeneity among isolates reveals that fitness in low oxygen correlates with *Aspergillus fumigatus* virulence. *mBio* **7**, e01515-16 (2016).
10. Alshareef, F. & Robson, G. D. Genetic and virulence variation in an environmental population of the opportunistic pathogen *Aspergillus fumigatus*. *Microbiology* **160**, 742–751 (2014).
11. Knox, B. P. et al. Characterization of *Aspergillus fumigatus* isolates from air and surfaces of the International Space Station. *mSphere* **1**, e00227-16.
12. Ries, L. N. A. et al. Nutritional heterogeneity among *Aspergillus fumigatus* strains has consequences for virulence in a strain- and host-dependent manner. *Front. Microbiol.* **10**, 854 (2019).
13. Steenwyk, J. L. et al. Variation among biosynthetic gene clusters, secondary metabolite profiles, and cards of virulence across *Aspergillus* species. *Genetics* **216**, 481–497 (2020).
14. Dos Santos, R. A. C. et al. Genomic and phenotypic heterogeneity of clinical isolates of the human pathogens *Aspergillus fumigatus*, *Aspergillus lentulus*, and *Aspergillus fumigatiaffinis*. *Front. Genet.* **11**, 459 (2020).
15. Fedorova, N. D. et al. Genomic islands in the pathogenic filamentous fungus *Aspergillus fumigatus*. *PLoS Genet.* **4**, e1000046 (2008).
16. Abdolrasouli, A. et al. Genomic context of azole-resistance mutations in *Aspergillus fumigatus* using whole-genome sequencing. *mBio* **6**, e00536 (2015).
17. Garcia-Rubio, R. et al. Genome-wide comparative analysis of *Aspergillus fumigatus* strains: the reference genome as a matter of concern. *Genes* **9**, 363 (2018).
18. Fan, Y., Wang Y. and Xu, J. Comparative genome sequence analyses of geographic samples of *Aspergillus fumigatus*—relevance for amphotericin B resistance. *Microorganisms* **8**, 1673 (2020).
19. Puértolas-Balint, F. et al. Revealing the virulence potential of clinical and environmental *Aspergillus fumigatus* isolates using whole-genome sequencing. *Front. Microbiol.* **10**, 1970 (2019).
20. Barber, A. E. et al. Effects of agricultural fungicide use on *Aspergillus fumigatus* abundance, antifungal susceptibility, and population structure. *mBio* **11**, e02213-20 (2020).
21. Arendrup, M. C. et al. Method for the determination of broth dilution minimum inhibitory concentrations of antifungal agents for conidia forming moulds. E.DEF 9.3.2. *EUCAST* https://www.eucast.org/fileadmin/src/media/PDFs/EUCAST_files/AFST/Files/EUCAST_E_Def_9.3.2_Mould_testing_definitive_revised_2020.pdf (2020).
22. Jombart, T., Devillard, S. & Balloux, F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet.* **11**, 94 (2010).
23. Nierman, W. C. et al. Genomic sequence of the pathogenic and allergenic filamentous fungus Aspergillus fumigatus. *Nature* **438**, 1151–1156 (2005).
24. Steenwyk, J. L. et al. Genomic and phenotypic analysis of COVID-19-associated pulmonary aspergillosis isolates of *Aspergillus fumigatus*. *Microbiol. Spectr.* **9**, e0001021 (2021).

25. Abad, A. et al. What makes *Aspergillus fumigatus* a successful pathogen? Genes and molecules involved in invasive aspergillosis. *Rev. Iberoam. Micol.* **27**, 155–82 (2010).
26. Bignell, E., et al. Secondary metabolite arsenal of an opportunistic pathogenic fungus. *Philos. Trans. R Soc. B* **371**, 20160023 (2016).
27. Kjaerbolling, I. et al. Linking secondary metabolites to gene clusters through genome sequencing of six diverse *Aspergillus* species. *Proc. Natl Acad. Sci. USA* **115**, E753–E761 (2018).
28. Urban, M. et al. PHI-base: the pathogen-host interactions database. *Nucleic Acids Res.* **48**, D613–D620 (2020).
29. O'Hanlon, K. A. et al. Targeted disruption of nonribosomal peptide synthetase pes3 augments the virulence of *Aspergillus fumigatus*. *Infect. Immun.* **79**, 3978–3992 (2011).
30. Sugui, J. A. et al. Genes differentially expressed in conidia and hyphae of *Aspergillus fumigatus* upon exposure to human neutrophils. *PLoS ONE* **3**, e2655 (2008).
31. Willger, S. D. et al. A sterol-regulatory element binding protein is required for cell polarity, hypoxia adaptation, azole drug resistance, and virulence in *Aspergillus fumigatus*. *PLoS Pathog.* **4**, e1000200 (2008).
32. Blatzer, M., et al. SREBP coordinates iron and ergosterol homeostasis to mediate triazole drug and hypoxia responses in the human fungal pathogen *Aspergillus fumigatus*. *PLoS Genet.* **7**, e1002374 (2011).
33. Bertuzzi, M. et al. The pH-responsive PacC transcription factor of *Aspergillus fumigatus* governs epithelial entry and tissue invasion during pulmonary aspergillosis. *PLoS Pathog.* **10**, e1004413 (2014).
34. Bignell, E. et al. The *Aspergillus* pH-responsive transcription factor PacC regulates virulence. *Mol. Microbiol.* **55**, 1072–1084 (2005).
35. Pongpom, M. et al. Divergent targets of *Aspergillus fumigatus* AcuK and AcuM transcription factors during growth in vitro versus invasive disease. *Infect. Immun.* **83**, 923–933 (2015).
36. Camps, S. M. T. et al. Molecular epidemiology of *Aspergillus fumigatus* isolates harboring the TR34/L98H azole resistance mechanism. *J. Clin. Microbiol.* **50**, 2674–2680 (2012).
37. McCarthy, C. G. P. & Fitzpatrick, D. A. Pan-genome analyses of model fungal species. *Microb. Genom.* **5**, e000243 (2019).
38. Lind, A. L. et al. Drivers of genetic diversity in secondary metabolic gene clusters within a fungal species. *PLoS Biol.* **15**, e2003583 (2017).
39. Rybak, J. Mutations in *hmg1*, challenging the paradigm of clinical triazole resistance in *Aspergillus fumigatus*. *mBio* **10**, e00437-19 (2019).
40. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
41. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
42. Cingolani, P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly* **6**, 80–92 (2012).
43. Varga, J. Mating type gene homologues in *Aspergillus fumigatus*. *Microbiology* **149**, 816–819 (2003).
44. Paoletti, M. et al. Evidence for sexuality in the opportunistic fungal pathogen *Aspergillus fumigatus*. *Curr. Biol.* **15**, 1242–1248 (2005).
45. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
46. Peng, Y., et al. IDBA—a practical iterative de Bruijn graph de novo assembler. In *Proc. Research in Computational Molecular Biology. RECOMB 2010.* (ed. Berger, B.) 426–440 (Springer, 2010).
47. Gurevich, A. et al. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).
48. Palmer, J & Stajich J. Funannotate v.1.5.3. *Zenodo* https://zenodo.org/record/2604804 (2019).
49. Smit, A., Hubley, R. & Green, P. RepeatMasker Open-4.0. (2013–2015); http://www.repeatmasker.org
50. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* **6**, 11 (2015).
51. Haas, B. J. et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).
52. Ter-Hovhannisyan, V. et al. Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. *Genome Res.* **18**, 1979–1990 (2008).
53. Stanke, M. et al. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637–644 (2008).
54. Lowe, T. M. & Chan, P. P. tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. *Nucleic Acids Res.* **44**, W54–W57 (2016).
55. Finn, R. D. et al. Pfam: clans, web tools and services. *Nucleic Acids Res.* **34**, D247–D251 (2006).
56. Rawlings, N. D., Barrett, A. J. & Bateman, A. MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Res.* **40**, D343–D350 (2012).
57. Zhang, H. et al. dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res.* **46**, W95–W101 (2018).
58. Simao, F. A. et al. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
59. Aramaki, T. et al. KofamKOALA: KEGG ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* **36**, 2251–2252 (2020).
60. Jones, P. et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
61. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
62. Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinform.* **5**, 113 (2004).
63. Suyama, M., Torrents, D. & Bork, P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* **34**, W609–W612 (2006).
64. Minh, B. Q. et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
65. Hoang, D. T. et al. UFBoot2: improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).
66. Didelot, X. & Wilson, D. J. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PLoS Comput. Biol.* **11**, e1004041 (2015).
67. Schliep, K. P. phangorn: Phylogenetic analysis in R. *Bioinformatics* **27**, 592–593 (2011).
68. Yu, G. et al. Two methods for mapping and visualizing associated data on phylogeny using Ggtree. *Mol. Biol. Evol.* **35**, 3041–3043 (2018).
69. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
70. Whelan, F. J., Rusilowicz, M. and McInerney, J. O. Coinfinder: detecting significant associations and dissociations in pangenomes. *Microb. Genom.* **6**, e000338 (2020).
71. Kang, H. M. et al. Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354 (2010).
72. Brynildsrud, O. et al. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biol.* **17**, 238 (2016).

Acknowledgements

This project was supported by The Federal Ministry for Education and Science (Bundesministerium für Bildung und Forschung) within the framework of InfectControl 2020 projects FINAR and FINAR 2.0 (grant nos 03ZZ0809 and 03ZZ0834 to O.K.). The NRZMyk is supported by the Robert Koch Institute with funds provided by the German Ministry of Health (grant no. 1369-240 to O.K.). A.E.B. and G.P. are supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany´s Excellence Strategy (EXC 2051, Project ID 390713860). G.P. thanks the Deutsche Forschungsgemeinschaft (DFG) CRC/Transregio 124 'Pathogenic fungi and their human host: Networks of interaction', subproject INF (project number 210879364) for intellectual input. The authors thank M. Blango for thoughtful discussions on this manuscript.

Author contributions

A.E.B., G.P. and O.K. conceptualized and designed the study. The experimental work was performed by A.E.B. and G.W. A.E.B., B.S., G.P., K.K., J.L., O.K., T.S.-O. and G.W. analysed the data and interpreted results. A.E.B. wrote the primary manuscript and all of the listed authors were involved in the editing and review of the manuscript. O.K. acquired the primary funding for this work.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41564-021-00993-x.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41564-021-00993-x.

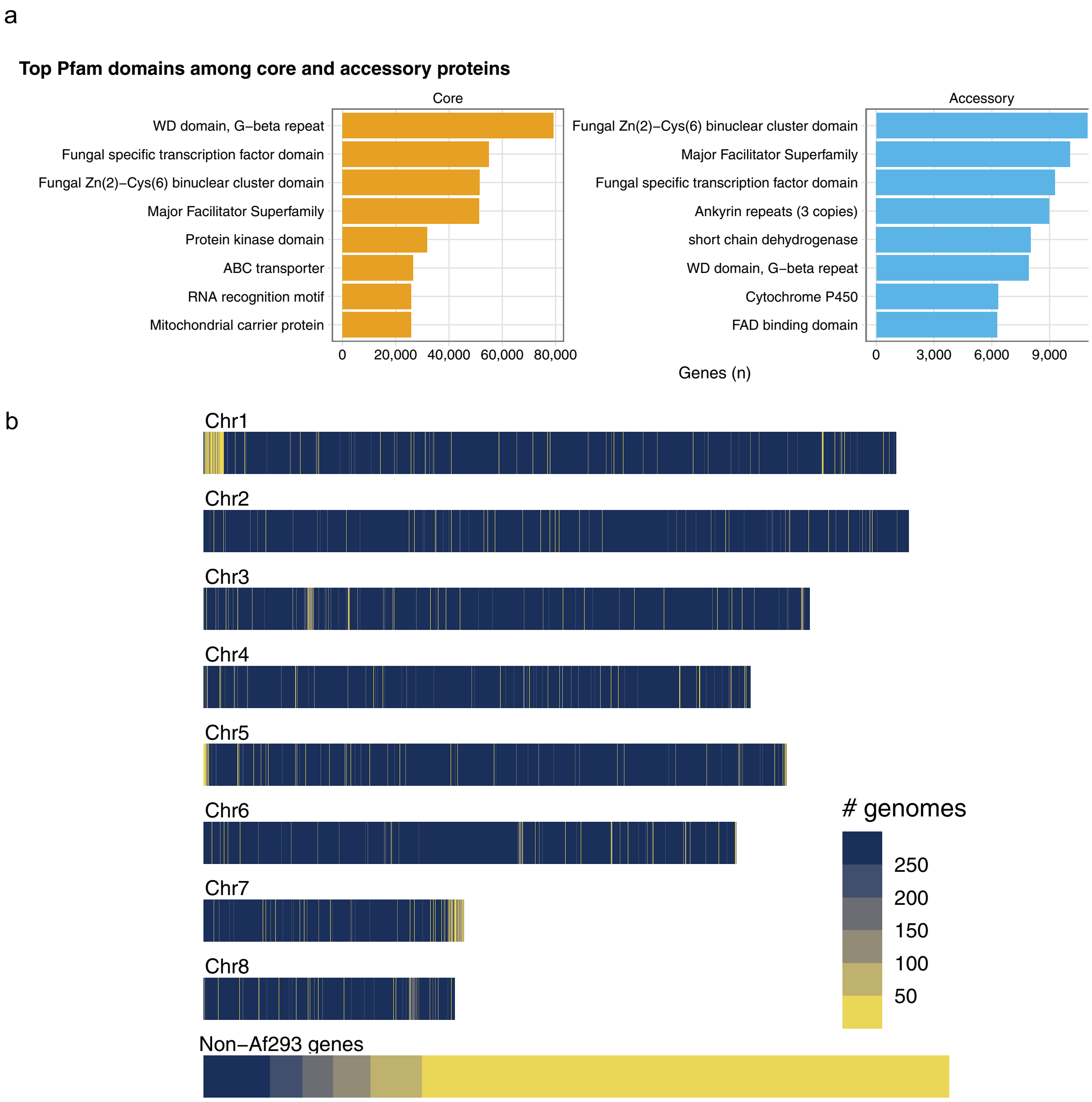
Correspondence and requests for materials should be addressed to Gianni Panagiotou or Oliver Kurzai.

Peer review information *Nature Microbiology* thanks Nancy Keller and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

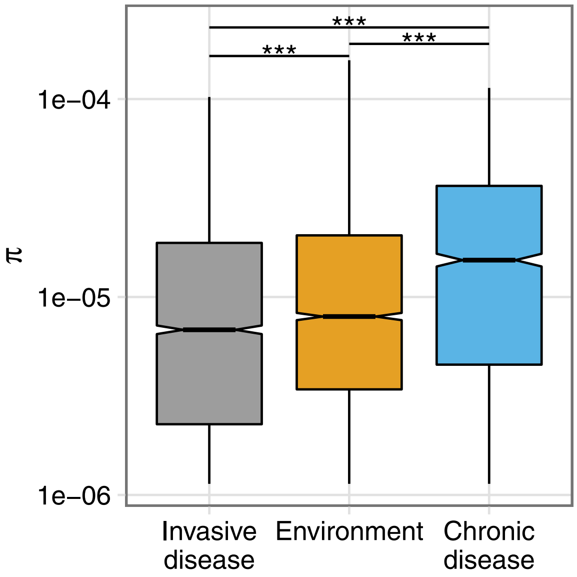
Reprints and permissions information is available at www.nature.com/reprints.

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

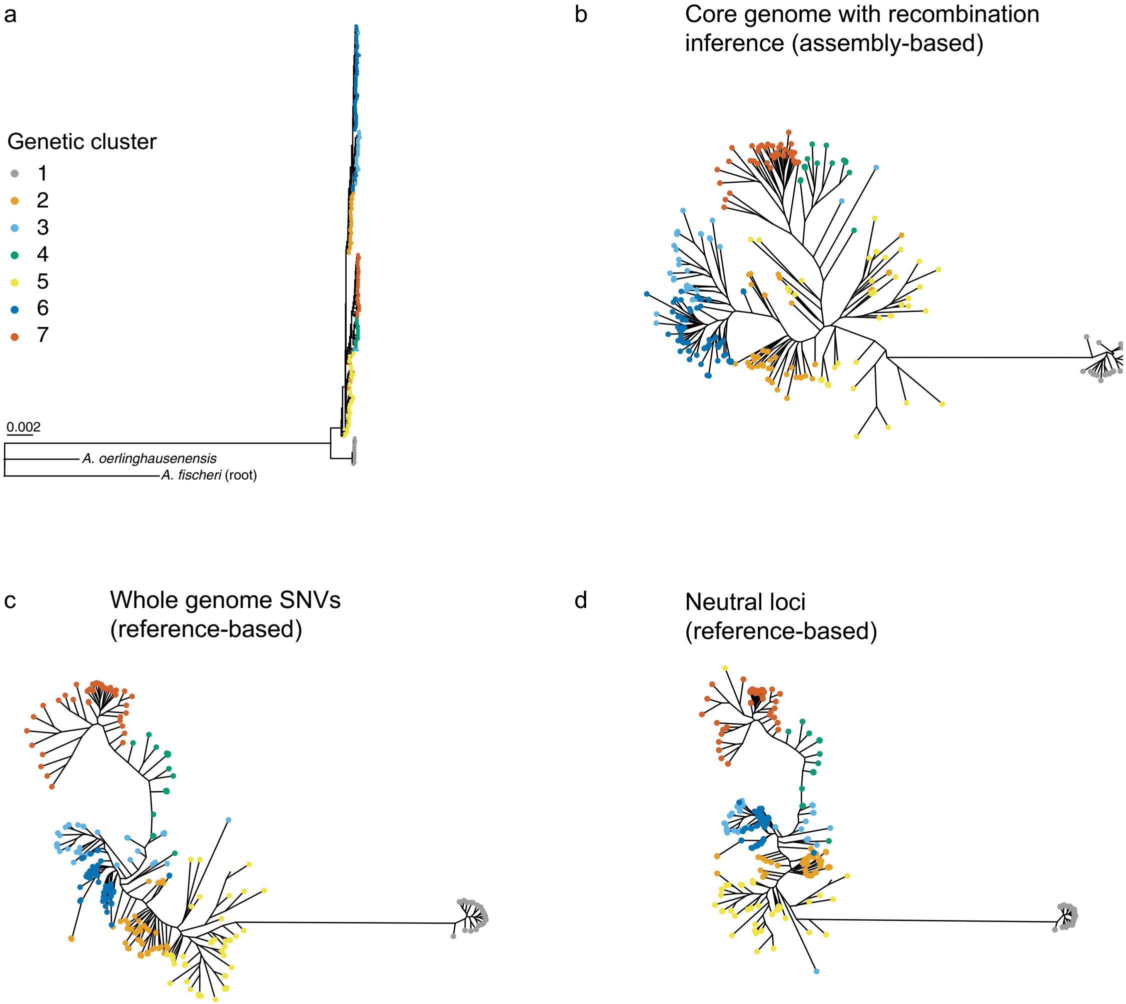
© The Author(s), under exclusive licence to Springer Nature Limited 2021



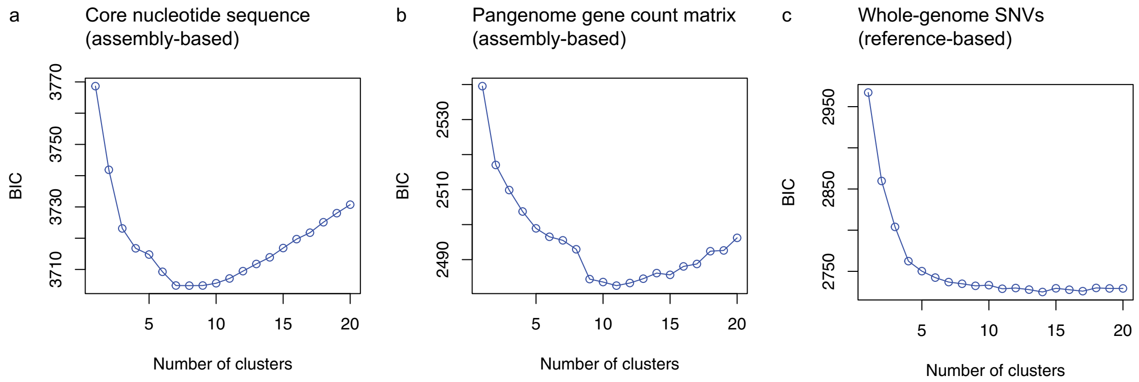
Extended Data Fig. 1 | The pan-genome of *A. fumigatus*. (A) Most frequently occurring Pfam domains among the core and accessory genomes. Values represent the total sum of domain-containing proteins among all 300 genomes. (B) Conservation of Af923 genes in the *A. fumigatus* pan-genome, arranged by chromosomal location in Af293. Each gene in Af293 is represented by a uniform-sized band that is coloured according to its prevalence among the 300 isolates analysed. Genes not in Af293 and their relative frequency are depicted at the bottom.



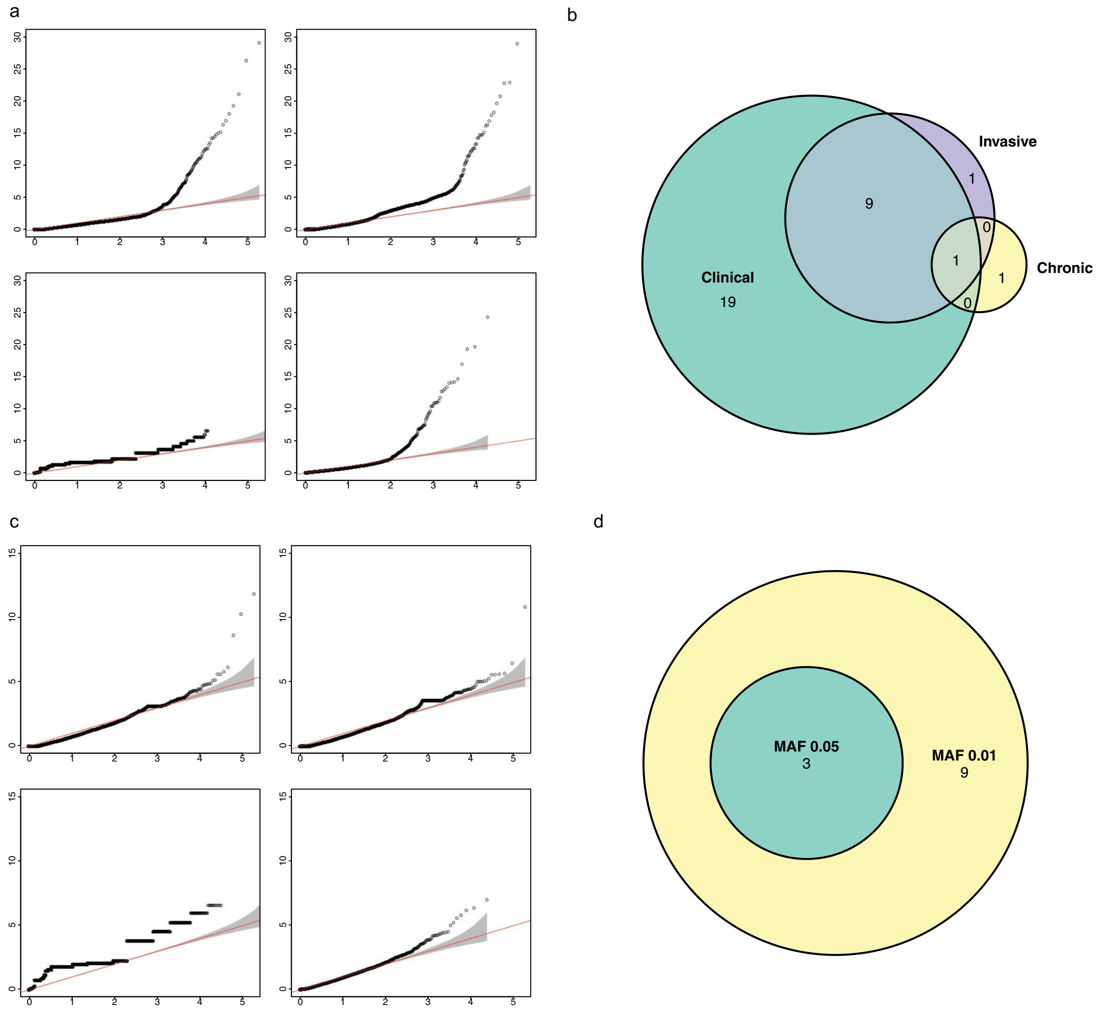
Extended Data Fig. 2 | Nucleotide diversity (π) of *A. fumigatus* isolates from the environment, invasive disease and chronic disease. π was calculated using 5 kb sliding windows across with genome with a 500 bp step size. Due to the underrepresentation of isolates from chronic disease in the dataset, isolates from the environment and invasive disease were downsampled to match the number of isolates from chronic disease (n = 19 isolates per group). The bold line in the box-and-whisker plot indicates the 50th percentile, and the box extends from the 25th to the 75th percentiles. The whiskers denote the lowest and highest values within 1.5 interquartile range. Statistical significance determined by two-sided Mann–Whitney *U* test with Bonferroni correction. *** represents $P < 0.001$. Exact P-values are: chronic vs. environmental: $P = 97e\text{-}39$; chronic vs invasive: $P = 1.6e\text{-}78$; invasive vs. environmental: P = 5.6e-14.


a
Genetic cluster
● 1
● 2
● 3
● 4
● 5
● 6
● 7
0.002
A. oerlinghausenensis
A. fischeri (root)
b
Core genome with recombination inference (assembly-based)
c
Whole genome SNVs (reference-based)
d
Neutral loci (reference-based)

Extended Data Fig. 3 | Phylogenies constructed from the genomes of 300 *A. fumigatus* using de novo assembled genomes and reference-base analyses. (a-b) Core genome phylogeny built from nucleotide coding sequence of 5,380 single-copy orthologous genes shared by all 300 *A. fumigatus* isolates, *A. oerlinghausenensis* and *A. fischeri* (alignment length = 9,178,893 bp). Panel a shows the phylogeny rooted with *A. fischeri* and depicts the scaled relationship between the two outgroups and the *A. fumigatus* samples. Panel b depicts this phylogeny unrooted and with outgroups removed for comparison to the other phylogenies. (c) Phylogeny from concatenated SNVs following read alignment to Af293 and variant calling (n = 341,031 base pair). Genomic positions with zero coverage in any sample were removed from the alignment. (d) SNV-based phylogeny constructed from 4-fold degenerate (neutral) loci (n = 35,052 base pair).



Extended Data Fig. 4 | Discriminant analysis of principle components of 300 *A. fumigatus* isolates. (a-c) Number of clusters vs. Bayesian information criteria (BIC) was used to assess the best supported number of genetic clusters for the dataset. The input for analysis was either (a) a non-gapped core nucleotide alignment from 5,830 single-copy orthologous genes (b) a gene count matrix from orthogroup-based clustering of pansequences or (c) or whole-genome SNV data with of positions with zero genomic coverage in any isolate in the dataset excluded.



Extended Data Fig. 5 | GWAS for variants associated with clinical isolates and triazole resistance. (a & c) Q–Q plots for association with isolate source (a; clinical vs. environmental) and triazole resistance (c; resistance to one or more triazole vs. susceptible to all examined; c). Four software were utilized: EMMAX (top, left), GEMMA (top, right), treeWAS (bottom, left) and ECAT (bottom, right). The resulting Q–Q plots were used to identify the tool that produced outputs where the expected p-value distribution (x-axis) best matched the observed p-values (y-axis). (b) Venn diagram showing the gene overlap between GWAS for all clinical strains relative to environmental and significant genes specific to acute and chronic disease. (d) Venn diagram showing the gene overlap for association with triazole resistance when minor allele frequencies (MAF) of 0.01 and MAF 0.05 were used.

Manuscript II

Transcriptomic analysis of presymptomatic sepsis patients reveals pathogen-specific host immune responses

Tongta Sae-Ong[1], Sascha Schäuble[1], Albert Garcia Lopez[1], Roman A Lukaszewski[2], Mervyn Singer[2,3,*], Gianni Panagiotou[1,4,*]

Overview

In manuscript II, we aimed to understand the different immune responses in sepsis patients to different types of pathogens, which will result in early detection based on gene regulation differences. Therefore, we applied the transcriptome data analysis method to study the gene expression profiles from pre-symptomatic sepsis blood. The results showed promising key genes regulating different immune mechanisms in sepsis patients infected by bacteria, fungi, and co-infection of bacteria and fungi.

FORM 1

Manuscript No. 2

Manuscript title: Transcriptomic analysis of presymptomatic sepsis patients reveals pathogen-specific host immune responses

Authors: Tongta Sae-Ong, Sascha Schäuble, Albert Garcia Lopez, Roman A Lukaszewski, Mervyn Singer, Gianni Panagiotou

Bibliographic information:

Sae-Ong, T., Schäuble S., Garcia Lopez, A., Lukaszewski R.A., Singer M., Panagiotou G., Transcriptomic analysis of presymptomatic sepsis patients reveals pathogen-specific host immune responses. (in preparation)

The candidate is (Please tick the appropriate box.)

☒ First author, ☐ Co-first author, ☐ Corresponding author, ☐ Co-author.

Status: in preparation

Authors' contributions (in %) to the given categories of the publication

Author	**Conceptual**	**Data analysis**	**Experimental**	**Writing the manuscript**	**Provision of material**
Sae-Ong, T.	50%	90%		40%	
Schäuble, S.				5%	
Garcia Lopez, A.		10%		5%	
Lukaszewski, R.A.			80%	5%	
Singer, M.			20%	5%	50%
Panagiotou, G.	50%			40%	50%
Others					
Total:	100%	100%	100%	100%	100%

Signature candidate

Signature supervisor (member of the Faculty)

Transcriptomic analysis of presymptomatic sepsis patients reveals pathogen-specific host immune responses

Tongta Sae-Ong[1], Sascha Schäuble[1], Albert Garcia Lopez[1], Roman A Lukaszewski[2], Mervyn Singer[2,3,*], Gianni Panagiotou[1,4,*]

*Corresponding authors
E-mails: m.singer@ucl.ac.uk or gianni.panagiotou@leibniz-hki.de

Affiliations:
[1]Research Group Systems Biology and Bioinformatics, Leibniz Institute for Natural Product Research and Infection Biology-Hans Knöll Institute, Jena, Germany
[2]Bloomsbury Institute of Intensive Care Medicine, Division of Medicine, University College London, London, UK
[3]Division of Critical Care and, NIHR University College London Hospitals Biomedical Research Centre, University College London Hospitals NHS Foundation Trust, London, UK
[4]Department of Medicine and State Key Laboratory of Pharmaceutical Biotechnology, University of Hong Kong, Hong Kong, China

Keywords:

Sepsis, pathogen type, WGCNA, PPI, hub genes

Abstract

Sepsis is a life-threatening organ dysfunction resulting from a dysregulated host immune response to infection. An early diagnosis and precise management within an hour significantly reduce mortality rates. Therefore, several studies aimed to identify biomarker genes for the effective diagnosis of sepsis. However, the diagnosis of the causative pathogens is still limited, which leads to inappropriate antimicrobial treatment. Our study aims to disentangle the heterogeneity of immune responses in presymptomatic sepsis patients infected by bacteria and/or fungi. Therefore, we investigated the differential expression of genes and pathways among 51 post-operative sepsis patients, which developed sepsis by bacterial, fungal, and co-infection, and 49 non-infected patients, using blood samples up to 3 days before clinical manifestation. Using weighted gene co-expression, protein-protein interaction networks, and machine learning, we were able to identify signatures of different pathogen-induced sepsis. *SH2D1B* gene involved in NK-cell activation was higher expressed in bacterial sepsis. Genes involved in transcriptional processes (*YBX*, *MAX*) and genes associated with the progression of various pathogen infections (*GFT2F2*, *RBM17*, *NEDD4,* and *AMPH*) were highly expressed in fungal and co-infection sepsis, respectively. These pathogen-type specific signature genes and pathways could serve as potential biomarkers for diagnosis and/or prognosis, allowing early targeted therapies to reduce sepsis mortality.

Introduction

Sepsis is a life-threatening organ dysfunction due to inappropriate host response to infections. Sepsis is the most common cause of death among hospitalized patients in the intensive care unit (ICU) (1, 2). The sepsis mortality rates are higher than 15% and can rise to 56% when patients present septic shock (3, 4). Besides, half of the survivors experience long-term outcomes by developing chronic critical illness (CCI) (5). Early recognition of sepsis and immediate management are able to reduce the mortality rate. Regarding the Surviving Sepsis Campaign (SSC) guidelines, empiric broad-spectrum antibiotic drugs should be given to sepsis patients within an hour of the disease recognition. The targeted antimicrobials should be applied after the pathogen has been identified (6). However, this nonspecific treatment has a low chance of successfully treating non-bacterial sepsis. At the same time, it can induce unnecessary disturbances of the gut microbiota, a vital component of our immune system (7).

Over 60% of sepsis is initiated by bacterial infection, followed by fungal and viral pathogens, and approximately 20% of patients are co-infected by multiple pathogens (8, 9). Blood culture is a “gold standard” method for pathogen identification in sepsis (10). Common bacteria such as *Escherichia coli* can be reported within 24 hours (11, 12). However, the turnaround time could take longer than 36 hours for the uncommon pathogens (11, 12). At the same time, in 30% of blood samples, it was not possible to identify the causative pathogen due to negative cultures (8, 12). Due to the delay in turnaround time, around 30% of sepsis patients received inappropriate treatments when initial empirical drugs did not cover the causative pathogens or resistant strains involved (13–15). The inappropriate treatments, such as the initial use of antibiotics for a fungal pathogen, could also increase the mortality rate of sepsis, the risk of colonization and the development of antibiotic-resistant pathogen infections (13, 14, 16). Moreover, up to 20% of patients receiving at least one of the broad-spectrum antibiotic drugs experienced adverse drug events (ADEs) (17, 18). Therefore, rapid pathogen identification is crucial for targeted therapy.

Expression levels of proteins such as Procalcitonin (PCT) and C-reactive protein (CRP) have been used as sepsis biomarkers (19, 20). However, pathogen detection is still challenging due to the lack of reliable pathogen-specific biomarkers (21, 22). Recently, gene expression profiling has been used to understand the immune response mechanisms and the pathogenesis of sepsis and to identify new sepsis biomarkers (23–25). In our study, we performed a meta-analysis of transcriptional data generated recently from sepsis patients and non-infected controls as part of a large prospective, multi-center study in patients undergoing elective major surgery, with daily blood sampling and data recording commencing pre-operatively and continuing up to two weeks after (Lukaszewski et al., 2022, in revision). We used here gene expression profiles from 100 patients obtained up to a week before patients were diagnosed with sepsis to identify bacterial-, fungal- and co-infection-specific host response signatures. Gene co-expression and protein-protein interaction networks were constructed to identify the different pathogen-induced sepsis-associated genes. We also performed functional annotation to understand the relationship between the immunity pathways and sepsis associated with the type of pathogens. Our results indicate that it is possible to identify early pathogen-specific biomarkers for targeted treatments that would improve the clinical outcome of sepsis.

Results

Patient recruitment with infection status

An overview of the study design is given in Figure 1. In brief, blood samples from 100 surgery patients were collected at pre- and post-operation time points. Pre-surgery blood samples were taken prior to surgery (PreSurgery). After the operation, blood samples were collected daily until seven days, or upon hospital discharge, or before the patient was diagnosed with sepsis (BeforeDx). Blood samples were collected daily from the day patients were diagnosed with sepsis up to seven days after diagnosis (AfterDx). In total, we obtained 427 blood samples (Table S1), including 216 samples from 51 patients diagnosed with sepsis and 211 blood samples from 49 non-infected controls. The 216 sepsis samples included 159 from bacterial sepsis (35 PreSurgery, 110 BeforeDx and 14 AfterDx), 9 fungal sepsis (3 PreSurgery and 6 BeforeDx), 17 co-infections with bacteria and fungi (4 PreSurgery, 11 BeforeDx and 2 AfterDx), and 31 samples that microbiological analysis could not identify the pathogen (8 PreSurgery, 20 BeforeDx and 3 AfterDx). Comparative blood samples were selected from non-infected surgery patients with whom they had matched clinical profiles with sepsis patients. Of total 211 comparators included 149 bacterial comparators (34 PreSurgery, 102 BeforeDx and 13 AfterDx), 9 fungal comparators (3 PreSurgery and 6 BeforeDx), 20 co-infection comparators (5 PreSurgery, 12 BeforeDx and 3 AfterDx) and 33 comparators for sepsis caused by un-identified pathogens (8 PreSurgery, 22 BeforeDx and 3 AfterDx).

Whole transcriptome profiles differentiate sepsis from non-infected patients but do not differentiate pathogen types

After data preprocessing and quality assessment, we obtained the expression matrix for 327 post-surgery samples (including before and after diagnosis of sepsis) to investigate the divergences of whole-transcriptome profiles in sepsis. A total of 30,646 transcripts corresponding to 18,010 genes (according to IlluminaHumanv4.db) were used in the downstream analysis. We first studied the differences between leukocytes in sepsis patients and comparators. Deconvolutions were performed to estimate leukocyte subpopulations between sepsis and comparators based on lymphocyte's signature genes (Figure 2a; Table S2). The proportion of immune cells in the sepsis compared to non-infected comparators showed significant differences (Wilcoxon rank-sum test, FDR = 0.05). Whereas CD8 T cells (FDR = 0.002), naive B cells (FDR

= 0.002), resting NK cells (FDR = 0.004), CD4 memory activated T cells (FDR = 0.004) and activated dendritic cells (FDR = 0.004) had higher proportion in sepsis patients, memory B cells (FDR = 0.001), γδ T cells (FDR = 0.005), CD4 memory resting T cells (FDR = 0.005) and neutrophils (FDR = 0.006) were relatively lower than comparators (Figure 2b).

Similarly, a PCA analysis of the gene expression profiles between sepsis and comparators also showed a significant separation of the two groups (PERMANOVA, *p*-value < 0.001) (Figure 2c). We subsequently explored the differences between the different types of pathogens -bacterial, fungal, co-infection- against their respective comparators. In all the sub-groups, a clear separation between the sepsis and controls was observed (*p*-value < 0.04, Figure 2d-2f). Interestingly, based on R^2 values, the differences in the transcriptional profile were more pronounced in the fungal pathogen-induced sepsis, followed by the co-infection and the bacterial-induced sepsis. Unidentified pathogen-induced sepsis also showed differences in the transcriptome profile compared to their comparators (*p*-value < 0.001, Figure 2g). Direct comparisons of the transcriptional profiles between the different types of pathogens showed no significant separation (Figure S2, *p*-value > 0.05). However, the sepsis samples in those groups were not matched in terms of other clinical characteristics, which may introduce bias.

Hierarchical clustering identified signature immune pathways of bacterial, fungal, and co-infection-induced sepsis

We used the transcriptomic profiles of 247 samples retrieved from patients after surgery but before sepsis diagnosis (BeforeDx) to identify pathogen-specific pre-diagnosis biomarkers of sepsis. Differentially expressed gene (DEG) analysis between sepsis and controls was performed using 110 bacterial infected sepsis samples versus 102 bacterial comparators, 6 fungal sepsis versus 6 fungal comparators, and 11 co-infection of bacterial and fungal sepsis patients versus 10 co-infection comparators. In total, 990 DEGs were obtained (limma, FDR = 0.05 and $|\log_2 FC| > 0.5$) including 270 bacterial-related DEGs, 400 fungal-related DEGs and 464 co-infection-related DEGs (Figure 3a-3d, Table S3). Interestingly, only *B4GALT5*, *IL4R,* and *KIF1B* were up-regulated, while *HLA-DRA* and *HLA-DRB1* were down-regulated in all sepsis patients, independently of the type of pathogen (Figure 3a). We further inspected the top overexpressed and specific DEGs to each pathogen-induced sepsis. *CD177*, a neutrophil-specific gene, was a gene specifically up-regulated in bacterial infection. High expression of this gene was reported in severe bacterial infections before (26). The top up-regulated gene in fungal infection is *SIGLEC1* (Sialic Acid Binding Ig-like Lectin 14). A recent study reported that *SIGLEC14* recognizes lipid compounds produced by the fungal pathogen *Trichophyton* spp and modulated innate immune response. (27). Moreover, the top unique up-regulated gene in co-infection sepsis is *TACSTD2* (transmembrane glycoprotein Trop2). A study by Lenárt et al., 2021 showed up-regulation of *TACSTD2* gene in lungs infected with viruses, bacteria, and fungi (28).

Unsupervised hierarchical clustering (HC; *k*-means) was performed to group the 990 DEGs by their expression patterns and identified five major gene clusters (Figure 3e, Table S4). We performed a functional enrichment analysis of the genes in each cluster (Table S4). The top five GO terms of each cluster are shown in Figure 3e. The results show that genes in clusters 1 and 2, associated with glycerolipid synthesis pathway and neutrophil-mediated immunity, respectively, were higher expressed in sepsis patients than comparators, independently of pathogen type. In contrast, genes in clusters 3, 4, and 5 were expressed higher in comparators. These three clusters were associated with DNA packaging, co-translational protein pathway, and IFN-γ-mediated inflammatory response. The five clusters in the heatmap showed corresponsive results to DEGs analysis (Figure S3, Table S5). The 196 bacterial up-regulated DEGs and 191 co-infection up-regulated DEGs were enriched in cluster 2 (hypergeometric distribution test, FDR < 0.001, Table S4) (Figure 3b,d,e), while the 274 down-regulated DEGs of co-infection induced sepsis were found in clusters 3 and 4 (FDR < 0.001; Figure 3d,e, Table S4). The 145 fungal up-regulated DEGs were enriched in cluster 1, while the 255 down-regulated DEGs were enriched in cluster 5 (FDR < 0.001; Figure 3c,e, Table S4).

We used up-regulated and down-regulated DEGs from each comparison (bacterial DEGs, fungal DEGs, and co-infection DEGs) to perform the Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis (ORA; Figure S3, Table S4). We found that T cell receptor (TCR) signaling pathway-related terms such as “MHC class II protein complex”, “T cell receptor signaling pathway”, “antigen receptor-mediated signaling pathway”, “MAPK signaling pathway”, and “Calcium signaling pathway”, were common enriched terms for all types of pathogens. When looking into pathogen-specific pathways, the results showed that the immune-related pathways involving reactive oxygen species (ROS) were uniquely enriched in bacterial up-regulated DEGs, while bacterial down-regulated DEGs were enriched for IFN-**α**-related immune responses. We also found that histone acetyltransferase-related terms involved in gene transcription were uniquely enriched in fungal up-regulated DEGs. The respiratory chain complex assembly biological processes were uniquely enriched in fungal down-regulated DEGs. Moreover, the interleukin-1 (IL-1) regulatory-related pathways were specifically enriched in co-infection up-regulated DEGs. Down-regulated DEGs of co-infection sepsis enriched for ribosomal RNA-related processes. The activation of these processes is involved in cellular proliferation and tissue repair in recovery sepsis patients (29). The co-expression and enrichment analysis results showed that each pathogen’s transcriptomic signature is associated with different immune-related pathways.

Weighted gene co-expression network analysis and protein-protein interaction network analysis reveal hub genes in bacterial, fungal, and co-infected-induced sepsis

In addition, we constructed a gene co-expression network based on the expression levels of DEGs (see Methods for details; Figure S4, Table S3) using only the 127 pre-diagnosis sepsis samples (BeforeDx). Using the "Weighted Gene Co-expression Network Analysis (WGCNA)", we detected eight gene modules (Figure S5-S6, Table S6). To highlight the signature gene modules of each pathogen group, we looked for DEGs unique to each pathogen in the gene co-expression modules (Table S6). WGCNA provided us with the higher infection group resolution with eight gene modules. However, the unique and shared gene modules among pathogens were consistent with the heatmap clusters. The 150 bacterial unique DEGs were significantly enriched in green (hypergeometric distribution test, FDR < 0.001, Figure S6a) and turquoise modules (FDR < 0.001, Figure S6b). Genes in these modules are functionally associated with neutrophil activation and ROS regulation. The 362 fungal unique DEGs were significantly enriched in black (FDR < 0.001, Figure S6c), yellow (FDR ≤ 0.001, Figure S6d), red (FDR ≤ 0.001, Figure S6e) and blue modules (FDR = 0.002, Figure S6f). The black module was enriched for antigen processing and presentation of exogenous antigen and MHC class II biological processes, whereas the yellow is functionally associated with the "nuclear receptor binding" pathway. The red and blue modules are enriched for co-translational related pathways such as "SRP-dependent co-translational protein targeting to membrane", "protein targeting to ER", and "establishment of protein localization to membrane" pathways. Moreover, the 339 unique DEGs of co-infection induced sepsis were significantly enriched in brown (FDR < 0.001, Figure S6g), pink (FDR = 0.002, Figure S6h) and blue modules (FDR = 0.015, Figure S6f). Genes in the brown module are related to DNA packaging pathways.

We subsequently looked for hub genes and pathogen-associated genes using a module membership (MM) of 0.2 (with the FDR adjusted *p*-value of 0.05) and a gene significance (GS) of 0.2 (with the FDR adjusted *p*-value of 0.05) as cutoffs. A total of 172 DEGs were identified as hub genes and were associated with at least one pathogen-induced sepsis (Table S6). We also constructed a protein-protein interaction (PPI) network of DEGs to predict the relationship between DEGs at the protein level. Node degree centrality of each node was further calculated by Network Analyzer in Cytoscape. Using a degree of more than 10 as the cutoff (30), we identified 255 hub genes in the PPI network (Table S7). A total of 42 genes were identified as hub genes in both PPI and WGCNA networks. PCA plots also showed that using the common hub genes improved the separation of the pathogen types (*p*-value < 0.001) (Figure 4a). However, the separation between bacterial- and fungal-induced sepsis was still not statistically significant (*P*-value = 0.134; Figure 4b) compared to co-infection vs. bacterial, and co-infection vs. fungal sepsis (*P*-values < 0.001; Figure 4a-d).

To evaluate the importance of common hub genes for discriminating the types of pathogens, we built classification models to distinguish (i) Bacterial vs. Non-bacterial sepsis, (ii) Fungal vs. Non-fungal sepsis, and (iii) Co-infection vs. Non-co-infection sepsis. Due to the small number of samples in fungal and co-infection groups, prediction performance was evaluated using "Leave-One-Out Cross-Validation" (LOOCV). Since the majority of samples in our dataset were bacterial-induced sepsis, we reweighted the samples in the dataset to compensate for the class imbalance problem so that each class had the same total weight. The classification results showed high performances for all models with an area under the curve (AUC) greater than 91% and accuracies greater than 97% (Figure 4e-f, Table S8). Significant features (genes) were examined by using the "sigFeaturePvalue" function in "sigFeature" package with a p-value ≤ 0.05. Moreover, we calculated AUC from the receiver operating characteristic (ROC) curve for each hub gene using the "pROC" package. Table 1 shows significant feature genes with high classification performances from each model (sigFeature p-value ≤ 0.05, AUC > 0.7).

Altogether, we found that the bacterial DEGs, *GNLY,* and *HLA-DRB1* were significant features in the bacterial classification model, and they were also bacterial-associated genes in the WGCNA network analysis. The fungal DEGs *HLA-DQB1*, *CLTA*, *ACSL,* and *RPS9* were significant feature genes of the fungal classification model and significantly associated with fungal sepsis in the WGCNA network analysis. A total of 16 co-infection DEGs were significant features for the co-infection classification model; among them, *AMPH*, *SH2D1B,* and *MAX* were associated with co-infection sepsis in the WGCNA network analysis.

To examine the possible value of the signature genes for pathogen identification in sepsis patients, we used the 42 hub gene-based classification models to predict the infection status of 20 "BeforeDx" samples from 8 unidentified pathogen sepsis patients. We separately applied the bacteria, fungal, and co-infection classification models to the set of patients with unknown pathogen sepsis. The bacterial models classified all sepsis patients as being induced by bacterial pathogens, whereas both the fungal and the co-infection models predicted those patients as non-fungal and non-co-infection patients (Table S9). Our results suggest that a combination of hub gene expressions is promising as early pathogen predictive biomarkers in sepsis patients.

Discussion

In this unique study in terms of clinical design, we found that the human immune system in presymptomatic sepsis responded differently to bacteria, fungi, or co-infection. Of 990 genes expressed differently in presymptomatic sepsis patients depending on the causative agent, only 5 responded significantly to all pathogen types. *B4GALT5* (Beta-1,4-galactosyltransferase 5), *IL4R* (Interleukin 4 Receptor), and *KIF1B* (Kinesin Family Member 1B) were commonly up-regulated in all pathogen-associated groups. At the same time, Human Leukocyte antigen-DR genes, including *HLA-DRA* and *HLA-DRB1,* were common down-regulated genes. Previous studies reported that *B4GALT5* and *KIF1B* were important up-regulated genes in sepsis (31, 32). *HLA-DRA* and *HLA-DRB1* genes are major histocompatibility complex (MHC) class II receptors, which are expressed by antigen-presenting cells (APCs) during infections (33). A recent study showed the significantly reduced expression of the classical HLA class II and MHC class II regulator genes in post-operative sepsis patients (34). In addition, we also found that T cell receptor (TCR) signaling pathway-related GO terms were low expressed in all sepsis patients. TCRs recognize the antigens presented by MHC molecules on antigen-presenting cells (APCs) (35). The TCR regulated pathway had been reported as a poor prognosis related to a high mortality rate and septic shock in sepsis patients (36, 37). Moreover, the results from immune cell deconvolution revealed that sepsis patients had less APCs such as memory B cells and macrophages than comparator samples. Our study showed strong evidence supporting the TCR signaling pathway interacting with MHC class II was reduced, suggesting a poor prognosis in post-operative sepsis patients.

The prospective nature of our clinical study allowed us to explore the transcriptomic signatures of the host induced by different types of pathogens before the clinical diagnosis of sepsis. The WGCNA and PPI analysis based on significant genes revealed potential pathogen-specific genes in presymptomatic sepsis that may serve as predictive biomarkers for immediate appropriate antimicrobial drug administrations. Based on those genes, we developed machine learning models that suggested a number of genes as highly predictive of the type of pathogen involved in the deterioration of those patients. Among those, genes of great interest were highly expressed specifically to each pathogen, including *SH2D1B*, *YBX1*, *MAX*, *GFT2F2*, *RBM17*, *NEDD4*, and *AMPH*.

SH2D1B (SH2 domain-containing 1B or Ewing's sarcoma-associated transcript 2: *EAT2*) was a higher expressed gene in bacterial than co-infection sepsis. *SH2D1B* plays a vital role in NK cell activation, responding to tumor and infected cells (38, 39). Recently, a study by Duffy et al. found up-regulation of *SH2D1B* in mice with *Mycobacterium tuberculosis* infection (40). Nevertheless, in the same study, the expression of *SH2D1B* was decreased in humans with tuberculosis (40). Another study found the significant dysregulation of *SH2D1B* in mental diseases (41). However, the association between *SH2D1B* and fungal infection has not been observed.

Y-box protein 1, or the nuclease-sensitive element-binding protein, is encoded by *YBX1* or *NSEP1* gene (42). It is an essential DNA and RNA binding protein that functions in several signaling pathways such as transcriptional and translational regulation, DNA reparation, pre-mRNA splicing, mRNA packaging, cell proliferation, cell differentiation, and apoptosis (43). Y-box protein 1 is considered an oncoprotein, as a higher expression of this protein is related to metastasis and poor prognosis in many cancer types (44, 45). This protein also showed a critical role in supporting viral replication, such as dengue virus (46) and human immunodeficiency virus (HIV) (42), which it was quantified as a new antiviral target. The up-regulation of this gene was found in antifungal drug (Amphotericin B; AMB) responsiveness (47). However, the effect of this protein on the antifungal drug is still unclear. Our study showed the higher expression of the *YBX1* gene in fungal sepsis while it was down-regulated in bacterial and co-infection sepsis. We also found that fungal DEGs were uniquely enriched in histone acetyltransferase-related pathways involved in transcriptional processes. While another transcriptional factor, *MAX* (MYC Associated

Factor X), was significantly decreased in co-infection sepsis patients compared to bacterial and fungal induced sepsis. The results suggested a high correlation of transcriptional processes in fungal sepsis patients.

We also found highly expressed genes in co-infection sepsis, including *GFT2F2*, *RBM17*, *NEDD4,* and *AMPH*. *NEDD4* (NEDD4 E3 Ubiquitin Protein Ligase) has an important role in pathogenesis. The high expression of *NEDD4* is related to viral replication and disease progression, including SARS-CoV-2/COVID-19 (48–50). The study between *NEDD4* and bacterial and fungal infection is still limited. However, recent studies found that the *NEDD4* functions involve bacterial infected cell clearance (51) and fungal killing (52). In contrast, another study reported that the increase of *NEDD4* supports the survival of bacteria in macrophage cells (53). In our study, *NEDD4* was up-regulated in sepsis-induced by all types of pathogens. However, it was significantly up-regulated in co-infection sepsis patients, supporting the relationship between this gene and pathogenesis.

GTF2F2 and *RBM17* were up-regulated in co-infection sepsis while down-regulated in bacterial and fungal sepsis compared to their comparators. *GTF2F2* (General Transcription Factor IIF, Polypeptide 2 or TFIIF) functions involved in transcription elongation by binding to RNA polymerase II (Pol II) (54). This gene has been widely studied for the interaction with viral proteins causing pathogenesis of viral infection and autoimmune diseases such as HIV-1 (55), dengue virus (56) and SARS-CoV-2 (57). Another study by Wu et al. used PPI network analysis, suggesting *GTF2F2* as a potential therapeutic target for sepsis (58). *RBM17* encodes RNA-binding motif protein 17 or SPF45, which is a part of the spliceosome complex with an important role in mRNA splicing (59). Several studies showed that the overexpression of *RBM17* indicated broad multidrug resistance to anticancer drugs (60). This gene was also found to support HIV-1 replication (61). Another study reported the overexpression of this gene in hepatitis B virus (HBV) infection (62). In conclusion, the results suggested that co-infection patients are more susceptible to viral infection than sepsis patients infected by one type of pathogen.

Furthermore, *AMPH* (Amphiphysin), a nerve terminal-enriched protein in BAR (*Bin-Amphiphysin-R*vsp) protein superfamily, has functions involved in clathrin-mediated endocytosis and phagocytosis (63–65). The reduction of AMPH1 protein level was found in neurodegenerative diseases such as Alzheimer’s disease (AD) (66) and several cancer progressions, such as breast cancer (67), lung cancer (68) and osteosarcoma (69). The overexpression of *AMPH* was also found in tuberculous meningitis infection in brain tissue (70). In this study, *AMPH* was up-regulated in all sepsis patients compared to their comparators. However, it was significantly up-regulated in co-infection sepsis. Our findings showed that the overexpression of genes in co-infection patients suggested a poorer prognosis for patients with multidrug resistance and higher susceptibility to various pathogens infection than sepsis induced by sole bacteria or fungi.

Our study also has limitations. As the transcriptomic data in presymptomatic sepsis were available here for the first time, the number of patients diagnosed with fungal or co-infection pathogens was significantly lower than bacteria-induced sepsis. Even though we could still identify with high statistical significance a large number of DEGs related to fungal and co-infection pathogens, the low number of patients posed challenges in developing machine learning models for simultaneous classification of the three types of pathogens. Nevertheless, the combination of potential genes/proteins and specific immune pathways identified in our study appear promising for differentiating as early as possible bacterial, fungal, and co-infection presymptomatic sepsis patients, and they should be further explored in more extensive clinical studies.

Materials and Methods

Microarray data generation and preprocessing

Microarray data of 427 blood samples were generated as described previously (Lukaszewski et al., 2022, in revision). Briefly, Globin-reduced RNA (GlobinClear^{TM}, ThermoFisher, Waltham, MA) was prepared from total RNA for each sample. RNA integrities were measured using a Bioanalyzer 2100 (Agilent, Santa Clara, CA), and concentrations were assessed using a NanoQuant^{TM} (Tecan, Männedorf, Switzerland). cRNA was prepared by amplification and labeling using the Illumina® TotalPrepTM RNA Amplification Kit (ThermoFisher) and hybridized to Human HT-12v4 Beadarrays (Illumina®, San Diego, CA). Expression levels of RNA samples were analyzed with Illumina® HighScanHQ^{TM}. The Illumina® HighScanHQ^{TM} then imaged each chip with resulting intensities indicating the expression level of each probe's corresponding gene. Finally, the chip data were preprocessed, and background corrected using GenomeStudio^{TM} Software v2011.1 (Illumina®). The data is under embargo and available upon request to primary authors.

A Principal Component Analysis (PCA) and boxplots of microarray data suggested a batch effect across datasets (ANOSIM; P-value < 0.001; Figure S1a-b). Therefore, batch deviation in the gene expression data from two different cohorts was removed by "ComBat" function of the "sva" package in R (71). A new PCA and boxplot indicated that batch effects had been removed (*p*-value = 0.99; Figure S1c-d). The transcriptome data were normalized using the "normalizeBetweenArrays" function of "limma" R package. The microarray probes were annotated to gene symbols and NCBI Entrez GeneIDs using the package 'illuminaHumanv4.db' (Version 1.26.0) (72).

Principal Component Analysis (PCA)

PCA method was applied to reduce the high-dimensional expression data into two-dimensional spaces using the "prcomp" function in the R package "stats" (R version 4.0.3) (73).

Immune cell deconvolution analysis

MySort tool (default version) implemented in an R function (74) was used to resolve the relative proportion of 21 lymphocytes based on the expression of signature genes using a linear regression model (62) (75).

Differential gene expression analysis

Normalized data were $\log_2$ transformed and assessed the differentially expressed genes (DEGs) by the Empirical Bayes method in the "limma" software package (Version 3.44.3) from Bioconductor in R (76). Then, a false discovery rate (FDR) adjusted p-value ≤ 0.05 and $|\log_2(\text{Fold change})| \geq 0.5$ were used as thresholds for identifying significant DEGs.

We performed DEG analysis between different pathogen-induced sepsis patients vs. their comparators (bacterial sepsis vs. bacterial comparators, fungal sepsis vs. fungal comparators, and co-infection sepsis vs. co-infection comparators) used for downstream analysis. In addition, to control DEG analysis between samples, we also performed DEG analysis between comparators among different types of pathogens. The results showed no significant DEGs between control samples.

Moreover, DEGs between different sepsis groups (bacterial vs. non-bacterial sepsis, fungal vs. non-fungal sepsis, and co-infection vs. bacterial/fungal sepsis) were also calculated as the additional DEGs for gene co-expression network, protein-protein interaction network, and machine learning analysis (Figure S4).

Weighted gene co-expression network analysis (WGCNA)

A total of 990 DEGs from bacterial/fungal/co-infection sepsis versus their comparators (Figure 3a-d) and 84 DEGs from direct comparisons between the different types of pathogens (Figure S4) were used to create a gene co-expression network. The network was built using the "weighted gene co-expression network analysis (WGCNA, Version 1.69)" (77) package. The soft-thresholding power (β) of 5 was chosen by applying the scale-free topology criterion of $r^2 = 0.8$ (Figure S5a-b). The unsigned co-expression network of sepsis samples was constructed using the "blockwiseModules" function in the WGCNA package. An adjacency matrix was calculated and transformed into a topological overlap matrix (TOM). TOM dissimilarity was used to perform hierarchical clustering resulting in the gene modules. Similar modules were merged based on the minimum module size of 30 genes and minimum height for identifying modules at 0.25, resulting in eight final modules (Figure S5c). Module eigengene (MEs) values of each module were calculated to explain the maximum variation of the gene expression profile of a module. Gene pairs with adjacency values of 0.03 or higher were exported to Cytoscape (Version 3.7.1) (78) for visualization.

We calculated the Pearson's correlation (*r*) between module eigengenes (MEs) and infection status to identify modules related to sepsis based on the types of pathogens. The sepsis-associated genes were defined using the gene significance test (GS) and module membership test (MM). GS evaluates the significance of genes and infection status, while MM evaluates the significance of genes and MEs. In addition, the significance of modules and genes were also calculated using FDR correction. As criteria for associated modules, we used |r| 0.2 and FDR 0.05. The associated genes were chosen using |GS| 0.2, |MM| 0.2, and FDR 0.05.

Protein-protein interaction (PPI)

To create the PPI network, all significant DEGs were uploaded to the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING; string-db.org/) database (79). The degree of each gene was calculated by the "NetworkAnalyzer" implemented tool in Cytoscape. Genes with 10 degrees or higher, were selected as hub genes from PPI analysis (30).

Machine learning and significant gene evaluation

We employed a supervised machine learning algorithm, the weighted kernel Nearest Neighbor (wKNN) from "kknn" package (version 1.3.1) in R (80), to assess the performances of hub genes to differentiate the sepsis based on pathogen types. We set the tuning parameters for the maximum number of neighbors (kmax) of 5, distance function (d) of 1, and kernel functions (kernel) including "rectangular", "biweight", and "optimal". The Leave-one-out cross-validation (LOOCV) was used to train and optimize sepsis classification models. To compensate for the imbalanced sizes of infection sepsis, we reweighted the classes of infection classes before training the models. The significant genes for each model were estimated using the support vector machine recursive feature elimination (SVM-RFE) algorithm and *t*-statistic from the "sigFeature" package (version 1.8.0) (81).

Receiver operating characteristic (ROC) curve analysis

ROC curves and area under the curves (AUC) were used for (i) investigating the performance of classification models and (ii) evaluating the predictive values of significant genes from classification models. The ROC curves and AUC were calculated using "pROC" package (Version 1.18.0) (82).

Functional annotation

The DEGs of different pathogen-induced sepsis, heatmap clusters, and gene modules from the gene co-expression network were evaluated by GO enrichment analysis and a Kyoto Encyclopedia of Genes and Genomes (KEGG) (83) pathway enrichment analysis using the "clusterProfiler" package (Version 3.16.1) (84) in R. GO terms and KEGG pathways for genes were obtained from the Bioconductor package “org.Hs.eg.db” (Version 3.11.4) (85). We selected the significant terms based on criteria of FDR ≤ 0.05 and two DEGs or more were involved.

Statistical analysis

ANOSIM and Adonis tests were performed based on the Bray Curtis metric using ‘vegan’ package (Version 2.5.7) (86). ANOSIM was employed to test statistical differences in gene expression profiles between two datasets. Adonis was used to testing for differences in gene expression levels between PCA clusters. All DEGs were clustered using the *K*-means algorithm. Heatmap was created using the ComplexHeatmap package (Version 2.4.3) (87). Figures were generated using the R package ‘ggplot2’ (Version 3.3.3), ‘VennDiagram’ (Version 1.6.20) (88), ComplexHeatmap, and enrichplot (Version 1.8.1) (89).

References:

1. Singer M, Deutschman CS, Seymour C, Shankar-Hari M, Annane D, Bauer M, Bellomo R, Bernard GR, Chiche JD, Coopersmith CM, Hotchkiss RS, Levy MM, Marshall JC, Martin GS, Opal SM, Rubenfeld GD, Poll T Der, Vincent JL, Angus DC. 2016. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA 315:801.
2. Rello J, Valenzuela-Sánchez F, Ruiz-Rodriguez M, Moyano S. 2017. Sepsis: A Review of Advances in Management. Adv Ther 2017 3411 34:2393–2411.
3. Bauer M, Gerlach H, Vogelmann T, Preissing F, Stiefel J, Adam D. 2020. Mortality in sepsis and septic shock in Europe, North America and Australia between 2009 and 2019- results from a systematic review and meta-analysis. Crit Care 24:1–9.
4. Nasir N, Jamil B, Siddiqui S, Talat N, Khan FA, Hussain R. 2015. Mortality in Sepsis and its relationship with Gender. Pakistan J Med Sci 31:1201.
5. Stortz JA, Mira JC, Raymond SL, Loftus TJ, Ozrazgat-Baslanti T, Wang Z, Ghita GL, Leeuwenburgh C, Segal MS, Bihorac A, Brumback BA, Mohr AM, Efron PA, Moldawer LL, Moore FA, Brakenridge SC. 2018. Benchmarking clinical outcomes and the immunocatabolic phenotype of chronic critical illness after sepsis in surgical intensive care unit patients. J Trauma Acute Care Surg 84:342.
6. Evans L, Rhodes A, Alhazzani W, Antonelli M, Coopersmith CM, French C, Machado FR, Mcintyre L, Ostermann M, Prescott HC, Schorr C, Simpson S, Wiersinga WJ, Alshamsi F, Angus DC, Arabi Y, Azevedo L, Beale R, Beilman G, Belley-Cote E, Burry L, Cecconi M, Centofanti J, Coz Yataco A, De Waele J, Dellinger RP, Doi K, Du B, Estenssoro E, Ferrer R, Gomersall C, Hodgson C, Møller MH, Iwashyna T, Jacob S, Kleinpell R, Klompas M, Koh Y, Kumar A, Kwizera A, Lobo S, Masur H, McGloughlin S, Mehta S, Mehta Y, Mer M, Nunnally M, Oczkowski S, Osborn T, Papathanassoglou E, Perner A, Puskarich M, Roberts J, Schweickert W, Seckel M, Sevransky J, Sprung CL, Welte T, Zimmerman J, Levy M. 2021. Surviving sepsis campaign: international guidelines for management of sepsis and septic shock 2021. Intensive Care Med 2021 4711 47:1181–1247.
7. Greenwood C, Morrow AL, Lagomarcino AJ, Altaye M, Taft DH, Yu Z, Newburg DS,

Ward D V., Schibler KR. 2014. Early Empiric Antibiotic Use in Preterm Infants Is Associated with Lower Bacterial Diversity and Higher Relative Abundance of Enterobacter. J Pediatr 165:23–29.

8. Prescott HC. 2018. The Epidemiology of Sepsis. Handb Sepsis 15–28.
9. Grondman I, Pirvu A, Riza A, Ioana M, Netea MG. 2020. Biomarkers of inflammation and the etiology of sepsis. Biochem Soc Trans 48:1–14.
10. Mancini N, Burioni R, Clementi M. 2015. Microbiological Diagnosis of Sepsis: The Confounding Effects of a “Gold Standard.” Methods Mol Biol 1237:1–4.
11. Weinbren MJ, Collins M, Heathcote R, Umar M, Nisar M, Ainger C, Masters P. 2018. Optimization of the blood culture pathway: a template for improved sepsis management and diagnostic antimicrobial stewardship. J Hosp Infect 98:232–235.
12. Tabak YP, Vankeepuram L, Ye G, Jeffers K, Gupta V, Murray PR. 2018. Blood culture turnaround time in U.S. acute care hospitals and implications for laboratory process optimization. J Clin Microbiol 56:500–518.
13. Ibrahim EH, Sherman G, Ward S, Fraser VJ, Kollef MH. 2000. The Influence of Inadequate Antimicrobial Treatment of Bloodstream Infections on Patient Outcomes in the ICU Setting. Chest 118:146–155.
14. Kollef MH. 2008. Broad-Spectrum Antimicrobials and the Treatment of Serious Bacterial Infections: Getting It Right Up Front. Clin Infect Dis 47:S3–S13.
15. Shehadeh F, Zacharioudakis IM, Zervou FN, Mylonakis E. 2019. Cost-effectiveness of rapid diagnostic assays that perform directly on blood samples for the diagnosis of septic shock. Diagn Microbiol Infect Dis 94:378–384.
16. Al-Sunaidar KA, Abd Aziz N, Hassan Y. 2020. Appropriateness of empirical antibiotics: risk factors of adult patients with sepsis in the ICU. Int J Clin Pharm 42:527–538.
17. Tamma PD, Avdic E, Li DX, Dzintars K, Cosgrove SE. 2017. Association of Adverse Events With Antibiotic Use in Hospitalized Patients. JAMA Intern Med 177:1308.
18. Hagiya H, Kokado R, Ueda A, Okuno H, Morii D, Hamaguchi S, Yamamoto N, Yoshida H, Tomono K. 2019. Association of Adverse Drug Events with Broad-spectrum Antibiotic Use in Hospitalized Patients: A Single-center Study. Intern Med 58:2621.
19. Karzai W, Oberhoffer M, Meier-Hellmann A, Reinhart K. 1997. Procalcitonin — A new indicator of the systemic response to severe infections. Infection 25:329.
20. Raveendran AV, Kumar A, Gangadharan S. 2019. Biomarkers and newer laboratory investigations the diagnosis of sepsis. J R Coll Physicians Edinb 49:207–216.
21. Vincent JL. 2016. The Clinical Challenge of Sepsis Identification and Monitoring. PLOS Med 13:e1002022.
22. Trzeciak A, Pietropaoli AP, Kim M. 2020. Biomarkers and Associated Immune Mechanisms for Early Detection and Therapeutic Management of Sepsis. Immune Netw 20:1–20.
23. Lu X, Xue L, Sun W, Ye J, Zhu Z, Mei H. 2018. Identification of key pathogenic genes of sepsis based on the Gene Expression Omnibus database. Mol Med Rep 17:3042–3054.
24. Sweeney TE, Perumal TM, Henao R, Nichols M, Howrylak JA, Choi AM, Bermejo-Martin JF, Almansa R, Tamayo E, Davenport EE, Burnham KL, Hinds CJ, Knight JC, Woods CW, Kingsmore SF, Ginsburg GS, Wong HR, Parnell GP, Tang B, Moldawer LL, Moore FE, Omberg L, Khatri P, Tsalik EL, Mangravite LM, Langley RJ. 2018. A community approach to mortality prediction in sepsis via gene expression analysis. Nat Commun 2018 91 9:1–10.

25. Wong HR. 2021. Pediatric sepsis biomarkers for prognostic and predictive enrichment. Pediatr Res 2021 912 91:283–288.
26. Göhring K, Wolff J, Doppl W, Schmidt KL, Fenchel K, Pralle H, Sibelius U, Bux J. 2004. Neutrophil CD177 (NB1 gp, HNA-2a) expression is increased in severe bacterial infections and polycythaemia vera. Br J Haematol 126:252–254.
27. Suematsu R, Miyamoto T, Saijo S, Yamasaki S, Tada Y, Yoshida H, Miyake Y. 2019. Identification of lipophilic ligands of Siglec5 and -14 that modulate innate immune responses. J Biol Chem 294:16776–16788.
28. Lenárt S, Lenárt P, Knopfová L, Kotasová H, Pelková V, Sedláková V, Čan V, Šmarda J, Souček K, Hampl A, Beneš P. 2021. TACSTD2 upregulation is an early reaction to lung infection. bioRxiv 2021.06.29.450320.
29. Cheng PL, Chen HH, Jiang YH, Hsiao TH, Wang CY, Wu CL, Ko TM, Chao WC. 2021. Using RNA-Seq to Investigate Immune-Metabolism Features in Immunocompromised Patients With Sepsis. Front Med 8:2621.
30. Cui H, Shan H, Miao MZ, Jiang Z, Meng Y, Chen R, Zhang L, Liu Y. 2020. Identification of the key genes and pathways involved in the tumorigenesis and prognosis of kidney renal clear cell carcinoma. Sci Reports 2020 101 10:1–10.
31. Xie K, Kong S, Li F, Zhang Y, Wang J, Zhao W. 2020. Bioinformatics-Based Study to Investigate Potential Differentially Expressed Genes and miRNAs in Pediatric Sepsis. Med Sci Monit 26:e923881-1.
32. Fan Y, Han Q, Li J, Ye G, Zhang X, Xu T, Li H. 2022. Revealing potential diagnostic gene biomarkers of septic shock based on machine learning analysis. BMC Infect Dis 22:1–16.
33. Kessal K, Liang H, Rabut G, Daull P, Garrigue JS, Docquier M, Parsadaniantz SM, Baudouin C, Brignole-Baudouin F. 2018. Conjunctival inflammatory gene expression profiling in dry eye disease: Correlations with HLA-DRA and HLA-DRB1. Front Immunol 9:2271.
34. Siegler BH, Altvater M, Thon JN, Neuhaus C, Arens C, Uhle F, Lichtenstern C, Weigand MA, Weiterer S. 2021. Postoperative abdominal sepsis induces selective and persistent changes in CTCF binding within the MHC-II region of human monocytes. PLoS One 16:e0250818.
35. Daniels MA, Teixeiro E. 2015. TCR signaling in T cell memory. Front Immunol 6:617.
36. Kim KS, Jekarl DW, Yoo J, Lee S, Kim M, Kim Y. 2021. Immune gene expression networks in sepsis: A network biology approach. PLoS One 16:e0247669.
37. Venet F, Filipe-Santos O, Lepape A, Malcus C, Poitevin-Later F, Grives A, Plantier N, Pasqual N, Monneret G. 2013. Decreased T-cell repertoire diversity in sepsis: A preliminary study. Crit Care Med 41:111–119.
38. Bagheri Y, Barati A, Aghebati-Maleki A, Aghebati-Maleki L, Yousefi M. 2021. Current progress in cancer immunotherapy based on natural killer cells. Cell Biol Int 45:2–17.
39. Miyairi S, Baldwin WM, Valujskikh A, Fairchild RL. 2021. Natural Killer Cells: Critical Effectors during Antibody-mediated Rejection of Solid Organ Allografts. Transplantation 284–290.
40. Duffy FJ, Olson GS, Gold ES, Jahn A, Aderem A, Aitchison JD, Rothchild AC, Diercks AH, Nemeth J. 2021. Use of a Contained Mycobacterium tuberculosis Mouse Infection Model to Predict Active Disease and Containment in Humans. J Infect Dis XX:1–9.
41. Moni MA, Lin PI, Quinn JMW, Eapen V. 2021. COVID-19 patient transcriptomic and

genomic profiling reveals comorbidity interactions with psychiatric disorders. Transl Psychiatry 2021 111 11:1–13.

42. Weydert C, van Heertum B, Dirix L, De Houwer S, De Wit F, Mast J, Husson SJ, Busschots K, König R, Gijsbers R, De Rijck J, Debyser Z. 2018. Y-box-binding protein 1 supports the early and late steps of HIV replication. PLoS One 13:e0200080.

43. Lyabin DN, Eliseeva IA, Ovchinnikov LP. 2014. YB-1 protein: functions and regulation. Wiley Interdiscip Rev RNA 5:95–110.

44. Goodarzi H, Liu X, Nguyen HCB, Zhang S, Fish L, Tavazoie SF. 2015. Endogenous tRNA-Derived Fragments Suppress Breast Cancer Progression via YBX1 Displacement. Cell 161:790–802.

45. Xu L, Li H, Wu L, Huang S. 2017. YBX1 promotes tumor growth by elevating glycolysis in human bladder cancer. Oncotarget 8:65946.

46. Diosa-Toro M, Kennedy DR, Chuo V, Popov VL, Pompon J, Garcia-Blanco MA. 2022. Y-box binding protein 1 interacts with dengue virus nucleocapsid and mediates viral assembly. bioRxiv 2022.01.25.477802.

47. Rogers PD, Pearson MM, Cleary JD, Sullivan DC, Chapman SW. 2002. Differential expression of genes encoding immunomodulatory proteins in response to amphotericin B in human mononuclear cells identified by cDNA microarray analysis. J Antimicrob Chemother 50:811–817.

48. Lin X, Yu S, Ren P, Sun X, Jin M. 2020. Human microRNA-30 inhibits influenza virus infection by suppressing the expression of SOCS1, SOCS3, and NEDD4. Cell Microbiol 22:e13150.

49. Vastrad B, Vastrad C, Tengli A. 2020. Bioinformatics analyses of significant genes, related pathways, and candidate diagnostic biomarkers and molecular targets in SARS-CoV-2/COVID-19. Gene Reports 21:100956.

50. Xu Q, Zhu N, Chen S, Zhao P, Ren H, Zhu S, Tang H, Zhu Y, Qi Z. 2017. E3 Ubiquitin Ligase Nedd4 Promotes Japanese Encephalitis Virus Replication by Suppressing Autophagy in Human Neuroblastoma Cells. Sci Reports 2017 71 7:1–12.

51. Pei G, Buijze H, Liu H, Moura-Alves P, Goosmann C, Brinkmann V, Kawabe H, Dorhoi A, Kaufmann SHE. 2017. The E3 ubiquitin ligase NEDD4 enhances killing of membrane-perturbing intracellular bacteria by promoting autophagy. Autophagy 13:2041–2055.

52. Nuro-Gyina PK, Tang N, Guo H, Yan C, Zeng Q, Waldschmidt TJ, Zhang J. 2021. HECT E3 Ubiquitin Ligase Nedd4 Is Required for Antifungal Innate Immunity. J Immunol 207:868–877.

53. Cui G, Wei P, Zhao Y, Guan Z, Yang L, Sun W, Wang S, Peng Q. 2014. Brucella infection inhibits macrophages apoptosis via Nedd4-dependent degradation of calpain2. Vet Microbiol 174:195–205.

54. Reinberg D, Horikoshi M, Roeder RG. 1987. Factors involved in specific transcription in mammalian RNA polymerase II. Functional analysis of initiation factors IIA and IID and identification of a new factor operating at sequences downstream of the initiation site. J Biol Chem 262:3322–3330.

55. Zhou M, Kashanchi F, Jiang H, Ge H, Brady JN. 2000. Phosphorylation of the RAP74 Subunit of TFIIF Correlates with Tat-Activated Transcription of the HIV-1 Long Terminal Repeat. Virology 268:452–460.

56. Miao M, Yu F, Wang D, Tong Y, Yang L, Xu J, Qiu Y, Zhou X, Zhao X. 2019. Proteomics Profiling of Host Cell Response via Protein Expression and Phosphorylation

upon Dengue Virus Infection. Virol Sin 34:549–562.

57. Pierzynowska K, Gaffke L, Węgrzyn G. 2020. Transcriptomic analyses suggest that mucopolysaccharidosis patients may be less susceptible to COVID-19. FEBS Lett 594:3363–3370.

58. Wu Y, Xia P, Zheng C. 2015. Bioinformatics analysis of transcription profiling of sepsis: http://dx.doi.org/101177/1721727X15590946 13:82–90.

59. Corsini L, Bonnal S, Basquin J, Hothorn M, Scheffzek K, Valcárcel J, Sattler M. 2007. U2AF-homology motif interactions are required for alternative splicing regulation by SPF45. Nat Struct Mol Biol 2007 147 14:620–629.

60. Perry WL, Shepard RL, Sampath J, Yaden B, Chin WW, Iversen PW, Jin S, Lesoon A, O'Brien KA, Peek VL, Rolfe M, Shyjan A, Tighe M, Williamson M, Krishnan V, Moore RE, Dantzig AH. 2005. Human Splicing Factor SPF45 (RBM17) Confers Broad Multidrug Resistance to Anticancer Drugs When Overexpressed— a Phenotype Partially Reversed By Selective Estrogen Receptor Modulators. Cancer Res 65:6593–6600.

61. König R, Zhou Y, Elleder D, Diamond TL, Bonamy GMC, Irelan JT, Chiang C yuan, Tu BP, De Jesus PD, Lilley CE, Seidel S, Opaluch AM, Caldwell JS, Weitzman MD, Kuhen KL, Bandyopadhyay S, Ideker T, Orth AP, Miraglia LJ, Bushman FD, Young JA, Chanda SK. 2008. Global Analysis of Host-Pathogen Interactions that Regulate Early-Stage HIV-1 Replication. Cell 135:49–60.

62. Chen J, Xu Q, Zhang Y, Zhang H. 2020. RNA Profiling Analysis of the Serum Exosomes Derived from Patients with Chronic Hepatitis and Acute-on-chronic Liver Failure Caused By HBV. Sci Reports 2020 101 10:1–9.

63. Zhang B, Zelhof AC. 2002. Amphiphysins: Raising the BAR for Synaptic Vesicle Recycling and Membrane Dynamics. Traffic 3:452–460.

64. Wigge P, Köhler K, Vallis Y, Doyle CA, Owen D, Hunt SP, McMahon HT. 1997. Amphiphysin heterodimers: Potential role in clathrin-mediated endocytosis. Mol Biol Cell 8:2003–2015.

65. Yamada H, Ohashi E, Abe T, Kusumi N, Li SA, Yoshida Y, Watanabe M, Tomizawa K, Kashiwakura Y, Kumon H, Matsui H, Takei K. 2007. Amphiphysin 1 Is Important for Actin Polymerization during Phagocytosis. Mol Biol Cell 18:4669.

66. De Jesús-Cortés HJ, Nogueras-Ortiz CJ, Gearing M, Arnold SE, Vega IE. 2012. Amphiphysin-1 protein level changes associated with tau-mediated neurodegeneration. Neuroreport 23:942.

67. Chen Y, Liu J, Li L, Xia H, Lin Z, Zhong T. 2018. AMPH-1 is critical for breast cancer progression. J Cancer 9:2175.

68. Yang H, Wan Z, Huang C, Yin H, Song D. 2019. AMPH-1 is a tumor suppressor of lung cancer by inhibiting Ras-Raf-MEK-ERK signal pathway. Lasers Med Sci 34:473–478.

69. Zhang H, Liu Y, Xu K, Mao K, Han W, Xu F, Wan W, Sun Y. 2019. AMPH-1 As A Critical Tumor Suppressor That Inhibits Osteosarcoma Progression. Cancer Manag Res 11:9913.

70. Kumar GSS, Venugopal AK, Mahadevan A, Renuse S, Harsha HC, Sahasrabuddhe NA, Pawar H, Sharma R, Kumar P, Rajagopalan S, Waddell K, Ramachandra YL, Satishchandra P, Chaerkady R, Prasad TSK, Shankar K, Pandey A. 2012. Quantitative proteomics for identifying biomarkers for tuberculous meningitis. Clin Proteomics 9:1–13.

71. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. 2012. The sva package for removing batch effects and other unwanted variation in high-throughput experiments.

Bioinformatics 28:882–883.

72. Dunning M, Lynch A, Eldridge M. 2015. illuminaHumanv4.db: Illumina HumanHT12v4 annotation data (chip illuminaHumanv4). R Packag version 1260. https://bioconductor.org/packages/release/data/annotation/html/illuminaHumanv4.db.html. Retrieved 28 March 2022.

73. R Core Team. 2020. R version 4.0. 3: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.r-project.org/.

74. Chen S-H, Yu B-Y, Kuo W-Y, Lin Y-B, Su S-Y, Lu I-H, Lin C-Y. 2020. mySORT: A Web Framework by using Deconvolution Approach to Estimating Immune Cell Composition from Complex Tissues https://doi.org/10.20944/PREPRINTS202011.0385.V1.

75. Chen SH, Kuo WY, Su SY, Chung WC, Ho JM, Lu HHS, Lin CY. 2018. A gene profiling deconvolution approach to estimating immune cell composition from complex tissues. BMC Bioinformatics 19:15–23.

76. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 43:e47–e47.

77. Langfelder P, Horvath S. 2008. WGCNA: An R package for weighted correlation network analysis. BMC Bioinformatics 9:1–13.

78. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Res 13:2498–2504.

79. von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, Jouffre N, Huynen MA, Bork P. 2005. STRING: known and predicted protein–protein associations, integrated and transferred across organisms. Nucleic Acids Res 33:D433–D437.

80. Hechenbichler K, Schliep K. 2004. Weighted k-Nearest-Neighbor Techniques and Ordinal Classification. Collab Res Cent 386.

81. Das P, Roychowdhury A, Das S, Roychoudhury S, Tripathy S. 2020. sigFeature: Novel Significant Feature Selection Method for Classification of Gene Expression Data Using Support Vector Machine and t Statistic. Front Genet 11:247.

82. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Müller M. 2011. pROC: An open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 12:1–8.

83. Kanehisa M, Goto S. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 28:27–30.

84. Yu G, Wang LG, Han Y, He QY. 2012. ClusterProfiler: An R package for comparing biological themes among gene clusters. Omi A J Integr Biol 16:284–287.

85. Carlson M. 2020. org.Hs.ed.db: Genome wide annotation for Human. R package version 3.11.4.

86. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Szoecs E, Wagner H. 2020. vegan: Community Ecology Package. R package verion 2.5-7.

87. Gu Z, Eils R, Schlesner M. 2016. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics 32:2847–2849.

88. Chen H, Boutros PC. 2011. VennDiagram: A package for the generation of highly-

customizable Venn and Euler diagrams in R. BMC Bioinformatics 12:1–7.
89. Yu G. 2021. enrichplot: Visualization of Functional Enrichment Result. R package version 1.12.3.

Supplementary materials

Table S1: Metadata
Table S2: Cell deconvolution results
Table S3: DEG analysis results
Table S4: Heatmap clustering
Table S5: Functional enrichment analysis
Table S6: WGCNA modules
Table S7: PPI network analysis
Table S8: Machine learning model from hub genes
Table S9: Unknown prediction

Table 1: Significant *p*-value and area under the curves of hub genes.

Illumina ID	Symbol	Module	P-value	AUC
Bacterial model				
ILMN_2096372	*ALDH1A1*	turquoise	<0.001	0.755
ILMN_1765796	*ENO2*	brown	<0.001	0.71
ILMN_1715169	*HLA-DRB1*	brown	<0.001	0.79
ILMN_1708779	*GNLY*	blue	0.001	0.735
ILMN_1701237	*SH2D1B*	blue	0.007	0.718
ILMN_2098126	*CCL5*	brown	0.019	0.708
ILMN_1661266	*HLA-DQB1*	blue	0.02	0.703
ILMN_2363426	*MAX*	brown	0.022	0.704
Fungal model				
ILMN_1661266	*HLA-DQB1*	blue	<0.001	0.961
ILMN_2374036	*CTSL*	brown	0.003	0.719
ILMN_2124769	*YBX1*	pink	0.004	0.876
ILMN_1809583	*CREBBP*	turquoise	0.004	0.738
ILMN_2096372	*ALDH1A1*	turquoise	0.005	0.8
ILMN_1771203	*SMAD2*	blue	0.007	0.921
ILMN_1796409	*C1QB*	brown	0.008	0.748
ILMN_1812995	*CTSL*	brown	0.008	0.747
ILMN_1781906	*RBM17*	pink	0.012	0.832
ILMN_2370882	*ACSL5*	blue	0.019	0.862
ILMN_3285611	*RPS9*	red	0.03	0.833
ILMN_1706546	*MAX*	brown	0.033	0.707
ILMN_1695420	*CLTA*	black	0.04	0.813
Co-infection model				
ILMN_2374036	*CTSL*	brown	<0.001	0.908
ILMN_1812995	*CTSL*	brown	<0.001	0.869
ILMN_2184184	*ANXA1*	turquoise	<0.001	0.805
ILMN_1715169	*HLA-DRB1*	brown	<0.001	0.815
ILMN_1706546	*MAX*	brown	<0.001	0.853
ILMN_1745798	*GTF2F2*	turquoise	0.002	0.806
ILMN_1654566	*HSPA1L*	yellow	0.002	0.791
ILMN_1703140	*NEDD4*	turquoise	0.002	0.772
ILMN_1701237	*SH2D1B*	blue	0.002	0.808
ILMN_1753468	*CD63*	turquoise	0.003	0.762
ILMN_1809583	*CREBBP*	turquoise	0.003	0.786
ILMN_1797341	*ARID1A*	yellow	0.003	0.776
ILMN_2363426	*MAX*	brown	0.003	0.827
ILMN_1708779	*GNLY*	blue	0.004	0.777
ILMN_1796409	*C1QB*	brown	0.005	0.788
ILMN_1765796	*ENO2*	brown	0.005	0.751
ILMN_1724718	*NCK2*	brown	0.007	0.765
ILMN_1693341	*SNRPN*	brown	0.008	0.748
ILMN_1704236	*MAX*	brown	0.009	0.775
ILMN_2096372	*ALDH1A1*	turquoise	0.01	0.704
ILMN_1673357	*SLA2*	brown	0.013	0.762
ILMN_2098126	*CCL5*	brown	0.013	0.79
ILMN_1661945	*SLIRP*	red	0.014	0.731
ILMN_1733324	*ITGB3*	brown	0.018	0.725
ILMN_1797074	*EMG1*	blue	0.019	0.731
ILMN_1785902	*C1QC*	brown	0.025	0.748
ILMN_1778143	*GRAP2*	brown	0.03	0.721
ILMN_1685834	*AMPH*	blue	0.031	0.722
ILMN_1773352	*CCL5*	brown	0.035	0.716
ILMN_1781906	*RBM17*	pink	0.037	0.713
ILMN_1748591	*ODC1*	pink	0.04	0.705
ILMN_1783621	*CMPK2*	blue	0.042	0.711

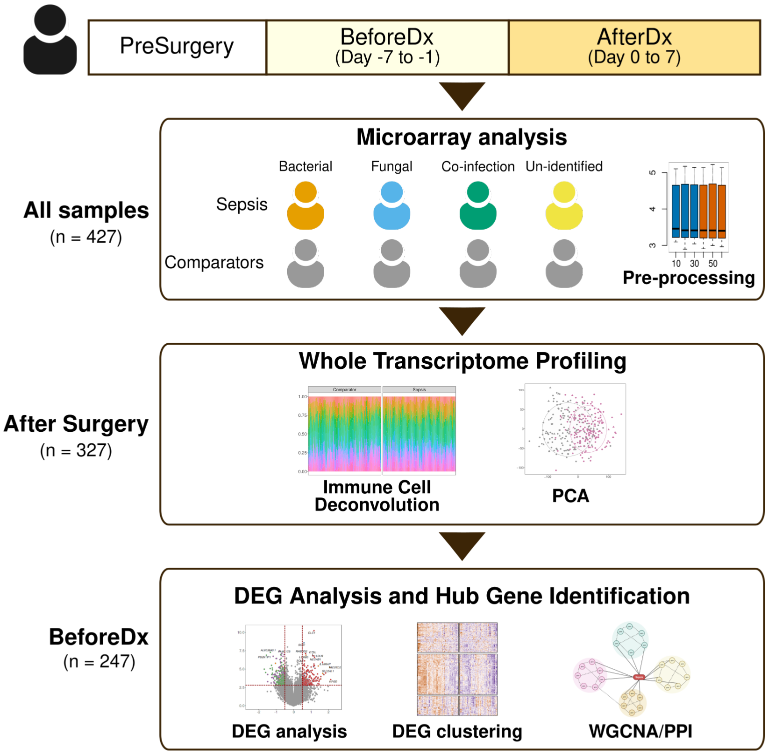


Figure 1: An overview of the study workflow Blood samples were taken from surgery patients at PreSurgery, BeforeDx, and AfterDx. A total of 427 blood samples were diagnosed as bacterial/fungal/co-infection/un-identified sepsis or their comparator. Microarray technology was used to examine gene expression patterns associated with surgery patients. All samples were used for quality assessment. The differences in transcriptomic profiles between sepsis and comparators were observed from post-operative samples. To identify the pathogen-specific gene(s)/pathway(s), we performed DEG analysis, HC clustering, WGCNA, and PPI network analysis.

PreSurgery, Pre-surgery samples; BeforeDx, samples were collected after surgery and before patients were diagnosed with sepsis; AfterDx, samples were collected after surgery and after patients were diagnosed with sepsis; PCA, principal component analysis; DEG, differentially expressed gene; HC, hierarchical clustering; WGCNA, Weighted Gene Co-expression Network Analysis; PPI, protein-protein interaction

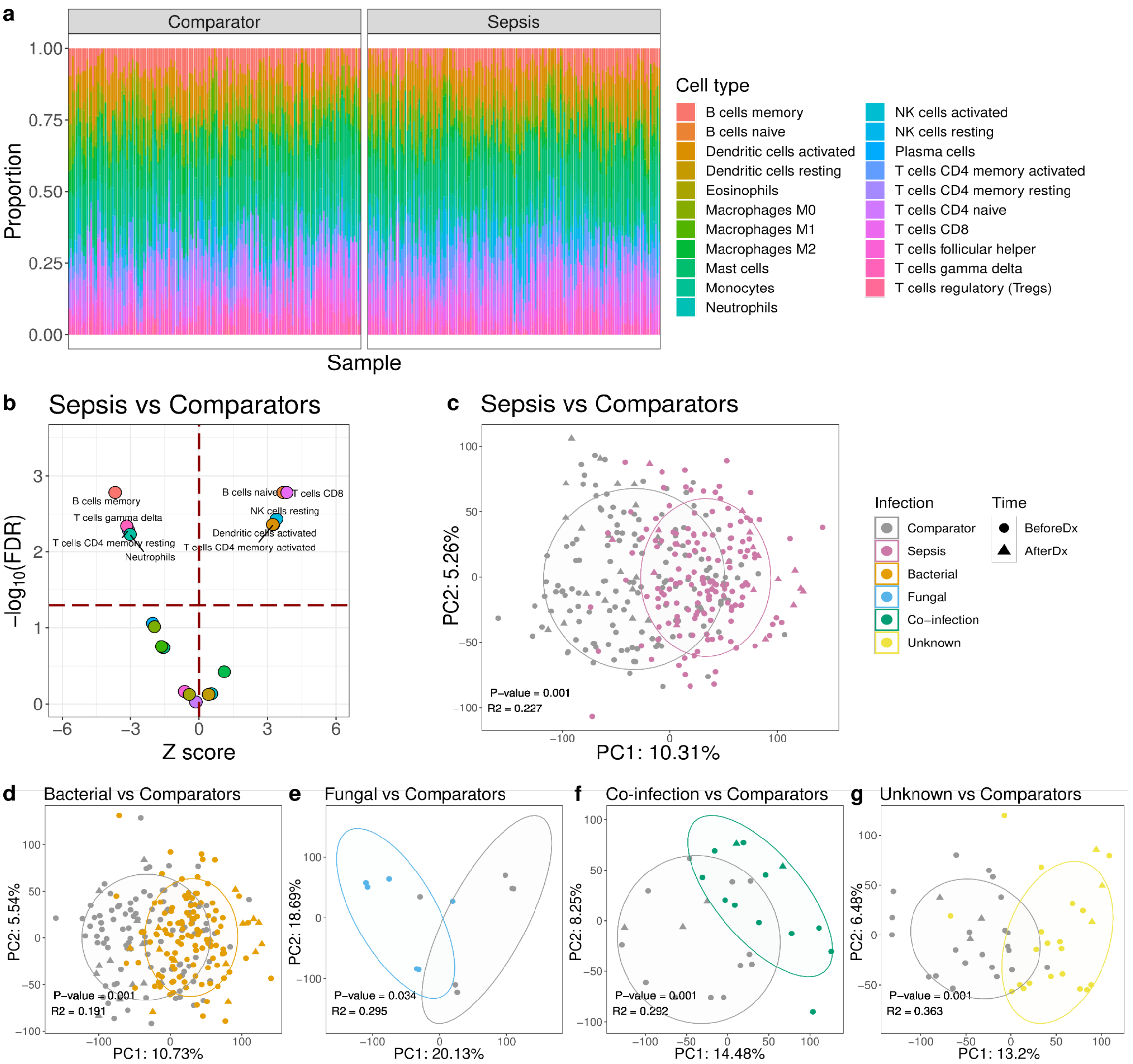


Figure 2: Whole transcriptome profiles (a) Bar plots of the proportion of immune cell types in sepsis patients and comparators. (b) Volcano Plot of the differentially infiltrated immune cells between sepsis and comparators. Colors indicate different cell types. (c-i) PCA plots of all transcripts from patients after surgery. (c) PCA plot of sepsis patients vs. comparators. (d) PCA plot of bacterial sepsis vs. comparators. (e) PCA plot of fungal sepsis vs. comparators. (f) PCA plot of co-infection sepsis vs. comparators. (g) PCA plot of un-identified pathogen sepsis vs. comparators.

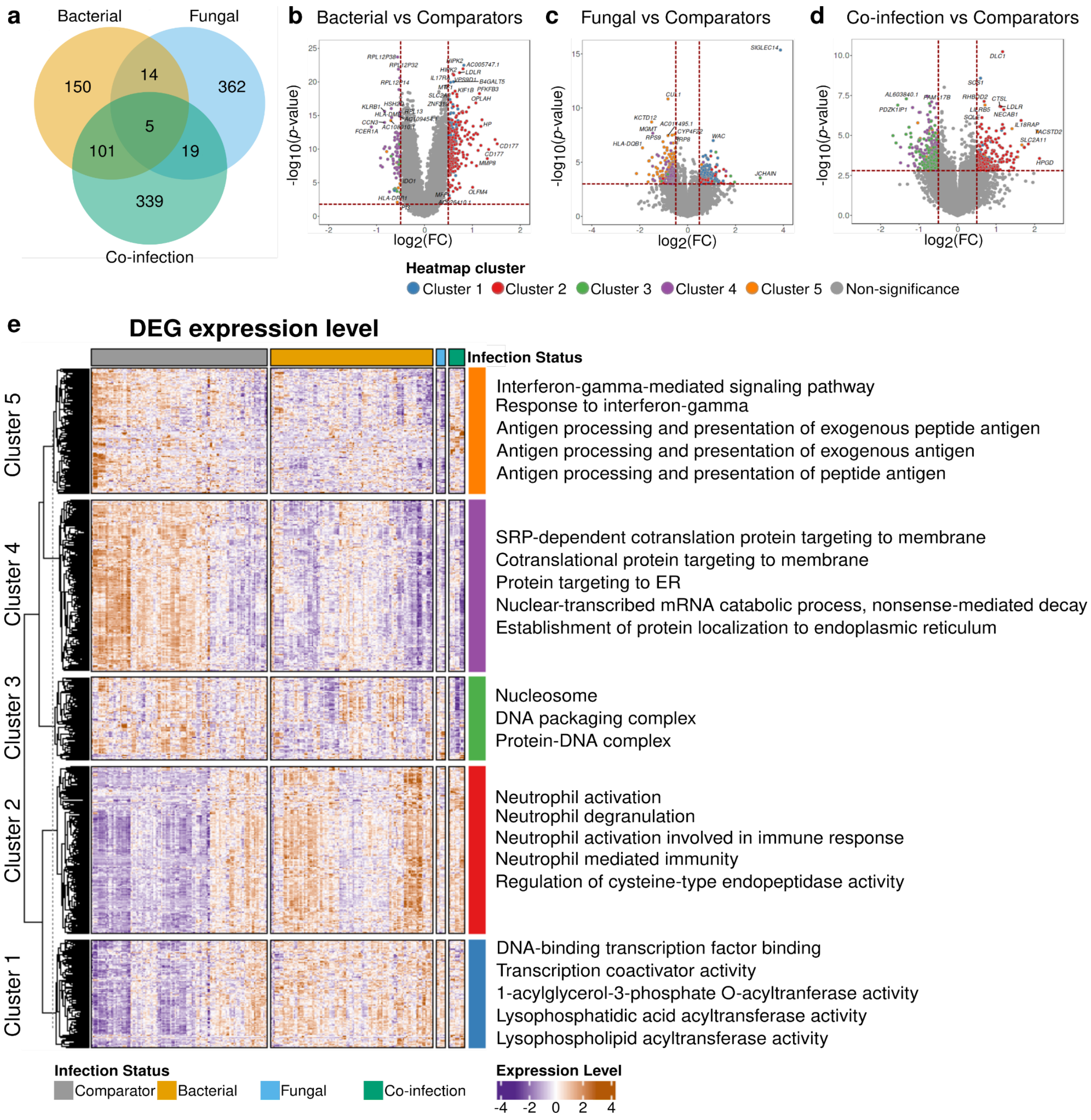
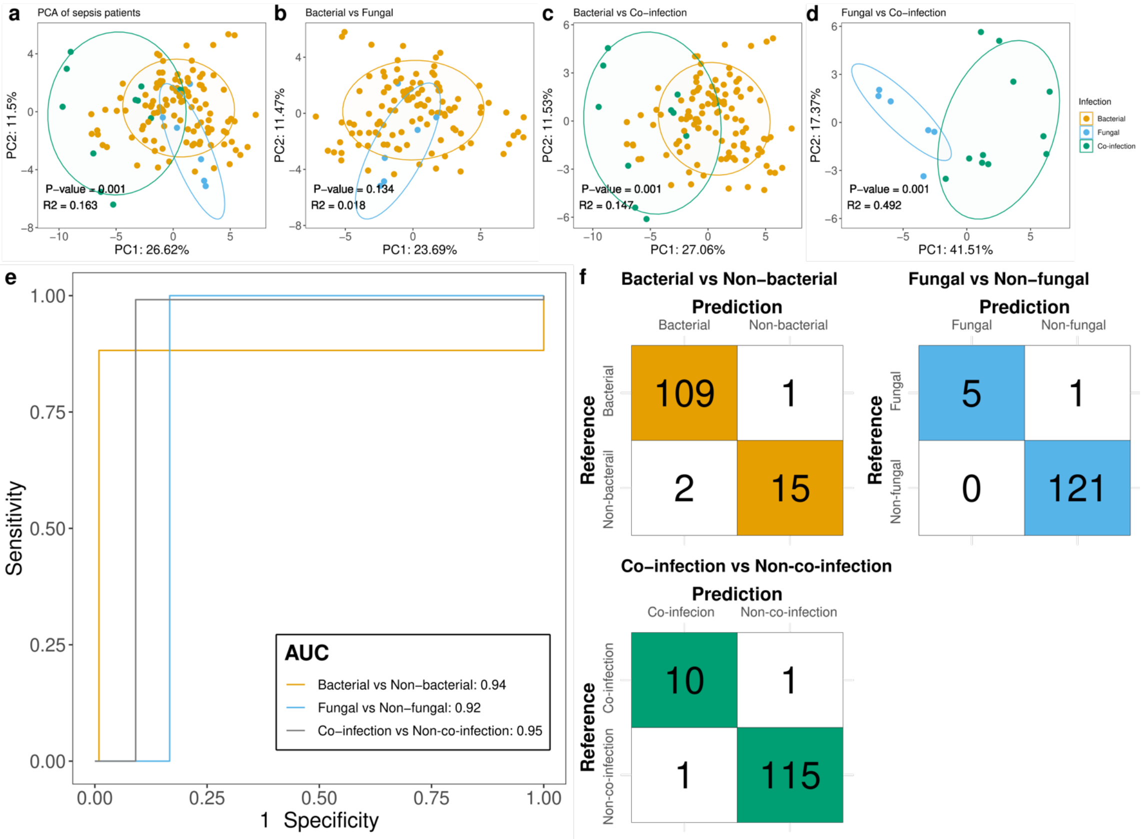


Figure 3: Differentially expressed gene analyses of BeforeDx samples (a) Heatmap of 990 significant DEGs of 247 BeforeDx samples classified into five clusters and their top five biological processes. (b-e) DEG analysis between different pathogen-induced sepsis vs. their comparators. Volcano plots showing DEGs between (b) bacterial sepsis vs. comparators, (c) fungal sepsis vs. comparators, and (d) co-infection sepsis vs. comparators. (e) The Venn diagram shows the overlapping DEGs among those three comparisons.



Bacterial vs Non-bacterial

Reference \ Prediction	Bacterial	Non-bacterial
Bacterial	109	1
Non-bacterail	2	15

Fungal vs Non-fungal

Reference \ Prediction	Fungal	Non-fungal
Fungal	5	1
Non-fungal	0	121

Co-infection vs Non-co-infection

Reference \ Prediction	Co-infecion	Non-co-infection
Co-infection	10	1
Non-co-infection	1	115

Figure 4: Hub gene performance evaluation (a-b) PCA plots of ‘BeforeDx’ sepsis patients based on 42 hub genes. (a) PCA plot of bacterial vs. fungal vs. co-infection sepsis. (b) PCA plot of bacterial vs. fungal sepsis (c) PCA plot of bacterial vs. co-infection sepsis (d) PCA plot of fungal vs. co-infection sepsis. (e) Receiver operating characteristic (ROC) curves and (f) confusion matrices of the binary classification model performance for three models, including bacterial vs. non-bacterial (orange), fungal vs. non-fungal (blue), and co-infection vs. non-infection (green) models in ‘BeforeDx’ samples.

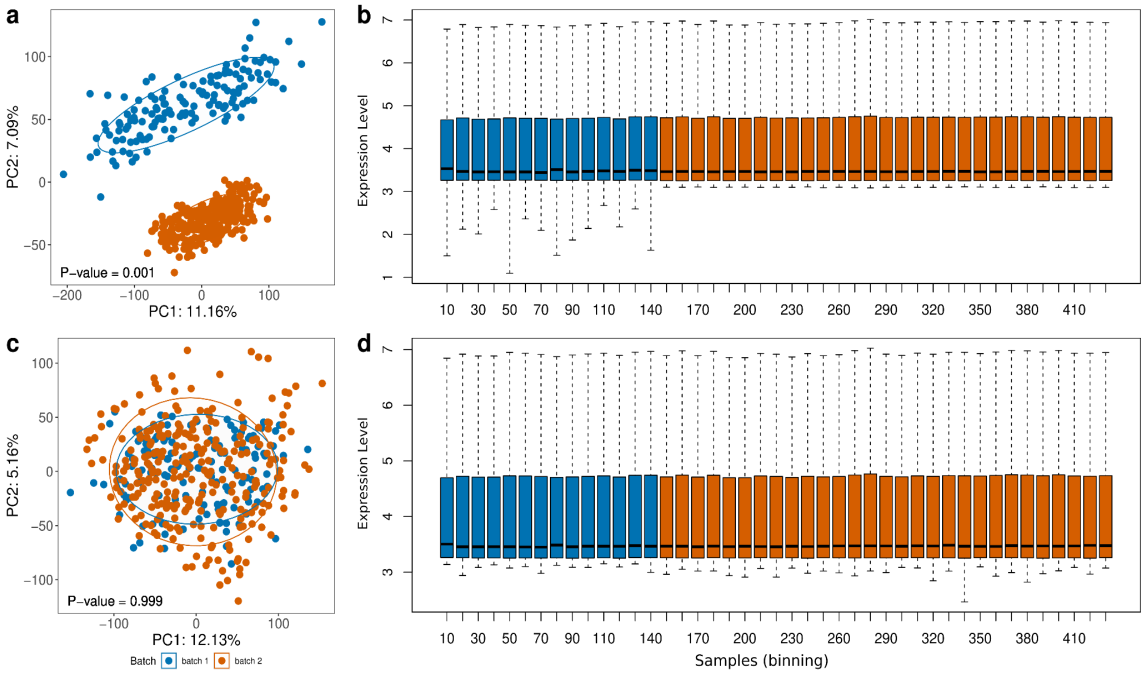


Figure S1: Microarray data before and after batch effect correction (a) PCA and (b) boxplots of microarray data before batch effect removal. (c) PCA and (d) boxplots of microarray data after batch effect removal.

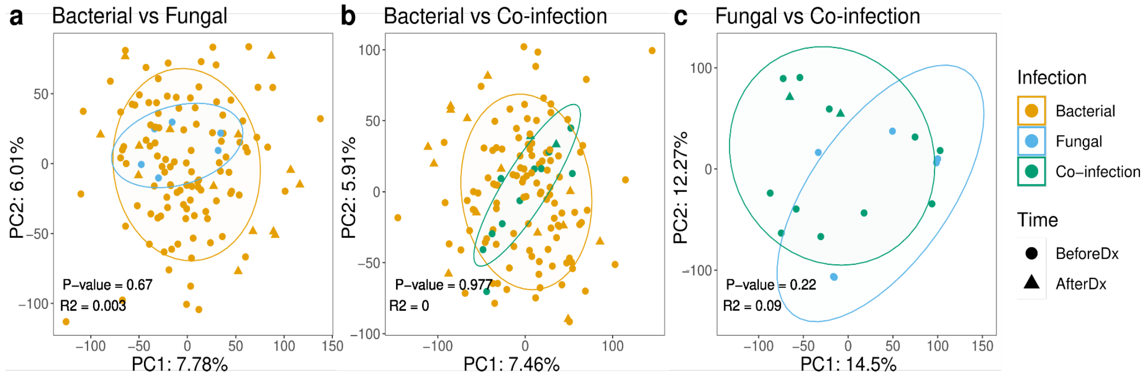


Figure S2: PCA plots of all transcripts from sepsis patients (a) PCA plot of bacterial vs. fungal sepsis. (b) PCA plot of bacterial vs. co-infection sepsis. (c) PCA plot of fungal vs. co-infection sepsis.

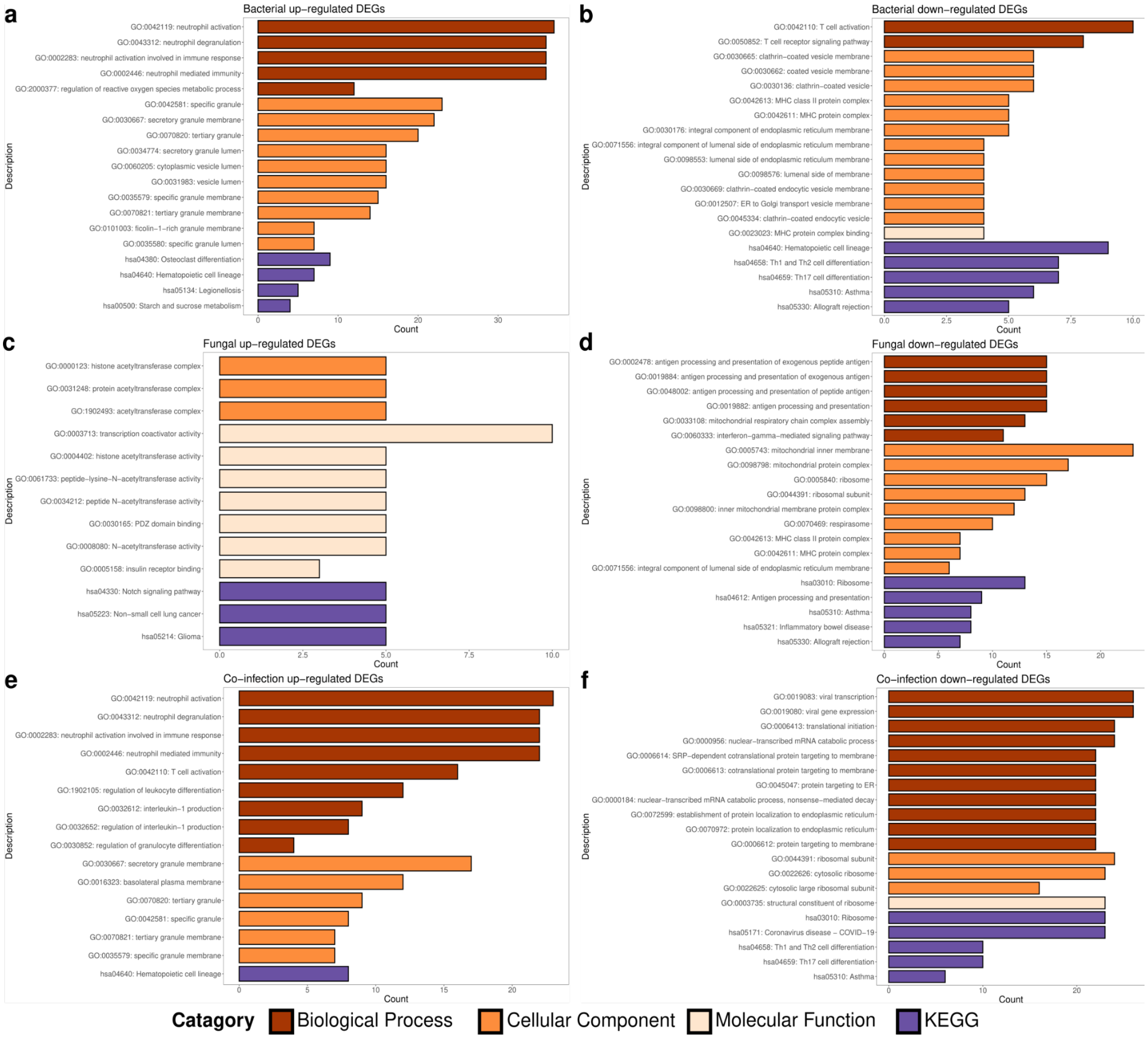


Figure S3: Functional annotation (a-f) Bar plots show the top 15 GO terms and top 5 KEGG pathways with the number of involved DEGs. (a) The top 20 enriched terms from up-regulated bacterial DEGs. (b) The top 20 enriched terms from down-regulated bacterial DEGs. (c) The top 20 enriched terms from up-regulated fungal DEGs. (d) The top 20 enriched terms from down-regulated fungal DEGs. (e) The top 20 enriched terms from up-regulated co-infection DEGs. (f) The top 20 enriched terms from down-regulated co-infection DEGs.

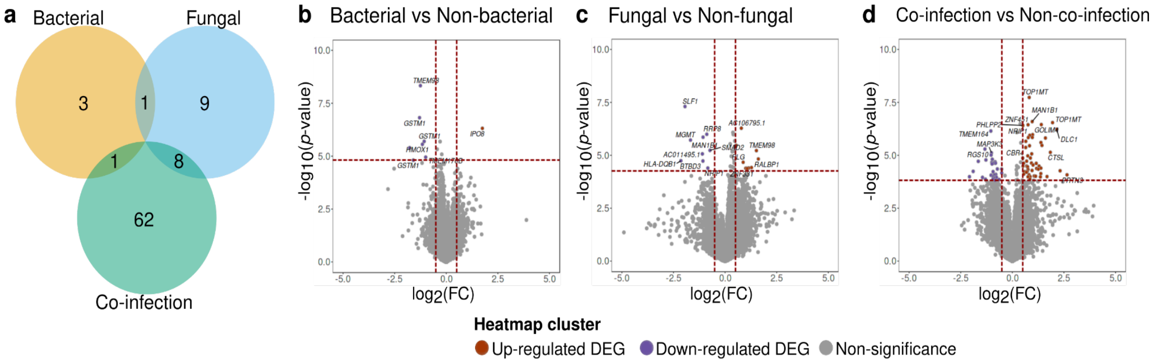


Figure S4: PCA plots of all transcripts from sepsis patients (a-d) DEG analyses between sepsis induced by different pathogens. (a) The Venn diagram shows the overlapping DEGs among three comparisons. Volcano plots showing DEGs between (b) bacterial vs. non-bacterial induced sepsis, (c) fungal vs. non-fungal induced sepsis, and (d) co-infection vs. non-co-infection induced sepsis.

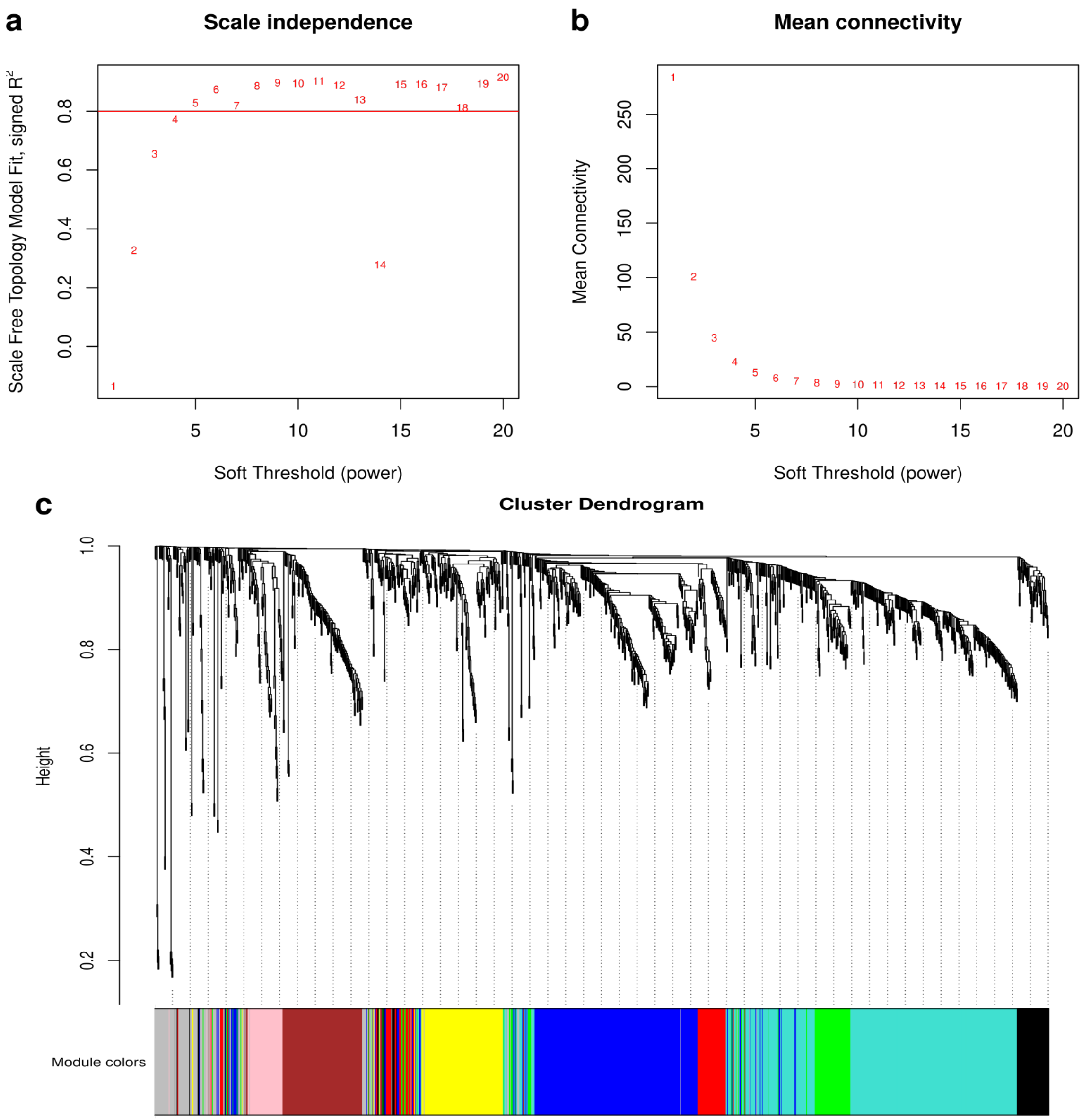


Figure S5: WGCNA-based identification of co-expression modules for sepsis patients (a) Analysis of the scale-free indices for various soft-threshold powers (β). (B) Mean connectivity analysis of various soft-thresholding powers. (c) Dendrogram of all differentially expressed genes clustering based on dissimilarity measurement (1-TOM). The branches and color bands represent the assigned module.

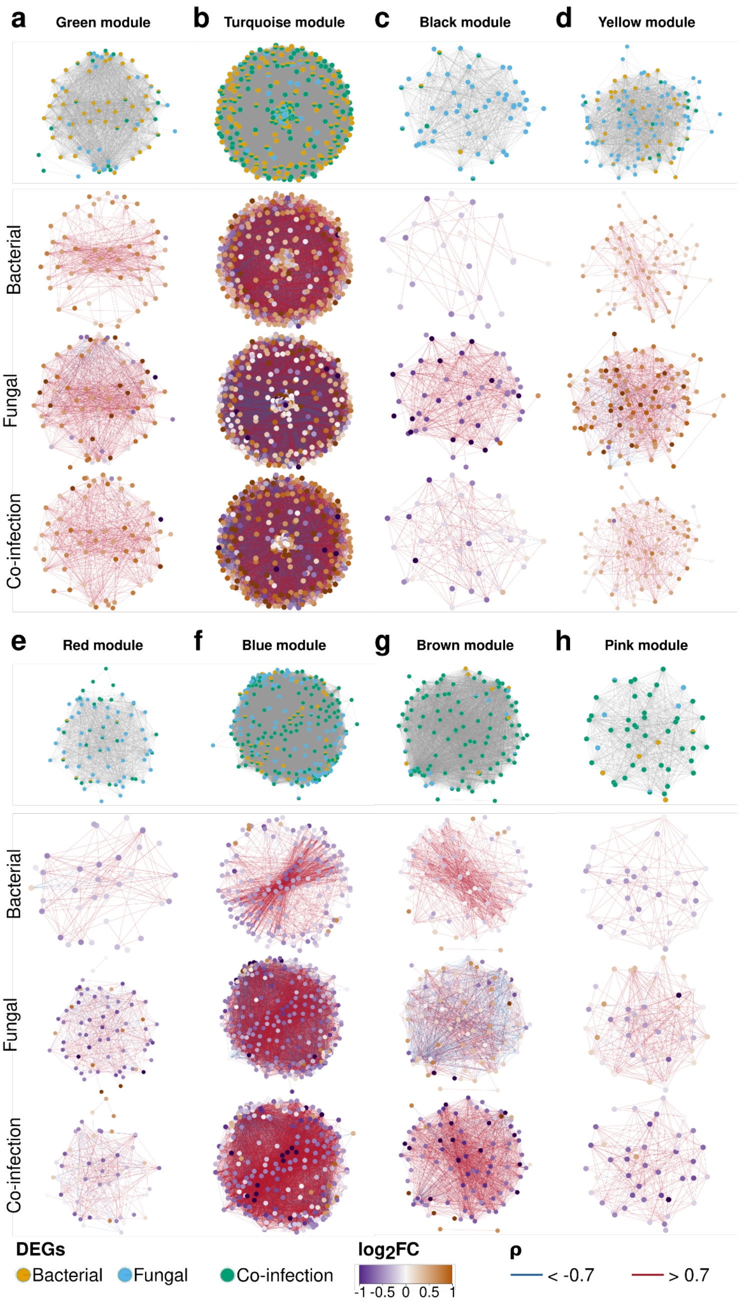


Figure S6: Eight gene modules from WGCNA Gene co-expression modules for (a) green, (b) turquoise, (c) black, (d) yellow, (f) red, (g) blue, (g) brown, and (h) pink. Node colors in the first row of each module indicate belonging DEGs to which pathogen-induced sepsis. For bacterial, fungal, and co-infection rows, visualize the gene expression changes and gene correlation between pathogen-induced sepsis compared to comparators. Nodes indicate the log$_2$FC of genes in each module. Edges indicate the |Spearman's correlation| ≥ 0.7 between genes in each module.

Manuscript III

AMERICAN SOCIETY FOR MICROBIOLOGY
mBio®

RESEARCH ARTICLE
Applied and Environmental Science

Effects of Agricultural Fungicide Use on *Aspergillus fumigatus* Abundance, Antifungal Susceptibility, and Population Structure

Amelia E. Barber,[a,b] Jennifer Riedel,[c] Tongta Sae-Ong,[a] Kang Kang,[a] Werner Brabetz,[d] Gianni Panagiotou,[a,e] Holger B. Deising,[c] Oliver Kurzai[a,b]

[a]Leibniz Institute of Natural Product Research and Infection Biology–Hans Knöll Institute, Jena, Germany
[b]Institute for Hygiene and Microbiology, University of Würzburg, Würzburg, Germany
[c]Institute for Agriculture and Nutritional Sciences, Martin Luther University Halle-Wittenberg, Halle (Saale), Germany
[d]Biotype GmbH, Dresden, Germany
[e]Department of Medicine, University of Hong Kong and State Key Laboratory of Pharmaceutical Biotechnology, University of Hong Kong, Hong Kong, China

Amelia E. Barber and Jennifer Riedel contributed equally. Order was determined by alphabetical order.

Overview

In manuscript III, we aimed to understand the effects of agricultural fungicide on *A. fumigatus* genomes, abundance, and antifungal susceptibility. Therefore, we re-assembled *A. fumigatus* genomes collected from organic and agricultural farms in Germany using a reference-based genome assembly method. The results showed that fungicide use in agriculture had no effect on genetic changes in *A. fumigatus* but reduced the population of susceptible *A. fumigatus* strains.

FORM 1

Manuscript No. 3

Manuscript title: Effects of Agricultural Fungicide Use on *Aspergillus fumigatus* Abundance, Antifungal Susceptibility, and Population Structure

Authors: Amelia E. Barber*, Jennifer Rieder*, **Tongta Sae-Ong**, Kang Kang, Warner Brabetz, Gianni Panagiotou, Holger B. Deising, Oliver Kurzai

Bibliographic information:

Barber A.E.*, Riedel J.*, **Sae-Ong T.**, Kang K., Brabetz W., Panagiotou G., Deising H.B., Kurzai O. 2020. Effects of agricultural fungicide use on *Aspergillus fumigatus* abundance, antifungal susceptibility, and population structure. mBio 11:e02213-20. https://doi.org/10.1128/mBio.02213-20

The candidate is (Please tick the appropriate box.)

☐ First author, ☐ Co-first author, ☐ Corresponding author, ☒ Co-author.

Status: published

Authors' contributions (in %) to the given categories of the publication

Author	**Conceptual**	**Data analysis**	**Experimental**	**Writing the manuscript**	**Provision of material**
Barber, A.E.*	30%	40%	45%	40%	
Riedel, J.*	30%	40%	45%	40%	
Sae-Ong, T		20%		5%	
Panagiotou, G.	10%			5%	20%
Deising, H.B.	15%			5%	40%
Kurzai, O.	15%			5%	40%
Others			10%		
Total:	100%	100%	100%	100%	100%

*Authors contributed equally

Signature candidate

Signature supervisor (member of the Faculty)

AMERICAN SOCIETY FOR MICROBIOLOGY mBio®

RESEARCH ARTICLE
Applied and Environmental Science

Effects of Agricultural Fungicide Use on *Aspergillus fumigatus* Abundance, Antifungal Susceptibility, and Population Structure

Amelia E. Barber,[a,b] Jennifer Riedel,[c] Tongta Sae-Ong,[a] Kang Kang,[a] Werner Brabetz,[d] Gianni Panagiotou,[a,e] Holger B. Deising,[c] Oliver Kurzai[a,b]

[a]Leibniz Institute of Natural Product Research and Infection Biology–Hans Knöll Institute, Jena, Germany
[b]Institute for Hygiene and Microbiology, University of Würzburg, Würzburg, Germany
[c]Institute for Agriculture and Nutritional Sciences, Martin Luther University Halle-Wittenberg, Halle (Saale), Germany
[d]Biotype GmbH, Dresden, Germany
[e]Department of Medicine, University of Hong Kong and State Key Laboratory of Pharmaceutical Biotechnology, University of Hong Kong, Hong Kong, China

Amelia E. Barber and Jennifer Riedel contributed equally. Order was determined by alphabetical order.

ABSTRACT Antibiotic resistance is an increasing threat to human health. In the case of *Aspergillus fumigatus*, which is both an environmental saprobe and an opportunistic human fungal pathogen, resistance is suggested to arise from fungicide use in agriculture, as the azoles used for plant protection share the same molecular target as the frontline antifungals used clinically. However, limiting azole fungicide use on crop fields to preserve their activity for clinical use could threaten the global food supply via a reduction in yield. In this study, we clarify the link between azole fungicide use on crop fields and resistance in a prototypical human pathogen through systematic soil sampling on farms in Germany and surveying fields before and after fungicide application. We observed a reduction in the abundance of *A. fumigatus* on fields following fungicide treatment in 2017, a finding that was not observed on an organic control field with only natural plant protection agents applied. However, this finding was less pronounced during our 2018 sampling, indicating that the impact of fungicides on *A. fumigatus* population size is variable and influenced by additional factors. The overall resistance frequency among agricultural isolates is low, with only 1 to 3% of isolates from 2016 to 2018 displaying resistance to medical azoles. Isolates collected after the growing season and azole exposure show a subtle but consistent decrease in susceptibility to medical and agricultural azoles. Whole-genome sequencing indicates that, despite the alterations in antifungal susceptibility, fungicide application does not significantly affect the population structure and genetic diversity of *A. fumigatus* in fields. Given the low observed resistance rate among agricultural isolates as well the lack of genomic impact following azole application, we do not find evidence that azole use on crops is significantly driving resistance in *A. fumigatus* in this context.

IMPORTANCE Antibiotic resistance is an increasing threat to human health. In the case of *Aspergillus fumigatus*, which is an environmental fungus that also causes life-threatening infections in humans, antimicrobial resistance is suggested to arise from fungicide use in agriculture, as the chemicals used for plant protection are almost identical to the antifungals used clinically. However, removing azole fungicides from crop fields threatens the global food supply via a reduction in yield. In this study, we survey crop fields before and after fungicide application. We find a low overall azole resistance rate among agricultural isolates, as well as a lack of genomic and population impact following fungicide application, leading us to conclude azole use on crops does not significantly contribute to resistance in *A. fumigatus*.

Citation Barber AE, Riedel J, Sae-Ong T, Kang K, Brabetz W, Panagiotou G, Deising HB, Kurzai O. 2020. Effects of agricultural fungicide use on *Aspergillus fumigatus* abundance, antifungal susceptibility, and population structure. mBio 11:e02213-20. https://doi.org/10.1128/mBio.02213-20.

Editor Michael Lorenz, University of Texas Health Science Center

Copyright © 2020 Barber et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International license.

Address correspondence to Amelia E. Barber, amelia.barber@leibniz-hki.de, or Oliver Kurzai, okurzai@hygiene.uni-wuerzburg.de.

Received 10 August 2020
Accepted 22 October 2020
Published 24 November 2020

KEYWORDS *Aspergillus fumigatus*, *Aspergillus*, antifungal resistance, azole, fungicide, population genomics, antibiotic resistance

Aspergillus fumigatus is a globally distributed fungus responsible for an estimated 300,000 cases of invasive disease and more than 10 million cases of chronic and allergic disease globally each year (1). Humans inhale the infectious particles, or spores, on a daily basis but are actually an accidental host for the fungus, whose primary niche is soil and decaying vegetation. Management and prophylaxis against aspergillosis relies largely on the azole class of antifungals, with voriconazole and isavuconazole recommended as the first-line therapy (2). Unfortunately, clinical resistance to the azoles in *A. fumigatus* is an increasing problem, with some medical centers reporting rates as high as 30% in specific patient populations and similarly high rates for environmentally isolated *A. fumigatus* (3, 4). Regrettably, the mortality for resistant infections is upwards of 90% in some patient populations (5–7). While resistance can evolve during patient therapy (8, 9), the emergence of resistance in *A. fumigatus* has mainly been linked to the use of azoles in agriculture, as structurally similar and mechanistically indistinguishable compounds are heavily used for plant protection (10). This resistance has been described as collateral damage, as *A. fumigatus* is not a plant pathogen that is being directly targeted by fungicide treatments (11). The triazoles were first released for agriculture in 1973, well before they were first introduced to human medicine in the early 1990s, and are currently the most widely utilized antifungal compound group in agriculture due to their systemic distribution in treated plants, high efficiency, and broad spectrum of target pathogens (12, 13). Crops, particularly cereals and fruits, are sprayed multiple times each growing season at a recommended dose of 100 g/hectare to control powdery mildew, rust, septoria leaf blotch, and other phytopathogenic fungi (14). Currently, there are 32 azoles commercially available for plant protection (15) but only five in regular use in human medicine (16).

The most common azole resistance mechanism in *A. fumigatus* occurs via mutations in the target protein of azole fungicides, sterol 14α-demethylase (CYP51A, also called ERG11), a key enzyme of the ergosterol biosynthesis pathway. In *A. fumigatus*, the dominant resistance mechanism among both environmental and clinical isolates is a 34-bp tandem repeat (TR_{34}) in the *cyp51a* promoter coupled with a leucine-to-histidine substitution (L98H) in the amino acid coding sequence, the net effect of which is an increase in gene expression as well as an alteration in both the stability of the target enzyme and the interaction between the protein heme cofactor of *cyp51a* and the azole ligand (17–19). Additional mutations that have been identified to confer azole resistance in *A. fumigatus* include other variations of the tandem repeat, such as TR_{46}/Y121F/T289A and TR_{53}, as well as other point mutations in the *cyp51a* coding sequence (20, 21).

Disease-causing fungi are responsible for roughly 20% of crop yield loss, with a further 10% loss postharvest (16), so the use of azoles is critical for securing the food supply. However, this must be balanced against the need to preserve the activity of the azoles for clinical use, and, as such, there is an urgent need to identify the contributions of azole fungicide use on food crops to the development of resistance in *A. fumigatus*. We address this through systematic soil sampling conducted on 10 agricultural sites in Germany over a 3-year period, including conventionally managed fields applying azoles fungicides as well as those practicing organic agriculture that do not use these compounds. In the largest published *A. fumigatus* sequencing effort to date, and the first to focus on the fungus in its natural niche, we also use whole-genome sequencing (WGS) to examine the impact of azole fungicides on the population genetics of 64 agriculturally isolated *A. fumigatus* isolates.

RESULTS

Variable abundance of *Aspergillus fumigatus* on agricultural sites in Germany. To examine the depth distribution of *A. fumigatus* in agricultural soils, we collected soil

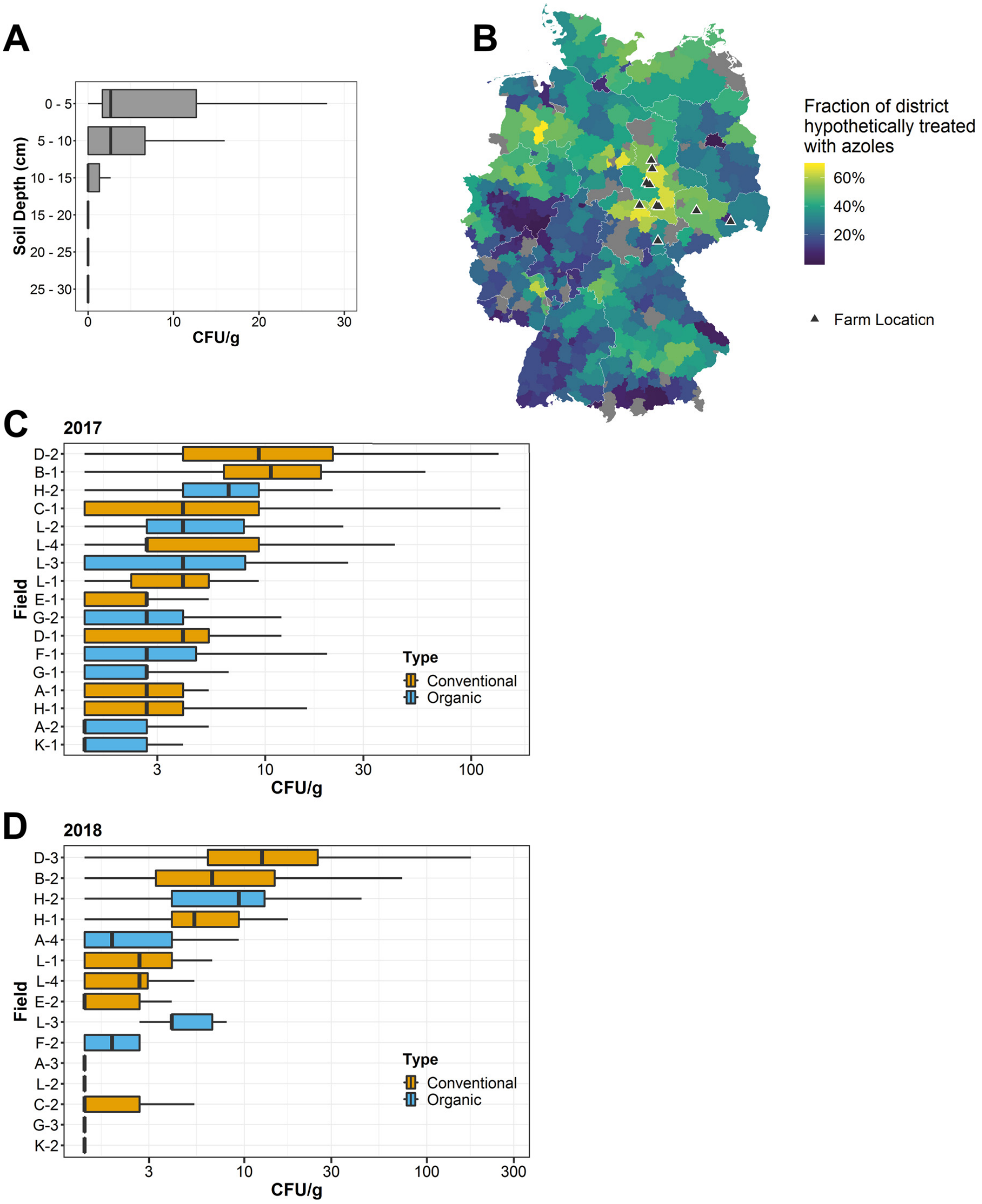


FIG 1 Abundance of *A. fumigatus* in the soil of conventional and organic farms. (A) *A. fumigatus* (CFU/g) at various soil depths. $n = 10$ samples per depth. (B) Estimated fungicide treatment rates and areas in Germany. The fraction of each district that is theoretically treated with fungicides was calculated using land use and organic agriculture share data reported by the Statistical Office of Germany in December 2016. Districts where no data on land use were available are shaded gray. (C and D) Abundance of *A. fumigatus* in the spring as measured by number of CFU/g soil in the spring of 2017 (C) and 2018 (D). For 2017, boxplots represent $n = 50$ soil samples per field for farms A, B, C, E, F, H, and K and $n = 25$ for farms D, G, and L. For 2018, $n = 50$ soil samples per field for farms A, B, C, D, E, G, H, and K and $n = 25$ for farms F and L. (E) Comparison of the mean number of CFU/g soil on farms between 2017 and 2018.

samples from a test field at 5-cm intervals down to a depth of 30 cm below the surface. *A. fumigatus* was most abundant in the top 5 cm of soil and was not significantly observed below a depth of 15 cm (Fig. 1A). As a result, only the top 5-cm layer of soil was collected in subsequent soil samples.

To identify whether our sampling areas are representative of azole and fungicide usage for Germany as a whole, we calculated the fraction of each administrative district that is theoretically treated with azoles in the context of agriculture using publicly available data. Overall, 51% of Germany is designated agricultural land (see Materials and Methods for data sources). However, not all this land is sprayed with azoles. For example, meadow or pastureland is rarely treated with fungicides. Additionally, approximately 7% of agriculture in Germany utilizes organic farming and does not apply azoles, with a range of 3 to 15% between the different federal states. Using land use data on arable farmland and permanent crop areas for each district in Germany, we calculated that the mean fraction of potentially azole-treated area in Germany is 32% (for the districts where data are available), with a range of 1 to 68% (Fig. 1B). The districts where we performed our soil sampling were almost all above the average for Germany, with a mean of 52% potentially azole-treated hectares and a range of 30 to 65%.

To examine the inter- and intrafield variability in the density of *A. fumigatus* on agricultural fields, soil samples were taken from nine conventional and eight organic fields before the growing season in 2017. Conventional fields were also sampled again after the vegetative period and application of azoles. In total, 2,875 soil samples were taken between 2016 and 2018 (see Table S1 in the supplemental material). The predominant crops being grown were cereals such as wheat and barley, but several apple orchards were also sampled. Of the fields sampled during this period in 2017, 67% of the soil samples taken were positive for *A. fumigatus*, with a large range between fields (28 to 100%) (Fig. 1C). We also observed a large degree of variation in the mean number of CFU per gram of soil between different fields, with some fields having 30× higher *A. fumigatus* density over others (0.7 to 18.8 CFU/g).

To examine the stability of *A. fumigatus* population sizes in agricultural soil, we investigated the same farms a year later in the spring of 2018 and repeated the soil sampling on eight conventional and seven organic fields. Due to crop rotation, it was not possible to sample the same fields as the previous year, except for the apple orchards on farms H and L. During the 2018 sampling period, an overall lower proportion of samples were positive for *A. fumigatus* (51% with a range of 10 to 96% between different fields) (Fig. 1D), but the mean number of CFU per gram for the 1,000 samples was similar to that of the previous year (5.18 CFU/g soil for 2017 compared to 5.74 CFU/g for 2018). As in the previous year, there was a large variability in the mean number of CFU of *A. fumigatus* present between fields. When comparing the six apple fields that were sampled over consecutive years, we did not observe a consistent trend in the stability of *A. fumigatus* population size. Two fields showed similar levels of *A. fumigatus* between 2017 and 2018, while two fields showed an increased abundance between the years, and the remaining two fields showed a significant reduction in abundance (Fig. S1A). To investigate potential factors that might support a higher abundance of *A. fumigatus*, we compared the total organic carbon (TOC) content of a random selection of samples with the number of CFU per gram for *A. fumigatus*, but we did not detect a clear relationship between the two (Fig. S1B). Altogether, we observed a nonuniform distribution of *A. fumigatus* in soil samples taken from the same field and a large degree of heterogeneity between fields.

Variable effects of fungicide application on *A. fumigatus* abundance. To examine the impact of fungicide application and the azoles on *A. fumigatus* in agricultural soil, we performed additional soil sampling on the conventional fields surveyed in the spring at the end of the vegetative period and after several months of fungicidal crop protection. A schematic illustration of the soil sampling and fungicide application timelines for 2017 and 2018 can be found in Fig. S2A and B. Unfortunately, the fungicide history for farms H and L was not available to us. Fields were typically treated with fungicides twice during the growing period, and azoles were by far the most dominant class of fungicide applied. Every application recorded contained at least one azole. However, fungicides are often applied as commercially available cocktails of different chemicals, so other classes were also present in 0 to 55% of applications in 2017 and 2018 (summarized in Fig. S2C).

When comparing the amount of *A. fumigatus* on fields before fungicide application to that after fungicide application and azole exposure, we detected a significant reduction in the number of CFU per gram of soil on the majority of fields in 2017 (Fig. 2A), even though it is not being directly targeted as a plant pathogen. To investigate whether this reduction in agricultural *A. fumigatus* populations was the result of fungicide application and not a seasonal effect from comparing April to July, monthly soil samples were taken from a conventional field and an organic field not treated with azoles or other nonnatural fungicides as a control. Samples were taken beginning in April, before azoles were applied to the conventional field, through the harvest period in July, and then additional samples were taken in October and November to allow for a period without fungicide application. From April to July, the conventional field was sprayed with azoles every 3 to 5 weeks. When analyzing the abundance of *A. fumigatus* on the organic field, we did not observe any significant differences between the abundance recorded monthly between April and July (Fig. 2B). However, the conventional field showed a significant reduction in abundance between April and May, corresponding to the beginning of the azole application period, and this reduction was maintained through the rest of the azole application period (Fig. 2C).

When comparing *A. fumigatus* density before and after fungicide treatment in 2018, we did not observe the same reduction in abundance, and most fields did not show significant changes between the time points (Fig. 2D). In fact, only one of eight fields sampled showed a statistically significant reduction in *A. fumigatus* abundance. Altogether, the impact of fungicide application on *A. fumigatus* abundance was variable between fields and more so between years, as other environmental factors also appear to influence *A. fumigatus* population size in agricultural soil.

Reduced susceptibility to agricultural azoles in populations isolated after the growing season and azole exposure. To assess the susceptibility of *A. fumigatus* to commonly applied agricultural azoles, we screened 435 isolates from 2017 and 342 isolates from 2018 for their ability to grow at a set concentration of difenoconazole and tebuconazole (approximately 20 isolates per field and sampling point). To limit potentially clonal isolates from skewing the results, a maximum of two isolates per soil sample were included for testing. As there are no established breakpoints for defining resistance to these compounds in *A. fumigatus*, we selected concentrations that mimicked MIC_{90} values for these azoles (1 mg/liter for difenoconazole and 2 mg/liter for tebuconazole). When examining conventional and organic farms in the spring, we observed a wide range in the fraction of isolates per field that were able to grow when challenged with agricultural azoles. For difenoconazole, this ranged from 10 to 55% per field in 2017 ($n = 17$ fields, 320 isolates in total) and 0 to 50% in 2018 ($n = 15$ fields, 261 isolates in total) (Fig. 3A and Table S2). For tebuconazole, the rates ranged from 0 to 25% in 2017 and 0 to 20% in 2018 (Fig. 3B and Table S2). We did not detect any significant differences in the rates between conventional and organic fields in the spring or between fields growing different crops (cereals or apples).

Given the reduction in the *A. fumigatus* population size observed on most conventional fields following fungicide application in 2017 and more variably on fields in 2018, we wanted to examine the effect of fungicides on the local azole susceptibility following the vegetative period and several months of fungicide application. We observed an increase in the proportion of isolates that were able to grow at the test concentrations of difenoconazole (1 mg/liter) and tebuconazole (2 mg/liter) for fields sampled after azole exposure compared to the same field in the spring prior to azole application (Fig. 3C and D and Table S3). We detected a 1.97-fold increase in the proportion of isolates that were resistant to our test concentration of difenoconazole in 2017, with a range from −1.1- to 4.5-fold for individual fields (Fig. 3C). In 2018, this increased to 2.84-fold, with a range of 1- to 4.5-fold increase for individual fields (Fig. 3D). We also saw a similar increase in the fraction of isolates with reduced

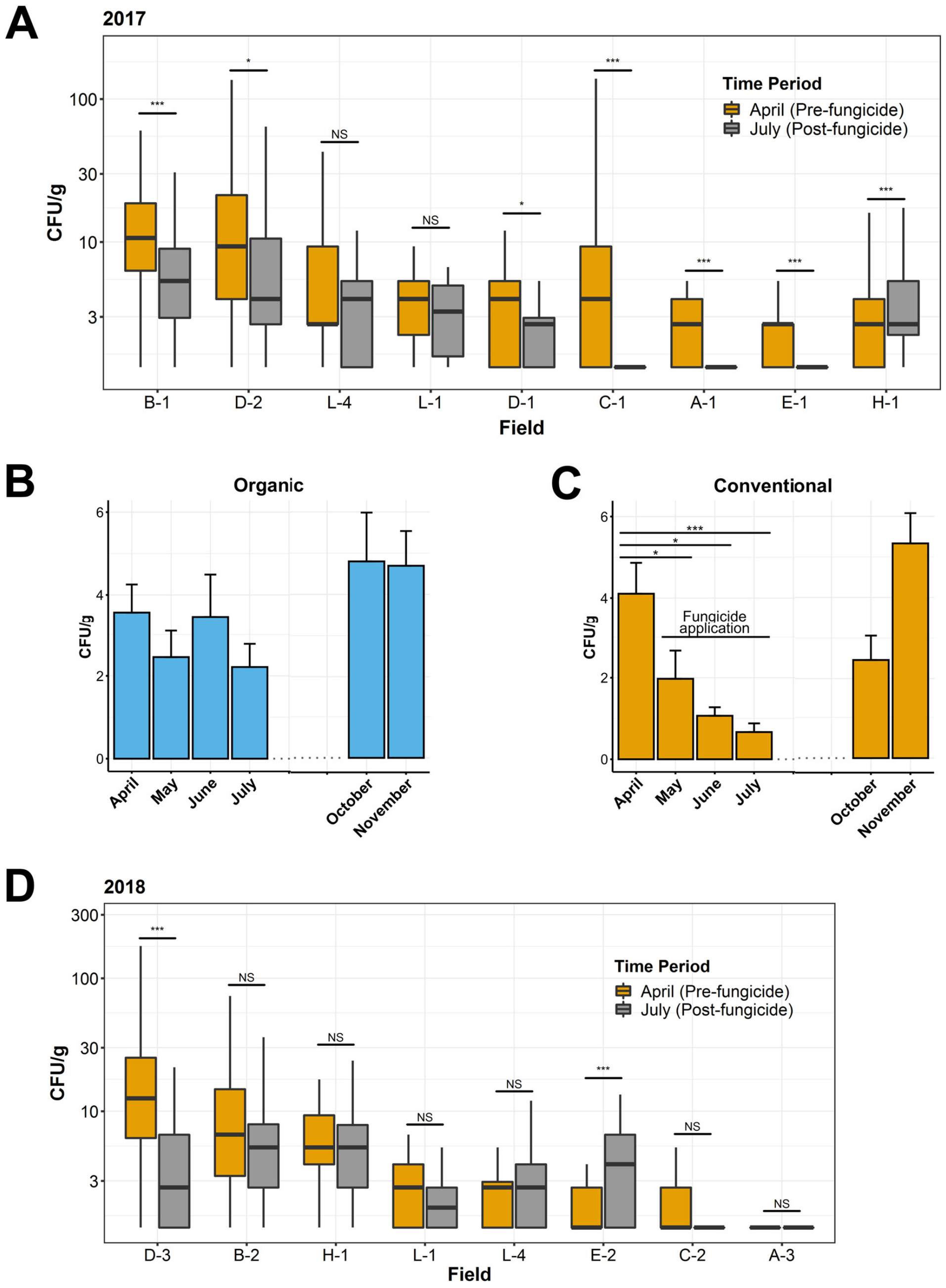


FIG 2 Abundance of *A. fumigatus* in the soil of conventional farms before and after the vegetative period and fungicide application. (A and D) *A. fumigatus* (CFU/g) on conventional fields sampled in April, prior to the vegetative period and fungicide application (orange), and in July, after the vegetative period and 3 months of fungicide application, including azole fungicides (gray) in 2017 (A) and 2018 (D). *, $P \leq 0.05$; **, $P \leq 0.01$; ***, $P \leq 0.001$; NS, not significant; determined by Mann-Whitney U test. (A) Boxplots represent $n = 50$ samples per field and time point for farms A, B, C, E, and H and $n = 25$ samples for farms D and L. (D) $n = 50$ samples per field and time point for farms A, B, C, D, E, and H and $n = 25$ samples for farm L. (B and C) *A. fumigatus* (CFU/g) during the months of April, May, June, July, October, and November of a conventional field applying fungicides from May to July (C) and an organic field not applying nonnatural fungicides (B). Bars represent means ± standard errors of the means from 50 soil samples per month. No significant difference was found in abundance between the months of April, May, June, and July for the organic field using a

(Continued on next page)

susceptibility to our test concentration of tebuconazole after azole exposure compared to that before. In 2017, we observed a 5.9-fold increase in the number of isolates that were able to grow at the test concentration after azole exposure versus before exposure, with a range of 0 to 25.0% for individual fields (Fig. 3E). In 2018, we detected a more modest 1.9-fold, with a range of −5 to 18.2% for individual fields (Fig. 3F).

For the fields growing cereals, we were not able to sample the same fields over subsequent years due to crop rotation and the fields not being in use the following year. However, we were able to compare the same apple fields in both 2017 and 2018. We tracked the local susceptibility to agricultural azoles in these fields over two consecutive years at two time points, in the spring prior to fungicide application and just prior to harvest after ≈3 months of fungicide application. The fraction of isolates that were able to grow in the presence of 1 mg/liter difenoconazole or 2 mg/liter tebuconazole was, in general, low in the spring and increased after fungicide application (Fig. 3G and Table S3). Interestingly, in the spring of 2018, the proportion had returned to a level comparable to what we observed in the spring of 2017, indicating that the reduced susceptibility is transient and recedes when the selective pressure imposed by fungicide is removed. In summary, we found a wide range of susceptibilities to agricultural azoles between different fields but a consistent decrease in susceptibility following the growing season, fungicide application, and azole exposure. However, this change is seemingly transient or reversible, and the *A. fumigatus* populations from fungicide-treated fields typically returned to what they were prior to fungicide application by the following spring.

Resistance to medical azoles in agricultural *A. fumigatus* isolates. We next examined our isolate collection for resistance to medical azoles and determined the proportion that would be considered clinically resistant. Using the VIPcheck agar-based screening method, followed by broth microdilution for isolates showing growth on agar-containing wells (22, 23), we determined that only a very small fraction of *A. fumigatus* organisms isolated from agriculture showed resistance to itraconazole, voriconazole, or posaconazole in 2016 to 2018 (Table 1). The overall resistance rate to itraconazole among all isolates collected was higher than that for other compounds, with 3.0% (11/333) of isolates being resistant in 2016, 0.7% in 2017 (4/460), and 0.6% (2/322) in 2018. We observed lower resistance rates for posaconazole and voriconazole, with only 2.1% (7/333) being resistant in 2016, 0.7% (4/460) in 2017, and 0.0% (0/322) in 2018. As there are no clinical breakpoints established for agricultural azoles, we calculated epidemiological cutoff values (ECOFFs) for difenoconazole and tebuconazole using MICs from 160 randomly selected isolates from 2017 and 2018. Using these values, we found that 1.3% of the 160 isolates had MICs above the ECOFF for difenoconazole (2 mg/liter), and 4.4% of isolates had MICs above the ECOFF for tebuconazole (2 mg/liter) (Table 2). As has been described previously (24), isolates resistant to one or more medical azoles often displayed elevated MICs to agricultural azoles, indicating cross-resistance (Table S4).

To quantify what mutations were responsible for azole resistance in the population analyzed, we genotyped the *cyp51a* locus encoding the azole target enzyme for all isolates resistant to one more medical azoles using Sanger sequencing ($n = 18$). We found that the most dominant DNA alteration observed was the well-characterized TR_{34}/L98H mutation (Fig. 3H and Table S4). This *cyp51a* genotype accounted for 6/12 (50%) of resistant isolates in 2016 and 3/4 (75%) in 2017. In 2018, we identified only two resistant isolates among the 322 screened, and both had wild-type *cyp51a* loci. However, both of these isolates were only weakly resistant to itraconazole, but not other azoles, with itraconazole MIC values right at the breakpoint of 2 to 4 mg/liter. For

FIG 2 Legend (Continued)
Kruskal-Wallis test. In contrast, we found a significant difference ($P = 0.004$) for the abundances in this time period on the conventional field. *P* values from subsequent pairwise comparisons between months are indicated. *, $P \leq 0.05$; **, $P \leq 0.01$; ***, $P \leq 0.001$; determined by Wilcoxon signed rank.

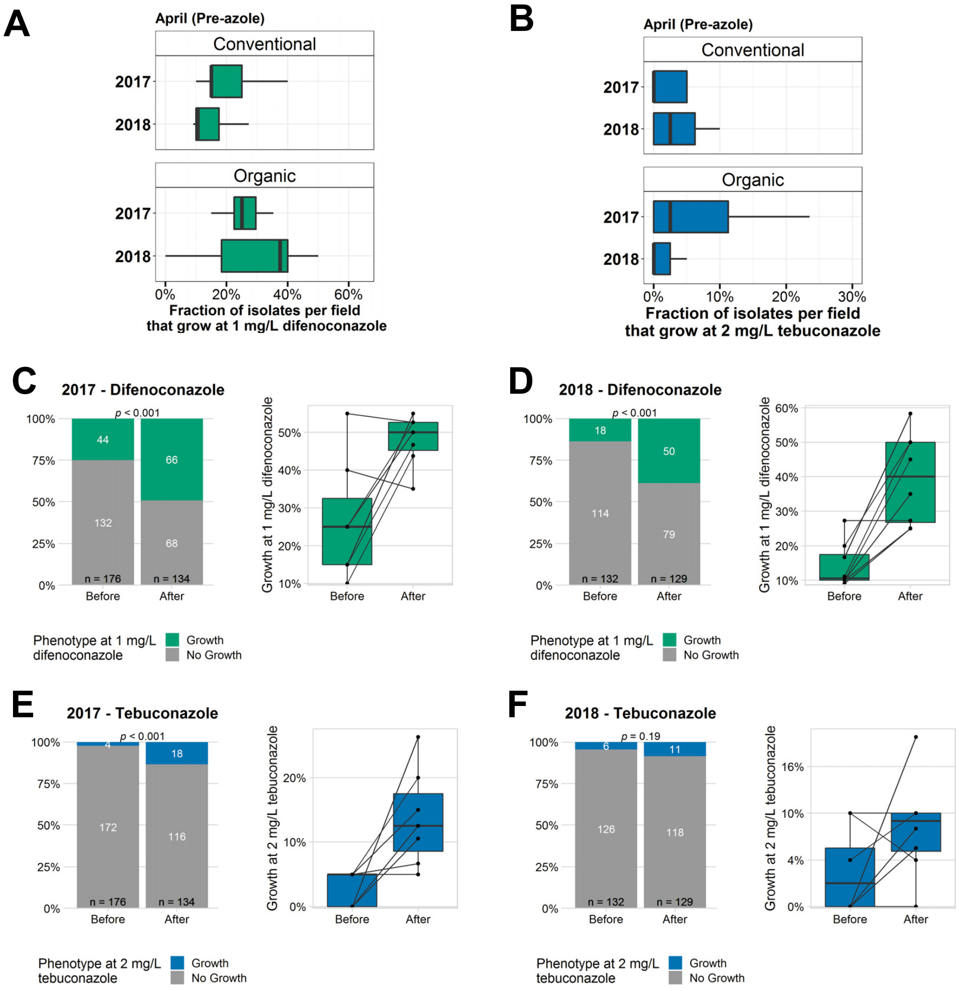


FIG 3 Azole resistance among agricultural *A. fumigatus*. (A and B) Fraction of the isolates per field that grow at 1 mg/liter difenoconazole (A) and 2 mg/liter tebuconazole (B). For 2017, $n = 340$ isolates from 9 conventional and 8 organic fields, ≈20 isolates per field, were used. For 2018, $n = 213$ isolates from 8 conventional and 7 organic fields, ≈20 isolates per field, were used (full summary in Tables S2 and S3). (C to F) Comparison of the proportion of isolates that grow at 1 mg/liter difenoconazole (C and D) and 2 mg/liter tebuconazole (E and F) before and after the vegetative period and fungicide application. For 2017 (C and E), $n = 275$ isolates from 7 fields were used, and for 2018, $n = 261$ isolates from 8 fields were used (full summary in Tables S4 and S5). (Left) Overall summary of all isolates tested that year. (Right) Within-field changes ($n \approx 40$ isolates per field; 20 before, 20 after). *P* values were calculated by Wilcoxon signed-rank test between before and after values. (G) Temporal changes in antifungal susceptibility of *A. fumigatus* on apple fields sampled before and after fungicide exposure over a 2-year period. Shown is the fraction of isolates that can grow at 1 mg/liter difenoconazole (top) and 2 mg/liter tebuconazole (bottom). $n = 12$ to 20 isolates/field and time point. (H) *cyp51a* genotypes of isolates resistant to one or more medical azole. (I) MICs of agricultural *A. fumigatus* isolated before and after azole exposure. $n = 159$ randomly selected isolates from 2017 and 2018; $n = 80$ before and $n = 79$ after. *P* values were calculated by Wilcoxon signed rank test between before and after values.

comparison, resistant isolates from other years had MIC values of >8 mg/liter. In total, the genetic cause of resistance remained unknown for 8 isolates from 2016 to 2018.

Since the majority of fields had no resistant isolates, it was not possible to effectively compare resistance rates among agricultural *A. fumigatus* organisms for medical azoles before and after fungicide treatment. In lieu of this, we examined the MIC distribution for isolates collected before and after the growing season and fungicide application. Examining 79 isolates from the before period and 79 isolates from after azole exposure, we observed a shift in the MIC distribution toward higher MICs for all azoles examined,

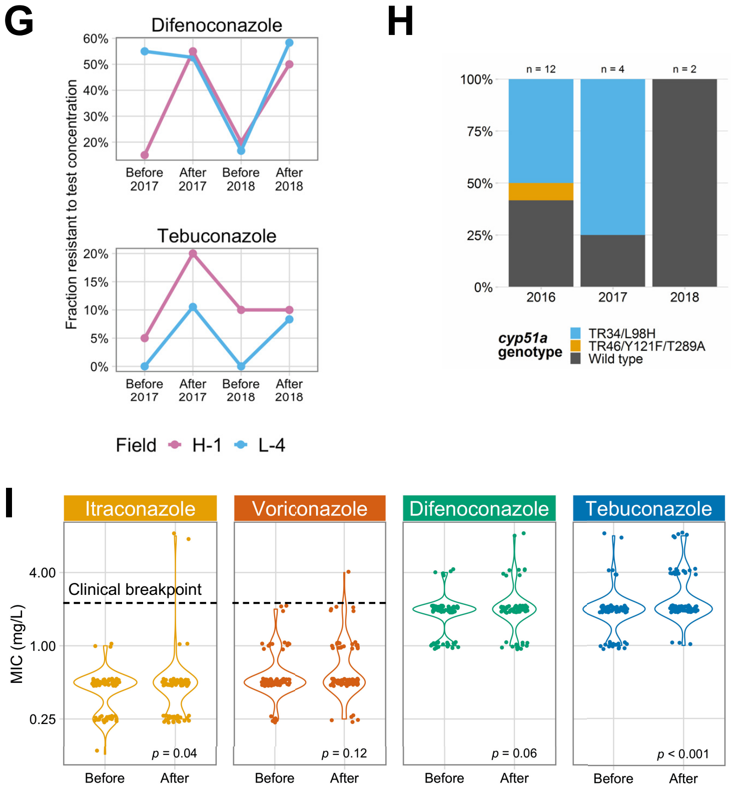


FIG 3 (Continued)

both medical and agricultural (Fig. 3I). However, the median MICs remained unchanged, indicating that the majority of the population following azole exposure does not exhibit a change in MIC. Taken together, these results indicate that the rate of resistance to clinical azoles among environmental *A. fumigatus* organisms isolated from agricultural environments is low overall and that exposure to agricultural azoles alone or in combination with other fungicides causes a minor increase in MIC values for some of the population, but the majority of isolates are left unchanged.

No distinct population structure for *A. fumigatus* isolates from different sampling sites. To better understand the population structure of *A. fumigatus* in the environment and how it is impacted by the azoles, we performed whole-genome sequencing on isolates from four conventional farms collected before and after azole exposure. Sixty-four isolates were sequenced by Illumina paired-end sequencing, representing eight isolates per farm and time point. Isolates from each field and time point were randomly selected, with a maximum of two per soil sample to avoid sequencing of clonal isolates. Raw reads were checked for quality and then aligned to the Af293 reference genome, resulting in a median depth of coverage after mapping of 31.5× (range, 10× to 90×) and a median genome coverage median of 94% (range, 92.8 to 96.7%) (Data Set S1). Consistent with what has been observed among sequenced clinical strains (25), isolates differed from the Af293 reference by a median of 84,690 single-nucleotide variants (SNVs) or 2.88 SNVs/kb, with a range of 65,854 to 146,055 SNVs. The analysis of copy number variations (CNVs) identified 8,277 unique CNVs in total, with a median of 3,115 CNVs per isolate (range, 1,666 to 4,532 CNVs). These CNVs

TABLE 1 Azole resistance rates among agriculturally isolated *A. fumigatus* isolates[a]

		Resistance rate (%)		
Year	*n*	ITR	POS	VOR
2016	333	3.0	2.1	2.1
2017	460	0.7	0.7	0.7
2018	322	0.6	0	0

[a] Isolates were screened for potential azole resistance to itraconazole (ITR), posaconazole (POS), and voriconazole (VOR) using VIPcheck agar-based screening. Resistance was confirmed and MICs determined via EUCAST broth microdilution testing.

were further delimited into a median of 2,589 deletions (range, 1,247 to 4,405) and 5,687 insertions (range, 3,872 to 7,030).

A maximum likelihood phylogeny based on SNVs indicated no population stratification among isolates from different farms and regions of Germany (Fig. 4). To more directly assess the association between genetic and geographic distance, we performed a Mantel test correlating a geographic distance matrix with the fixation index (F_{ST}) genetic distance matrix and observed no significant association between the two. When considering the azole-resistance status of the isolates, the two itraconazole-resistant TR_{34}/L98H isolates clustered next to each other, despite originating from separate farms, while the third itraconazole-resistant isolate with an undefined resistance mechanism was on a distinct branch. Despite being the nearest sequenced neighbors, the two TR_{34}/L98H isolates were genetically distinct, each possessing 19,439 and 60,841 unique SNVs not shared by the other isolate, along with 67,196 common SNVs relative to Af293.

Comparative analysis of molecular variance (AMOVA) indicated that the majority of the variation seen among the 64 isolates came from the population as a whole (within sample) (94.8%) and between samples (5.0%) (Table 3). There was no significant molecular variance between farms (0.2%), with the exception of modest variation between farm B and farm C (1.2% of variation observed). Weighted Weir and Cockerham's fixation indexes (F_{ST}) for each farm were essentially zero, indicating an interbreeding, panmimetic population with no separation between farms (Fig. S3A). Analysis of copy number variation (V_{ST}), estimating population differentiation based on copy number variation, also indicated no subdivision among the farms (Fig. S3B). To examine the genetic diversity within farms, nucleotide diversity (the average number of nucleotide differences per site for all possible pairs in the population or π) and the number of polymorphic sites (Watterson estimator, or θ) were calculated along 5-kb windows with a 500-bp step size for each farm population. Farm C showed the greatest intrafarm diversity, while farm E showed the smallest (Fig. S3C and D). Taken together, there was no population differentiation between *A. fumigatus* isolates from different farms in Germany, a finding in line with the fungus' capacity for aerosol dispersal.

Changes in population genetics following azole exposure vary by field. Given the observed reduction in overall *A. fumigatus* abundance after azole treatment, we examined the populations for changes in genetic diversity and evidence of selective sweeps in *A. fumigatus* field populations. Using 5-kb sliding windows with a 500-bp step size, nucleotide diversity (π) was calculated for the individual farms before and

TABLE 2 MIC_{50} for medical and agricultural azoles calculated from 160 randomly selected isolates as well as $ECOFF_{95}$ and the fraction of isolates with MICs above this value

Azole	MIC_{50}[a]	ECOFF[b]	Fraction of isolates above ECOFF (%)
Itraconazole	0.5	2	1.3
Voriconazole	0.5	2	0.6
Difenoconazole	2	8	1.3
Tebuconazole	2	8	4.4

[a] Calculated from 160 randomly selected isolates.
[b] Rounded up to next dilution.

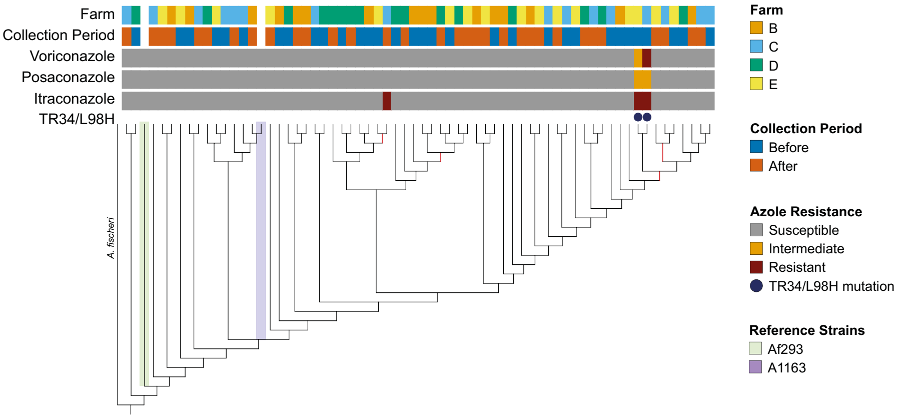


FIG 4 Phylogeny of agricultural *A. fumigatus* isolates from before and after the vegetative period and azole application. From top to bottom, the colored bars indicate the farm where the isolate was collected, the collection period, voriconazole resistance (susceptible, intermediate, or resistant, according to EUCAST definitions), posaconazole resistance, itraconazole resistance, and the presence of the TR_{34}/L98H allele in *cyp51a*. *A. fischeri* is indicated as an outgroup, and the two *A. fumigatus* reference strains, Af293 and A1163 (CEA10), are also marked. Branches with support values of less than 0.9 are marked in red.

after azole treatment. No clear trend was seen regarding changes in nucleotide diversity between the time points. The isolates from farms B and E showed increased diversity following azole application, while nucleotide diversity decreased on farms C and D (Fig. 5A). We also calculated the number of segregating sites (θ) for the same 5-kb windows and observed the same lack of consensus. Farms B and C showed similar values of θ before compared to after azole application, while farm D showed a dramatic decrease in θ following azole application (Fig. 5B). AMOVA on isolates from before and after the vegetative period and azole exposure did not indicate any significant molecular variation between the time periods, with the majority of the variation being between and within samples (Table 4). Finally, we measured Tajima's D to test neutrality along 5-kb windows, where negative values indicate less variation than expected and are indicative of a selective sweep. Positive values denote a population that is more heterogenous than would be expected and suggest either a sudden population contraction or balancing selection. Overall, the bulk of the Tamija's D values were close to neutral, and there was no clear trend between farms, indicating that there was no genomic signature of a population bottleneck or selective sweep following azole exposure (Fig. 5C). The median Tajima's D was roughly zero for farm B and increased slightly to 0.37 following azole exposure prior to azole application, and the same direction shift was seen for farm D, but the starting Tajima's D was negative at the time point before azole exposure (−0.53 to 0.53) (Fig. 5C). Conversely, the median Tajima's D for farms C and E shifted from positive to negative after the growing season and azole exposure (0.75 to −0.57 for field C and 0.61 to −0.10 for field E). Taken together, these results indicate that despite the reduction in abundance of *A. fumigatus* on agricultural

TABLE 3 AMOVA between farms[a]

Farm	Total variation	Variation within samples (%)	*P* value	Variation between samples (%)	*P* value	Variation between farms (%)	*P* value
All farms	289,735.7	274,622.7 (94.8)	<0.001	14,629 (5)	<0.001	484 (0.2)	0.303
B vs C	147,358.0	139,257.7 (94.5)	<0.001	6,386.2 (4.3)	<0.001	1,714.1 (1.2)	0.028
B vs D	190,425.2	181,470 (95.3)	<0.001	9,002.4 (4.7)	<0.001	−47.2 (0.0)	0.429
B vs E	162,874.7	155,745.5 (95.6)	<0.001	6,879.3 (4.2)	<0.001	249.9 (0.2)	0.211
C vs D	130,625.7	122,692.8 (93.9)	<0.001	7,932 (6.1)	<0.001	1 (0.0)	0.355
C vs E	105,529.1	99,521.4 (94.3)	<0.001	6,098.3 (5.8)	<0.001	−90.6 (−0.1)	0.414
D vs E	143,684.1	135,847.0 (94.5)	<0.001	8,382.7 (5.8)	<0.001	−545.6 (0.4)	0.890

[a]Calculated using isolates from before and after azole application. *n* = 16 isolates per farm.

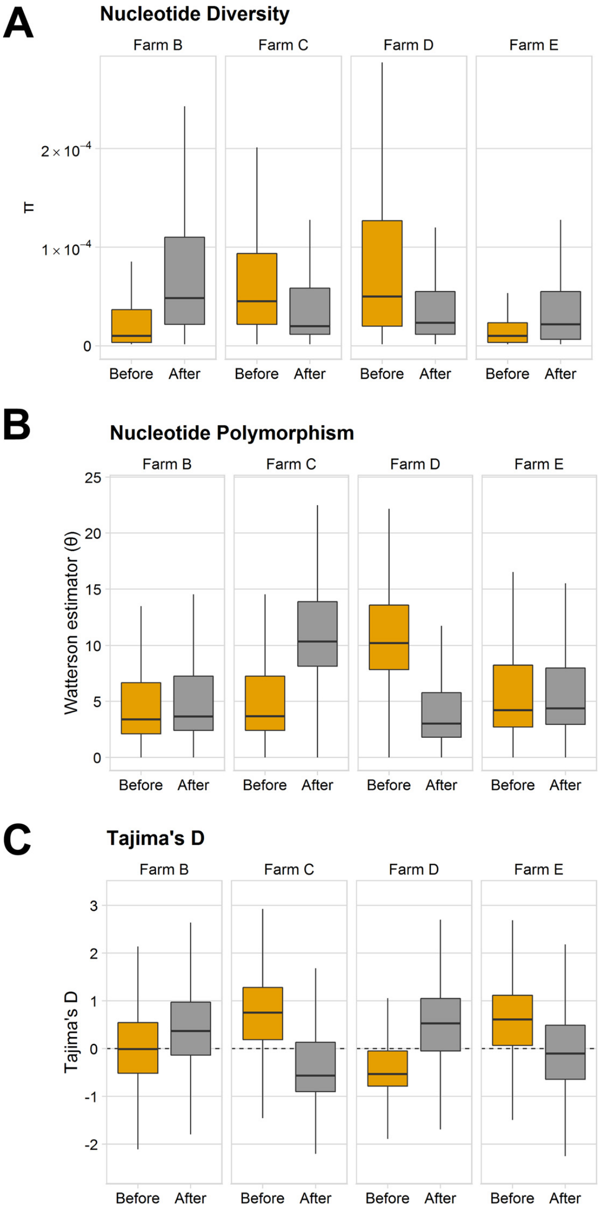


FIG 5 Genetic diversity among isolates from before and after the vegetative period and azole exposure. (A to C) Nucleotide diversity (π) (A), nucleotide polymorphism (Watterson estimator, or θ) (B), and Tajima's D (C) along 5-kb windows with a 500-bp step size before and after the vegetative period and azole exposure. $n = 8$ isolates per farm and time point, 64 in total.

fields following azole application, we were unable to detect any marked changes at the population level.

DISCUSSION

The use of azole fungicides for plant protection has been previously suggested as a driver of clinical resistance in the environmental saprobe and human pathogen *A.*

TABLE 4 Analysis of molecular variance between isolates collected before and after azole exposure[a]

Farm	Total variation	Variation within samples (%)	*P* value	Variation between samples (%)	*P* value	Variation between time periods (%)	*P* value
B	106,712.9	103,295.3 (96.8)	<0.001	3,729.5 (3.5)	<0.001	−311.9 (−0.3)	0.617
C	48,438.0	45,393.1 (93.7)	<0.001	3,341.7 (6.9)	<0.001	−296.8 (−0.6)	0.593
D	85,724.6	80,522.3 (93.9)	<0.001	5,592 (6.5)	<0.001	−389.7 (−0.5)	0.564
E	59,366.5	55,941.4 (94.2)	<0.001	3,001.2 (5.1)	<0.001	423.9 (0.7)	0.149

[a] $n = 16$ isolates per farm field (eight pre-azole and eight post-azole).

fumigatus, as well as the emergence of new fungal pathogens, such as *Candida auris* (11, 26, 27). However, direct evidence linking azole use in agriculture and clinical resistance is missing. Additionally, delineation of specific roles for the use of azoles in crops versus ornamental plants, such as flower bulbs, has not been defined and is important to make, as limiting azole fungicide use would significantly impact disease control and yield in many crops. In this study, we provide the first systematic investigation of the impact of fungicide use on the ecology and azole resistance status of a human pathogen, *A. fumigatus*. Through analysis of 2,875 soil samples over a 3-year period in central Germany, we found an overall low incidence of *A. fumigatus* isolates that would be considered clinically resistant (1 to 3%). However, we observed a modest, but consistent, decrease in azole susceptibility following the growing season and azole exposure, as well as a more variable reduction in fungal abundance following fungicide application. Interestingly, this change in susceptibility was transient and reset by the following spring. We also assessed the influence of fungicide application on *A. fumigatus* population dynamics by WGS and were unable to find a clear impact on the population structure or genetic diversity.

Despite sampling on fields that were actively treated with azole fungicides and in regions of Germany with above average fungicide exposure, only 1 to 3% of *A. fumigatus* isolates collected were resistant to medical azoles. This incidence is in agreement with clinically reported frequencies in Germany and other European countries, where the rate of triazole resistance ranged from 0.6% to 4.2% (3.2% in Germany) (28, 29). The environmental resistance rates reported for Europe, however, have been much broader. Some studies have found environmental resistance frequencies approaching 20%, while others have reported virtually no incidence of resistant environmental isolates (30–35). Part of these differences could be attributed to differences in methodology, such as the use of azole selection during the isolation procedure or inclusion of potentially clonal isolates from a given sample to skew the data. In the current study, we avoided azole selection during the isolation procedure and allowed only two isolates per soil sample to avoid potentially clonal isolates that influence the results. Another potentially contributing factor could be that while they are all technically “environmental” isolates, rural or agricultural settings could be a completely different niche than an urban flower garden, a supposition supported by a recent study where rural areas in the United Kingdom had much lower resistance rates (1.1%) than urban locations (13.8%) (34).

Another important and novel finding from this study is that isolates from after the vegetative period and azole exposure consistently showed decreased azole susceptibilities to difenoconazole and tebuconazole as well as a subtle MIC shift to both agricultural and medical azoles. However, one shortcoming to our study is that we did not collect isolates from organic fields at a matching time point for susceptibility testing, so we cannot exclude that the changes observed in azole susceptibility are not also influenced by seasonal changes. The observation that changes in susceptibility are transient and reset in the period between the end of the growth period and the following spring on the two fields is intriguing and worthy of further study. This transformation could either be the result of a naive population coming in via aerosol dispersal or a consequence of isolates acquiring unstable, epigenetic-mediated resistance, a phenomenon previously observed in the environmental saprobe and human pathogen *Mucor circinelloides* as well as the plant-pathogenic fungus *Monilinia fructi-*

cola following azole exposure (36, 37). Either scenario would be in agreement with our finding that fungicide application does not alter the population structure or genetic diversity of *A. fumigatus* in agricultural fields.

We observed a wide range in the number of CFU per gram of soil within and between farms during annual spring sampling. Our mean number of *A. fumigatus* CFU per gram of soil was in line with what was reported recently for abundance in wheat grain, maize silage, and fruit waste (38). However, this abundance is several magnitudes lower than that reported for *A. fumigatus* in flower bulb waste and green material waste in this same study, where it was not uncommon to isolate 10^4 CFU/g. We demonstrated a reduction in the *A. fumigatus* population size on most fields sampled in 2017 following the vegetative period and fungicide application. However, this finding was not strongly observed in 2018, indicating that other environmental factors also influence the abundance of *A. fumigatus* in agricultural soil. One example of such a potential factor is that Germany experienced extreme heat and drought during the 2018 growing season; in fact, one field had to be removed from analysis this year because it caught fire during the vegetative period.

Our study also provides the first WGS-based study focused on *A. fumigatus* in its natural niche, the environment. Previous studies have primarily concentrated on clinical isolates, with particular priority given to resistant strains (25, 39). Even while sampling within the same field, we found a large degree of genetic diversity, where the majority of diversity came from within samples. We also did not observe a defined population structure or separation between farms or regions. The degree to which this environmental diversity is recapitulated in clinical isolates, and whether there are enrichments for particular subgroups in the transition from environment to clinic, is an interesting question for further study.

Given our low observed resistance rate among agricultural isolates and the lack of discernible impact on the population structure and genomic diversity of *A. fumigatus* following fungicide application, our study does not find evidence that azole fungicide use in crop agriculture significantly contributes to resistance in *A. fumigatus*. Azoles should not necessarily be removed from use in this context due to their crucial role in global food production. Instead, our field study provides empirical support for the model that azole resistance in *A. fumigatus* is being driven not by the use of these compounds for crop agriculture but in settings such as the cultivation of flowers or ornamental plants as well as the storage of green waste. Both of these settings have been identified as hot spots for resistance development, owing to their higher overall fungal colony counts and higher fungicide concentrations (38, 40–42). Unfortunately, due to massive aerial dispersal of *A. fumigatus* conidia, the use of azoles in any hot spot can lead to the worldwide distribution of resistant strains.

MATERIALS AND METHODS

Site selection and soil sampling. During 2016 to 2018, soil sampling was conducted on agricultural sites in the federal states of Thuringia, Saxony-Anhalt, and Saxony, with the approval of the land owner and/or relevant ministries. The majority of fields sampled were growing cereals such as wheat or barley, but some apple orchards were sampled as well (see Table S1 in the supplemental material for full details). Farms were arbitrarily assigned an alphabetic identifier (A to L) and specific fields a numeric identifier, as described in Text S1. Due to crop rotation, the same field could not be surveyed over subsequent years, with the exception of the apple orchards. Soil samples were collected at the beginning of the vegetation period on the conventional and organic cultivated sites, and after azole application an additional sampling was carried out on the conventional sites. In general, 50 soil samples per site and type of farming (conventional or organic) were collected, with a total of approximately 1,000 soil samples per year. Soil samples were selected to best cover the field with a minimum distance of 1 m between samples. For each sample, the top layer of soil was collected by a metal spatula into a sterile sample cup and refrigerated until processing.

Soil processing and isolation of *A. fumigatus*. Three grams of soil from each sample cup was weighed out and resuspended in 8 ml 0.2 M NaCl containing 1% Tween 20. Samples were vortexed vigorously and then left to settle until a phase separation became apparent. Two milliliters of the upper phase was transferred to a new tube for plating onto Sabouraud glucose agar (SGA) containing 50 μg/ml chloramphenicol (Sigma-Aldrich, Taufkirchen, Germany). Of this 2 ml, 150 μl was plated onto one plate and the remaining volume was plated onto a second plate to adjust for variable fungal concentrations in samples. Plates were then incubated at 50°C for 5 days to select for *A. fumigatus*, which is unique among *Aspergillus* spp. in its ability to grow at this temperature. On day 5, the incubator temperature was reduced to 42°C to allow for sporulation, and plates were grown for another 2 days. The number of *A. fumigatus* colonies was counted, and up to two colonies per soil sample were transferred to new plates for isolation.

Antifungal susceptibility testing. Quick screening of susceptibility to itraconazole, voriconazole, and posaconazole was assessed using the agar-based VIPcheck assay (Mediaproducts BV, Groningen, Netherlands) by following the manufacturer's directions. For testing, isolates were grown on SGA for 2 to 4 days at 37°C. Plates were then swabbed with a damp, sterile cotton swab to prepare a conidial suspension of 0.5 to 2 McFarland. Twenty-five microliters of this suspension was then plated onto the 4 wells of the VIPcheck plate containing 4 mg/liter itraconazole, 2 mg/liter voriconazole, or 0.5 mg/liter posaconazole or a control well containing no drug. To assess susceptibility to agricultural azoles for isolates collected during 2017 and 2018, RPMI plus 2% glucose agar plates containing either 1 mg/liter difenoconazole or 2 mg/liter tebuconazole was prepared as described in reference 43. Resistance was defined as significant inhibition of germination and hyphal growth compared to the no drug control. Isolates that showed resistance to any of the medical azoles in the VIPcheck assay were subject to broth microdilution following EUCAST methodology (protocol E.DEF 9.3). ECOFFs were calculated using the ECOFFinder program available from EUCAST.

DNA extraction. For WGS and PCR-based amplification, isolates were grown shaking in SG broth at 37°C, and genomic DNA was isolated using the Quick-DNA fungal/bacterial miniprep kit (Zymo Research, Irvine, CA) according to the manufacturer's suggested protocol.

***cyp51a* genotyping.** The *cyp51a* coding sequence and upstream region containing the tandem repeat was amplified using the primers described in Table S5. Cleaned-up PCR products were sequenced and the *cyp51a* genotype determined using FunResDB (https://elbe.hki-jena.de/FunResDb/index.php).

Genome sequencing, quality assessment, and alignment. Library preparation and 2 × 150-bp paired-end sequencing were performed on a NextSeq 500 v2 by LGC Genomics (Berlin, Germany) by following the manufacturer's recommended protocols. Sequence data quality control and filtering were performed using an in-house script and FastQC (v0.11.5). Quality reads were mapped to the *A. fumigatus* Af293 reference genome (version 2015-09-27; retrieved from FungiDB [44]) using BWA-MEM (version 0.7.8-r779-dirty) (45). PCR duplicates were marked using MarkDuplicate from Picard version 2.18.25 embedded in the Genome Analysis Toolkit (GATK; version 4.1.0.0). All WGS samples included for analysis possessed greater than 10-fold genome coverage after mapping, and more than 90% of reads mapped to the reference genome.

Variant identification and SNV-based phylogeny. Short variants, including single-nucleotide variants (SNVs) and short insertions and deletions (InDels), were detected using GATK Haplotype Caller by following the recommended best practices for single calling (46). Copy number variants (CNVs) were identified using Control-FREEC (47). For the phylogenetic analysis, nucleotide consensus sequences were extracted from vcf files using VCFtools (48), and an in-house script was used to translate nucleotides to protein-coding sequences. Multiple-sequence alignment was performed using MUSCLE v3.8.31 (49) with 7,771 conserved core genes and an approximately maximum likelihood phylogeny constructed using FastTree2 (version 2.1.10) (50). The Interactive Tree of Life (iTOL) v4 was used for visualization (51).

Genetic diversity analyses. Analysis of molecular variance (AMOVA) was determined using the R package ade4 (nrep = 999). Nucleotide diversity (π) was calculated by VCFtools (version 0.1.6) using 5-kbp windows with a step size of 500 bp. Nucleotide polymorphism (θ) and Tajima's D were calculated using ANGSD (52) with a window size of 5 kbp and a step size of 500 bp. Weighted Weir and Cockerham's F_{ST} values were calculated using VCFtools, while V_{ST} values were calculated as in reference 53. A Mantel test correlating geographic distance matrices with pairwise F_{ST} matrices was performed using the R package ade4 (nrep = 9999).

Estimation of fungicide treatment areas and rates in Germany. The fraction of each district theoretically treated with fungicides was calculated using publicly reported data from 2016, available from the German Statistical Offices (https://statistikportal.de), using the following equation:

$$\frac{(\mathrm{Ha}_{\mathrm{Ackerland}} + \mathrm{Ha}_{\mathrm{Dauerkulturen}}) \times \mathrm{Fraction}_{\mathrm{Conventional}}}{\mathrm{Ha}_{\mathrm{Total}}}$$

The total area per district was obtained from Table 33111-01-02-4, ground area by actual use. The sum of arable farmland and permeant crop areas was calculated for each administrative district using Table 41141-01-01-4, farms and their agricultural use area by crop type. To accommodate that some percentage of this area is cultivated under organic agriculture methods, the hectares of cropland were then multiplied by the fraction of nonorganic agriculture for the federal state in which the district is located to estimate the number of hectares potentially treated with fungicides (as calculated using data on farms, agricultural areas, and workers; accessed on 1 November 2019 from https://www.statistikportal.de/node/254). Unfortunately, no data were available on the breakdown of agricultural methods at the district level to allow for more exact estimation. Finally, this value of estimated treated area per district was divided by the total area of the district for visualization as a choropleth map.

Box and whisker plots. Box and whisker plots presented in this paper are in the style of Tukey, where the boldface line indicates the 50th percentile and the hinges represent the 25th and 75th percentiles. The lower whisker extends from the lower hinge to the lowest datum within a 1.5 interquartile range (IQR), while the upper whisker represents the highest datum still within 1.5 IQR. Outliers are marked with points.

Data and isolate availability. Isolates generated within this study were submitted to and are publicly available in the Jena Microbial Resource Collection. Raw FASTQ files were uploaded to the NCBI Sequence Read Archive and are publicly available under BioProject number PRJNA595552.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

TEXT S1, DOCX file, 0.02 MB.

FIG S1, EPS file, 1 MB.

FIG S2, EPS file, 1.1 MB.

FIG S3, EPS file, 1.4 MB.

TABLE S1, DOCX file, 0.02 MB.

TABLE S2, DOCX file, 0.02 MB.

TABLE S3, DOCX file, 0.02 MB.

TABLE S4, DOCX file, 0.02 MB.

TABLE S5, DOCX file, 0.01 MB.

DATA SET S1, XLSX file, 0.02 MB.

ACKNOWLEDGMENTS

We thank Ivo Schliebner for helpful scientific discussions and his input in establishing the network of farms sampled in this study, Anja Kunert for scientific advice, and Matthew Blango for thoughtful comments and discussions regarding the manuscript. We also acknowledge Antonia Lange for her experimental support in the isolation of environmental strains from 2016.

G.P. thanks the Deutsche Forschungsgemeinschaft (DFG) CRC/Transregio 124 "Pathogenic fungi and their human host: Networks of interaction," subproject INF and B5.

O.K. and H.B.D. conceived the study. A.E.B. and J.B. collected samples and performed experiments. W.B. provided reagents and experimental advice. A.E.B., J.B., T.S.E., K.K., G.P., H.B.D., and O.K. analyzed the data and interpreted the results. A.E.B. wrote and revised the manuscript with input from all authors.

This project was supported by The Federal Ministry for Education and Science (Bundesministerium für Bildung und Forschung) within the framework of InfectControl 2020 (Projects FINAR and FINAR 2.0, grants 03ZZ0809 and 03ZZ0834). The work of the German National Reference Center for Invasive Fungal Infections is supported by the Robert Koch Institute from funds provided by the German Ministry of Health (grant 1369-240).

REFERENCES

1. Leading International Fungal Education. Invasive aspergillosis. http://www.life-worldwide.org/fungal-diseases/invasive-aspergillosis. Accessed 1 November 2019.
2. Patterson TF, Thompson GR, Denning DW, Fishman JA, Hadley S, Herbrecht R, Kontoyiannis DP, Marr KA, Morrison VA, Nguyen MH, Segal BH, Steinbach WJ, Stevens DA, Walsh TJ, Wingard JR, Young J-AH, Bennett JE. 2016. Practice guidelines for the diagnosis and management of aspergillosis: 2016 update by the Infectious Diseases Society of America. Clin Infect Dis 63:433–442. https://doi.org/10.1093/cid/ciw444.
3. Buil JB, Snelders E, Denardi LB, Melchers WJG, Verweij PE. 2019. Trends in azole resistance in Aspergillus fumigatus, the Netherlands, 1994–2016. Emerg Infect Dis 25:176–178. https://doi.org/10.3201/eid2501.171925.
4. Steinmann J, Hamprecht A, Vehreschild MJ, Cornely OA, Buchheidt D, Spiess B, Koldehoff M, Buer J, Meis JF, Rath PM. 2015. Emergence of azole-resistant invasive aspergillosis in HSCT recipients in Germany. J Antimicrob Chemother 70:1522–1526. https://doi.org/10.1093/jac/dku566.
5. Chowdhary A, Kathuria S, Xu J, Meis JF. 2013. Emergence of azole-resistant Aspergillus fumigatus strains due to agricultural azole use creates an increasing threat to human health. PLoS Pathog 9:3–7. https://doi.org/10.1371/annotation/4ffcf1da-b180-4149-834c-9c723c5dbf9b.
6. Chowdhary A, Kathuria S, Xu J, Sharma C, Sundar G, Singh PK, Gaur SN, Hagen F, Klaassen CH, Meis JF. 2012. Clonal expansion and emergence of environmental multiple-triazole-resistant Aspergillus fumigatus strains carrying the TR(3)(4)/L98H mutations in the cyp51A gene in India. PLoS One 7:e52871. https://doi.org/10.1371/journal.pone.0052871.
7. Meis JF, Chowdhary A, Rhodes JL, Fisher MC, Verweij PE. 2016. Clinical implications of globally emerging azole resistance in Aspergillus fumigatus. Philos Trans R Soc Lond B Biol Sci 371:20150460. https://doi.org/10.1098/rstb.2015.0460.
8. Ballard E, Melchers WJG, Zoll J, Brown AJP, Verweij PE, Warris A. 2018. In-host microevolution of Aspergillus fumigatus: a phenotypic and genotypic analysis. Fungal Genet Biol 113:1–13. https://doi.org/10.1016/j.fgb.2018.02.003.
9. Hare RK, Gertsen JB, Astvad KMT, Degn KB, Lokke A, Stegger M, Andersen PS, Kristensen L, Arendrup MC. 2019. In vivo selection of a unique tandem repeat mediated azole resistance mechanism (TR120) in Aspergillus fumigatus cyp51A, Denmark. Emerg Infect Dis 25:577–580. https://doi.org/10.3201/eid2503.180297.
10. Berger S, El Chazli Y, Babu AF, Coste AT. 2017. Azole resistance in Aspergillus fumigatus: a consequence of antifungal use in agriculture? Front Microbiol 8:1024. https://doi.org/10.3389/fmicb.2017.01024.
11. Verweij PE, Snelders E, Kema GHJ, Mellado E, Melchers WJG. 2009. Azole resistance in Aspergillus fumigatus: a side-effect of environmental fungicide use? Lancet Infect Dis 9:789–795. https://doi.org/10.1016/S1473-3099(09)70265-8.
12. Morton V, Staub T. 2008. A short history of fungicides. APSnet Features https://doi.org/10.1094/APSnetFeature-2008-0308.
13. Maertens JA. 2004. History of the development of azole derivatives. Clin

Microbiol Infect 10(Suppl 1):1–10. https://doi.org/10.1111/j.1470-9465.2004.00841.x.

14. Price CL, Parker JE, Warrilow AG, Kelly DE, Kelly SL. 2015. Azole fungicides– understanding resistance mechanisms in agricultural fungal pathogens. Pest Manag Sci 71:1054–1058. https://doi.org/10.1002/ps.4029.
15. FRAC Code List. 2019. Fungal control agents sorted by cross resistance pattern and mode of action. Fungicide Resistance Action Committee, Basel, Switzerland.
16. Fisher MC, Hawkins NJ, Sanglard D, Gurr SJ. 2018. Worldwide emergence of resistance to antifungal drugs challenges human health and food security. Science 360:739–742. https://doi.org/10.1126/science.aap7999.
17. Mellado E, Garcia-Effron G, Alcazar-Fuoli L, Melchers WJ, Verweij PE, Cuenca-Estrella M, Rodriguez-Tudela JL. 2007. A new Aspergillus fumigatus resistance mechanism conferring in vitro cross-resistance to azole antifungals involves a combination of cyp51A alterations. Antimicrob Agents Chemother 51:1897–1904. https://doi.org/10.1128/AAC.01092-06.
18. Liu M, Zheng N, Li D, Zheng H, Zhang L, Ge H, Liu W. 2016. cyp51A-based mechanism of azole resistance in Aspergillus fumigatus: illustration by a new 3D structural model of Aspergillus fumigatus CYP51A protein. Med Mycol 54:400–408. https://doi.org/10.1093/mmy/myv102.
19. Nash A, Rhodes J. 2018. Simulations of CYP51A from Aspergillus fumigatus in a model bilayer provide insights into triazole drug resistance. Med Mycol 56:361–373. https://doi.org/10.1093/mmy/myx056.
20. Zhang J, Snelders E, Zwaan BJ, Schoustra SE, Meis JF, van Dijk K, Hagen F, van der Beek MT, Kampinga GA, Zoll J, Melchers WJG, Verweij PE, Debets AJM. 2017. A novel environmental azole resistance mutation in Aspergillus fumigatus and a possible role of sexual reproduction in its emergence. mBio 8:e00791-17. https://doi.org/10.1128/mBio.00791-17.
21. Vermeulen E, Maertens J, Schoemans H, Lagrou K. 2012. Azole-resistant Aspergillus fumigatus due to TR46/Y121F/T289A mutation emerging in Belgium, July 2012. Euro Surveill 17:20326.
22. Arendrup MC, Verweij PE, Mouton JW, Lagrou K, Meletiadis J. 2017. Multicentre validation of 4-well azole agar plates as a screening method for detection of clinically relevant azole-resistant Aspergillus fumigatus. J Antimicrob Chemother 72:3325–3333. https://doi.org/10.1093/jac/dkx319.
23. Buil JB, van der Lee HAL, Rijs A, Zoll J, Hovestadt J, Melchers WJG, Verweij PE. 2017. Single-center evaluation of an agar-based screening for azole resistance in Aspergillus fumigatus by using VIPcheck. Antimicrob Agents Chemother 61:e01250-17. https://doi.org/10.1128/AAC.01250-17.
24. Snelders E, Camps SMT, Karawajczyk A, Schaftenaar G, Kema GHJ, van der Lee HA, Klaassen CH, Melchers WJG, Verweij PE. 2012. Triazole fungicides can induce cross-resistance to medical triazoles in Aspergillus fumigatus. PLoS One 7:e31801. https://doi.org/10.1371/journal.pone.0031801.
25. Garcia-Rubio R, Monzon S, Alcazar-Fuoli L, Cuesta I, Mellado E. 2018. Genome-wide comparative analysis of Aspergillus fumigatus strains: the reference genome as a matter of concern. Genes 9:363. https://doi.org/10.3390/genes9070363.
26. Bowyer P, Denning DW. 2014. Environmental fungicides and triazole resistance in Aspergillus. Pest Manag Sci 70:173–178. https://doi.org/10.1002/ps.3567.
27. Rhodes J. 2019. Rapid worldwide emergence of pathogenic fungi. Cell Host Microbe 26:12–14. https://doi.org/10.1016/j.chom.2019.06.009.
28. Resendiz Sharpe A, Lagrou K, Meis JF, Chowdhary A, Lockhart SR, Verweij PE, ISHAM/ECMM Aspergillus Resistance Surveillance Working Group. 2018. Triazole resistance surveillance in Aspergillus fumigatus. Med Mycol 56:83–92. https://doi.org/10.1093/mmy/myx144.
29. Bader O, Weig M, Reichard U, Lugert R, Kuhns M, Christner M, Held J, Peter S, Schumacher U, Buchheidt D, Tintelnot K, Gross U, MykoLabNet DP, MykoLabNet-D Partners. 2013. cyp51A-based mechanisms of Aspergillus fumigatus azole drug resistance present in clinical samples from Germany. Antimicrob Agents Chemother 57:3513–3517. https://doi.org/10.1128/AAC.00167-13.
30. Bader O, Tünnermann J, Dudakova A, Tangwattanachuleeporn M, Weig M, Groß U, Hoberg N, Geibel S, Vogel E, Büntzel J, Springer J, Lehning LY, Schädel C, Antweiler E, Metzger L, Zautner A, Buchheidt D, Spiess B, Hamprecht A, Steinmann J, Rößler S, Wiegmann S, Klingebiel S, Loock AC, Hegewald J, Hassenpflug M, Aurin A, Szymczak J, Diffloth N, Kuhns M. 2015. Environmental isolates of azole-resistant Aspergillus fumigatus in Germany. Antimicrob Agents Chemother 59:4356–4359. https://doi.org/10.1128/AAC.00100-15.
31. Alvarez-Moreno C, Lavergne RA, Hagen F, Morio F, Meis JF, Le Pape P. 2019. Fungicide-driven alterations in azole-resistant Aspergillus fumigatus are related to vegetable crops in Colombia, South America. Mycologia 111:217–224. https://doi.org/10.1080/00275514.2018.1557796.
32. Prigitano A, Venier V, Cogliati M, De Lorenzis G, Esposto MC, Tortorano AM. 2014. Azole-resistant aspergillus fumigatus in the environment of Northern Italy, May 2011 to June 2012. Eurosurveillance 19:1–7. https://doi.org/10.2807/1560-7917.ES2014.19.12.20747.
33. Mortensen KL, Mellado E, Lass-Flörl C, Rodriguez-Tudela JL, Johansen HK, Arendrup MC. 2010. Environmental study of azole-resistant Aspergillus fumigatus and other aspergilli in Austria, Denmark, and Spain. Antimicrob Agents Chemother 54:4545–4549. https://doi.org/10.1128/AAC.00692-10.
34. Sewell TR, Zhang Y, Brackin AP, Shelton JMG, Rhodes J, Fisher MC. 2019. Elevated prevalence of azole resistant Aspergillus fumigatus in urban versus rural environments in the United Kingdom. Antimicrob Agents Chemother 63:e00548-19. https://doi.org/10.1128/AAC.00548-19.
35. Jeanvoine A, Rocchi S, Reboux G, Crini N, Crini G, Millon L. 2017. Azole-resistant Aspergillus fumigatus in sawmills of eastern France. J Appl Microbiol 123:172–184. https://doi.org/10.1111/jam.13488.
36. Cox KD, Bryson PK, Schnabel G. 2007. Instability of propiconazole resistance and fitness in Monilinia fructicola. Phytopathology 97:448–453. https://doi.org/10.1094/PHYTO-97-4-0448.
37. Calo S, Shertz-Wall C, Lee SC, Bastidas RJ, Nicolas FE, Granek JA, Mieczkowski P, Torres-Martinez S, Ruiz-Vazquez RM, Cardenas ME, Heitman J. 2014. Antifungal drug resistance evoked via RNAi-dependent epimutations. Nature 513:555–558. https://doi.org/10.1038/nature13575.
38. Schoustra SE, Debets AJM, Rijs A, Zhang J, Snelders E, Leendertse PC, Melchers WJG, Rietveld AG, Zwaan BJ, Verweij PE. 2019. Environmental hotspots for azole resistance selection of Aspergillus fumigatus, the Netherlands. Emerg Infect Dis 25:1347–1353. https://doi.org/10.3201/eid2507.181625.
39. Abdolrasouli A, Rhodes J, Beale MA, Hagen F, Rogers TR, Chowdhary A, Meis JF, Armstrong-James D, Fisher MC. 2015. Genomic context of azole resistance mutations in Aspergillus fumigatus determined using whole-genome sequencing. mBio 6:e00536. https://doi.org/10.1128/mBio.00536-15.
40. Dunne K, Hagen F, Pomeroy N, Meis JF, Rogers TR. 2017. Intercountry transfer of triazole-resistant Aspergillus fumigatus on plant bulbs. Clin Infect Dis 65:147–149. https://doi.org/10.1093/cid/cix257.
41. Alvarez-Moreno C, Lavergne RA, Hagen F, Morio F, Meis JF, Le Pape P. 2017. Azole-resistant Aspergillus fumigatus harboring TR34/L98H, TR46/Y121F/T289A and TR53 mutations related to flower fields in Colombia. Sci Rep 7:45631. https://doi.org/10.1038/srep45631.
42. Gisi U. 2014. Assessment of selection and resistance risk for demethylation inhibitor fungicides in Aspergillus fumigatus in agriculture and medicine: a critical review. Pest Manag Sci 70:352–364. https://doi.org/10.1002/ps.3664.
43. Guinea J, Verweij PE, Meletiadis J, Mouton JW, Barchiesi F, Arendrup MC, Subcommittee on Antifungal Susceptibility Testing of the EECfAST. 2018. How to: EUCAST recommendations on the screening procedure E.Def 10.1 for the detection of azole resistance in Aspergillus fumigatus isolates using four-well azole-containing agar plates. Clin Microbiol Infect 25:681–687. https://doi.org/10.1016/j.cmi.2018.09.008.
44. Basenko EY, Pulman JA, Shanmugasundram A, Harb OS, Crouch K, Starns D, Warrenfeltz S, Aurrecoechea C, Stoeckert CJ, Jr, Kissinger JC, Roos DS, Hertz-Fowler C. 2018. FungiDB: an integrated bioinformatic resource for fungi and oomycetes. J Fungi 4:39. https://doi.org/10.3390/jof4010039.
45. Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760. https://doi.org/10.1093/bioinformatics/btp324.
46. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20:1297–1303. https://doi.org/10.1101/gr.107524.110.
47. Boeva V, Popova T, Bleakley K, Chiche P, Cappo J, Schleiermacher G, Janoueix-Lerosey I, Delattre O, Barillot E. 2012. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. Bioinformatics 28:423–425. https://doi.org/10.1093/bioinformatics/btr670.
48. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Hand-

saker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, Genomes Project Analysis Group. 2011. The variant call format and VCFtools. Bioinformatics 27:2156–2158. https://doi.org/10.1093/bioinformatics/btr330.

49. Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797. https://doi.org/10.1093/nar/gkh340.
50. Price MN, Dehal PS, Arkin AP. 2010. FastTree 2–approximately maximum-likelihood trees for large alignments. PLoS One 5:e9490. https://doi.org/10.1371/journal.pone.0009490.
51. Letunic I, Bork P. 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res 47:W256–W259. https://doi.org/10.1093/nar/gkz239.
52. Korneliussen TS, Moltke I, Albrechtsen A, Nielsen R. 2013. Calculation of Tajima's D and other neutrality test statistics from low depth next-generation sequencing data. BMC Bioinformatics 14:289. https://doi.org/10.1186/1471-2105-14-289.
53. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, Cho EK, Dallaire S, Freeman JL, Gonzalez JR, Gratacos M, Huang J, Kalaitzopoulos D, Komura D, MacDonald JR, Marshall CR, Mei R, Montgomery L, Nishimura K, Okamura K, Shen F, Somerville MJ, Tchinda J, Valsesia A, Woodwark C, Yang F, Zhang J, Zerjal T, Zhang J, Armengol L, Conrad DF, Estivill X, Tyler-Smith C, Carter NP, Aburatani H, Lee C, Jones KW, Scherer SW, Hurles ME. 2006. Global variation in copy number in the human genome. Nature 444:444–454. https://doi.org/10.1038/nature05329.

Manuscript IV

A pan-genome resembling genome-scale metabolic model platform of 252 *Aspergillus fumigatus* strains reveals growth dependencies from the lung microbiome

Mohammad H. Mirhakkak[*,1], Xiuqiang Chen[*,1], Tongta Sae-Ong[1], Lin Lin Xu[1], Thorsten Heinekamp[2], Oliver Kurzai[3,4,5], Amelia Barber[6], Axel Brakhage[2,7], Sebastien Boutin[8,9], Sascha Schäuble[#,1], Gianni Panagiotou[#,1,10]

Overview

In manuscript IV, we aimed to understand the similarities and differences of *A. fumigatus* genomes at the metabolic level. Therefore, genome-scale metabolic models of *A. fumigatus* were constructed based on their *de novo* assembled genomes. The results showed the strong point of metabolic differences between environmental and clinical strains, which improved our understanding of the evolutionary fitness of *A. fumigatus* clinical strains in human lungs.

FORM 1

Manuscript No. 4

Manuscript title: A pan-genome resembling genome-scale metabolic model platform of 252 *Aspergillus fumigatus* strains reveals growth dependencies from the lung microbiome

Authors: Mohammad H. Mirhakkak*, Xiuqiang Chen*, **Tongta Sae-Ong**, Lin Lin Xu, Thorsten Heinekamp, Oliver Kurzai, Amelia Barber, Axel Brakhage, Sebastien Boutin, Sascha Schäubl, Gianni Panagiotou

Bibliographic information:

Mirhakkak, M.H.*, Chen, X.*, **Sae-Ong, T.**, Xu, L., Heinekamp, T., Kurzai, O., Barber, A., Brakhage, A., Boutin, S., Schäuble, S., Panagiotou, G., A pan-genome resembling genome-scale metabolic model platform of 252 *Aspergillus fumigatus* strains reveals growth dependencies from the lung microbiome. (in preparation)

The candidate is (Please tick the appropriate box.)

☐ First author, ☐ Co-first author, ☐ Corresponding author, ☒ Co-author.

Status: in preparation

Authors' contributions (in %) to the given categories of the publication

Author	Conceptual	Data analysis	Experimental	Writing the manuscript	Provision of material
Mirhakkak, M.H.*	15%	40%		25%	
Chen, X.*	15%	40%		25%	
Sae-Ong, T.		10%		5%	
Xu, L.L.		10%		5%	
Schäuble, S.	35%			20%	20%
Panagiotou, G.	35%			20%	40%
Others			100%		40%
Total:	100%	100%	100%	100%	100%

*Authors contributed equally

Signature candidate

Signature supervisor (member of the Faculty)

A pan-genome resembling genome-scale metabolic model platform of 252 *Aspergillus fumigatus* strains reveals growth dependencies from the lung microbiome

Mohammad H. Mirhakkak*[,1], Xiuqiang Chen*[,1], Tongta Sae-Ong[1], Lin Lin Xu[1], Thorsten Heinekamp[2], Oliver Kurzai[3,4,5], Amelia Barber[6], Axel Brakhage[2,7], Sebastien Boutin[8,9], Sascha Schäuble#[,1], Gianni Panagiotou#[,1,10]

Affiliations

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

[2] Department of Molecular and Applied Microbiology, Leibniz Institute for Natural Product Research and Infection Biology, Jena 07745, Germany

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

[4] Research Group Fungal Septomics, Leibniz Institute of Natural Product Research and Infection Biology–Hans Knöll Institute, Jena, Germany

[5] National Reference Center for Invasive Fungal Infections (NRZMyk), Leibniz Institute of Natural Product Research and Infection Biology–Hans Knöll Institute, Jena, Germany

[6] Junior Research Group Fungal Informatics, Leibniz Institute of Natural Product Research and Infection Biology–Hans Knöll Institute, Jena, Germany

[7] Institute of Microbiology, Friedrich Schiller University Jena, Jena 07745, Germany

[8] Department of Infectious Diseases, Medical Microbiology and Hygiene, University Hospital Heidelberg, Im Neuenheimer Feld 324, 69120, Heidelberg, Germany

[9] Translational Lung Research Center Heidelberg (TLRC), German Center for Lung Research (DZL), University of Heidelberg, Heidelberg, Germany

[10] Department of Medicine and State Key Laboratory of Pharmaceutical Biotechnology, University of Hong Kong, Hong Kong, China

Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft (DFG - www.dfg.de) CRC/Transregio 124 "'Pathogenic fungi and their human host: Networks of interaction" (DFG project number 210879364, subproject INF to MM, SS and GP) and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2051 – Project-ID 390713860 (AB, LR, AB and GP). We thank Yueqiong Ni for fruitful discussions and proofreading our manuscript.

*These authors have contributed equally

#Correspondence to Sascha.Schaeuble@leibniz-hki.de and Gianni.Panagiotou@leibniz-hki.de

Working title: Aspergillus fumigatus strain-GEM collection

Abstract

The saprotrophic fungus *Aspergillus fumigatus* is an opportunistic human fungal pathogen and one of the most common causes of infectious death in immunocompromised patients. Here, we present a collection of 252 strain-specific, genome-scale metabolic models (GEMs) of this important fungal pathogen to study and better understand the metabolic component of its pathogenic versatility. Metabolism showed a notable accessory reactome of 22.7%, which was mainly associated to amino acid, but also nucleotide, and nitrogen metabolism. Presence of reactions and feasible reaction fluxes supporting fungal growth were sufficient to differentiate environmental from clinical strain origin. In addition, shotgun metagenomics of sputum from 40 cystic fibrosis patients before and after they were diagnosed with an *A. fumigatus* infection suggests that the fungus shapes the lung microbiome towards a more beneficial fungal growth environment associated with glycine/serine biosynthesis and the shikimate pathway. Taken together, the here presented first collection of *A. fumigatus* strain-GEMs highlights metabolic differences between different strains of environmental and clinical origin and improves our understanding of fungal survival in the non-native environment of the human lung. These may serve as starting points for the development of alternative clinical intervention strategies targeting the fungal metabolic needs for survival and colonization by diet modification or microbiome intervention..

Introduction

Fungal infections are an emerging public concern for both human health care and economics (*1*, *2*). *Aspergillus fumigatus* is a globally occurring environmental saprotrophic mold that poses a serious threat to hospitalized, particularly immunocompromised patients (*3*). It affects more than 1 million people annually with invasive aspergillosis (IA) and 3 million with chronic pulmonary aspergillosis, both of which show high mortality rates especially in vulnerable cohorts, while diagnostics remain challenging (https://gaffi.org/why/fungal-disease-frequency/, April 2022). On top, the prevalence of chronic obstructive pulmonary disease (COPD) appears to be much higher than estimated with IA contributing substantially to fatal disease progression (*4*), while *A. fumigatus* is related to as many as half of the worldwide cystic fibrosis cases (*5*).

It remains largely unknown to which extent environmental and clinical *A. fumigatus* isolates possess distinct characteristics to cope with external stresses or accessible nutrient profiles in challenging environments such as the human lung. Recently, we explored the genetic diversity of *A. fumigatus* to reveal a remarkably low fraction of core genes shared by all members of the species (69% of the total genes identified) (*6*). However, how the genetic diversity of *A. fumigatus* influences phenotypic and metabolic heterogeneity, particularly in their ability to thrive in the non-native niche of the human lung, has not been addressed yet.

One promising approach to study metabolic capabilities and growth dependencies of pathogens is the application of genome-scale metabolic model (GEM) reconstruction and analysis(*7*). We have previously applied GEM analysis to reveal gut microbiome species that influence colonization levels of the opportunistic fungal pathogen *Candida albicans* (*8*). Given the exponentially increasing number of available genome sequences, the reconstruction of multi-strain genome-scale metabolic models is now possible. The first multi-strain-GEM collection of *Escherichia coli* enabled the definition of strain-specific adaptation to nutrition availability and the prediction of nutritional auxotrophies in some strains (*9*). Protocols and databases were consequently updated to allow for bacterial GEM reconstruction at strain resolution (*10*, *11*), while reconstructions of multi-strain-GEMs remain to be explored in eukaryotes.

In this study we provide the first non-bacterial, multi-strain-GEM reconstruction using *A. fumigatus* as a fungal model organism. Defining metabolic differences between 252 environmental and clinical strain-specific GEMs allowed us to identify metabolic reactions that differ between the two populations. Subsequently, we performed shotgun metagenomics on sputum from 40 cystic fibrosis patients before and after they were diagnosed with *A. fumigatus* infection. By computationally defining the metabolic output of the lung microbiome, we propose that the presence of *A. fumigatus* shapes the metabolic landscape of the lung microbiome in a manner favorable for fungal growth. Resolving the impact of genetic diversity on *A. fumigatus* metabolism appears important to extend our understanding of adaptation mechanisms particularly with respect to aromatic amino acid metabolism involving the Shikimate pathway that can ultimately guide the development of new antifungal therapies.

Results

Reconstruction of a most comprehensive *Aspergillus fumigatus* pan-GEM

To create a template for subsequent strain-specific GEM design, we first derived a comprehensive pan-GEM for *A. fumigatus* metabolism (Fig. 1A). To start, we combined available draft reconstructions for *A. fumigatus* with seven automatically derived draft reconstructions for different *Aspergillus spp.* (see Methods for details) (*12*, *13*). This approach allowed us to acquire as many as possible *Aspergillus*-associated reactions in the core metabolism of *A. fumigatus* (i.e. metabolic reactions present in all strains). It also allowed us to acquire a more comprehensive catalogue of optional accessory metabolic reactions by defining strain subset diversity enabling the subsequent strain-specific gap filling curation steps (Fig. 1A). In total, this first draft model was comprised of 7,606 reactions (of which 3,233 are responsible for metabolite exchange with simulated environment) and 3,578 metabolites.

Next, we adapted 62 metabolic components based on fungal and particularly *A. fumigatus* specific literature information to create the biomass objective function essential to simulate *A. fumigatus* growth rates (see Methods) (*14*). The largest fractions of the derived biomass function included carbohydrates and proteins (42.8% and 30%, respectively). Additional essential components included lipids, DNA, and energetic co-factors (Fig. 1B, Supplementary Table S1).

Subsequently, we screened available *A. fumigatus* gene information relevant for metabolism and added 1,453 genes and 2,003 corresponding gene to reaction rules for metabolic reactions, as defined by KEGG (https://www.kegg.jp/) or MetaCyc (https://metacyc.org/, Methods). The remaining 2,370 metabolic reactions (excluding exchange reactions) could not be mapped to any gene in our pan-GEM draft model and were removed from the generic pan-GEM accordingly. However, these reactions were retained for subsequent strain-specific refinement steps, which require accessory information including gap-filling of fragmented metabolic pathways (Fig. 1A). Concurrently, we incorporated reaction-to-pathway association information from both KEGG and MetaCyc. The broadest pathway categories included amino acids and carbohydrates (Fig. 1C). For the 2,003 metabolic reactions with gene annotation, we predicted nine compartments in our pan-GEM using WoLF PSORT (Fig. 1D, see Methods) (*15*). In parallel, we identified and resolved erroneous energy-generating cycles (*16*) by correcting or removing thermodynamically implausible reactions, such as cases where free energy dissipation was diminished.

For the final curation of our pan-GEM, we generated phenotypic growth data for *A. fumigatus* wild type (Af293 strain) and five mutant strains affecting nitrogen or carbon metabolic components, and considered publicly available gene essentiality information (*17*) (Supplementary Table S2). The initial agreement of our pan-GEM to our metabolite specific growth data was already good (Fig. 1E). To optimize the simulation accuracy of our pan-GEM, we manually resolved any incompatibility between our growth data, available gene essentiality data, and our *in silico* model predictions. These curation efforts improved growth simulation accuracy from 58% to 84% for all tested carbon sources and improved nitrogen growth simulation accuracy from 55% to 85% (Fig. 1E). The pan-GEM achieved 79% and 65% compatibility for the tested phosphorus and sulfur sources respectively and reached 83% if we neglected sulfur source growth data for *ΔniaD* and *ΔlysF* (Fig. 1E) (see Methods). This final model also reached 75% agreement with the available gene essentiality data (Fig. 1F). Altogether, our final pan-GEM of *A. fumigatus* was comprised of 1,453 genes, 3,882 reactions and 4,170 metabolites distributed across 9 compartments. Of these, 3,051 metabolic reactions and 1,957 metabolites were unique across all compartments.

A. fumigatus strains show notable accessory reaction content

Using a genomic dataset of 252 *A. fumigatus* strains from Germany (203 environmental and 49 clinical strains) that we generated previously (*6*), we mapped strain-specific gene profiles to the reference pan-GEM and subsequently derived strain-specific GEMs (Supplementary Table S3). For all strain-specific GEMs, we ensured viable growth was predicted in minimal media with glucose as the carbon source by identifying and resolving minimal sets of essential reactions (*18*) and crosschecking against blocked reactions with

FASTCC (see Methods) (*19*). Model size varied in the different *A. fumigatus* strains from 1,366 to 1,455 reactions (mean 1,413).

Although all strain-specific GEMs are derived from *A. fumigatus*, we found a strikingly low number of core metabolic components shared by all GEMs. In line with the considerably high genome diversity of this organism (*6*), only 984 metabolic genes (69.8%) and 1,150 metabolic reactions (77.3%) were shared by all strain-specific GEMs, resulting in a large degree of metabolic variation across all GEMs (426 accessory genes and 338 accessory reactions). Most accessory content was involved in nucleotide, energy (including oxidative phosphorylation and nitrogen metabolism) and amino acid metabolic pathways (Fig. 2A). Only 56%, 63% and 68% of all reactions included in the strain-GEMs for these pathways, respectively, were conserved across all strain models, demonstrating considerable metabolic pathway variation between strains (Fig. 2A). The majority of the accessory content (70% of accessory genes, 77% of accessory reactions, Table 1, Fig. 2B) was shared by more than 80% of all strain-GEMs. We previously observed that one genetic lineage of *A. fumigatus* possessed significantly fewer accessory genes than the other lineages, including notably fewer metabolic accessory genes (*6*) (Supplementary Fig. S1). In contrast, metabolic reaction content in the strain-GEMs did not show a reduced number of metabolic reactions in this lineage, demonstrating the presence of redundancy among metabolic accessory genes (Fig. 2B). Finally, a small, but notable amount of reactions appeared in at most 40% of all strain-GEMs (Table 1, Fig. 2A) comprising mostly reactions of amino acid metabolism, but also of lipid and energy metabolism including but not limited to nitrogen dependent chorismate pyruvate-lyase or nicotinamidase and acyl-CoA dependent acyltransferases.

Taken together, our generated 252 strains showed notable accessory content and therefore potential metabolic diversity among the strains as well as metabolic robustness despite reduced accessory metabolically relevant genes.

Metabolic activity of 25 reactions allows differentiation between environmental and clinical strain-GEMs

When calculating the pairwise Jaccard distance, we found that strain-specific GEMs differed by at most 15% (Fig. 2C). Neither accessory reaction information nor Jaccard distance allowed discriminating metabolic capabilities between environmental and clinical strains (Fig. 2B, C). However, we identified eight metabolic reactions present primarily in either environmental or clinical strain-GEMs that, when taken together, were able to significantly differentiate the two populations (exact Fisher-test, $p<0.05$, Fig. 2D). In agreement with the statistical significance of these eight reactions, decision tree machine learning (ML) using the presence or absence of these metabolic reactions, as well as the capability of the strains to grow on different minimal media compositions required only a few steps to correctly categorize 216 out of 252 strains (86%, Fig. 2E). Four of these reactions were chorismate dependent and involved chorismate lyase activity that generates 4-hydroxybenzoate and pyruvate from chorismate. Chorismate lyase activity is linked to differential activity in the shikimate pathway, which has been associated with virulence in *A. fumigatus* (*20*, *21*). Interestingly, the ability to convert chorismate and glutamine to anthranilate, pyruvate and glutamate, as well as, of utilizing carbon metabolites including methionine, acetate, succinate and also thioredoxin was sufficient for the strain origin classification and yielded complementary metabolic discriminators to the sole presence/absence statistical analysis of metabolic reactions in our strain-GEM collection (Fig. 2C, D).

Given that only a few metabolic reactions were sufficient to differentiate strains from clinical and environmental origin by statistical and decision tree analysis, we further explored whether reaction fluxes between strain-GEMs could be used to further improve differentiation of strain origin. We analyzed feasible reaction flux ranges for all strain-GEMs by simulating each on minimal media including glucose as a carbon source and calculating growth supporting flux ranges using flux variability analysis (FVA) (*22*). The derived flux ranges were subsequently used as the input for ML-based classification (see Methods). Classifying environmental from clinical strains achieved an accuracy of 0.80 (AUC = 0.72) with information from only 25 reactions (Fig. 2F, Supplementary Table S4). In addition to previously highlighted chorismate-associated reactions, the ML-model also selected features associated with amino acid reactions, such as homoserine succinate-lyase or L-methionine:oxidized-thioredoxin S-oxidoreductase or cystathionine gamma-lyase, suggesting aromatic amino acid metabolism as a differentiating factor of clinical and environmental *A. fumigatus* strains.

Taken together, we did not observe major differences in strain origin given the strain's accessory gene or reaction content (Fig. 2B, Supplementary Fig. S1) or complete metabolic reaction presence (Fig. 2C). In contrast, we identified a small defined set of reactions that mainly associate to amino acid and chorismate metabolic acitivity which are sufficient to differentiate clinical from environmental origin to a large extend (Fig. 2D-F).

The structure of the lung microbiome changes upon *A. fumigatus* colonization

To investigate the applicability of the GEMs from clinical origin for the prediction of metabolic components supporting *A. fumigatus* growth in the human lung, we analyzed sputum samples from 40 cystic fibrosis patients in Germany (*cf.* Methods for cohort description). For all patients, we had an initially culture-negative sample and a subsequent sample that was positive for *A. fumigatus* growth. To investigate the changes to the lung microbiota after *A. fumigatus* colonization, we performed shotgun metagenomic sequencing for all 80 sputum samples (*A. fumigatus*-negative and positive), generating an average of 5.59 Gbp of sequencing data per sample (*s.d.* 0.80 Gbp). Using Kraken for taxonomic profiling, we identified 228 genera and 598 species from all samples using a relative abundance cut-off of 0.1%. Despite the differences in patient cohort, starting biomaterial, and sequencing method, the taxonomic annotation of the top 10 most abundant genera (Supplementary Table S5) showed striking similarities to two recent studies, where the lung microbiome of *A. fumigatus* infected and control patients was investigated using either sputum samples and or bronchoalveolar lavage and 16S rRNA sequencing (*23*, *24*).

The prevalence of the top abundant genera was consistently high (Fig. 3A). Notably, from the top 10 abundant genera, *Sphingomonas*, *Burkholderia, Stenotrophomonas* and *Pseudomonas* were detected as highly abundant (within the 10 most abundant genera) in 8, 12, 17 and 36 out of 80 samples, (10%, 15%, 21%, and 45%, respectively) showing an uneven distribution in the population (Supplementary Table S5). Similarly, the most prevalent species, making up between 2.5% and 51.3% relative abundance were present in most samples (70%) with the exception of *Burkholderia multivorans* and *Sphingomonas sp. FARSPH*, which had a prevalence of 38.8% and 41.3%, respectively (Supplementary Table S5). Intriguingly, *Pseudomonas aeruginosa* was among the top10 most abundant species in 16 samples before and only 11 samples after *A. fumigatus* infection in the same patients, although this species has been described to commonly outgrow in cystic fibrosis patients and co-occur frequently with *A. fumigatus* colonization (*25*).

Since we did not find statistically different alpha- and beta-diversity (Supplementary Fig. 2A, B), we analyzed species co-abundance networks to further examine the compositional changes of the lung microbiome following *A. fumigatus* colonization. Using differential gene correlation analysis (DGCA), we generated networks from differentially correlated microbial pairs in *A. fumigatus* negative versus *A. fumigatus* positive patients' paired samples (Fig. 3B). We then analyzed the resulting networks using MEGENA (*26*) and identified two notable modules in the global network that contained four differentially abundant species (Metagenomseq, zero-inflated gaussian mixture model p<0.05) between the *A. fumigatus* negative and subsequently positive patients' samples. The interactions of these six species — *Schaalia meyeri*, *Abiotrophia defectiva*, *Pseudomonas fulva, Pseudomonas resinovorans, Pseudomonas sp. S1-A32-2,* and *Haemophilus parahaemolyticus* — were highlighted to emphasize discussion (Fig. 3B). Existing edges (class +/0 in Fig. 3B) of *Pseudomonas sp. S1-A32-2, Abiotrophia defectiva and Haemophilus parainfluenzae* with *Clostridium intestinale, Actinomyces sp. Oral taxon 171* and *Streptococcus sp. oral taxon 431* in *A. fumigatus*-negative samples were lost upon colonization with *A. fumigatus*. Similar patterns were observed in the negative association (-/0) between *Schaalia meyeri* with *Clostridium intestinale, Rhizobium leguminosarum* and *Pseudomonas sp. DY-1*, and also *for Pseudomonas resinovorans, Pseudomonas fulva* and *Pseudomonas sp. S1-A32-2* with *Haemophilus influenzae, Streptococcus sp. Oral taxon 064* and *Streptococcus sp. Oral taxon 431* (magenta edges, Fig. 3E). The associations of *Pseudomonas resinovorans* with *Streptococcus salivarius, and Schaalia odontolytica, Schaalia meyeri, Haemophilus parainfluenzae, Streptococcus intermedius* with *Capnocytophaga endodontalis, Streptococcus sp. oral taxon 064, Veillonella dispar* and *Schaalia meyeri*, respectively, also changed direction in the presence of *A. fumigatus* (0/-, -/0, respectively).

To evaluate the functional implications of microbiome restructuring following *A. fumigatus* colonization, we performed KEGG orthology (KO) enrichment analysis in the identified four modules in our co-abundance networks (see details in Methods). Interestingly, both module 1 and 2 were enriched in amino acid metabolism (e.g. phenylalanine, tyrosine, and tryptophan, but also valine and (iso-)leucine). Further enrichments included propanoate and butanoate metabolism (module 1), folate biosynthesis (module 2), glycan biosynthesis and fatty acid metabolism (module 3) and cyanoamino acid metabolism (module 4, Supplementary Table S5).

In summary, albeit not significantly different with alpha or beta diversity we identified a distinct set of co-abundance differences in the lung microbiome upon *A. fumigatus* colonization. The associated enriched metabolic functions pinpointed again towards amino acid, particularly aromatic amino acid pathways, but also fatty acid, nitrogen and sulfur metabolic pathways, suggesting that lung microbiome metabolic activity is reshaped in the presence of *A. fumigatus*.

A. fumigatus is predicted to contribute to the shaping of the lung microbiome and its metabolic activity to support its own growth

We subsequently investigated whether the changes in the lung microbial community triggered by the presence of *A. fumigatus* were accompanied by changes in the metabolic output of the microbiome. To integrate not only the metabolic output of the microbiome, but also the host's, the pathogen's, as well as additional factors such as dietary molecules, we opted for *in silico* prediction. Towards this aim we derived the most likely lung microbiome metabolic profile supporting the relative abundances of our metagenomics species by growth rate using the MAMBO algorithm (*27*). Briefly, MAMBO iteratively calculates growth rates of bacterial models corresponding to a samples' metagenomic profile and infers a metabolome profile. We found significant differences in the beta-diversity of derived metabolite profiles between patient samples before and after *A. fumigatus* infection (Euclidean distance; PERMANOVA, p=0.03, Fig. 4A, Supplementary Table S6).

We next quantified how the changes in the metabolic output of the lung microbiome following *A. fumigatus* colonization might alter the predicted growth of the *A. fumigatus* clinical strain-GEMs. Using the MAMBO-derived metabolite profiles present after *A. fumigatus* colonization, we observed that the GEMs of the 49 clinical strains showed a significant increase in the predicted growth rate compared to GEMs simulated on the metabolic outputs from before *A. fumigatus* colonization (14% increase, Wilcoxon signed rank test, p=3.55e-15, Fig. 4B) suggesting that the changes induced by *A. fumigatus* in the lung microbiome led to a nutritional profile supporting its own growth.

To explore next whether we can identify a connection between the altered lung microbiome and the metabolic capacity of *A. fumigatus* we analyzed feasible flux ranges of reactions that were associated to enriched metabolic subsystems, which we identified before in the *A. fumigatus* affected CF lung microbiome (Fig. 3B, Supplementary Table S4). We identified 54 metabolic reactions across all *A. fumigatus* clinical GEMs that showed significantly altered lower or upper flux ranges to support fungal growth simulated with FVA on MAMBO derived media before compared to after *A. fumigatus* confirmed colonization (FDR corrected paired Wilcoxon test, p≤0.05, Fig. 4C, Supplementary Table S6). Most filtered reactions showed significant differences in the upper range, which suggests increased metabolic activity of *A. fumigatus* (Fig. 4C). Affected pathways mainly included (aromatic) amino acid metabolism, but also nitrogen, sulfur, butanoate or steroid metabolic pathways (Supplementary Table S6). Although predicted flux ranges overlapped between *A. fumigatus* negative and positive samples, the change of direction is mostly consistent on a per-strain-GEM level (Fig. 4D). Interestingly, the reactions

Tryptamine:oxygen oxidoreductase (EC 1.4.3.4) and Chorismate pyruvate-lyase (EC 4.1.3.27) were identified before as major discriminators between environmental and clinical strains simulated on minimal media (Fig. 2E). Only 17 reactions like showed significant differences in both, lower and upper flux bounds (Supplementary Table S6) with NADPH:oxidized-thioredoxin oxidoreductase (EC 1.8.1.9) showing notably constrained flux bound variability across all simulated strain GEMs.

Finally, reinvestigating our phenotypic microarray data for the clinical strain Af293 wild type we found also a positive growth effect of most amino acids tested as carbon source including the aromatic amino acids phenylalanine, tyrosine and tryptophan (Supplementary Table S2). Intermediates such as chorismate, anthralinate and cholines are not part of commercial microarray platforms, but appeared in multiple minimal media based and cystic fibrosis associated ML classification models. These metabolites pinpoint to an elevated role of the Shikimate pathway and warrant further investigation.

Discussion

In this study, we built the first suite of *A. fumigatus* genome-scale strain-specific metabolic reconstructions originating from 252 environmental and clinical isolates from Germany (*6*). We (i) reconstructed a comprehensive pan-GEM of *A. fumigatus* metabolism in a data-driven manner, which we validated against phenotypic microarray and gene essentiality data; (ii) derived 252 strain-specific GEM models by considering respective genome assemblies and manually curating the strain-specific GEMs towards growth feasibility and minimal fractioned network topologies; and (iii) determined metabolic differences differentiating clinical from environmental strains, such as metabolic reactions involving several amino acids, particularly aromatic amino acids as well as chorismate or thioredoxin. Chorismate is an important precursor for aromatic amino acids and formed in the Shikimate pathway. This seven step pathway is not present in animals and enables the synthesis of aromatic amino acids tyrosine, phenylalanine and tryptophan. Thioredoxin is an important factor for DNA synthase metabolism and was associated to *A. fumigatus* virulence before (*28*, *29*).

Multi-strain-GEMs have been utilized previously to elucidate the metabolic diversity of human-pathogenic bacteria. For example, they have defined the pan metabolic capabilities of *Pseudomonas putida* (*30*), loss of fitness relevant pathways for survival in the gastrointestinal environment in extraintestinal *Salmonella spp.* (*31*), and strain-specific metabolic capabilities in *Staphylococcus aureus* linked to pathogenic traits and virulence acquisitions (*32*). Here we bring this strategy to exploring metabolic diversity in a eukaryotic fungal pathogen for the first time. This strain-specific *A. fumigatus* GEMs platform is publicly available (Biomodels repository) as a platform for investigating the metabolic diversity influencing growth rate capabilities, metabolic adaptation and pathogenicity in this important human fungal pathogen. As a proof-of-concept for the applicability of our fungal GEM collection, we investigated sputum samples from a cohort of 40 cystic fibrosis patients for which samples from before and after confirmed *A. fumigatus* infection were collected. Clinical isolate specific simulations and analysis interestingly showed significantly increased growth rates in the patient samples after a confirmed *A. fumigatus* infection suggesting that the fungus influences the lung microbiome composition towards a more favorable fungal growth. Given these findings, our analyses of strain-resolution GEMs showed that we can recapitulate metabolic cues important for *A. fumigatus* growth that were reported before. Here, we showed that particularly fungal metabolic activity associated to aromatic amino acid metabolism and the Shikimate pathway are not only also important for discriminating environmental from clinical strains, but also to differentiate metabolic activity in clinical *A. fumigatus* strains given metabolite profiles shaped by the lung microbiome in cystic fibrosis patients. Our data-driven analysis highlighted 54 metabolic reactions, for which we predicted significant different flux ranges after *A. fumigatus* colonization of the lung. Our insights suggest that *A. fumigatus* influences its microbiome environment towards a more favorable growth environment. In addition, these reactions do not only appear in aromatic amino acid metabolism, but also sulfur, nitrogen and lipid metabolic pathways, highlighting the advantage of including topological pathway information when analyzing metabolic activity. Together with the 25 metabolic reactions, which we identified as important features of our ML driven classification to differentiate environmental from clinical strains, these reactions represent primarily novel metabolic targets for *A. fumigatus* growth modulation, which need to be investigated further to confirm their potential as biomarker, diagnostic or treatment target with respect to *A. fumigatus* colonization.

As a potential caveat to our study, there may be genomic differences between the clinical strain collection used to build the strain-specific GEMs and the clinical strains present in the cystic fibrosis patients. Though both datasets originate from Germany, the majority of the clinical strains in our GEM collection were from a different patient cohort (invasive aspergillosis). Although genetic diversity differs between strains from invasive and chronic *A. fumigatus* associated disease, we have also shown in our previous study that genomic similarities of clinical strains are relatively high even when the strains originate from different countries (*6*). We used the Af293 *A. fumigatus* genome annotation as a reference, which precludes potentially metabolically relevant genes present in further *A. fumigatus* annotation. In lack of an existing model system we used MAMBO, which relies on taxonomic species annotation to infer the contribution of the lung microbiome to the most likely nutritional profile in the human lung.

Altogether, the presented analyses demonstrated that strain level fungal genome-scale metabolic modeling is feasible and contributes towards our mechanistic understanding of the genome diversity impact on phenotype of *A. fumigatus*. Moreover we could show a pronounced impact of the lung microbiome profile on available nutrition, which appeared to foster *A. fumigatus* colonization levels. Targeting towards patient stewardships involving diets with particularly suboptimal fungal growth compositions and drugging against fungal specific metabolic routes in the context of (aromatic) amino acid biosynthesis and also the Shikimate pathway (*33*), which is unavailable to the human host, appear promising targets that warrant further analyses in patients suffering from pulmonary diseases involving *A. fumigatus* colonization.

Methods

Biomass formulation

We adapted a specific *A. fumigatus* biomass composition according to several literature sources. First, we assigned proportions of main biomass components as described before (*34*) to 38.8% carbohydrates, 9.9% lipids, 30% proteins as well as 0.6% DNA and 3.7% RNA. Since this resource neglected polyols we added 4% polyols as were reported before for *Aspergillus oryzae* (*35*) to a total of 42.8% carbohydrates. After adding a fraction of 6.6% co-factors these main components made up the total biomass composition together with reported 6.4% ash fraction (*34*). Next, we screened the literature to further specify fractions of subcategories for carbohydrates, e.g. glucans or trehalose (*36–38*), lipids, including sterols, phospholipid, neutral lipid and free fatty acid compositions (*39*, *40*), amino acid composition of the protein content (*41*, *42*) and co-factor content including energy carriers, such as NADH or vitamins like riboflavin (*43*, *44*). After calculating the mmol/g content for each fraction, we added the ATP demand according to prior developed models from *Saccharomyces cerevisae* and *Aspergillus niger* (*45*, *46*) as well as added the non-growth associated ATP maintenance value as reported for the curated *Saccharomyces cerevisae* GEM (Supplementary Table S1)(*46*). Finally, we modified the proportion of all components to resemble 1g dry weight (Supplementary Table S1).

Pan-GEM reconstruction

All reconstruction and analysis efforts were done with COBRApy (v0.17.1)(*47*) in python 3.6.8 and the academic version of the IBM CPLEX solver (v12.8.0.0).

We gathered and combined information from automatically generated draft reconstructions based on the CoRoCo pipeline (*12*). We downloaded the Aspergillus CoReCo model for *A. fumigatus* (Biomodels ID MODEL1604280029) and further Aspergillus models from the CoReCo repository including *Aspergillus oryzae* (Biomodels ID MODEL1604280012), *Aspergillus nidulans* (Biomodels ID MODEL1604280008), *Aspergillus niger* (Biomodels ID MODEL1604280021), *Aspergillus clavatus* (Biomodels ID MODEL1604280016), *Aspergillus terreus* (Biomodels ID MODEL1604280019), and *Aspergillus gossypii* (Biomodels ID MODEL1604280044) from the BioModels repository (https://www.ebi.ac.uk/biomodels/). In addition we adapted metabolite and reaction information from a recently published *A. fumigatus* central metabolism model (*13*). Combined together, this yielded a base model consisting of 7,606 reactions of which 3,233 were exchange reactions and 3,578 metabolites. All subsequent curation efforts were tailored towards keeping only reactions, for which annotation information was available or which were necessary to keep the model feasible. By filtering duplicate reactions and metabolites we reduced the model by 73 reactions and 201 metabolites. The biomass formation was modified to given literature on *Aspergillus fumigatus* metabolism and enriched with information from closely related species when we did not find *A. fumigatus* specific information as described in the *Biomass formulation* section. Next, we screened the KEGG (https://www.kegg.jp/) and MetaCyc (https://metacyc.org/) database for gene annotation to *A. fumigatus* metabolism and added 1,453 genes to the model. Whenever available we adopted AND and OR relationships of the genes for metabolic reaction encoding and cross-checked with gene to reaction encoding in the yeast consensus model (*46*). During this step 2,370 reactions could not be mapped to any annotated gene and were therefore removed from the template model. Instead these reactions were kept aside for subsequent gap-filling procedures. Further curation efforts were run in parallel, since any modification influenced different aspects of the curation efforts. This step included compartmentalization, resolving erroneous energy generating cycles (EGCs) (*16*) and gene essentiality information (*17*) as well as adaptation to phenotypic growth assays (cf. Methods section *Biolog phenotypic microarray*, Supplementary Table S2). To add compartment information for all reactions we applied WoLF PSORT subcellular localization prediction (*15*). A reaction was allocated to a particular compartment if more than 50% of the associated genes are predicted to be located in that compartment with more than 50% probability. Reactions, including exchange reaction, were associated to nine compartments accordingly. These included cytoplasm, mitochondrion, nucleus, peroxisome, endoplasmic reticulum, lipid particles, vacuole, golgi and extracellular space. In cases where the prediction was ambiguous or precluded a viable model as measured by biomass production based on defined minimal media we adapted concurrent alternative compartment localization as either predicted by WoLF PSORT or included in the curated *S. cerevisae* GEM (*46*). Compartment-connecting transport reactions were adapted from the yeast consensus model (*46*). A minimal set of additional necessary transport reactions were added by using gap-filling functionality as provided by COBRApy in order to allow biomass precursor production based on minimal media with glucose.

In parallel we resolved again (EGCs)(*16*) and adapted our GEM model to publicly available gene essentiality data (*17*). EGCs are metabolic reactions running in a potentially non-trivial circle without a net-flux except for generating energy carriers. ATP, CTP, GTP, UTP, ITP, NADH, NADPH, FADH2, FMNH2, Acetyl-CoA, L-Glutamate, ubiquinol-8, demethylmenaquinol-8, menaquinol-8 were part in at least one EGC (Supplementary

Table S7). The directions of 44 reactions were refined considering the reaction directionality according to the BiGG (*48*) and BRENDA (*49*) databases and Gibbs free energy of the reactions stated in the MetaCyc database (*50*). Incompatible gene essentiality information, were resolved by either correcting feasible thermodynamically reaction direction or removal of erroneously present reactions without gene annotation.

Finally, we ran several phenotypic microarrays with the *A. fumigatus* reference strain Af293 and five mutant strains (detailed out in the next section) and identified essential carbon, nitrogen, sulfur and phosphorus components (Supplementary Table S2). This step included again plausible correction of thermodynamically feasible reaction directions and removal of present reactions without gene annotation. In case incompatibilities could not be resolved in this way we screened our catalogue of initially removed reactions without gene annotation using gap-filling procedures using the COBRApy gap-fill functionality. Resolving growth compatibility for two of our mutants (*ΔniaD* and *ΔlysF*, cf. *Biolog phenotypic microarray*) on sulfur would have caused a notable performance drop in the overall growth prediction for all investigated growth media and gene essentiality performance. Since growth accuracy on sulfur was very good for the remaining wild type and two mutant strains and because optimizing growth on carbon and nitrogen sources was very good over all mutant data, we refrained from resolving *ΔniaD* and *ΔlysF* sulfur growth accuracy (Fig. 1B).

Strain-GEM reconstruction and curation

Recently, the pan-genome of *A. fumigatus* was derived for 300 environmental and clinical strains from a global distribution (*6*). Mapping the genomes for 252 of these strains to the Af293 *A. fumigatus* reference genome annotation, we identified metabolically relevant genes by requiring at least 95% sequence identity (small deviations from that threshold did not change the results) under the rational that high sequence identity preserves metabolic function. To ensure that all strain-specific GEMs were showing non-zero growth capabilities based on minimal media with glucose as carbon source we identified and resolved minimal sets of essential reactions that needed to operate in adaptation to the minimal cut set concept (*18*). Finally, we guaranteed a consistent network property by identifying and discarding blocked reactions per isolate with FASTCC (*19*).

Biolog phenotypic microarray

Fungal strains were grown at 25°C for 7 days prior to experimental assay on Malt agar supplemented with 5 mM uracil. Mature conidia were harvested by rubbing plates with sterile distilled water and the resulting solution filtered through a 30 µm cell strainer to remove mycelial fragments. The spore solutions were then adjusted to a transmittance of 75%. Phenotypic microarrays were performed using Biolog Phenotypic Microarray plates PM1, PM2, PM3, and PM4 and plates prepared following the manufacture's protocol for filamentous fungi with the modification of 0.16 ml of Biolog Redox Dye D added to the master mix of each plate for the quantification of metabolic activity. The plates were incubated at 37°C for three days and the metabolic activity measured colorimetrically using an OmniLog microplate reader with readings taken every 15 minutes. Experiments were performed in biological duplicates or triplicates (Supplementary Table S2). The phenotypic microarray results were analyzed in R, and statistical comparison was done using Dunnett-type comparison of growth signals of negative control against all the other wells in one plate. All the wells with greater signals than the negative control and p-value < 0.05 were considered as growth cases.

Cystic Fibrosis sample acquisition

This study was approved by the ethics committee of the University of Heidelberg and written informed consent was obtained from all patients or their parents/legal guardians (S-370/2011). Patients were treated according to standard of care (*51*). The diagnosis of cystic fibrosis was verified by established diagnostic criteria (*52*, *53*). Spontaneously expectorated sputum was collected during visits at the Cystic Fibrosis Center at the University Hospital Heidelberg and frozen in liquid nitrogen on the day of visit. Pulmonary function testing was performed on the same day of sputum collection according to ATS/ERS (European Respiratory Society) guidelines (*54*, *55*) and FEV1 (Forced expiratory value in 1s) values were normalized according to the global lung function initiative (*56*).

Metagenomics and subsequent MAMBO analysis

Sputum samples of 40 cystic fibrosis patients were collected before and after they had positive *A. fumigatus* colonization. The cohort comprised 15 females and 25 males (80 samples in total) with age=23.6±4.96 (mean±standard deviation) before *A. fumigatus* infection.

Trimmomatic was used to clip adapter and low-quality bases (v0.36, ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10:1:TRUE, LEADING:3, TRAILING:3, SLIDINGWINDOW:4:15, MINLEN:30). Remaining reads with less than 30 base pairs length were discarded. BWA (v07.17) was used to align quality filtered reads to the human reference genome (hg38). From originally 1.9e+07±2.7e+06 (mean±standard deviation) metagenomic reads 7.8e+05±8.9e+05 remained after preprocessing per sample. To estimate the taxonomic composition of the non-human reads, Kraken2 (v2.0.7,, default parameters) was used with its standard database as reference. Low abundance species were removed at a cut-off=0.1% (Supplementary Table S5). For the functional composition annotation, the MG-RAST (v4.0.3) pipeline was used to assign non-human reads to KEGG pathways. R packages vegan (v2.5) and picante (v1.8.2) were used to calculate alpha diversity with Shannon and Phylogenetic diversity index for each samples on the read counts of species, Statistical differences between samples before (*A. fumigatus* -) and after (*A. fumigatus* -) infection with *A. fumigatus* were obtained by Wilcoxon signed rank test. For beta diversity, R package coda.base (v0.3.1) was used to calculate the pairwise Aitchison distance for samples on the relative abundance of species. Statistical difference between samples before (*A. fumigatus* -) and after (*A. fumigatus* -) infection with *A. fumigatus* was calculated by PERMANOVA.

The abundance network was constructed based on relative abundance values of all detected species (prevalence filter: 10%, abundance filter: 0.1%). DGCA (v2.0.0) was applied to construct the network from differentially correlated microbial pairs in paired cystic fibrosis samples before compared to after *A. fumigatus* infection (empirical p value<0.05). Subsequently, MEGENA (v1.3.7) was used to identify co-expressed modules in the constructed network using significant differing microbial pairs (module p value<0.05). To identify molecular functions, we investigated enrichment of KEGG pathway information (https://www.genome.jp/kegg/pathway.html) by permutation testing to determine whether correlations between modules and KOs were possible by chance or not (*57*). Firstly, for a given module all correlation values and p values between a particular KEGG Orthology and all species in this module were obtained using the spearman correlation methodThe sum of absolute correlation values in this module was then calculated. Following that, the same number of species in the module were chosen at random 1000 times from all species, and the sum of absolute values of every correlation was calculated for each set. Finally, the sum of significant correlation values in a given module was evaluated whether it was higher than in 95% of the sums of significant correlation values in the repeated random selected species.

To associate the most likely metabolite abundance profile to our metagenomic samples we applied the MAMBO algorithm (*27*). In brief, MAMBO optimizes a high correlating metabolic profile to a given metagenomic relative abundance profile based on bacterial GEMs associated to a given metagenomic sample. We opted for using only GEMs associated to species from the metagenomics profile at an abundance threshold of

0.5% to ensure that the abstracted media is associated to dominant species in the lung. 53 bacteria species were present beyond this more rigid threshold for which we found and downloaded 51 matching bacterial GEMs from the AGORA (https://vmh.life)(*58*) and CarveMe collection (https://github.com/cdanielmachado/embl_gems/tree/master/models) (*59*). Optimizations were run in a python environment (v3.7) using a HPC (192 cores, 1TB RAM). After removing metabolites that appeared in less than 80% of the samples, missing metabolite abundance values in any remaining sample were imputed with MICE (miceRanger, v1.4.0 with m=1 and maxiter = 50) resulting in a final list of 357 metabolites (Supplementary Table S6). Metabolites differing significantly in the MAMBO associated media for samples before compared to after *A. fumigatus* infected were identified in three steps. Firstly, candidate metabolites were selected using $p<0.2$ as cut-off from a Wilcoxon signed rank test (*60*). Secondly, the identified metabolites in the first step were investigated with an adaptive Lasso statistical design using R package glmnet (v4.1) to identify important metabolites for group differentiation (family="binomial", type.measure="class")(*61*). Finally, we used a fixed Lasso design using R package selectiveInference (v1.2.5) as post-selection inference method to identify significance for each of the important metabolites ($p\leq0.05$) (*62*).

Machine learning approach

Unless otherwise noted we used the following machine learning methodology. In cases where the group sizes were unbalanced (e.g. environmental and clinical origin labels) we randomly sampled 50% of the majority group and oversampled samples of the minority group using ADASYN implemented in R package imbalance (v1.0.2.1). Subsequently, feature selection was performed using Boruta (v7.0.0), VSURF (v1.1.0), MUVR (v0.0.973) and sPLS-DA (mixOmics, v6.16.0). These steps were repeated 50 times and selected features as well as their selection frequency recorded. Finally, the Extra Trees algorithm from PyCaret (v2.3.2) was run for different feature sets scanning different frequency cut-offs to optimize the best cut-off value for ML performance. The best hyperparameters of the Extra Trees model was automatically selected by scikit-optimize (v0.8.1, bayesian optimization).

References

1. M. C. Fisher, D. A. Henk, C. J. Briggs, J. S. Brownstein, L. C. Madoff, S. L. McCraw, S. J. Gurr, Emerging fungal threats to animal, plant and ecosystem health. *Nature*. **484**, 186–194 (2012).
2. M. C. Fisher, S. J. Gurr, C. A. Cuomo, D. S. Blehert, H. Jin, E. H. Stukenbrock, J. E. Stajich, R. Kahmann, C. Boone, D. W. Denning, N. A. R. Gow, B. S. Klein, J. W. Kronstad, D. C. Sheppard, J. W. Taylor, G. D. Wright, J. Heitman, A. Casadevall, L. E. Cowen, Threats posed by the fungal kingdom to humans, wildlife, and agriculture. *MBio*. **11** (2020), doi:10.1128/MBIO.00449-20.
3. A. Arastehfar, A. Carvalho, J. Houbraken, L. Lombardi, R. Garcia-Rubio, J. D. Jenks, O. Rivero-Menendez, R. Aljohani, I. D. Jacobsen, J. Berman, N. Osherov, M. T. Hedayati, M. Ilkit, D. James-Armstrong, T. Gabaldón, J. Meletiadis, M. Kostrzewa, W. Pan, C. Lass-Flörl, D. S. Perlin, M. Hoenigl, Aspergillus fumigatus and aspergillosis: From basics to clinics. *Stud. Mycol.* **100**, 100115 (2021).
4. E. E. Hammond, C. S. McDonald, J. Vestbo, D. W. Denning, The global impact of Aspergillus infection on COPD. *BMC Pulm. Med.* **20**, 1–10 (2020).
5. J. Armstead, J. Morris, D. W. Denning, Multi-Country Estimate of Different Manifestations of Aspergillosis in Cystic Fibrosis. *PLoS One*. **9**, e98502 (2014).
6. A. E. Barber, T. Sae-Ong, K. Kang, B. Seelbinder, J. Li, G. Walther, G. Panagiotou, O. Kurzai, Aspergillus fumigatus pan-genome analysis identifies genetic variants associated with human infection. *Nat. Microbiol. 2021 612*. **6**, 1526–1536 (2021).
7. A. Bordbar, J. M. Monk, Z. A. King, B. O. Palsson, Constraint-based models predict metabolic and associated cellular functions. *Nat. Rev. Genet. 2014 152*. **15**, 107–120 (2014).
8. M. H. Mirhakkak, S. Schäuble, T. E. Klassert, S. Brunke, P. Brandt, D. Loos, R. V. Uribe, F. Senne de Oliveira Lino, Y. Ni, S. Vylkova, H. Slevogt, B. Hube, G. J. Weiss, M. O. A. Sommer, G. Panagiotou, Metabolic modeling predicts specific gut bacteria as key determinants for Candida albicans colonization levels. *ISME J.* **15**, 1257–1270 (2021).
9. J. M. Monk, P. Charusanti, R. K. Aziz, J. A. Lerman, N. Premyodhin, J. D. Orth, A. M. Feist, B. Ø. Palsson, Genome-scale metabolic reconstructions of multiple Escherichia coli strains highlight strain-specific adaptations to nutritional environments. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 20338–43 (2013).
10. C. J. Norsigian, N. Pusarla, J. L. McConn, J. T. Yurkovich, A. Dräger, B. O. Palsson, Z. King, BiGG Models 2020: multi-strain genome-scale models and expansion across the phylogenetic tree. *Nucleic Acids Res.* **48**, D402–D406 (2020).
11. C. J. Norsigian, X. Fang, Y. Seif, J. M. Monk, B. O. Palsson, A workflow for generating multi-strain genome-scale metabolic models of prokaryotes. *Nat. Protoc.* **15**, 1 (2020).
12. S. Castillo, D. Barth, M. Arvas, T. M. Pakula, E. Pitkänen, P. Blomberg, T. Seppanen-Laakso, H. Nygren, D. Sivasiddarthan, M. Penttilä, M. Oja, Whole-genome metabolic model of Trichoderma reesei built by comparative reconstruction. *Biotechnol. Biofuels*. **9**, 252 (2016).
13. M. Srivastava, E. Bencurova, S. K. Gupta, E. Weiss, J. Löffler, T. Dandekar, Aspergillus fumigatus Challenged by Human Dendritic Cells: Metabolic and Regulatory Pathway Responses Testify a Tight Battle. *Front. Cell. Infect. Microbiol.* **0**, 168 (2019).
14. A. M. Feist, B. O. Palsson, The biomass objective function. *Curr. Opin. Microbiol.* **13**, 344–349 (2010).
15. P. Horton, K.-J. Park, T. Obayashi, N. Fujita, H. Harada, C. J. Adams-Collier, K. Nakai, WoLF PSORT: protein localization predictor. *Nucleic Acids Res.* **35**, W585 (2007).

16. C. J. Fritzemeier, D. Hartleb, B. Szappanos, B. Papp, M. J. Lercher, Erroneous energy-generating cycles in published genome scale metabolic networks: Identification and removal. *PLOS Comput. Biol.* **13**, e1005494 (2017).

17. W. Hu, S. Sillaots, S. Lemieux, J. Davison, S. Kauffman, A. Breton, A. Linteau, C. Xin, J. Bowman, J. Becker, B. Jiang, T. Roemer, Essential Gene Identification and Drug Target Prioritization in Aspergillus fumigatus. *PLoS Pathog.* **3** (2007), doi:10.1371/JOURNAL.PPAT.0030024.

18. S. Klamt, E. D. Gilles, Minimal cut sets in biochemical reaction networks. *Bioinformatics*. **20**, 226–234 (2004).

19. N. Vlassis, M. P. Pacheco, T. Sauter, Fast Reconstruction of Compact Context-Specific Metabolic Network Models. **10**, e1003424 (2014).

20. A. Sasse, S. N. Hamer, J. Amich, J. Binder, S. Krappmann, Mutant characterization and in vivo conditional repression identify aromatic amino acid biosynthesis to be essential for Aspergillus fumigatus virulence. *Virulence*. **7**, 56 (2016).

21. J. S. Brown, A. Aufauvre-Brown, J. Brown, J. M. Jennings, H. Arst, D. W. Holden, H. DW, Signature-tagged and directed mutagenesis identify PABA synthetase as essential for Aspergillus fumigatus pathogenicity. *Mol. Microbiol.* **36**, 1371–1380 (2000).

22. R. Mahadevan, C. H. Schilling, The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab. Eng.* **5**, 264–76 (2003).

23. A. Hérivaux, J. R. Willis, T. Mercier, K. Lagrou, S. M. Goncąlves, R. A. Goncąles, J. Maertens, A. Carvalho, T. Gabaldón, C. Cunha, Lung microbiota predict invasive pulmonary aspergillosis and its outcome in immunocompromised patients. *Thorax* (2021), doi:10.1136/THORAXJNL-2020-216179.

24. O. G. G. de Almeida, C. P. da C. Capizzani, L. Tonani, P. H. Grizante Barião, A. F. da Cunha, E. C. P. De Martinis, L. A. G. M. M. Torres, M. R. von Zeska Kress, The Lung Microbiome of Three Young Brazilian Patients With Cystic Fibrosis Colonized by Fungi. *Front. Cell. Infect. Microbiol.* **10**, 668 (2020).

25. S. M. Gonçalves, K. Lagrou, C. Duarte-Oliveira, J. A. Maertens, C. Cunha, A. Carvalho, The microbiome-metabolome crosstalk in the pathogenesis of respiratory fungal diseases. *Virulence*. **8**, 673–684 (2017).

26. W. M. Song, B. Zhang, Multiscale Embedded Gene Co-expression Network Analysis. *PLOS Comput. Biol.* **11**, e1004574 (2015).

27. D. R. Garza, M. C. Verk, M. A. Huynen, B. E. Dutilh, Towards predicting the environmental metabolome from metagenomics with a mechanistic model. *Nat. Microbiol. 2018 34*. **3**, 456 (2018).

28. A. C. Marshall, S. E. Kidd, S. J. Lamont-Friedrich, G. Arentz, P. Hoffmann, B. R. Coad, J. B. Bruning, Structure, mechanism, and inhibition of aspergillus fumigatus thioredoxin reductase. *Antimicrob. Agents Chemother.* **63** (2019), doi:10.1128/AAC.02281-18/SUPPL_FILE/AAC.02281-18-S0001.PDF.

29. J. Binder, Y. Shadkchan, N. Osherov, S. Krappmann, The Essential Thioredoxin Reductase of the Human Pathogenic Mold Aspergillus fumigatus Is a Promising Antifungal Target. *Front. Microbiol.* **11**, 1383 (2020).

30. J. Nogales, J. Mueller, S. Gudmundsson, F. J. Canalejo, E. Duque, J. Monk, A. M. Feist, J. L. Ramos, W. Niu, B. O. Palsson, High-quality genome-scale metabolic modelling of Pseudomonas putida highlights its broad metabolic capabilities. *Environ. Microbiol.* **22**, 255–269 (2020).

31. Y. Seif, E. Kavvas, J.-C. Lachance, J. T. Yurkovich, S.-P. Nuccio, X. Fang, E. Catoiu, M. Raffatellu, B. O. Palsson, J. M. Monk, Genome-scale metabolic reconstructions of multiple Salmonella strains reveal serovar-specific metabolic traits. *Nat. Commun.* **9**, 3771 (2018).

32. E. Bosi, J. M. Monk, R. K. Aziz, M. Fondi, V. Nizet, B. Ø. Palsson, Comparative

genome-scale modelling of *Staphylococcus aureus* strains identifies strain-specific metabolic capabilities linked to pathogenicity. *Proc. Natl. Acad. Sci.* **113**, E3801–E3809 (2016).

33. K. Jastrzębowska, I. Gabriel, Inhibitors of amino acids biosynthesis as antifungal agents. *Amino Acids*. **47**, 227–249 (2015).
34. T. Tsukahara, Changes in Chemical Composition of Conidia of Aspergillus fumigatus during Maturation and Germination. *Microbiol. Immunol.* **24**, 747–751 (1980).
35. H. Pedersen, M. Carlsen, J. Nielsen, Identification of Enzymes and Quantification of Metabolic Fluxes in the Wild Type and in a Recombinant Aspergillus oryzae Strain. *Appl. Environ. Microbiol.* **65**, 11 (1999).
36. D. Maubon, S. Park, M. Tanguy, M. Huerre, C. Schmitt, M. C. Prévost, D. S. Perlin, J. P. Latgé, A. Beauvais, AGS3, an α(1–3)glucan synthase gene family member of Aspergillus fumigatus, modulates mycelium growth in the lung of experimentally infected mice. *Fungal Genet. Biol.* **43**, 366–375 (2006).
37. M. J. Lee, D. C. Sheppard, in *The mycota : a comprehensive treatise on fungi as experimental systems for basic and applied research*, K. Esser, D. Hoffmeister, Eds. (Cham: Springer, ed. 3, 2016).
38. D. Hagiwara, S. Suzuki, K. Kamei, T. Gonoi, S. Kawamoto, The role of AtfA and HOG MAPK pathway in stress tolerance in conidia of Aspergillus fumigatus. *Fungal Genet. Biol.* **73**, 138–149 (2014).
39. B. Ghfir, J. L. Fonvieille, Y. Koulali, R. Ecalle, R. Dargent, Effect of essential oil ofHyssopus officinalis on the lipid composition ofAspergillus fumigatus. *Mycopathol. 1994 1263*. **126**, 163–167 (1994).
40. L. Alcazar-Fuoli, E. Mellado, G. Garcia-Effron, J. F. Lopez, J. O. Grimalt, J. M. Cuenca-Estrella, J. L. Rodriguez-Tudela, Ergosterol biosynthesis pathway in Aspergillus fumigatus. *Steroids*. **73**, 339–347 (2008).
41. M. Schrettl, N. Beckmann, J. Varga, T. Heinekamp, I. D. Jacobsen, C. Jöchl, T. A. Moussa, S. Wang, F. Gsaller, M. Blatzer, E. R. Werner, W. C. Niermann, A. A. Brakhage, H. Haas, HapX-Mediated Adaption to Iron Starvation Is Crucial for Virulence of Aspergillus fumigatus. *PLOS Pathog.* **6**, e1001124 (2010).
42. E. Barreto-Bergter, P. A.J. Gorin, L. R. Travassos, Cell constituents of mycelia and conidia of Aspergillus fumigatus. *Carbohydr. Res.* **95**, 205–217 (1981).
43. A.-M. Dietl, Z. Meir, Y. Shadkchan, N. Osherov, H. Haas, Riboflavin and pantothenic acid biosynthesis are crucial for iron homeostasis and virulence in the pathogenic mold Aspergillus fumigatus. *https://doi.org/10.1080/21505594.2018.1482181*. **9**, 1036–1049 (2018).
44. S. W. Chocklett, P. Sobrado, Aspergillus fumigatus SidA Is a Highly Specific Ornithine Hydroxylase with Bound Flavin Cofactor. *Biochemistry*. **49**, 6777–6783 (2010).
45. M. R. Andersen, M. L. Nielsen, J. Nielsen, Metabolic model integration of the bibliome, genome, metabolome and reactome of Aspergillus niger. *Mol. Syst. Biol.* **4**, 178 (2008).
46. H. Lu, F. Li, B. J. Sánchez, Z. Zhu, G. Li, I. Domenzain, S. Marcišauskas, P. M. Anton, D. Lappa, C. Lieven, M. E. Beber, N. Sonnenschein, E. J. Kerkhoven, J. Nielsen, A consensus S. cerevisiae metabolic model Yeast8 and its ecosystem for comprehensively probing cellular metabolism. *Nat. Commun. 2019 101*. **10**, 1–13 (2019).
47. A. Ebrahim, J. A. Lerman, B. O. Palsson, D. R. Hyduke, COBRApy: COnstraints-Based Reconstruction and Analysis for Python. *BMC Syst. Biol.* **7**, 74 (2013).
48. Z. A. King, J. Lu, A. Dräger, P. Miller, S. Federowicz, J. A. Lerman, A. Ebrahim, B. O. Palsson, N. E. Lewis, BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Res* (2015), doi:10.1093/nar/gkv1049.

49. S. Placzek, I. Schomburg, A. Chang, L. Jeske, M. Ulbrich, J. Tillack, D. Schomburg, BRENDA in 2017: new perspectives and new tools in BRENDA. *Nucleic Acids Res.* **45**, D380–D388 (2017).
50. R. Caspi, R. Billington, C. A. Fulcher, I. M. Keseler, A. Kothari, M. Krummenacker, M. Latendresse, P. E. Midford, Q. Ong, W. K. Ong, S. Paley, P. Subhraveti, P. D. Karp, The MetaCyc database of metabolic pathways and enzymes. *Nucleic Acids Res.* **46**, D633–D639 (2018).
51. C. Castellani, A. J. A. Duff, S. C. Bell, H. G. M. Heijerman, A. Munck, F. Ratjen, I. Sermet-Gaudelus, K. W. Southern, J. Barben, P. A. Flume, P. Hodková, N. Kashirskaya, M. N. Kirszenbaum, S. Madge, H. Oxley, B. Plant, S. J. Schwarzenberg, A. R. Smyth, G. Taccetti, T. O. F. Wagner, S. P. Wolfe, P. Drevinek, ECFS best practice guidelines: the 2018 revision. *J. Cyst. Fibros.* **17**, 153–178 (2018).
52. K. De Boeck, N. Derichs, I. Fajac, H. R. de Jonge, I. Bronsveld, I. Sermet, F. Vermeulen, D. N. Sheppard, H. Cuppens, M. Hug, P. Melotti, P. G. Middleton, M. Wilschanski, New clinical diagnostic procedures for cystic fibrosis in Europe. *J. Cyst. Fibros.* **10 Suppl 2** (2011), doi:10.1016/S1569-1993(11)60009-X.
53. P. M. Farrell, T. B. White, C. L. Ren, S. E. Hempstead, F. Accurso, N. Derichs, M. Howenstine, S. A. McColley, M. Rock, M. Rosenfeld, I. Sermet-Gaudelus, K. W. Southern, B. C. Marshall, P. R. Sosnay, Diagnosis of Cystic Fibrosis: Consensus Guidelines from the Cystic Fibrosis Foundation. *J. Pediatr.* **181S**, S4-S15.e1 (2017).
54. P. D. Wagner, The physiological basis of pulmonary gas exchange: implications for clinical interpretation of arterial blood gases. *Eur. Respir. J.* **45**, 227–243 (2015).
55. M. R. Miller, Defining airflow obstruction. *Eur. Respir. J.* **45**, 560 (2015).
56. P. H. Quanjer, S. Stanojevic, T. J. Cole, X. Baur, G. L. Hall, B. H. Culver, P. L. Enright, J. L. Hankinson, M. S. M. Ip, J. Zheng, J. Stocks, C. Schindler, Multi-ethnic reference values for spirometry for the 3-95-yr age range: the global lung function 2012 equations. *Eur. Respir. J.* **40**, 1324–1343 (2012).
57. Permutation, Parametric and Bootstrap Tests of Hypotheses. *Permut. Parametr. Bootstrap Tests Hypotheses* (2005), doi:10.1007/B138696.
58. S. Magnúsdóttir, A. Heinken, L. Kutt, D. A. Ravcheev, E. Bauer, A. Noronha, K. Greenhalgh, C. Jäger, J. Baginska, P. Wilmes, R. M. T. Fleming, I. Thiele, Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. *Nat. Biotechnol.* **35**, 81–89 (2016).
59. D. Machado, S. Andrejev, M. Tramontano, K. R. Patil, Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. *Nucleic Acids Res.* **46**, 7542–7553 (2018).
60. D. W. Hosmer, S. Lemeshow, R. X. Sturdivant, Applied Logistic Regression: Third Edition. *Appl. Logist. Regres. Third Ed.*, 1–510 (2013).
61. H. Zou, The Adaptive Lasso and Its Oracle Properties. *https://doi.org/10.1198/016214506000000735*. **101**, 1418–1429 (2012).
62. J. D. Lee, D. L. Sun, Y. Sun, J. E. Taylor, Exact post-selection inference, with application to the lasso. *Ann. Stat.* **44**, 907–927 (2013).

Figure captions

Figure 1: General reconstruction workflow and *A. fumigatus* pan-genome-scale metabolic model (GEM) statistics. (**A)** Workflow towards *A. fumigatus* strain-specific GEM reconstructions. Colors indicate different strains and associated metabolic models. (**B-F)** Characteristics of pan-GEM reconstruction for *A. fumigatus*. **(B)** Contribution of macromolecules in comprising one unit of biomass (Supplementary Table S1). **(C)** Distribution of pan-GEM reactions across major pathway categories (Supplementary Table S7). **(D)** Distribution of pan-GEM reactions across nine compartments (Supplementary Table S7). **(E)** Growth prediction accuracy of pan-GEM for *A. fumigatus* wild-type (Af293) and five mutant strains using phenotypic microarray data. Growth accuracy for phenotypic microarrays on sulfur (S) are also indicated for neglecting *ΔniaD* and *ΔlysF* mutants (see Results for details, Supplementary Table S2). C: carbon, N: nitrogen, P: phosphor, S: sulfur. **(F)** Confusion matrix of pan-GEM accuracy in predicting the essentiality of 20 genes according to the literature (see Results and Methods).

Figure 2: Core and accessory metabolic capabilities of all *A. fumigatus* strain-specific genome-scale metabolic models (GEMs). The core and accessory metabolic content was determined for 252 unique *A. fumigatus* strains with environmental and clinical origin. (A) Summary of the core and accessory reactome across higher level metabolic pathway categories. Pathway categories are according to the KEGG pathway definition (https://www.kegg.jp/kegg/pathway.html). **(B)** The distribution of the accessory reactome across all isolate models. Indicated percentage ranges correspond to accessory reaction presence across all strain-GEMs. **(C)** Heatmap with pairwise Jaccard distance values for isolate GEM pairs based on presence or absence of metabolic reactions. **(D)** Fisher-test based most statistically significant reactions enriched in the indicated isolate subsets. Presence frequency indicates the fraction of reaction presence over all investigated models (color associates to sample origin). **(E)** Decision tree optimized towards showing best separation into clinical and environmental isolate origin. The decision tree is based on absence/presence of metabolic reactions and growth capability on different nutrients across all isolate GEMs. **(F)** Machine learning mean AUC performance based on FVA derived flux ranges for all reactions (objective function: biomass) of strain models with clinical vs. environmental origin.

Figure 3: Metagenomics sequencing of 80 paired sputum samples from cystic fibrosis (N=40) patients before (*A. fumigatus* -) and after *A. fumigatus* infection (*A. fumigatus* +). (A) Relative abundances of the top 10 genera and species over all 80 samples. X axis is ordered by patient sample. **(B)** Differential correlation analysis of species in *A. fumigatus*+ relative to *A. fumigatus*- to reveal changes in the interactome of the lung microbiome upon *A. fumigatus* colonization. Edge colors and different class information indicate direction of correlation in *A. fumigatus-*/*A. fumigatus*+. The associated count indicates the number of species pairs in the network exhibiting this pattern of change. Only species pairs with significant differential correlations were included (permutation test, p<0.05), Species with orange background indicate significant differentially abundant species between *A. fumigatus-* vs *A. fumigatus*+ samples (metagenomeSeq, zero inflated gaussian mixture model, p<0.05). Labels: *A. fumigatus*-: samples before infection; *A. fumigatus*+: samples after infection.

Figure 4: Statistics for MAMBO-derived metabolite profiles in cystic fibrosis patient samples. (A) Beta diversity (Euclidean distance) of MAMBO derived media. PERMANOVA was used to assess the statistical significance of beta diversity comparisons. **(B)** Growth rate differences of genome-scale metabolic models corresponding to clinical *A. fumigatus* strains based on MAMBO derived media compositions associated to cystic fibrosis samples before and after *A. fumigatus* infection. **(C)** Significantly different flux ranges (either in lower or upper bound) of clinical strain models simulated with FVA on MAMBO derived media before and after *A. fumigatus* infection. Significance was tested according to paired Wilcoxon signed rank test and adjusted by FDR. **(D)** Three selected enzymatic reactions with significant flux bound differences in either lower or upper bound as displayed in **(C)**. Both bounds are indicated. EC1.4.3.4: Tryptamine:oxygen oxidoreductase; EC4.1.3.27: Chorismate pyruvate-lyase; EC1.8.1.9: NADPH:oxidized-thioredoxin oxidoreductase**.** *A.f.*-/+: *A. fumigatus* positive and negative samples, respectively.

Tables

Table 1: Number of accessory genes and reactions across all isolate GEMs. The occurrence column refers to the number of reactions occurring in a certain fraction range of isolate specific GEMs based on 338 accessory reactions in total.

	Genes		Reactions	
Occurrence (in %)	Mean	Standard deviation	Mean	Standard deviation
[1,20]	1.9	1.8	4.7	7.7
[21,40]	6.9	2.5	1.3	1.2
[41,60]	16.7	3.2	17.4	6.0
[61,80]	17.8	2.4	20.2	6.3
[81,99]	296.9	8.1	219.3	6.8

Supplementary material

Supplementary Table S1 – Biomass composition

Supplementary Table S2 – Phenotypic growth information

Supplementary Table S3 – *A. fumigatus* strain metadata

Supplementary Table S4 – ML model using minimal media

Supplementary Table S5 – Metagenomics

Supplementary Table S6 – Metabolite analysis

Supplementary Table S7 – Detailed GEM information

Supplementary Figure S1 – Accessory genome conservation among 252 Aspergillus fumigatus strains. Indicated percentage ranges denote accessory gene presence across the genomes of all strain-GEMs. Cluster with gray background denotes genetic lineage of *A. fumigatus* with significantly fewer accessory genes than other lineages as previously published (6). Associates to Fig. 2B.

Supplementary Figure S2 – Alpha and beta diversity of CF lung microbiome. **(A)** Statistical significance according to Shannon and Chao diversity index for alpha diversity). **(B)** Beta diversity (Aitchison distance) based on Kraken derived taxonomic profiles. Wilcoxon signed rank test was used for alpha diversity comparisons; PERMANOVA was used to assess the statistical significance of beta diversity comparisons.

Figure 1

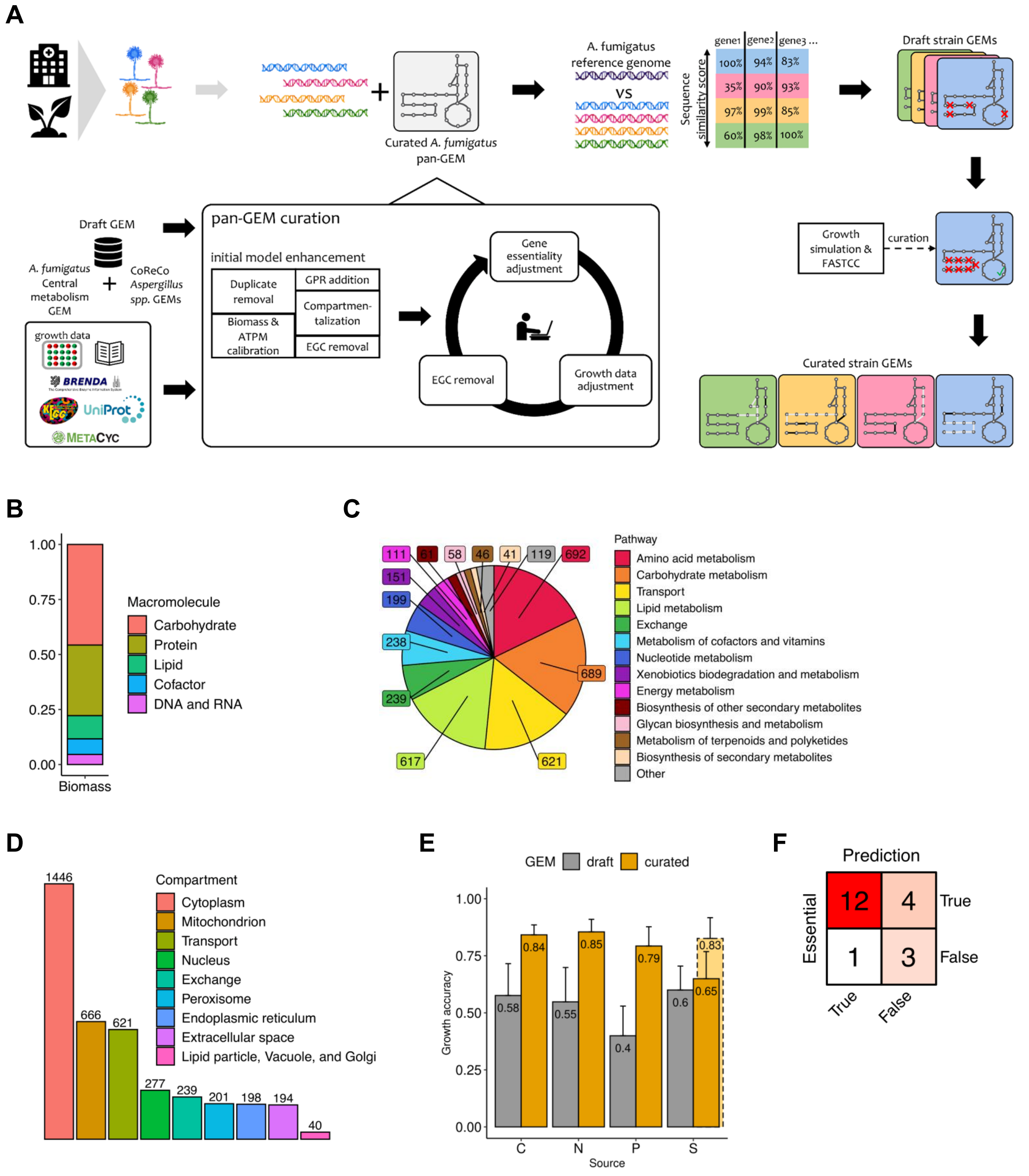

A
Curated A. fumigatus pan-GEM
A. fumigatus reference genome
VS
Sequence similarity score
gene1 gene2 gene3 ...
100% 94% 83%
35% 90% 93%
97% 99% 85%
60% 98% 100%
Draft strain GEMs
Draft GEM
A. fumigatus Central metabolism GEM
CoReCo Aspergillus spp. GEMs
growth data
BRENDA
KEGG
UniProt
METACYC
pan-GEM curation
initial model enhancement
Duplicate removal
GPR addition
Compartmentalization
Biomass & ATPM calibration
EGC removal
Gene essentiality adjustment
EGC removal
Growth data adjustment
Growth simulation & FASTCC
curation
Curated strain GEMs
B
1.00
0.75
0.50
0.25
0.00
Biomass
Macromolecule
Carbohydrate
Protein
Lipid
Cofactor
DNA and RNA
C
111
61
58
46
41
119
692
151
199
238
239
617
621
689
Pathway
Amino acid metabolism
Carbohydrate metabolism
Transport
Lipid metabolism
Exchange
Metabolism of cofactors and vitamins
Nucleotide metabolism
Xenobiotics biodegradation and metabolism
Energy metabolism
Biosynthesis of other secondary metabolites
Glycan biosynthesis and metabolism
Metabolism of terpenoids and polyketides
Biosynthesis of secondary metabolites
Other
D
1446
666
621
277
239
201
198
194
40
Compartment
Cytoplasm
Mitochondrion
Transport
Nucleus
Exchange
Peroxisome
Endoplasmic reticulum
Extracellular space
Lipid particle, Vacuole, and Golgi
E
GEM draft curated
Growth accuracy
1.00
0.75
0.50
0.25
0.00
0.58
0.84
0.55
0.85
0.4
0.79
0.6
0.65
0.83
C
N
P
S
Source
F
Prediction
Essential
12
4
1
3
True
False
True
False

Figure 2

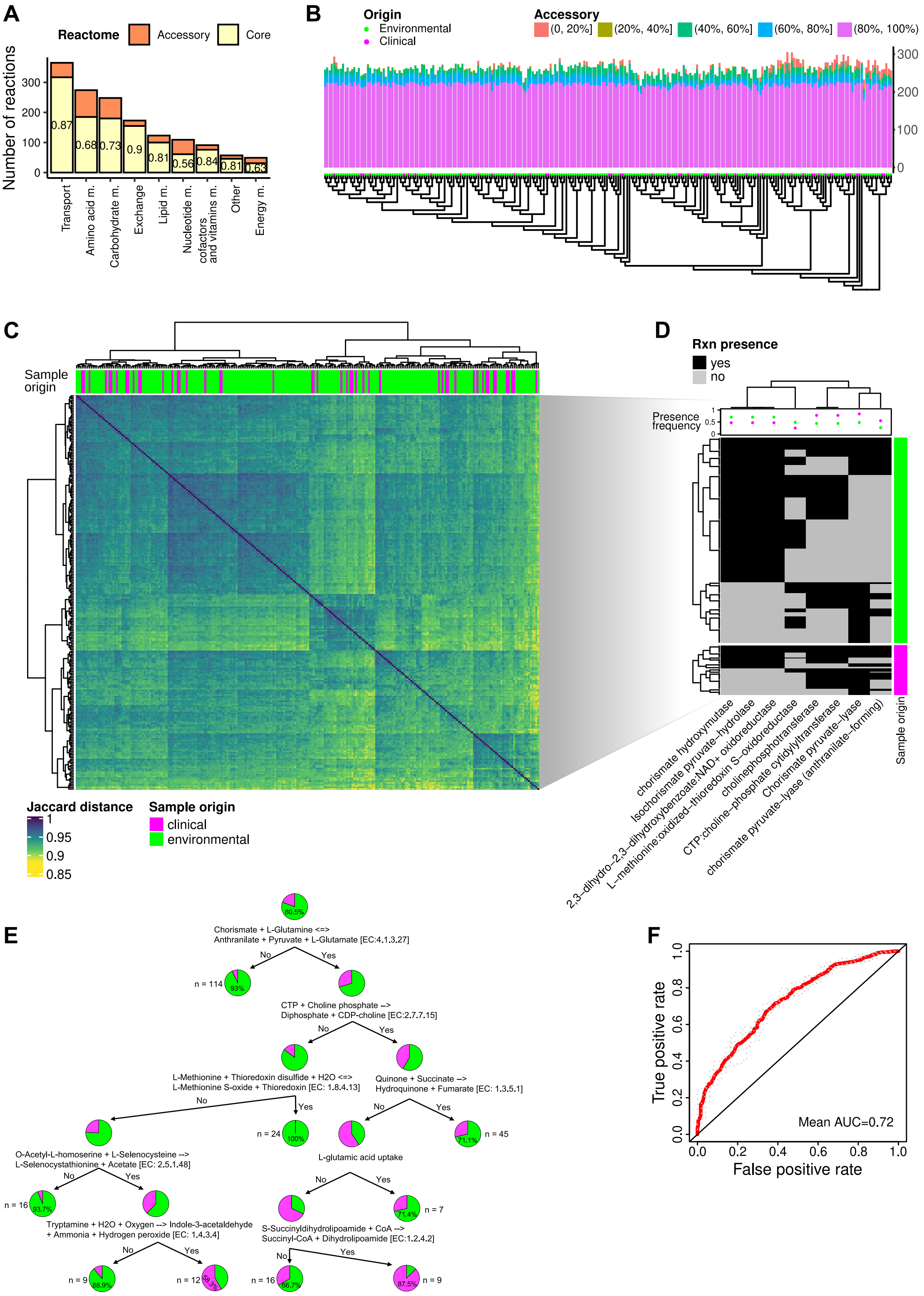

A
Reactome
Accessory
Core
Number of reactions
0
100
200
300
0.87
0.68
0.73
0.9
0.81
0.56
0.84
0.81
0.63
Transport
Amino acid m.
Carbohydrate m.
Exchange
Lipid m.
Nucleotide m.
cofactors and vitamins m.
Other
Energy m.
B
Origin
Environmental
Clinical
Accessory
(0, 20%]
(20%, 40%]
(40%, 60%]
(60%, 80%]
(80%, 100%)
300
200
100
0
C
Sample origin
Jaccard distance
1
0.95
0.9
0.85
Sample origin
clinical
environmental
D
Rxn presence
yes
no
Presence frequency
1
0.5
0
chorismate hydroxymutase
Isochorismate pyruvate–hydrolase
2,3–dihydro–2,3–dihydroxybenzoate:NAD+ oxidoreductase
L–methionine:oxidized–thioredoxin S–oxidoreductase
cholinephosphotransferase
CTP:choline–phosphate cytidylyltransferase
Chorismate pyruvate–lyase
chorismate pyruvate–lyase (anthranilate–forming)
Sample origin
E
80.5%
Chorismate + L-Glutamine <=>
Anthranilate + Pyruvate + L-Glutamate [EC:4.1.3.27]
No
Yes
n = 114
93%
CTP + Choline phosphate –>
Diphosphate + CDP-choline [EC:2.7.7.15]
No
Yes
L-Methionine + Thioredoxin disulfide + H2O <=>
L-Methionine S-oxide + Thioredoxin [EC: 1.8.4.13]
Quinone + Succinate –>
Hydroquinone + Fumarate [EC: 1.3.5.1]
No
Yes
n = 24
100%
No
Yes
71.1%
n = 45
O-Acetyl-L-homoserine + L-Selenocysteine -->
L-Selenocystathionine + Acetate [EC: 2.5.1.48]
No
Yes
n = 16
93.7%
L-glutamic acid uptake
No
Yes
71.4%
n = 7
Tryptamine + H2O + Oxygen --> Indole-3-acetaldehyde
+ Ammonia + Hydrogen peroxide [EC: 1.4.3.4]
No
Yes
n = 9
88.9%
n = 12
58.3%
S-Succinyldihydrolipoamide + CoA -->
Succinyl-CoA + Dihydrolipoamide [EC:1.2.4.2]
No
Yes
n = 16
66.7%
87.5%
n = 9
F
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
Mean AUC=0.72

Figure 3

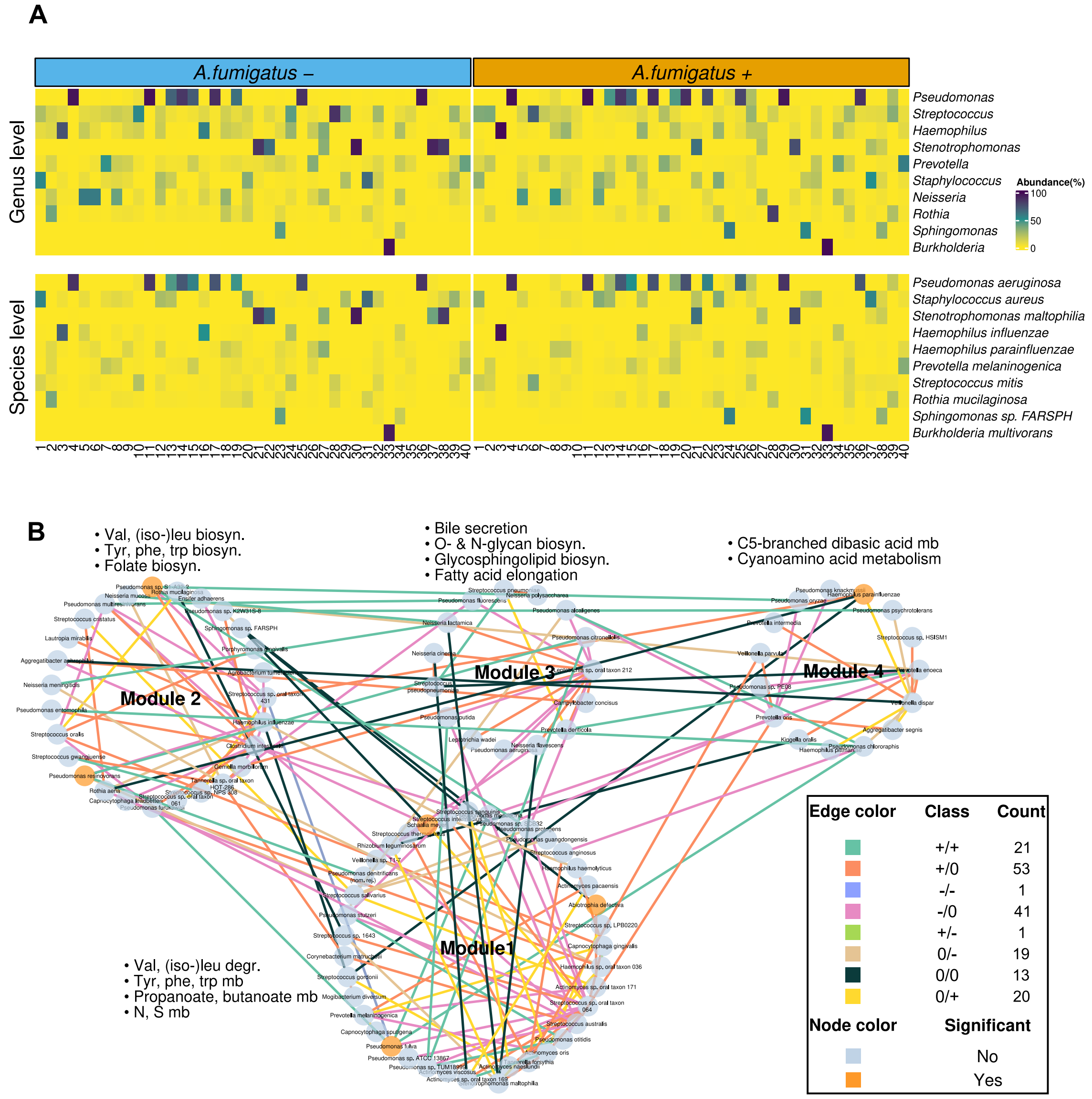

A
A.fumigatus −
A.fumigatus +
Genus level
Species level
Pseudomonas
Streptococcus
Haemophilus
Stenotrophomonas
Prevotella
Staphylococcus
Neisseria
Rothia
Sphingomonas
Burkholderia
Abundance(%)
100
50
0
Pseudomonas aeruginosa
Staphylococcus aureus
Stenotrophomonas maltophilia
Haemophilus influenzae
Haemophilus parainfluenzae
Prevotella melaninogenica
Streptococcus mitis
Rothia mucilaginosa
Sphingomonas sp. FARSPH
Burkholderia multivorans
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40
B
• Val, (iso-)leu biosyn.
• Tyr, phe, trp biosyn.
• Folate biosyn.
• Bile secretion
• O- & N-glycan biosyn.
• Glycosphingolipid biosyn.
• Fatty acid elongation
• C5-branched dibasic acid mb
• Cyanoamino acid metabolism
Module 2
Module 3
Module 4
Module1
• Val, (iso-)leu degr.
• Tyr, phe, trp mb
• Propanoate, butanoate mb
• N, S mb
Edge color
Class
Count
+/+ 21
+/0 53
-/- 1
-/0 41
+/- 1
0/- 19
0/0 13
0/+ 20
Node color
Significant
No
Yes

Figure 4

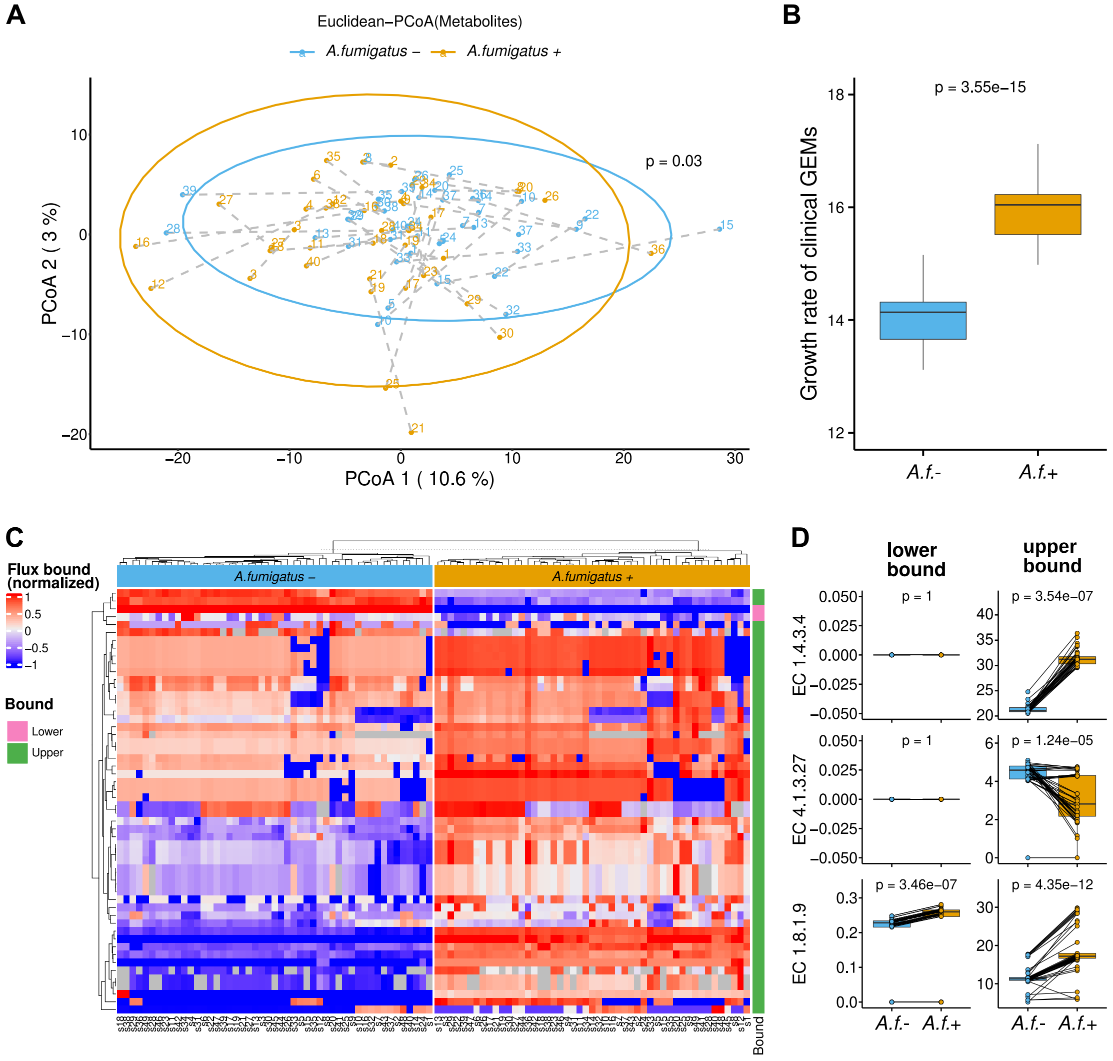

A
Euclidean−PCoA(Metabolites)
A.fumigatus −
A.fumigatus +
p = 0.03
PCoA 2 (3 %)
PCoA 1 (10.6 %)
B
p = 3.55e−15
Growth rate of clinical GEMs
A.f.-
A.f.+
C
Flux bound (normalized)
1
0.5
0
−0.5
−1
Bound
Lower
Upper
A.fumigatus −
A.fumigatus +
D
lower bound
upper bound
EC 1.4.3.4
p = 1
p = 3.54e−07
EC 4.1.3.27
p = 1
p = 1.24e−05
EC 1.8.1.9
p = 3.46e−07
p = 4.35e−12
A.f.-
A.f.+

Figure S1

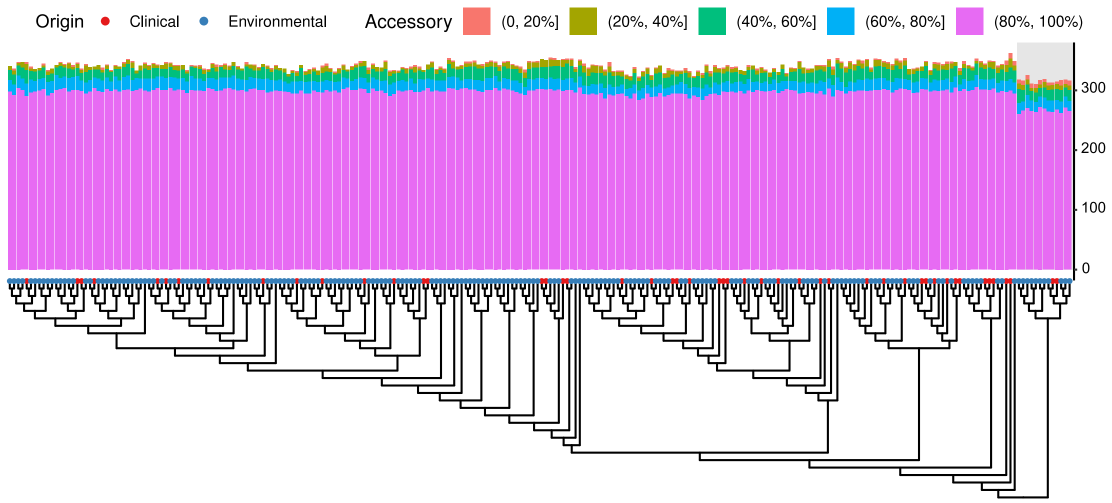
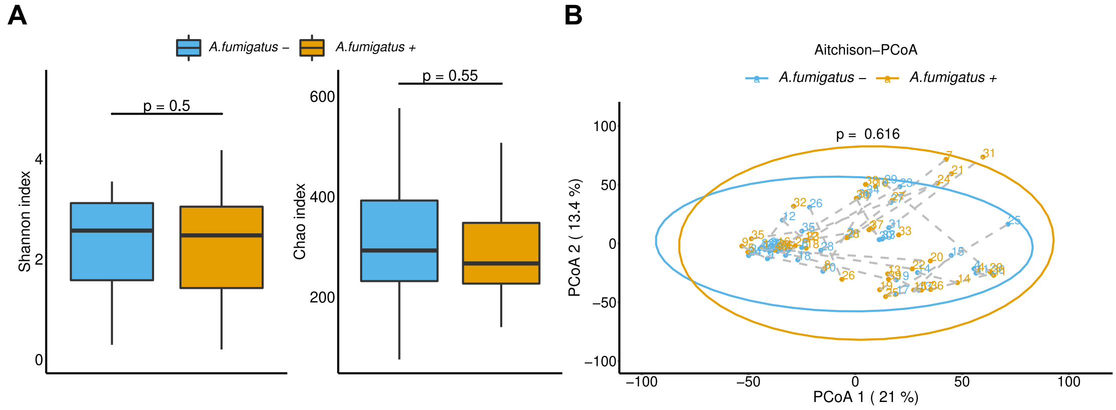


Figure S2



CHAPTER IV DISCUSSION

In this thesis, a collection of state-of-art bioinformatics approaches for fungal genome analysis were applied to *A. fumigatus* genomes, a deadly airborne fungal pathogen. Based on the biological motivation to elucidate the fungal virulence mechanisms of *A. fumigatus*, the whole genome sequences of 252 strains from environmental and clinical origins were collected and were combined with 48 more strains globally distributed. Besides genetic determinants of antifungal resistance, I contributed to the analysis of the metabolic potential of these strains in relation to adaptation in the human lung.

In addition, another series of transcriptome data analysis methods were applied to study gene expression (transcriptome) profiles from pre-symptomatic sepsis patients. This part was driven by the motivation to understand host immune response mechanisms leading to early pathogen detection (bacterial, fungal, and co-infection) in pre-diagnosis sepsis patients to administrate timely the appropriate treatment. This cumulative dissertation comprises four research papers belonging to four main topics:

1. Using *de novo* and reference-based genome assemblies provided a powerful tool for genome reconstruction, variant detection, and strain-specific metabolic network analysis (**Manuscript I, III, IV**)
2. WGS-based phylogenetic tree and genome clustering are highly reliable methods for studying evolutionary relationships (**Manuscript I, III**)
3. Genome-wide association study revealed the genomic variants associated with *A. fumigatus* pathogenicity (**Manuscript I**)
4. Network-based analyses and machine learning facilitated the discovery of potential bacterial, fungal, and co-infection host biomarkers for pre-symptomatic sepsis patients (**Manuscript II**)

Using de novo and reference-based genome assemblies provided a powerful tool for genome reconstruction, variant detection, and strain-specific metabolic network analysis

In the **manuscript I**, I re-assembled 300 *A. fumigatus* genomes using reference-based and *de novo* assembly methods. *De novo* assembled genomes of *A. fumigatus* contained genome sizes ranging from 26.9 Mb to 30.8 Mb. The pan-genome of these strains was constructed and composed of 10,907 orthologous genes. Of all orthologous genes, 69% were found in all strains (core genes), and 31% were accessory genes. Compared to a recent study, Lofgren *et al*. built the pan-genome of 260 *A. fumigatus* strains, by approximately 67% of pan-genomes were core genes [187]. In **manuscript IV**, the *de novo* assembled genomes of 252 *A. fumigatus* isolates collected in Germany were used to construct genome-scale metabolic models (GEMs) by Mirhakkak and Chen. This pan-GEM was composed of 1,453 metabolic genes and 3,882 metabolic reactions. Of all metabolic genes, 69.8% were presented in all strains resulting in 77.3% of all metabolic reactions. These results showed consistency of high diversity among *A. fumigatus* genomes at genetic and phenotypic levels.

In another study by McCarthy and Fitzpatrick, the authors showed that 83% of the pan-genome of 12 *A. fumigatus* strains were shared in all genomes [188]. These results suggested that the diversity of *A. fumigatus* genomes was increased concerning the size of the population. However, the distribution of pan-genome diversity was leveled after more than 250 genomes were added, as shown in Lofgren's and this studies [187]. Compared with other fungal pathogens, *Neonectria neomacrospora* has a genome size of approximately 40 Mb. The pan-genome of 66 *N. neomacrospora* strains comprised 13,069 orthologous genes with 64% core genes [189]. While *Saccharomyces cerevisiae* has an approximate genome size of 12 Mb, 85% of the pan-genome was core genes among 100 strains [188]. McCarthy and Fitzpatrick also analyzed the pan-genome of *Cryptococcus neoformans* and *Candida albicans,* which possess genome sizes <20Mb; these pathogens shared their core genes for 80% and 90% of the pan-genomes. In conclusion, a positive correlation between genome size and pan-genome diversity has been observed in fungal genomes.

On the other hand, the *A. fumigatus* Af293 genome, with a genome size of 29.4, was used as the reference genome for the read mapping method (**manuscripts I and III**). Genome sequences were mapped on 90.8 – 97.9% of the Af293 reference genome, with 39.8 – 98.3% of reads per genome. Compared to a study by Garcia-Rubio *et al*., they assembled 101 *A. fumigatus* genomes against Af293 and A1163 reference genomes [86]. The sequences were aligned on 84 – 99% of reference genomes, with more than 88% of reads were mapped on the references. Abdolrasouli *et al.* assembled 24 *A. fumigatus* genomes using Af293 as a reference genome. Their genomes can map to more than 92% of the reference genome [190]. The results from the reference-based method based on the Af293 genome suggested that approximately 90% of *A. fumigatus* genomes were conserved among strains. However, there were high percentages of reads that failed to be located on the reference genome. Furthermore, over 20% of pan-genome from the *de novo* assembly method were not found in the *A. fumigatus* Af293 reference strain. Altogether, the *de novo* assembly method can overcome the limitations of reference-based genome assembly, which provides more completed genomes for novel sequenced strains.

However, to identify genetic variations, a common reference genome is still required for comparing and searching nucleotide variants that might cause protein changes. Based on the Af293 reference, an average of 78,692 SNVs and 7,383 insertions/deletions per genome were identified in this study. Compared to other studies, Abdolrasouli *et al.* can identify 78,960 SNVs by average per genome [190], while Garcia-Rubio *et al.* can identify 93,609 SNVs [86]. Although there was a high number of identified variants from previous studies, 33% of identified SNVs in this study were not reported on the database. The results of pan-genome and pan-GEM analysis suggested a surprisingly high genomic diversity and metabolic uniqueness of *A. fumigatus* genomes.

In conclusion, reference-based and *de novo* genome assembly methods provided the most comprehensive reconstruction of *A. fumigatus* genomes, small variant identification, and genome-scale metabolic model construction. These works provide an extensive resource of *A. fumigatus* genomes to support further *A. fumigatus* genome studies.

Future perspectives

Genome assembly at the chromosomal level will provide more complete genomes, including improved localization of predicted genes on genomes. In addition, it will allow us to detect more complex variants, such as transversion/translocation. The third-generation sequencing technology, which generates ultra-long sequences using single-molecule technology, will enable assemblies at the chromosome level and significantly improve the accuracy of complex variant detection in fungal genomes, including *A. fumigatus*.

WGS-based phylogenetic tree and genome clustering are highly reliable methods for studying evolutionary relationship

In order to investigate the diversity of *A. fumigatus* genomes, phylogenetic tree and clustering techniques were used to study the evolutionary and genomic relationship among *A. fumigatus* genomes. In **manuscript I**, phylogenetic trees of all 300 *A. fumigatus* genomes were reconstructed using reference-based analyses (whole-genome SNVs), and *de novo* assembled genomes (core-genes). The trees revealed a total of 7 genetic clusters among all strains. Compared to phylogenetic trees of *A. fumigatus* from Garcia-Rubio *et al.* and Fan *et al.* studies, they performed tree reconstructing and clustering based on SNVs of 101 and 196 *A. fumigatus* strains, respectively [86,191]. In their study, Garcia-Rubio *et al.* found 4 clusters among their strains, while Fan *et al.* found the optimal clusters of 3.

Interestingly, the most distant cluster (cluster 1) was observed in reference-based and *de novo* assembled genome trees, including 13 novel strains from Germany and 6 strains from Spain (n = 4), Peru (n = 1), and Canada (n = 1). These 6 strains were also found in the notable most distant clusters in the tree from Garcia-Rubio and Fan studies. Garcia-Rubio *et al*. found that *A. fumigatus* strains in this cluster possess the highest number of SNVs, resulting in the most distant of these strains from other clusters [86]. However, they were the most homogeneous cluster [86]. This study also found that *A. fumigatus* strains in this cluster possessed fewer accessory genes but contained more high-impact variants than other clusters. The highly homogenous *A. fumigatus* strains in this cluster suggested the recent spread of *A. fumigatus* across geographic regions [191]. Statistical testing confirmed that *A. fumigatus* genomes were distributed regardless of geographical areas. Phylogenetic tree analysis from WGS showed robust and concordant results among three different studies.

The genome similarities of *A. fumigatus* harbor mutations of *cyp51a* and *cyp51b*, the triazole targeted genes, were further observed on the phylogenetic tree. In **manuscripts I and III**, the results showed the closed genetic relationship among *A. fumigatus* harbored TR_{34}/L98H mutation (tandem repeats of copies of 34-bp sequence resulting in a substitution of leucine 98 on *cyp51a* gene). These results strongly supported a previous study by Camps *et al*. [192]. They studied 142 *A. fumigatus* isolated from Europe, in which 80 strains were azole-resistant carrying TR_{34}/L98H mutation. They found that azole-resistant *A. fumigatus* strains carried TR_{34}/L98H mutation had identical microsatellite loci and putative cell surface protein (CSP) CSP typing [192]. In addition, I reconstructed neighbor-net trees based on WGS-SNVs and SNVs on *cyp51a* and *cyp51b* genes for investigating recombination events among isolates (**manuscript I**). The neighbor-net trees also showed that most *A. fumigatus* strains harboring TR_{34}/L98H mutation were highly related. Another study by Abdolrasouli *et al.* constructed a whole-genome SNP-based phylogenetic tree of 24 *A. fumigatus* from India, the Netherlands, and the United Kingdom, with 17 strains harboring TR_{34}/L98H mutation. Their tree also showed *A. fumigatus* genomes carrying the TR_{34}/L98H mutation diverse independently from the source of origin. The results in this part suggested the common TR_{34}/L98H mutation mechanism, which supports the distribution event of *A. fumigatus* across the globe. The non-synonymous-to-synonymous substitution (dN/dS) ratios in *cyp51a* and *cyp51b* genes also suggested the selective sweep under purifying selection for both genes (**manuscript I**).

In conclusion, phylogenetic tree analysis can provide meaningful results related to *A. fumigatus* evolutionary relationship across regions, reflecting the distribution events of *A. fumigatus* strains. Moreover, the trees also showed a common azole-resistant mechanism corresponding to *A. fumigatus* distribution and selective sweep of genotype associated with the predominant TR_{34}/L98H mutation.

Future perspective

The more accurate and complete genomes from long-read sequences can enhance more biological insight of phylogenomic models. However, phylogenetic tree reconstruction based on whole-genome can have computational burdens. Therefore, improving more phylogeny reconstructing models for the complex tree to reduce systematic errors and computational sources and support the sequencing data from the novel sequencing technologies will further improve the phylogenetic tree for closely representing biological evolution.

Genome-wide association study revealed the genomic variants associated with A. fumigatus pathogenicity

TR_{34}/L98H mutation on *cyp51a* has been reported as a predominant azole-resistance mechanism of *A. fumigatus* and primarily found in patients who had no experience with triazole-treatment through environmental *A. fumigatus* strains [193,194]. Therefore, in **manuscript I**, the genomic variations between clinical and environmental strains were studied through several statistical methods, including a genome-wide association study (GWAS).

Over two decades ago, we realized that *A. fumigatus* strains in hospitals differed from environmental *A. fumigatus* by using traditional DNA polymorphism comparison techniques [195]. Debeaupuis *et al*. studied the genetic diversity among 879 *A. fumigatus* strains from clinical and environmental sources using the Southern blot technique. Their results showed no prominent cluster related to geographical regions or environmental/clinical origins [196]. However, Aufauvre-Brown *et al.* studied *A. fumigatus* virulence between clinical and environmental strains using a mixed infection model. Their results showed that *A. fumigatus* strains from clinical origin were more virulent than those from environmental isolates [197].

In the study by Fan *et al*., which built the phylogenetic tree of 196 *A. fumigatus* strains based on genome-wide SNV, with a limited number of environmental compared to clinical strains (29:167), they also could not identify the genome differences between environmental and clinical isolates. In my study (**manuscript I**), with a large number of environmental strains (217 environmental:83 clinical) and a higher resolution of clustering, *A. fumigatus* clinical strains were significantly over-represented in one cluster (cluster 5) of the phylogenetic tree. Furthermore, at the nucleotide level, *A. fumigatus* environmental strains showed higher nucleotide diversity (π) than clinical strains.

In **manuscript I**, genomic variants associated with clinical and environmental *A. fumigatus* strains were further observed at the nucleotide level (GWAS) and gene level (panGWAS). GWAS indicated that 68 genomic variants were associated with clinical *A. fumigatus* strains, including variants on 27 protein-coding genes. Of those, some genes had been reported as *A. fumigatus* virulent genes, for example, *pacC,* which is a transcription factor involved in several fungal-host interaction processes [198,199]. In addition, *AcuK,* another transcription factor, and *srbA* genes involved in fungal growth and invasion were also identified in this study [200,201]. In addition, panGWAS reported accessory genes were significantly presented in clinical strains, such as a gene with predicted selenium binding activity and role in cell redox homeostasis (*Afu1g05220*).

In conclusion, with a large dataset of *A. fumigatus* strains and more balanced strains from the environmental and clinical origin, a clinical *A. fumigatus* cluster was detected in this study. Moreover, GWAS and panGWAS allowed the identification of clinical *A. fumigatus*-associated genes, and we hypothesized that some of these genes hold essential roles related to pathogenic versatility.

Future perspective

To better understand the azole-resistant mechanisms, more azole-resistant *A. fumigatus* strains will power the GWAS analysis between azole-susceptibility and azole-resistant strains. Moreover, with the advantages of long-read sequenced genome assemblies, we will be able to perform GWAS for structural variants such as CNV. They will provide the most comprehensive genome variant association studies in the future.

Network-based analyses and machine learning facilitated the discovery of potential bacterial, fungal, and co-infection host biomarkers for pre-symptomatic sepsis patients

This part will discuss the analyses from **manuscript II**, which focused on identifying the potential biomarkers for pre-symptomatic sepsis patients induced by bacterial, fungal, and co-infection based on their gene expression data. Sepsis is a complex disease that activates the innate immune system through several processes such as inflammatory, coagulation, and metabolism processes [202]. Recently, over 200 sepsis biomarkers were identified by more than 5,000 different studies [203]. However, there is still a lack of biomarkers to identify causative pathogens that will lead to appropriate treatments. Since the majority of causative pathogens were bacteria (60%-90% of identified pathogens) [204], several previous studies tried to develop biomarkers specific to bacteria-initiated sepsis patients. For example, Ramilo *et al*. and Cernada *et al*. analyzed transcriptome data to discriminate between gram-positive and gram-negative bacteria infection in sepsis patients [205,206]. However, identifying fungal pathogens in sepsis patients mainly relies on blood culture, of which 20% showed negative results [204].

Dix *et al*. recently used transcriptome data from microarrays to build a classifier to discriminate between bacterial and fungal blood infections. The classifier was built based on biomarker genes selected by differentially expressed gene analyses [207]. However, their blood samples were obtained from healthy volunteers. They were subsequently infected *in vitro* by sepsis causative microorganisms, which showed a limitation in representing the immune responses in sepsis patients. This study collected transcriptome data from sepsis patients infected by bacteria, fungi, and co-infection of both pathogen types. I also used weighted gene co-expression and protein-protein interaction networks to identify hub genes, which play essential roles in driving biological processes through several gene/protein interactions. `Weighted gene co-expression network analysis (WGCNA)` was successfully used to identify candidate biomarker genes for many cancers such as breast cancer [208], glioblastoma [209,210] as well as sepsis [211,212]. Based on network analyses and machine learning that was used to evaluate the pathogen discriminating performances of each gene, I could identify host genes and pathways that responded differently to bacteria, fungi, and co-infection of both pathogens in sepsis patients. Genes and pathways involved in NK-cell activation were highly associated with bacterial sepsis, while genes involved in transcriptional processes were highly associated with fungal and co-infection sepsis. Moreover, co-infection sepsis also showed highly expressed genes that are poor prognosis genes related to various pathogen infections.

This study provided valuable hypotheses of the host immune system specific to different types of pathogens that will narrow the choice of antimicrobial use in sepsis patients. However, the main limitation of this study was significantly fewer fungal and co-infected sepsis samples.

Future perspective

A more significant number of sepsis samples infected by fungal and co-infection is necessary to advance the machine learning model building for more reliable results. Furthermore, transcriptome data generated by sequencing technologies (RNA-seq) can provide broader dynamic ranges of gene-expression profiles. This technology will be an advantage in identifying more comprehensive DEGs and biomarker genes. More importantly, as mentioned that sepsis is one of the complex diseases, integration of multi-omics data, for example, genomics, transcriptomics, and proteomics, will allow us to identify more comprehensive and robust biomarkers. Once a handful of biomarkers are identified, the new portable sequencing technology such as MinION will have a critical role in routine sepsis diagnosis and prognosis. Altogether, they will improve the diagnostic turnaround times and personalized treatments for sepsis patients, including patients infected by non-bacterial pathogens.

REFERENCES

[1] Muñoz JF, Gallo JE, Misas E, Priest M, Imamovic A, Young S, et al. Genome Update of the Dimorphic Human Pathogenic Fungi Causing Paracoccidioidomycosis. PLoS Negl Trop Dis 2014;8:e3348.

[2] Kisand V, Lettieri T. Genome sequencing of bacteria: Sequencing, de novo assembly and rapid analysis using open source tools. BMC Genomics 2013;14:1–11.

[3] Zender L, Villanueva A, Tovar V, Sia D, Chiang DY, Llovet JM. Cancer gene discovery in hepatocellular carcinoma. J Hepatol 2010;52:921–9.

[4] Lander ES. Initial impact of the sequencing of the human genome. Nat 2011 4707333 2011;470:187–97.

[5] Blackwell GA, Hunt M, Malone KM, Lima L, Horesh G, Alako BTF, et al. Exploring bacterial diversity via a curated and searchable snapshot of archived DNA sequences. BioRxiv 2021:2021.03.02.433662.

[6] Grigoriev I V., Nikitin R, Haridas S, Kuo A, Ohm R, Otillar R, et al. MycoCosm portal: gearing up for 1000 fungal genomes. Nucleic Acids Res 2014;42:D699–704.

[7] Basenko EY, Pulman JA, Shanmugasundram A, Harb OS, Crouch K, Starns D, et al. FungiDB: An Integrated Bioinformatic Resource for Fungi and Oomycetes. J Fungi 2018, Vol 4, Page 39 2018;4:39.

[8] Kersey PJ, Lawson D, Birney E, Derwent PS, Haimel M, Herrero J, et al. Ensembl Genomes: Extending Ensembl across the taxonomic space. Nucleic Acids Res 2010;38:D563–9.

[9] Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 1977;74:5463.

[10] Craig Venter J, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science (80-) 2001;291:1304–51.

[11] Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nat 2001 4096822 2001;409:860–921.

[12] Morozova O, Marra MA. Applications of next-generation sequencing technologies in functional genomics. Genomics 2008;92:255–64.

[13] Gupta N, Verma VK. Next-Generation Sequencing and Its Application: Empowering in Public Health Beyond Reality. Microb Technol Welf Soc 2019;17:313.

[14] Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 2016 176 2016;17:333–51.

[15] Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nat 2005 4377057 2005;437:376–80.

[16] Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate Whole Human Genome Sequencing using Reversible Terminator Chemistry. Nature 2008;456:53.

[17] Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, et al. Accurate multiplex polony sequencing of an evolved bacterial genome. Science 2005;309:1728–32.

[18] Clarke L, Zheng-Bradley X, Smith R, Kulesha E, Xiao C, Toneva I, et al. The 1000 Genomes Project: data management and community access. Nat Methods 2012 95 2012;9:459–62.

[19] Delaneau O, Marchini J, McVeanh GA, Donnelly P, Lunter G, Marchini JL, et al. Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. Nat Commun 2014 51 2014;5:1–9.

[20] Nikpay M, Goel A, Won HH, Hall LM, Willenborg C, Kanoni S, et al. A comprehensive 1000 Genomes-based genome-wide association meta-analysis of coronary artery disease. Nat Genet 2015;47:1121.

[21] Keaton JM, Gao C, Guan M, Hellwege JN, Palmer ND, Pankow JS, et al. Genome-wide interaction with the insulin secretion locus MTNR1B reveals CMIP as a novel type 2 diabetes susceptibility gene in African Americans. Genet Epidemiol 2018;42:559–70.

[22] Chande AT, Chande AT, Chande AT, Rishishwar L, Rishishwar L, Conley AB, et al. Ancestry effects on type 2 diabetes genetic risk inference in Hispanic/Latino populations. BMC Med Genet 2020;21:1–14.

[23] Kasela S, Daniloski Z, Bollepalli S, Jordan TX, tenOever BR, Sanjana NE, et al. Integrative approach identifies SLC6A20 and CXCR6 as putative causal genes for the COVID-19 GWAS signal in the 3p21.31 locus. Genome Biol 2021;22:1–10.

[24] Ronholm J, Nasheri N, Petronella N, Pagotto F. Navigating Microbiological Food Safety in the Era of Whole-Genome Sequencing. Clin Microbiol Rev 2016;29:837.

[25] Voelkerding K V., Dames SA, Durtschi JD. Next-Generation Sequencing: From Basic Research to Diagnostics. Clin Chem 2009;55:641–58.

[26] Ardui S, Ameur A, Vermeesch JR, Hestand MS. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. Nucleic Acids Res 2018;46:2159–68.

[27] Harris TD, Buzby PR, Babcock H, Beer E, Bowers J, Braslavsky I, et al. Single-molecule DNA sequencing of a viral genome. Science (80-) 2008;320:106–9.

[28] Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. Real-time DNA sequencing from single polymerase molecules. Science (80-) 2009;323:133–8.

[29] Mizuguchi T, Toyota T, Adachi H, Miyake N, Matsumoto N, Miyatake S. Detecting a long insertion variant in SAMD12 by SMRT sequencing: implications of long-read whole-genome sequencing for repeat expansion diseases. J Hum Genet 2018 643 2018;64:191–7.

[30] Jain M, Olsen HE, Paten B, Akeson M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biol 2016 171 2016;17:1–11.

[31] Deamer D, Akeson M, Branton D. Three decades of nanopore sequencing. Nat Biotechnol 2016;34:518.

[32] Cordero P, Ashley EA. Whole-Genome Sequencing in Personalized Therapeutics. Clin Pharmacol Ther 2012;91:1001–9.

[33] Chinen J, Lawrence M, Dorsey M, Kobrynski LJ. Practical approach to genetic testing for primary immunodeficiencies. Ann Allergy, Asthma Immunol 2019;123:433–9.

[34] Han SM, Park J, Lee JH, Lee SS, Kim H, Han H, et al. Targeted Next-Generation Sequencing for Comprehensive Genetic Profiling of Pharmacogenes. Clin Pharmacol Ther 2017;101:396–405.

[35] Hadd AG, Houghton J, Choudhary A, Sah S, Chen L, Marko AC, et al. Targeted, High-Depth, Next-Generation Sequencing of Cancer Genes in Formalin-Fixed, Paraffin-Embedded and Fine-Needle Aspiration Tumor Specimens. J Mol Diagnostics 2013;15:234–47.

[36] Bartoletti-Stella A, Baiardi S, Stanzani-Maserati M, Piras S, Caffarra P, Raggi A, et al. Identification of rare genetic variants in Italian patients with dementia by targeted

gene sequencing. Neurobiol Aging 2018;66:180.e23-180.e31.

[37] Clark MJ, Chen R, Lam HYK, Karczewski KJ, Chen R, Euskirchen G, et al. Performance comparison of exome DNA sequencing technologies. Nat Biotechnol 2011;29:908.

[38] Qi XP, Du ZF, Ma JM, Chen XL, Zhang Q, Fei J, et al. Genetic diagnosis of autosomal dominant polycystic kidney disease by targeted capture and next-generation sequencing: Utility and limitations. Gene 2013;516:93–100.

[39] Antoniadi T, Buxton C, Dennis G, Forrester N, Smith D, Lunt P, et al. Application of targeted multi-gene panel testing for the diagnosis of inherited peripheral neuropathy provides a high diagnostic yield with unexpected phenotype-genotype variability. BMC Med Genet 2015;16:1–11.

[40] Bewicke-Copley F, Arjun Kumar E, Palladino G, Korfi K, Wang J. Applications and analysis of targeted genomic sequencing in cancer studies. Comput Struct Biotechnol J 2019;17:1348.

[41] Klein HG, Bauer P, Hambuch T. Whole genome sequencing (WGS), whole exome sequencing (WES) and clinical exome sequencing (CES) in patient care. LaboratoriumsMedizin 2014;38:221–30.

[42] Nicastro E, D'Antiga L. Next generation sequencing in pediatric hepatology and liver transplantation. Liver Transplant 2018;24:282–93.

[43] Keepers KG, Pogoda CS, White KH, Anderson Stewart CR, Hoffman JR, Ruiz AM, et al. Whole Genome Shotgun Sequencing Detects Greater Lichen Fungal Diversity Than Amplicon-Based Methods in Environmental Samples. Front Ecol Evol 2019;7:484.

[44] Pereira R, Oliveira J, Sousa M. Bioinformatics and Computational Tools for Next-Generation Sequencing Analysis in Clinical Genetics. J Clin Med 2020, Vol 9, Page 132 2020;9:132.

[45] Lüth S, Kleta S, Al Dahouk S. Whole genome sequencing as a typing tool for foodborne pathogens like Listeria monocytogenes – The way towards global harmonisation and data exchange. Trends Food Sci Technol 2018;73:67–75.

[46] Cacho A, Smirnova E, Huzurbazar S, Cui X. A Comparison of Base-calling Algorithms for Illumina Sequencing Technology. Brief Bioinform 2016;17:786–95.

[47] Ewing B, Green P. Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities. Genome Res 1998;8:186–94.

[48] Peng X, Wang J, Zhang Z, Xiao Q, Li M, Pan Y. Re-alignment of the unmapped reads with base quality score. BMC Bioinformatics 2015;16:1–10.

[49] Andrews S. FASTQC. A quality control tool for high throughput sequence data 2016.

[50] Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014;30:2114–20.

[51] Krueger F. Trim galore. A Wrapper Tool around Cutadapt FastQC to Consistently Apply Qual Adapt Trimming to FastQ Files 2015;516.

[52] Wilton R, Szalay AS. Performance optimization in DNA short-read alignment. Bioinformatics 2022.

[53] Sangiovanni M, Granata I, Thind AS, Guarracino MR. From trash to treasure: Detecting unexpected contamination in unmapped NGS data. BMC Bioinformatics 2019;20:1–12.

[54] Ku CS, Loy EY, Salim A, Pawitan Y, Chia KS. The discovery of human genetic variations and their use as disease markers: past, present and future. J Hum Genet 2010 557 2010;55:403–15.

[55] Sheynkman GM, Shortreed MR, Frey BL, Scalf M, Smith LM. Large-scale mass

spectrometric detection of variant peptides resulting from nonsynonymous nucleotide differences. J Proteome Res 2014;13:228–40.

[56] Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. Nat Rev Genet 2006 72 2006;7:85–97.

[57] Flchant GA, Quentln Y. A frameshift error detection algorithm for DNA sequencing projects. Nucleic Acids Res 1995;23:2900–8.

[58] Guan X, Uberbacher EC. Alignments of DNA and protein sequences containing frameshift errors. Comput Appl Biosci 1996;12:31–40.

[59] Lupski JR. Structural variation mutagenesis of the human genome: Impact on disease and evolution. Environ Mol Mutagen 2015;56:419–36.

[60] Fonseca NA, Rung J, Brazma A, Marioni JC. Tools for mapping high-throughput sequencing data. Bioinformatics 2012;28:3169–77.

[61] Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 2009;25:1754–60.

[62] Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012 94 2012;9:357–9.

[63] Burrows M, Burrows M, Wheeler D. A Block-Sorting Lossless Data Compression Algorithm. Digit SRC Res Rep 1994:12--4.

[64] Cox AJ, Bauer MJ, Jakobi T, Rosone G. Large-scale compression of genomic sequence databases with the Burrows–Wheeler transform. Bioinformatics 2012;28:1415–9.

[65] Keel BN, Snelling WM. Comparison of Burrows-Wheeler transform-based mapping algorithms used in high-throughput whole-genome sequencing: Application to illumina data for livestock genomes 1. Front Genet 2018;9:35.

[66] Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet 2011 126 2011;12:443–51.

[67] McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010;20:1297–303.

[68] Talwalkar A, Liptrap J, Newcomb J, Hartl C, Terhorst J, Curtis K, et al. SM a SH: a benchmarking toolkit for human genome variant calling. Bioinformatics 2014;30:2787–95.

[69] Pirooznia M, Kramer M, Parla J, Goes FS, Potash JB, McCombie WR, et al. Validation and assessment of variant calling pipelines for next-generation sequencing. Hum Genomics 2014;8:14.

[70] Bauer DC. Variant calling comparison CASAVA1.8 and GATK. Nat Preced 2011 2011:1–1.

[71] Hwang S, Kim E, Lee I, Marcotte EM. Systematic comparison of variant calling pipelines using gold standard personal exome variants. Sci Rep 2015;5.

[72] Kim BY, Park JH, Jo HY, Koo SK, Park MH. Optimized detection of insertions/deletions (INDELs) in whole-exome sequencing data. PLoS One 2017;12:e0182272.

[73] Andreu-Sánchez S, Chen L, Wang D, Augustijn HE, Zhernakova A, Fu J. A Benchmark of Genetic Variant Calling Pipelines Using Metagenomic Short-Read Sequencing. Front Genet 2021;12:537.

[74] Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 2012;6:80–92.

[75] Hurles ME, Dermitzakis ET, Tyler-Smith C. The functional impact of structural

variation in humans. Trends Genet 2008;24:238–45.

[76] Buchanan JA, Scherer SW. Contemplating effects of genomic structural variation. Genet Med 2008;10:639–47.

[77] Weischenfeldt J, Symmons O, Spitz F, Korbel JO. Phenotypic impact of genomic structural variation: insights from and for human disease. Nat Rev Genet 2013 142 2013;14:125–38.

[78] Zhao M, Wang Q, Wang Q, Jia P, Zhao Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: Features and perspectives. BMC Bioinformatics 2013;14:1–16.

[79] Boeva V, Popova T, Bleakley K, Chiche P, Cappo J, Schleiermacher G, et al. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. Bioinformatics 2012;28:423–5.

[80] Alkodsi A, Louhimo R, Hautaniemi S. Comparative analysis of methods for identifying somatic copy number alterations from deep sequencing data. Brief Bioinform 2015;16:242–54.

[81] Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 2010;38.

[82] McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics 2010;26:2069.

[83] Park K-J, Park J-H. Variations in Nomenclature of Clinical Variants between Annotation Tools. Lab Med 2021.

[84] Menon S. COMPARISON OF HIGH-THROUGHPUT NEXT GENERATION SEQUENCING DATA PROCESSING PIPELINES. Int Res J Mod Eng Technol Sci 2021;03:125–36.

[85] Sherman RM, Salzberg SL. Pan-genomics in the human genome era. Nat Rev Genet 2020 214 2020;21:243–54.

[86] Garcia-Rubio R, Monzon S, Alcazar-Fuoli L, Cuesta I, Mellado E. Genome-Wide Comparative Analysis of Aspergillus fumigatus Strains: The Reference Genome as a Matter of Concern. Genes 2018, Vol 9, Page 363 2018;9:363.

[87] Liao X, Li M, Zou Y, Wu FX, Yi-Pan, Wang J. Current challenges and solutions of <strong>*de novo*</strong> assembly. Quant Biol 2019;7:90–109.

[88] Chaisson MJP, Wilson RK, Eichler EE. Genetic variation and the de novo assembly of human genomes. Nat Rev Genet 2015 1611 2015;16:627–40.

[89] Paszkiewicz K, Studholme DJ. De novo assembly of short sequence reads. Brief Bioinform 2010;11:457–72.

[90] Zhang W, Chen J, Yang Y, Tang Y, Shang J, Shen B. A Practical Comparison of De Novo Genome Assembly Software Tools for Next-Generation Sequencing Technologies. PLoS One 2011;6:e17915.

[91] Flicek P, Birney E. Sense from sequence reads: methods for alignment and assembly. Nat Methods 2009 611 2009;6:S6–12.

[92] Lischer HEL, Shimizu KK. Reference-guided de novo assembly approach improves genome reconstruction for related species. BMC Bioinformatics 2017;18:1–12.

[93] Baker M. De novo genome assembly: what every biologist should know. Nat Methods 2012 94 2012;9:333–7.

[94] Narzisi G, Mishra B. Comparing De Novo Genome Assembly: The Long and Short of It. PLoS One 2011;6:e19175.

[95] Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics 2013;29:1072.

[96] Waterhouse RM, Seppey M, Simao FA, Manni M, Ioannidis P, Klioutchnikov G, et

al. BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. Mol Biol Evol 2018;35:543–8.

[97] Wang Z, Chen Y, Li Y. A Brief Review of Computational Gene Prediction Methods. Genomics Proteomics Bioinformatics 2004;2:216–21.

[98] Sleator RD. An overview of the current status of eukaryote gene prediction strategies. Gene 2010;461:1–4.

[99] Stanke M, Waack S. Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics 2003;19 Suppl 2.

[100] Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, Borodovsky M. Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. Genome Res 2008;18:1979–90.

[101] Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol 2008;9:1–22.

[102] Palmer J, Stajich J. Funanntoate: automated eukaryotic genome annotation pipeline 2019.

[103] Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic Acids Res 2014;42:D222–30.

[104] Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, et al. UniProt: the Universal Protein knowledgebase. Nucleic Acids Res 2004;32:D115–9.

[105] Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res 2016;44:D457–62.

[106] Kanehisa M, Sato Y. KEGG Mapper for inferring cellular functions from protein sequences. Protein Sci 2020;29:28–35.

[107] Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 1989;123:585–95.

[108] Nordborg M, Tavaré S. Linkage disequilibrium: what history has to tell us. Trends Genet 2002;18:83–90.

[109] Aguadé M. Nucleotide sequence variation at two genes of the phenylpropanoid pathway, the FAH1 and F3H genes, in Arabidopsis thaliana. Mol Biol Evol 2001;18:1–9.

[110] Wright S. ISOLATION BY DISTANCE. Genetics 1943;28:114–38.

[111] Gabaldón T, Koonin E V. Functional and evolutionary implications of gene orthology. Nat Rev Genet 2013 145 2013;14:360–6.

[112] Vernikos G, Medini D, Riley DR, Tettelin H. Ten years of pan-genome analyses. Curr Opin Microbiol 2015;23:148–54.

[113] Nichio BTL, Marchaukoski JN, Raittz RT. New tools in orthology analysis: A brief review of promising perspectives. Front Genet 2017;8:165.

[114] Emms DM, Kelly S. OrthoFinder: Phylogenetic orthology inference for comparative genomics. Genome Biol 2019;20:1–14.

[115] Sul JH, Martin LS, Eskin E. Population structure in genetic studies: Confounding factors and mixed models. PLOS Genet 2018;14:e1007309.

[116] Korte A, Farlow A. The advantages and limitations of trait analysis with GWAS: A review. Plant Methods 2013;9:1–9.

[117] Brachi B, Faure N, Horton M, Flahauw E, Vazquez A, Nordborg M, et al. Linkage and Association Mapping of Arabidopsis thaliana Flowering Time in Nature. PLOS Genet 2010;6:e1000940.

[118] Marchini J, Cardon LR, Phillips MS, Donnelly P. The effects of human population structure on large genetic association studies. Nat Genet 2004 365 2004;36:512–7.

[119] Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, et al. Variance

component model to account for sample structure in genome-wide association studies. Nat Genet 2010 424 2010;42:348–54.

[120] Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nat Genet 2012 447 2012;44:821–4.

[121] Brynildsrud O, Bohlin J, Scheffer L, Eldholm V. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. Genome Biol 2016;17:1–9.

[122] Rizzo J, Rouchka EC. Review of Phylogenetic Tree Construction. Kenucky, USA: 2007.

[123] Kapli P, Yang Z, Telford MJ. Phylogenetic tree building in the genomic age. Nat Rev Genet 2020 217 2020;21:428–44.

[124] Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 1987;4:406–25.

[125] Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 2003;52:696–704.

[126] Lake JA, Moore JE. Phylogenetic analysis and comparative genomics. Trends Biotechnol 1998;16:22–3.

[127] Minh BQ, Nguyen MAT, Von Haeseler A. Ultrafast Approximation for Phylogenetic Bootstrap. Mol Biol Evol 2013;30:1188–95.

[128] Dong ZC, Chen Y. Transcriptomics: Advances and approaches. Sci China Life Sci 2013 5610 2013;56:960–7.

[129] Abdel-Aziz MI, Neerincx AH, Vijverberg SJ, Kraneveld AD, Maitland-van der Zee AH. Omics for the future in asthma. Semin Immunopathol 2020;42:111–26.

[130] Clark NR, Hu KS, Feldmann AS, Kou Y, Chen EY, Duan Q, et al. The characteristic direction: A geometrical approach to identify differentially expressed genes. BMC Bioinformatics 2014;15:1–16.

[131] Wong HR. Clinical review: Sepsis and septic shock - the potential of gene arrays. Crit Care 2012;16:1–8.

[132] Barrett T, Edgar R. [19] Gene Expression Omnibus: Microarray Data Storage, Submission, Retrieval, and Analysis. Methods Enzymol 2006;411:352–69.

[133] Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 2002;30:207–10.

[134] Mantione KJ, Kream RM, Kuzelova H, Ptacek R, Raboch J, Samuel JM, et al. Comparing Bioinformatic Gene Expression Profiling Methods: Microarray and RNA-Seq. Med Sci Monit Basic Res 2014;20:138.

[135] Leinonen R, Sugawara H, Shumway M, Collaboration on behalf of the INSD. The Sequence Read Archive. Nucleic Acids Res 2011;39:D19–21.

[136] Stark R, Grzelak M, Hadfield J. RNA sequencing: the teenage years. Nat Rev Genet 2019 2011 2019;20:631–56.

[137] RNA Sequencing VS Microarray - Otogenetics n.d. https://www.otogenetics.com/rna-sequencing-vs-microarray/ (accessed 20 April 2022).

[138] Evans C, Hardin J, Stoebel DM. Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. Brief Bioinform 2018;19:776.

[139] Liu X, Li N, Liu S, Wang J, Zhang N, Zheng X, et al. Normalization Methods for the Analysis of Unbalanced Transcriptome Data: A Review. Front Bioeng Biotechnol 2019;7:358.

[140] Smyth GK, Speed T. Normalization of cDNA microarray data. Methods 2003;31:265–73.

[141] Casamassimi A, Federico A, Rienzo M, Esposito S, Ciccodicola A. Transcriptome

Profiling in Human Diseases: New Advances and Perspectives. Int J Mol Sci 2017, Vol 18, Page 1652 2017;18:1652.

[142] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015;43:e47–e47.

[143] R Core Team. R version 4.0. 3: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria 2020. https://www.r-project.org/.

[144] Liang J, Kachalo S. Computational analysis of microarray gene expression profiles: clustering, classification, and beyond. Chemom Intell Lab Syst 2002;62:199–216.

[145] Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol 2005;4.

[146] Hao T, Peng W, Wang Q, Wang B, Sun J. Reconstruction and Application of Protein–Protein Interaction Network. Int J Mol Sci 2016, Vol 17, Page 907 2016;17:907.

[147] Kadarmideen HN, Watson-haigh NS. Building gene co-expression networks using transcriptomics data for systems biology investigations: Comparison of methods using microarray data. Bioinformation 2012;8:855.

[148] Keller EF. Revisiting “scale-free” networks. BioEssays 2005;27:1060–8.

[149] van Dam S, Võsa U, van der Graaf A, Franke L, de Magalhães JP. Gene co-expression analysis for functional classification and gene–disease predictions. Brief Bioinform 2018;19:575–92.

[150] Brazma A, Vilo J. Gene expression data analysis. FEBS Lett 2000;480:17–24.

[151] Ghosh A, Barman S. Application of Euclidean distance measurement and principal component analysis for gene identification. Gene 2016;583:112–20.

[152] Mason MJ, Fan G, Plath K, Zhou Q, Horvath S. Signed weighted gene co-expression network analysis of transcriptional regulation in murine embryonic stem cells. BMC Genomics 2009;10:1–25.

[153] Langfelder P, Horvath S. WGCNA: An R package for weighted correlation network analysis. BMC Bioinformatics 2008;9:1–13.

[154] Nibbe RK, Chowdhury SA, Koyutürk M, Ewing R, Chance MR. Protein–protein interaction networks and subnetworks in the biology of disease. Wiley Interdiscip Rev Syst Biol Med 2011;3:357–67.

[155] von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, et al. STRING: known and predicted protein–protein associations, integrated and transferred across organisms. Nucleic Acids Res 2005;33:D433–7.

[156] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genet 2000 251 2000;25:25–9.

[157] Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 2000;28:27–30.

[158] Geistlinger L, Csaba G, Santarelli M, Ramos M, Schiffer L, Turaga N, et al. Toward a gold standard for benchmarking gene set enrichment analysis. Brief Bioinform 2021;22:545–56.

[159] Hung JH, Yang TH, Hu Z, Weng Z, DeLisi C. Gene set enrichment analysis: performance evaluation and usage guidelines. Brief Bioinform 2012;13:281–91.

[160] Simillion C, Liechti R, Lischer HEL, Ioannidis V, Bruggmann R. Avoiding the pitfalls of gene set enrichment analysis with SetRank. BMC Bioinformatics 2017;18:1–14.

[161] Kwok AJ, Mentzer A, Knight JC. Host genetics and infectious disease: new tools, insights and translational opportunities. Nat Rev Genet 2020 223 2020;22:137–53.

[162] Fauci AS. Infectious diseases: Considerations for the 21st century. Clin Infect Dis 2001;32:675–85.

[163] Kumar V, Cheng SC, Johnson MD, Smeekens SP, Wojtowicz A, Giamarellos-Bourboulis E, et al. Immunochip SNP array identifies novel genetic variants conferring susceptibility to candidaemia. Nat Commun 2014 51 2014;5:1–8.

[164] Méric G, Mageiros L, Pensar J, Laabei M, Yahara K, Pascoe B, et al. Disease-associated genotypes of the commensal skin bacterium Staphylococcus epidermidis. Nat Commun 2018 91 2018;9:1–11.

[165] Reid SD, Herbelin CJ, Bumbaugh AC, Selander RK, Whittam TS. Parallel evolution of virulence in pathogenic Escherichia coli. Nat 2000 4066791 2000;406:64–7.

[166] Kiss E, Hegedüs B, Virágh M, Varga T, Merényi Z, Kószó T, et al. Comparative genomics reveals the origin of fungal hyphae and multicellularity. Nat Commun 2019 101 2019;10:1–13.

[167] Kim CY, Lee M, Lee K, Yoon SS, Lee I. Network-based genetic investigation of virulence-associated phenotypes in methicillin-resistant Staphylococcus aureus. Sci Reports 2018 81 2018;8:1–12.

[168] Dissanayake TK, Schäuble S, Mirhakkak M, Wu WL, Ng ACK, Yip CCY, et al. Comparative Transcriptomic Analysis of Rhinovirus and Influenza Virus Infection. Front Microbiol 2020;11.

[169] Parnell GP, McLean AS, Booth DR, Armstrong NJ, Nalos M, Huang SJ, et al. A distinct influenza infection signature in the blood transcriptome of patients with severe community-acquired pneumonia. Crit Care 2012;16:1–12.

[170] Sweeney TE, Azad TD, Donato M, Haynes WA, Perumal TM, Henao R, et al. Unsupervised analysis of transcriptomics in bacterial sepsis across multiple datasets reveals three robust clusters. Crit Care Med 2018;46:915.

[171] Zeng X, Feng J, Yang Y, Zhao R, Yu Q, Qin H, et al. Screening of Key Genes of Sepsis and Septic Shock Using Bioinformatics Analysis. J Inflamm Res 2021;14:829.

[172] Zhai J, Qi A, Zhang Y, Jiao L, Liu Y, Shou S. Bioinformatics Analysis for Multiple Gene Expression Profiles in Sepsis. Med Sci Monit 2020;26:e920818-1.

[173] Tong DL, Kempsell KE, Szakmany T, Ball G. Development of a Bioinformatics Framework for Identification and Validation of Genomic Biomarkers and Key Immunopathology Processes and Controllers in Infectious and Non-infectious Severe Inflammatory Response Syndrome. Front Immunol 2020;11:380.

[174] Latgé JP. The pathobiology of Aspergillus fumigatus. Trends Microbiol 2001;9:382–9.

[175] Kuster S, Stampf S, Gerber B, Baettig V, Weisser M, Gerull S, et al. Incidence and outcome of invasive fungal diseases after allogeneic hematopoietic stem cell transplantation: A Swiss transplant cohort study. Transpl Infect Dis 2018;20:e12981.

[176] Harrison N, Mitterbauer M, Tobudic S, Kalhs P, Rabitsch W, Greinix H, et al. Incidence and characteristics of invasive fungal diseases in allogeneic hematopoietic stem cell transplant recipients: A retrospective cohort study. BMC Infect Dis 2015;15:1–9.

[177] Arastehfar A, Carvalho A, Houbraken J, Lombardi L, Garcia-Rubio R, Jenks JD, et al. Aspergillus fumigatus and aspergillosis: From basics to clinics. Stud Mycol 2021;100:100115.

[178] Resendiz-Sharpe A, group on behalf of the D-BM study, Mercier T, group on behalf of the D-BM study, Lestrade PPA, group on behalf of the D-BM study, et al. Prevalence of voriconazole-resistant invasive aspergillosis and its impact on mortality in haematology patients. J Antimicrob Chemother 2019;74:2759–66.

[179] Nywening A V., Rybak JM, Rogers PD, Fortwendel JR. Mechanisms of triazole resistance in Aspergillus fumigatus. Environ Microbiol 2020;22:4934–52.

[180] Singer M, Deutschman CS, Seymour C, Shankar-Hari M, Annane D, Bauer M, et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA 2016;315:801.

[181] Rello J, Valenzuela-Sánchez F, Ruiz-Rodriguez M, Moyano S. Sepsis: A Review of Advances in Management. Adv Ther 2017 3411 2017;34:2393–411.

[182] Bauer M, Gerlach H, Vogelmann T, Preissing F, Stiefel J, Adam D. Mortality in sepsis and septic shock in Europe, North America and Australia between 2009 and 2019-results from a systematic review and meta-analysis. Crit Care 2020;24:1–9.

[183] Nasir N, Jamil B, Siddiqui S, Talat N, Khan FA, Hussain R. Mortality in Sepsis and its relationship with Gender. Pakistan J Med Sci 2015;31:1201.

[184] Vincent JL. The Clinical Challenge of Sepsis Identification and Monitoring. PLOS Med 2016;13:e1002022.

[185] Trzeciak A, Pietropaoli AP, Kim M. Biomarkers and Associated Immune Mechanisms for Early Detection and Therapeutic Management of Sepsis. Immune Netw 2020;20:1–20.

[186] Casadevall A. Fungal Diseases in the 21st Century: The Near and Far Horizons. Pathog Immun 2018;3:183.

[187] Lofgren LA, Ross BS, Cramer RA, Stajich JE. Combined Pan-, Population-, and Phylo-Genomic Analysis of Aspergillus fumigatus Reveals Population Structure and Lineage-Specific Diversity. BioRxiv 2022:2021.12.12.472145.

[188] McCarthy CGP, Fitzpatrick DA. Pan-genome analyses of model fungal species. Microb Genomics 2019;5:e000243.

[189] Nielsen KN, Sirén K, Petersen B, Sicheritz-Pontén T, Gilbert MTP, Korneliussen TS, et al. The pangenome of the fungal pathogen Neonectria neomacrospora. BioRxiv 2021:2021.03.11.434922.

[190] Abdolrasouli A, Rhodes J, Beale MA, Hagen F, Rogers TR, Chowdhary A, et al. Genomic context of azole resistance mutations in Aspergillus fumigatus determined using whole-genome sequencing. MBio 2015;6.

[191] Fan Y, Wang Y, Xu J. Comparative Genome Sequence Analyses of Geographic Samples of Aspergillus fumigatus—Relevance for Amphotericin B Resistance. Microorg 2020, Vol 8, Page 1673 2020;8:1673.

[192] Camps SMT, Rijs AJMM, Klaassen CHW, Meis JF, O'Gorman CM, Dyer PS, et al. Molecular epidemiology of Aspergillus fumigatus isolates harboring the TR34/L98H azole resistance mechanism. J Clin Microbiol 2012;50:2674–80.

[193] Risum M, Hare RK, Gertsen JB, Kristensen L, Johansen HK, Helweg-Larsen J, et al. Azole-Resistant Aspergillus fumigatus Among Danish Cystic Fibrosis Patients: Increasing Prevalence and Dominance of TR34/L98H. Front Microbiol 2020;11:1850.

[194] Snelders E, Van Der Lee HAL, Kuijpers J, Rijs AJMM, Varga J, Samson RA, et al. Emergence of Azole Resistance in Aspergillus fumigatus and Spread of a Single Resistance Mechanism. PLOS Med 2008;5:e219.

[195] Tang CM, Cohen J, Rees AJ, Holden DW. Molecular epidemiological study of invasive pulmonary aspergillosis in a renal transplantation unit. Eur J Clin Microbiol Infect Dis 1994 134 1994;13:318–21.

[196] Debeaupuis JP, Sarfati J, Chazalet V, Latgé JP. Genetic diversity among clinical and environmental isolates of Aspergillus fumigatus. Infect Immun 1997;65:3080–5.

[197] Aufauvre-Brown A, Brown JS, Holden DW. Comparison of Virulence between Clinical and Environmental Isolates of Aspergillus fumigatus. Eur J Clin Microbiol

Infect Dis 1998 1711 1998;17:778–80.
[198] Bultman KM, Kowalski CH, Cramer RA. Aspergillus fumigatus virulence through the lens of transcription factors. Med Mycol 2017;55:24–38.
[199] Bertuzzi M, Schrettl M, Alcazar-Fuoli L, Cairns TC, Muñoz A, Walker LA, et al. The pH-Responsive PacC Transcription Factor of Aspergillus fumigatus Governs Epithelial Entry and Tissue Invasion during Pulmonary Aspergillosis. PLOS Pathog 2014;10:e1004413.
[200] Pongpom M, Liu H, Xu W, Snarr BD, Sheppard DC, Mitchell AP, et al. Divergent targets of Aspergillus fumigatus AcuK and AcuM transcription factors during growth in vitro versus invasive disease. Infect Immun 2015;83:923–33.
[201] Willger SD, Puttikamonkul S, Kim KH, Burritt JB, Grahl N, Metzler LJ, et al. A Sterol-Regulatory Element Binding Protein Is Required for Cell Polarity, Hypoxia Adaptation, Azole Drug Resistance, and Virulence in Aspergillus fumigatus. PLOS Pathog 2008;4:e1000200.
[202] Pierrakos C, Vincent JL. Sepsis biomarkers: A review. Crit Care 2010;14:1–18.
[203] Pierrakos C, Velissaris D, Bisdorff M, Marshall JC, Vincent JL. Biomarkers of sepsis: time for a reappraisal. Crit Care 2020 241 2020;24:1–15.
[204] Prescott HC. The Epidemiology of Sepsis. Handb Sepsis 2018:15–28.
[205] Ramilo O, Allman W, Chung W, Mejias A, Ardura M, Glaser C, et al. Gene expression patterns in blood leukocytes discriminate patients with acute infections. Blood 2007;109:2066–77.
[206] Cernada M, Pinilla-González A, Kuligowski J, Morales JM, Lorente-Pozo S, Piñeiro-Ramos JD, et al. Transcriptome profiles discriminate between Gram-positive and Gram-negative sepsis in preterm neonates. Pediatr Res 2021 913 2021;91:637–45.
[207] Dix A, Hünniger K, Weber M, Guthke R, Kurzai O, Linde J. Biomarker-based classification of bacterial and fungal whole-blood infections in a genome-wide expression study. Front Microbiol 2015;6:171.
[208] Jia R, Zhao H, Jia M. Identification of co-expression modules and potential biomarkers of breast cancer by WGCNA. Gene 2020;750:144757.
[209] Yang Q, Wang R, Wei B, Peng C, Wang L, Hu G, et al. Candidate Biomarkers and Molecular Mechanism Investigation for Glioblastoma Multiforme Utilizing WGCNA. Biomed Res Int 2018;2018.
[210] Zhou J, Guo H, Liu L, Hao S, Guo Z, Zhang F, et al. Construction of co-expression modules related to survival by WGCNA and identification of potential prognostic biomarkers in glioblastoma. J Cell Mol Med 2021;25:1633–44.
[211] Xu C, Xu J, Lu L, Tian W, Ma J, Wu M. Identification of key genes and novel immune infiltration-associated biomarkers of sepsis. Innate Immun 2020;26:666–82.
[212] Li Z, Huang B, Yi W, Wang F, Wei S, Yan H, et al. Identification of Potential Early Diagnostic Biomarkers of Sepsis. J Inflamm Res 2021;14:621.

APPENDICES

FORM 2

Manuscript No. 1

Short reference Barber *et al*., (2021), Nature microbiology

Contribution of the doctoral candidate

Contribution of the doctoral candidate to figures reflecting experimental data (only for original articles):

Figure(s) # 3, S2, S5	☒	**100%**	
Figure(s) # 1-2, S3	☒	Approximate contribution:	**90%**
Figure(s) #	☐	Approximate contribution:	**80%**
Figure(s) #	☐	Approximate contribution:	**70%**
Figure(s) #	☐	Approximate contribution:	**60%**
Figure(s) # 4, S1	☒	Approximate contribution:	**50%**
Figure(s) # 5	☒	Approximate contribution:	**40%**
Figure(s) # S4	☒	Approximate contribution:	**30%**
Figure(s) #	☐	Approximate contribution:	**20%**
Figure(s) #	☐	Approximate contribution:	**10%**
Figure(s) #	☐	**0%**	

(Add more table boxes depending on the number of figures)

Signature candidate

Signature supervisor (member of the Faculty)

FORM 2

Manuscript No. 2

Short reference Sae-Ong *et al.*, in preparation

Contribution of the doctoral candidate

Contribution of the doctoral candidate to figures reflecting experimental data (only for original articles):

Figure(s) # all	☒	**100%**
Figure(s) #	☐	Approximate contribution: **90%**
Figure(s) #	☐	Approximate contribution: **80%**
Figure(s) #	☐	Approximate contribution: **70%**
Figure(s) #	☐	Approximate contribution: **60%**
Figure(s) #	☐	Approximate contribution: **50%**
Figure(s) #	☐	Approximate contribution: **40%**
Figure(s) #	☐	Approximate contribution: **30%**
Figure(s) #	☐	Approximate contribution: **20%**
Figure(s) #	☐	Approximate contribution: **10%**
Figure(s) #	☐	**0%**

(Add more table boxes depending on the number of figures)

Signature candidate

Signature supervisor (member of the Faculty)

FORM 2

Manuscript No. 3

Short reference Barber *et al*., (2020), mBio

Contribution of the doctoral candidate

Contribution of the doctoral candidate to figures reflecting experimental data (only for original articles):

Figure(s) # 4-5	☒	**100%**	
Figure(s) #	☐	Approximate contribution:	**90%**
Figure(s) #	☐	Approximate contribution:	**80%**
Figure(s) #	☐	Approximate contribution:	**70%**
Figure(s) #	☐	Approximate contribution:	**60%**
Figure(s) #	☐	Approximate contribution:	**50%**
Figure(s) #	☐	Approximate contribution:	**40%**
Figure(s) #	☐	Approximate contribution:	**30%**
Figure(s) #	☐	Approximate contribution:	**20%**
Figure(s) #	☐	Approximate contribution:	**10%**
Figure(s) # 1-3	☒	**0%**	

(Add more table boxes depending on the number of figures)

Signature candidate

Signature supervisor (member of the Faculty)

FORM 2

Manuscript No. 4

Short reference Mirhakkak *et al.*, in preparation

Contribution of the doctoral candidate

Contribution of the doctoral candidate to figures reflecting experimental data (only for original articles):

Figure(s) #	☐	**100%**
Figure(s) #	☐	Approximate contribution: **90%**
Figure(s) #	☐	Approximate contribution: **80%**
Figure(s) #	☐	Approximate contribution: **70%**
Figure(s) #	☐	Approximate contribution: **60%**
Figure(s) #	☐	Approximate contribution: **50%**
Figure(s) #	☐	Approximate contribution: **40%**
Figure(s) #	☐	Approximate contribution: **30%**
Figure(s) #	☐	Approximate contribution: **20%**
Figure(s) # 2, S2	☒	Approximate contribution: **10%**
Figure(s) # 1, 3-4, S1	☒	**0%**

(Add more table boxes depending on the number of figures)

Signature candidate

Signature supervisor (member of the Faculty)

DECLARATION

I, Tongta Sae-Ong, as a doctoral student, hereby to confirm that:

- I am familiar with the valid doctoral examination regulations.
- I produced this doctoral thesis myself, I neither used any text passages from third parties nor their own previous final theses without citing them.
- I cited the tools, personal information, and sources having been used in this thesis.
- I provide names of the persons who assisted the applicant in selecting and analyzing materials and supported them in writing the manuscript.
- I did not receive any assistance from specialized consultants and that any third party did not receive either direct or indirect financial benefits from me for the work connected to the doctoral thesis submission.
- I have not already submitted the doctoral thesis project as my final thesis for a state examination or other scientific examination.
- I did not submit the same, a substantially similar, or another scientific paper to any other institution of higher education or to any other faculty and, if I did it, which mark I might have achieved.

21.04.2022, Jena

Tongta Sae-Ong

CURRICULUM VITAE

Personal Details

Name: Miss Tongta Sae-Ong

Date of birth: 15.11.1990

E-mail address: tongta.saeong@uni-jena.de

Education

2018 – present:	PhD Student, Leibniz -Institute for Natural Product Research and Infection Biology - Hans-Knöll-Institute (Leibniz-HKI), System Biology and Bioinformatics (SBI), Jena, Germany
2013 – 2016	M.Sc. in Bioinformatics and Systems Biology, King Mongkut's University of Technology Thonburi, Bangkok, Thailand
2009 – 2012	B.Sc. in Medical Technology, Prince of Songkhla University, Songkhla, Thailand

Scholarship

2018 – 2022	Graduate School Scholarship Program, German Academic Exchange Service (DAAD), Bonn, Germany
2013 – 2015	Full scholarship, M.Sc. in Bioinformatics, King Mongkut's University of Technology Thonburi, National Center for Genetic Engineering and Biotechnology (BIOTEC), Thailand

Experiences

2017	Research assistant, King Mongkut's University of Technology Thonburi, Bangkok, Thailand
June – Nov 2015	Internship Student, Computational Biology Unit, Department of Informatics, University of Bergen (UiB), Bergen, Norway

Publications

Barber AE*, Sae-Ong T*, Kang K, Seelbinder S, Li J, Walther G, Panagiotou G# & Kurzai O# (2021) ***Aspergillus fumigatus* pan-genome analysis identifies genetic variants associated with human infection** *Nat Microbiol* 6, 1526-1536.

Barber AE*, Riedel J*, Sae-Ong T, Kang K, Brabetz W, Panagiotou G, Deising HB, Kurzai O (2020) **Effects of agricultural fungicide use on *Aspergillus fumigatus* abundance, antifungal susceptibility, and population structure**. mBio 11(6), e02213-20.