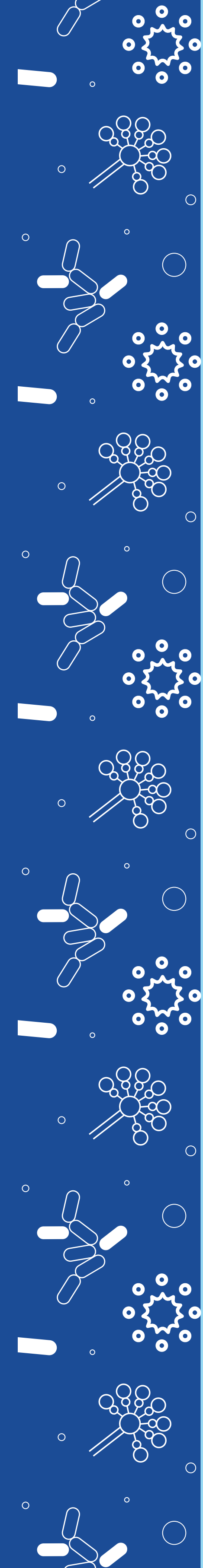


DETECTION OF PATHOGENS through Liquid Biopsy

SHOTGUN SEQUENCING

EMMY WESDORP

Detection of Pathogens through Liquid Biopsy Shotgun Sequencing

Emmy Wesdorp

Cover Illustration:	Josefina Quiroga | josefina.framer.ai
Layout Design:	Ridderprint | ridderprint.nl
Printing:	Ridderprint | ridderprint.nl
ISBN:	978-94-6522-589-0

© Emmy Wesdorp 2025
All rights reserved. No part of this thesis may be reproduced or transmitted in any form or by any means without prior written permission from the author.

Detection of Pathogens through Liquid Biopsy Shotgun Sequencing

Detectie van pathogenen door middel van liquid biopsy shotgun sequencing

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de
rector magnificus, prof. dr. ir. W. Hazeleger,
ingevolge het besluit van het College voor Promoties
in het openbaar te verdedigen op

donderdag 30 oktober 2025 des ochtends te 10.15 uur

door

Adriana Emma Wesdorp

geboren op 25 maart 1994
te Goes

Promotor:

Prof. dr. J. de Ridder

Copromotor:

Dr. M. Jager

Beoordelingscommissie:

Prof. dr. J.A.M. Borghans
Prof. dr. L.H. Franke
Prof. dr. M.M. Maurice
Dr. M.F. Seidl (voorzitter)
Prof. dr. W.J.E. Tissing

Dit proefschrift werd mede mogelijk gemaakt met financiële steun van Claret Bioscience.

Table of contents

Chapter 1	General introduction	7
Chapter 2	NGS-based *Aspergillus* detection in plasma and lung lavage of children with invasive pulmonary aspergillosis	29
Chapter 3	Exploring cell-free DNA fragmentomics to improve *Aspergillus* detection in invasive mold infections	71
Chapter 4	Bacterial cell-free DNA profiling reveals co-elevation of multiple bacteria in newborn foals with suspected sepsis	103
Chapter 5	Discussion	139
Addendum	Nederlandse samenvatting	156
	Author contributions per chapter	161
	List of publications	162
	Dankwoord	163
	Curriculum vitae	170

CHAPTER 1

General introduction

Emmy Wesdorp [1,2], Myrthe Jager [1,2], Jeroen de Ridder [1,2]

[1] Center for Molecular Medicine, University Medical Center Utrecht, Utrecht University, 3584 CX Utrecht, The Netherlands

[2] Oncode Institute, 3521 AL, Utrecht, The Netherlands

Detecting the invisible: understanding the impact of microscopic pathogens

Microscopic organisms, as their name suggests, are miniscule forms of life, requiring specialized instruments for observation and study. In 1665, Robert Hooke provided the first description of microorganisms following examination with light microscopy. However, it was not until the late 19th century, with the work of Robert Koch, that the link between these tiny organisms, observed through microscopes, and disease was definitively established. Today, we know that thousands of pathogens, including bacteria, fungi, parasites and viruses, infect both humans and animals.

The diversity among these pathogens is vast, both in terms of their types and the severity of the diseases they cause. Common bacterial pathogens in human medicine include *Streptococcus pneumoniae* and *Escherichia coli*, which cause pneumonia and urinary tract infections, as well as *Staphylococcus aureus* (including MRSA) and *Mycobacterium tuberculosis*, which are linked to more severe infections[1–4]. Fungal pathogens like *Candida* and *Aspergillus* are dangerous for immunocompromised individuals[5–7], while parasites such as *Plasmodium* cause malaria[8]. Additionally, well-known viruses like HIV, influenza and SARS-CoV-2 contribute significantly to global health challenges.

Infections have a profound impact on both human and veterinary medicine, leading to significant mortality, reduced quality of life, high healthcare costs, and a wide range of health complications. Sepsis, for example, is a life-threatening condition caused by the body's dysregulated response to infection by pathogens or their toxins in the bloodstream and is one of the leading causes of death in both humans and animals[9,10]. In 2017, almost 50 million incident cases of sepsis were recorded, and over 10 million sepsis-related deaths occurred, accounting for almost 20% of global deaths[10], with low- and middle-income countries disproportionately affected[10,11]. These numbers are staggering, and a clear testament to the enormous impact of microscopic pathogens on health.

Pathogen identification in clinical practice

Good tools to identify pathogens are crucial in reducing the disease burden by facilitating the timely and accurate detection of infectious diseases, guiding appropriate treatments and preventing further transmission. For instance, accurate bacterial identification supports the appropriate use of antibiotics, while rapid sepsis detection allows for prompt interventions that significantly improve survival rates while reducing healthcare costs[12–15]. The COVID-19 pandemic has also underscored the crucial role of molecular diagnostics, not as a point-of-care, but for managing infections and addressing broader public health challenges[16,17].

In both human and veterinary medicine, clinics use different methods available for resolving diagnostic challenges, including: traditional culture methods, microscopy and molecular techniques. Each method offers distinct advantages and challenges in terms of turnaround time, accuracy and/or interpretive limitations (as provisionally summarized in Table 1). For instance, microscopy can provide rapid identification of certain pathogens; however, its accuracy may be influenced by the technician's experience, the quality of the sample and the staining techniques employed. Traditional culture methods often require specific growth conditions and can take several days or even weeks to yield results. Additionally, certain pathogens are inherently difficult to culture, leading to false negatives and missed diagnoses. Furthermore, targeted methods like ELISA, PCR and certain next-generation sequencing (NGS) techniques typically focus on a limited range of pathogens. Despite the array of available diagnostic tools, limitations in accuracy, speed, and scope persist, creating significant challenges in pathogen identification. This highlights the urgent need for diagnostic tests that are accurate, timely, and unbiased, as delays in precise infectious disease diagnosis can profoundly impact patient outcomes, particularly in critical cases.

While molecular techniques for detecting microbial DNA, RNA, or proteins have significantly transformed microbiological diagnostics — enabling faster and more accurate pathogen identification than traditional culture and microscopy — there remains a critical need for innovative diagnostic, pathogen identification solutions. Specifically, there is a demand for tools that combine speed and accuracy with broad-spectrum pathogen identification, ideally offering additional guidance on susceptibility through the detection of resistance markers or by identifying species, strains, or subtypes.

Table 1: Overview of diagnostic approaches for pathogen identification: general strengths and limitations

Diagnostic Test	Turnaround time	Sensitivity	Specificity	Bias towards	Notes
Culture	Days to Weeks	Low to Moderate	Moderate to High	Organisms capable of growth under specific lab conditions	May miss organisms with special growth requirements; enables antimicrobial resistance profiling
Microscopy	Immediate to Hours	Low to Moderate	Moderate	Organisms visible and identifiable based on morphology and staining	Rapid results but dependent on staining methods and technician skill; limited by visible organisms
PCR	Hours	High	High	Specific to targeted pathogens or known genetic markers	Targets specific genes, including resistance markers; high sensitivity, but limited to known genes
ELISA	Hours	Moderate	Variable	Presence of specific antigens detectable by antibodies	Specificity and sensitivity depend on antibody quality
Next-Generation Sequencing (NGS)	Days to Weeks	High	High	Comprehensive, covering broad or specific genetic profiles	Versatile for broad or detailed genetic analysis; can detect unknown pathogens; higher complexity

Advances in sequencing technologies for pathogen detection

Both second- and third-generation high-throughput sequencing technologies (see Box 1) have made it possible to comprehensively sequence entire microbial genomes and conduct targeted and shotgun metagenomic sequencing with unprecedented speed and accuracy. Consequently, these innovations have significantly enhanced our understanding of microbial diversity and function, paving the way for novel approaches to microbial identification and characterization (see Box 2 for an overview of these methods).

Box 1: a short history on sequencing technologies

Over the past sixty years, researchers have developed and applied a diverse array of techniques and technologies to read the nucleic acid code of terrestrial life. As a result of this ongoing evolution, the field now has a spectrum of techniques available, spanning from first to third-generation sequencing.

First-generation sequencing techniques were pioneered by Frederick Sanger, Allan Maxam and Walter Gilbert. The Maxam-Gilbert method relied on the chemical breakdown of nucleic acids in polynucleotide chains[61], while Sanger introduced chain termination through chemically modified nucleotides[62]. Initially, sequencing efforts were concentrated on relatively pure RNA samples, such as microbial ribosomal RNA, and the genomes of single-stranded RNA bacteriophages. Advancements in Sanger's chain-termination methods and in base detection methodologies subsequently laid the fundament for the development of first-generation automated DNA sequencing instruments[63,64], culminating in 1995 in the first bacterial genome of *Haemophilus influenzae* to be sequenced[65]. However, at the time, throughput was limited. Sequencing of a single genome could therefore take several years and require substantial resources.

In the mid- and late 1990s, a new wave of **second-generation technologies** emerged, dramatically changing the landscape of biological and biomedical research. By the early 2000s, these advancements enabled the sequencing of hundreds of thousands to billions of bases within a timeframe of just a few hours to several days. These second-generation sequencing (NGS) methods require the fragmentation of long genomic and high-molecular-weight DNA/RNA sequences into shorter fragments. Solexa sequencing, later acquired by Illumina, emerged as the most successful. **Illumina** uses a technique known as bridge amplification, which involves adapter binding (to a chip) and solid-phase polymerase chain reaction (PCR) to generate clonal clusters before proceeding with sequencing by synthesis[66]. Sequencing-by-synthesis (i.e. Illumina) achieves exceptionally high accuracy through the use of reversible terminator nucleotides, which prevent subsequent incorporation of homopolymers. Illumina sequencing also facilitates paired-end readout, enhancing the mapping of reads to reference sequences and improving the detection of DNA rearrangements. The initial surge in large-scale genomics research initiatives led to a substantial increase in the number of sequenced microbial genomes. Illumina sequencing of bacteria transitioned from research settings to public health practice during the cholera epidemic in Haiti in 2010[67].

Although second-generation sequencing provides enormous volumes of data at limited cost-per-base, its primary limitation lies in generating only short-read sequences. This restricts its ability to resolve complete microbial genomes, accurately detect horizontal gene transfers, and may lead to ambiguous species-level identification. Notably, second-generation sequencing has proven particularly valuable for detecting the presence and mutations in antimicrobial resistance (AMR) genes, which are crucial for monitoring the evolution and spread of resistance patterns in pathogens. However, challenges persist in accurately mapping complex genomic regions associated with AMR, such as gene clusters or resistance islands, which often contain multiple co-located resistance genes. **Third-generation long-read platforms**, such as PacBio and Oxford Nanopore Technologies (ONT), provide key advantages and can effectively complement second-generation technologies in these contexts. Both the PacBio and ONT third-generation sequencing methods are fundamentally distinct in their operational principles and exhibit notable differences in error rates, throughput, portability, ability to detect modified bases, potential to get real-time access to the sequencing data and costs. PacBio sequencers produce low numbers of reads with high accuracy on large sequencing machines, while ONT sequencers produce higher numbers of reads with lower accuracy on portable devices. ONT sequencing is rapidly advancing and offers real-time analysis (as exemplified by recent work of our group[68]), yet challenges such as lower throughput (compared to i.e. Illumina), lower accuracy, evolving software and the lack of refined protocols persist. Nevertheless, the benefits of long-read sequencing are tremendous. For example, ONT sequencing has finally enabled the complete assembly of difficult-to-sequence areas in the human reference genome, including sequencing of the entire X chromosome[69], followed by the telomer-to-telomere human reference genome (T2T-CHM13v1.1)[70] as well as its ultimate completion of by sequencing of the human Y chromosome (resulting in the T2T-CHM13v2)[71]. ONT nanopore sequencing offers long-read capabilities that enable comprehensive microbial genome analysis, including AMR gene clusters. It effectively detects structural variations and complex genomic features, such as antimicrobial resistance gene islands, which are challenging for short-read methods[72]. Additionally, ONT enables the detection of epigenetic modifications with unparalleled detail and, since recently, enables real-time sequencing of ultra-short DNA fragments (as short as 20 bases). ONT sequencing in short (cell-free) DNA fragment mode also promises new research and diagnostic possibilities[73,74].

Box 1: *Continued*

The future of genome sequencing is poised for significant disruption, driven by emerging technologies from platforms such as Ultima Genomics, BGI's CycloneSeq, and Roche's Expandamer, all aimed at dramatically reducing costs. These innovations are reshaping the landscape of genomic research and clinical diagnostics by making high-quality sequencing more accessible than ever. Ultima Genomics is targeting the ambitious goal of a $100 genome by optimizing various aspects of the sequencing process, including advancements in sequencing chemistry and a unique spinning-disk flow cell design. This approach facilitates high-throughput and low-cost genomic sequencing. Meanwhile, BGI's CycloneSeq employs DNA nanoball sequencing, a pioneering method that amplifies DNA into highly compact nanoballs via rolling circle amplification. These nanoballs are immobilized on a patterned array, enabling ultra-high-density sequencing at significantly reduced costs. Roche's Expandamer technology takes a different approach by incorporating expanded nucleotide analogs into its sequencing-by-synthesis process. These chemically modified nucleotides enhance base discrimination and reduce sequencing errors, resulting in improved accuracy. Collectively, advancements by Ultima Genomics, BGI, and Roche are poised to further revolutionize genome sequencing, making it even more cost-effective, efficient, and accurate, and pave the way for further adoption in various fields, including personalized medicine, genomics research, as well as diagnostics.

In this thesis, **we focus on second-generation shotgun metagenomics sequencing of cell-free DNA (cfDNA) for pathogen identification** in mammalians. This method analyzes free-floating DNA molecules in bodily fluids, such as plasma, offering a minimally invasive approach for monitoring microscopic organisms and detecting pathogenic microbes. cfDNA is released from various sources, including tumors, fetal tissue, and microbes, and its short half-life allows for real-time monitoring of biological processes, including early disease detection. With shotgun sequencing of cfDNA it is possible to simultaneously analyze all cfDNA, predominantly sourced from host chromosomes, and thereby detect a wide range of microbes, including bacteria, fungi, and viruses. This approach has proven successful in various clinical settings. For instance, it has been employed to identify bacterial pathogens in patients suspected of bloodstream infections[18–22], and fungal pathogens in immunocompromised (cancer) patients and those with COVID-19[23–27]. Furthermore, shotgun cfDNA sequencing has been employed to map the virome[28–32] and might hold potential for identifying parasitic infections[33].

Given the need for diagnostic tools that combine speed and accuracy with broad-spectrum pathogen identification, cfDNA sequencing holds promise as an emerging tool for pathogen identification. cfDNA sequencing allows for the simultaneous detection of multiple pathogens without prior knowledge of which pathogen(s) may be present. This untargeted approach thereby enables the identification of unexpected or rare pathogens, which is crucial in clinical settings where the causative agent is unknown. Results can typically be obtained within 2 to 3 days, significantly faster than some traditional culture methods, which are often considered the gold standard but can take much longer to yield results. While resistance profiling can be challenging due to the low abundance of pathogen- and resistance gene-derived reads in untargeted cfDNA sequencing, this approach still allows for clinically relevant pathogen identification.

Despite its potential, detecting, quantifying, and evaluating low-abundance microbial cfDNA in clinical samples presents significant challenges. These challenges greatly depend on the source of infection — being viral, bacterial, or fungal — and the risk of contamination. Additionally, the lack of a standardized workflow for sample collection, processing, and data analysis hampers clinical validation. The variability in protocols complicates the identification of the most effective workflows for cfDNA NGS-based diagnostics in specific contexts. Consequently, the field remains in flux, with ongoing efforts to establish optimal methodologies, of which this thesis represents one contribution. While still evolving, cfDNA sequencing for pathogen identification is regarded as a promising strategy for enhancing microbiological diagnostics.

Pathogen identification using cfDNA from liquid biopsies represents just a small subset of the broader field of cfDNA-based diagnostics, enabling timely interventions in conditions such as **cancer**, **organ transplant rejection** and **infectious diseases**. A prominent example is non-invasive prenatal testing (NIPT), which has rapidly transitioned from concept (late 2000s) to clinical practice[34,35]. Initially employed in high-risk pregnancies, it is now widely used as a primary screening tool due to its high accuracy, allowing millions worldwide to detect chromosomal aneuploidies like trisomy 21[35,36]. Additionally, cfDNA sequencing to identify tumor-derived molecules plays a vital role in cancer care, and has shown to be greatly advantageous for classification, prognosis, and monitoring disease progression and therapeutic response. For example, the detection of EGFR mutations in cfDNA can indicate lung adenocarcinoma, while other mutations help assess treatment efficacy or resistance[37–39]. Furthermore, the detection of donor-derived cfDNA, which refers to circulating cell-free DNA originating from a donor, is gaining traction for organ transplant monitoring[24,40–42]. Moreover, commercial ventures, such as Karius Inc., have begun offering pathogen detection through cfDNA shotgun sequencing, highlighting its growing relevance in infectious disease diagnostics. Overall, cfDNA serves as a compelling biomarker for real-time monitoring, particularly through non-invasive liquid biopsies like urine and blood plasma. As sequencing technologies advance and costs decrease, the potential applications of cfDNA in healthcare continue to grow.

In the following sections we first offer an overview of cfDNA biology. Next, workflow choices for microbial cfDNA analysis are discussed, covering key steps such as library preparation and the identification of cfDNA origins, whether through taxonomic labeling or alternative approaches.

cfDNA biology: release & degradation mechanisms

cfDNA molecules are short, extracellular DNA fragments that circulate in various bodily fluids, including blood, urine, lung lavage, saliva, breast milk, nipple aspirate fluid, ascites, synovial fluid and more. These DNA fragments can arise from diverse cell-types, including (healthy) tissue, tumors, fetal sources, as well as microbes, and are thought to be released either during normal cellular functions, such as secretion and export in exosomes, as well as during cell death processes like necrosis, phagocytosis and apoptosis (Figure 1a)[43–45]. In healthy individuals, plasma cfDNA primarily comes from hematopoietic cells[46–49], with neutrophils being the dominant source [48,49], while erythrocyte progenitors and vascular endothelial cells contribute smaller fractions (Figure 1b) [47]. In cancer patients, tumor-derived nuclear DNA can become a significant source of cfDNA, with tumor fractions typically ranging from 0.1% to 10%, increasing with disease progression. Similarly, fetal cfDNA in maternal plasma rises to 10-30% by the later stages of pregnancy[50]. Non-nuclear cfDNA, such as mitochondrial and microbial cfDNA (of which bacterial cfDNA is generally the most abundant) usually represents less than 1% of the total cfDNA (Figure 1b)[40]. However, also these levels can rise significantly in the context of disease such as infection or inflammation[51–53].

Accurately determining the origin of cfDNA is crucial for its use in clinical applications. When sequencing is employed, this identification is often achieved through the analysis of distinct (epi)genomic sequences that differentiate cfDNA from normal host background DNA (see also '*Identifying cfDNA origin*'). Beyond the examination of nucleotide base sequences (i.e., the order of bases), the characteristics of cfDNA, such as fragment length, mapping location, end motifs and jagged ends (collectively referred to as **fragmentomics** features; see Figure 1c), provide additional insights into its tissue of origin. These fragmentomics features can vary substantially based on the physiological or pathological conditions under which the cfDNA is released.

The predominant size of cfDNA molecules is centered around 167 base pairs (bp), although shorter and longer fragments are also present, with fragment size varying depending on the type of liquid biopsy[54,55]. In analyses of eukaryotic host-derived cfDNA, longer fragments of over 10,000 base pairs are typically linked to necrosis, whereas those under 1,000 base pairs, including the 167 bp fragments, are linked to apoptosis[43–45,56]. Variation in cfDNA fragment length is a result of deoxyribonucleases (DNases) as well as other degradation processes that come into play after release of the DNA molecules from the cellular compartment. According to current models, DNA fragmentation factor B and deoxyribonuclease 1 like 3 (DNASE1L3), along with apoptotic nucleases, produce DNA fragments within the cell. Meanwhile, plasma DNASE1L3 also shows nuclease activity outside the cell[57]. Importantly, different DNases seem to produce distinct cfDNA ends. Intracellular degradation leads to an enrichment of adenines at the ends of cfDNA molecules, while a bias for cfDNA ending with cytosine is more commonly observed in circulation[57]. Additionally, DNase activity can create both double-stranded and single-stranded nicks in the DNA. These nicks can result in cfDNA ends displaying single-stranded overhangs, commonly referred to as jagged ends[58].

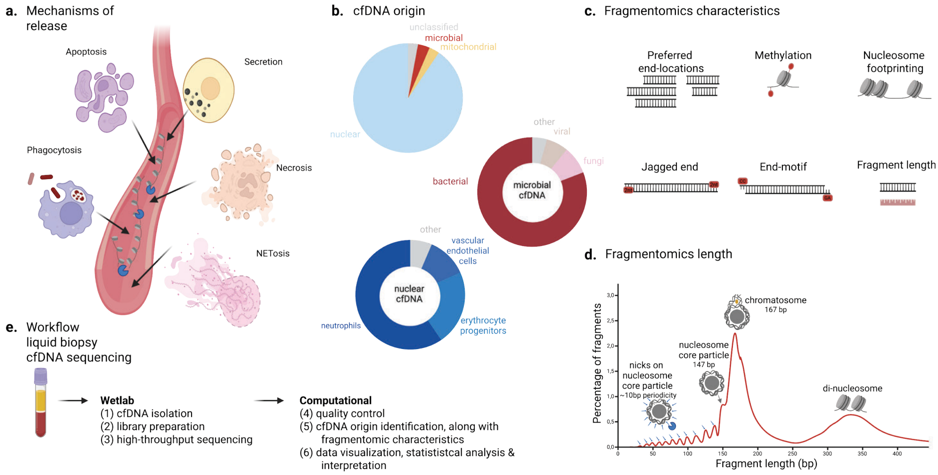


Figure 1. Mechanisms of cfDNA release, origins, fragmentomic signatures, and a standard workflow for comprehensive sequencing data analysis. a. The release of cfDNA fragments into bodily fluids through various endogenous mechanisms involves DNases in both their formation and degradation. **b.** A standardized workflow is generally followed in (microbial) liquid biopsy cfDNA sequencing, including key steps to identify cfDNA origin and obtain fragmentomic characteristics. **c.** Sources of cfDNA in blood samples. **d.** cfDNA fragmentomic characteristics. **e.** Length profile of cfDNA in blood samples (adapted from Thierry 2023[60]). Figure created with BioRender.com (accessed on 22-10-2024).

The primary structure that protects cfDNA from DNAse-mediated degradation is its wrapping around a histone octamer, to form a nucleosome structure. Specifically, the chromatosome — where double-stranded DNA is tightly wrapped around a histone octamer with an H1 histone linker — shields DNA from nuclease activity, resulting in its modal size of 167 bp (Figure 1d). A high abundance of 147 bp DNA fragments is also observed, likely due to DNA wrapping around the histone core without the H1 linker. Shorter fragments, typically under 143 bp, often exhibit a distinct ~10 bp periodicity, which arises from nicks in the DNA that occur while the DNA remains wrapped around the histone core. DNA exposed on the surface of the nucleosome is thereby accessible to DNases, allowing for cleavage at the minor groove[46,58]. Furthermore, regulatory proteins such as transcription factors that are bound to cfDNA can offer protection against degradation, typically resulting in sub-100bp fragments[46].

Interestingly, not all types of cfDNA are thought to have the same level of protection from DNase-mediated breakdown, primarily because the chromatin structure found in eukaryotic cells is not universal across all domains of life. For example, prokaryotes lack the histones present in eukaryotes, resulting in a simpler, more condensed organization of their DNA, which is supported by other proteins instead of nucleosomes. Additionally, mitochondrial DNA in eukaryotes is organized as histone-free (circular) nucleoids. As a result, mitochondrial DNA and bacterial DNA is generally more fragmented than nuclear genomic DNA[40]. Additionally, the release mechanisms of microbial cfDNA may differ from those of host-derived cfDNA. Microbial cfDNA is believed to be released in response to antimicrobial treatments or immune responses, such as phagocytosis and the formation of neutrophil extracellular traps (NETs)[59]. This difference in release mechanisms could also contribute to the shorter length of microbial cfDNA compared to host genomic cfDNA.

Box 2: Sequencing for microbial identification and characterization

To date, WGS methods have facilitated assembly of numerous microbial genomes. Genomic microbial sequencing is undertaken primarily to identify and characterize microbial species, analyze their genetic content, and assess their potential pathogenicity. Accurate genome assemblies further facilitate effective monitoring and detection of pathogens, which is vital for public health, diagnostics, and treatment strategies. Microbial assemblies are typically achieved through 1) reference-based assembly, 2) *de novo* assembly, or 3) a combined method. Reference-based assembly aligns DNA fragments to a closely related reference genome, while *de novo* assembly identifies overlaps among sequence reads and typically requires subsequent mapping to a reference for greater accuracy.

Microbial sequencing can be divided into culture-dependent and culture-independent approaches, each with its own benefits, applications, and challenges. Viral WGS can leverage clinical samples directly for culture-independent sequencing because viral DNA is often present in relatively high quantities in these samples, allowing for accurate genome assembly without prior culture. This is especially important for rapidly identifying and responding to emerging viral pathogens, monitoring outbreaks, and assessing viral evolution and resistance. In contrast, **culture-dependent sequencing** is frequently necessary for WGS of fungi and bacteria[77]. In the latter, microbes are first cultured to increase the microbial DNA yield and purity to improve quality of genome assembly and detection after sequencing. **Culture-independent sequencing** methods for microbial identification, characterization and monitoring offer a direct analysis of clinical samples, bypassing the time-consuming culture processes while providing a comprehensive view of microbial diversity. This approach minimizes bias and ensures a more accurate representation of the entire microbial community present within the sample. However, sequencing directly from clinical samples introduces unique challenges compared to culture-dependent techniques. In particular, the overwhelming presence of host DNA necessitates the use of tailored bioinformatics.

Box 2: *Continued*

Culture-independent sequencing methods can be broadly categorized into targeted (amplicon) sequencing and shotgun sequencing. **Targeted (amplicon) sequencing** focuses on enriching specific pathogens or genes before sequencing, typically through PCR amplification or probe- and bead-based enrichment. For instance, 16S rRNA (Illumina) sequencing is widely employed for bacterial identification[78], while 28S rRNA and ribosomal Internal Transcribed Spacer (ITS) genes are used for fungal identification[79,80]. Additionally, deep amplicon sequencing has been developed for viral diagnostics, including the detection of HIV and CMV resistance markers[81,82]. Despite its revolutionary impact on pathogen detection, amplicon sequencing has inherent limitations. These include the frequent lack of species-level resolution, which refers to the challenge of accurately distinguishing between closely related species within a microbial community. This limitation arises partly due to the design of the technique itself, which often amplifies only a specific region of the genome that may not contain sufficient unique identifiers for every species[83]. Moreover, it is well-documented that the choice of specific amplicon regions, particularly hypervariable regions within the 16S rRNA gene, significantly influences taxonomic classifications[83,84]. These hypervariable regions exhibit considerable genetic diversity, making them valuable for distinguishing closely related organisms. Amplifying regions that are too conserved may fail to differentiate species, while selecting more variable regions enhances the accuracy of microbial identification. This misalignment can lead to omitted taxa in microbial profiling and result in undetected pathogens, compromising the effectiveness of microbial analysis.

Shotgun sequencing aims for direct readout of all extracted nucleic acids, enabling the simultaneous detection of a broad range of microbes, including bacteria, fungi, parasites and, depending on the method, DNA- and/or also RNA viruses. In this approach, high-abundance host DNA can sometimes be depleted in the wet lab before sequencing. Alternatively, host DNA can be computationally identified and removed afterward. Shotgun sequencing offers a comprehensive multi-microbial approach and has been successfully applied in various clinical scenarios, such as detecting infections in blood by cfDNA readout[20,23], characterizing complex microbiomes in gastrointestinal disorders[85] and in respiratory infections[86,87].

Microbial cfDNA sequencing: workflow, challenges and considerations

To obtain data of the highest possible quality and reproducibility, microbial liquid biopsy cfDNA sequencing typically follows a standardized **workflow** (Figure 1e), consisting of key wetlab steps: (1) DNA isolation, (2) library preparation with quality control (and pooling), and (3) high-throughput sequencing. Subsequent computational analysis follows a similar structure with: (4) quality control and read preprocessing, (5) identifying cfDNA origin, along with fragmentomic characteristics, followed by (6) data visualization, statistical analysis and interpretation. While cfDNA sequencing offers various options at each step, no single method is universally optimal. Each application presents unique biological and technical challenges, making the best approach context-dependent, for instance on the type of liquid biopsy or pathogen.

Microbial cfDNA levels are generally low, making detection challenging. However, these levels can vary significantly based on the source and the individual's health status. In healthy individuals, viral and fungal cfDNA levels typically range from zero to only a few reads per million sequenced reads. In contrast, bacterial cfDNA can fluctuate considerably, potentially reaching hundreds of reads per million (primarily due to commensal bacteria in mammalian hosts that significantly contribute to the overall microbial cfDNA pool). Furthermore, the fragmentation of cfDNA can complicate its detection; microbial cfDNA often exists as short, fragmented sequences that may be overlooked during library preparation or sequencing, leading to incomplete data and reduced sensitivity in pathogen detection.

Wet lab challenges mainly focus on capturing and enriching the rare cfDNA types of interest. On the computational side, the main **computational challenge** lies in accurately identifying the origin of cfDNA, which is crucial for distinguishing microbial DNA from the host's DNA and identifying (pathogenic) microbes correctly — a process whose significance has been highlighted again in recent studies by Gihawi *et al.* 2023[97] and Sepich-Poore *et al.* 2024[98]. As a result, the optimal combination of protocol variations — including how to best capture cfDNA, process it, and analyze the resulting data — remains an open question. Research is ongoing to refine and improve the methodologies to suit different scenarios, and no single approach has yet proven ideal for all potential applications.

Library preparation methods

Conventional shotgun sequencing of cfDNA using short-read methods, such as Illumina, typically enables readout of fragments shorter than 750 bp, which represent the majority of the cfDNA pool in healthy individuals, particularly in plasma samples. This process involves ligating an adapter to the DNA molecules, followed by the untargeted readout of tens of millions of fragments. To enable adapter ligation two strategies are possible. In **double-stranded DNA library preparation** (ddLP) double-stranded breaks in DNA are targeted, while **single-stranded DNA library preparation** (ssLP) focuses on single-stranded nicks in either of the strands.

Historically, cfDNA analysis has centered on double-stranded DNA, reflected by the fact that most literature is based on dsLP. The fundamental of dsLP involves ligating a double-strand adapter to a double-stranded DNA molecule. Such methods are generally accessible, cost-effective for individual samples and have undergone continuous improvements. Additionally, their biases are well understood and extensively documented, making comparisons across studies more straightforward.

In contrast, single-stranded library preparation (ssLP) — initially developed for ancient DNA analysis[88,89] and recently refined — employs heat denaturation of double-stranded DNA to generate two single-stranded templates before adapter ligation. Various strategies are utilized to later attach the initial adapter to these denatured single-stranded molecules (see Cheng *et al.*, 2024 for review[90]), all allowing for the incorporation of both blunt-end and nicked double-stranded DNA, as well as single-stranded DNA. As a result, ssLP can encompass both single-stranded and converted double-stranded DNA molecules.

The choice between dsLP and ssLP significantly affects the resulting data. Each method offers distinct insights into cfDNA degradation patterns, particularly in terms of fragment size distribution and end-motifs. Notably, ssLP has been shown to capture a higher proportion of short cfDNA molecules compared to ddLP. This advantage is particularly relevant in the context of cfDNA shotgun sequencing for infectious disease scenarios, as ssLP facilitates the detection of these shorter molecules and enriches the analysis of both host mitochondrial, bacterial and viral cfDNA, as demonstrated by Burnham *et al.* 2014[40]. The impact of ssLP on readout is underscored by improved recovery efficiency of microbial cfDNA reads, leading to a 71-fold increase in the relative genomic coverage of microbial species over dsLP. However, it is essential to recognize that there is no one-size-fits-all solution, as cfDNA length can vary significantly between taxa within the same kingdom, as for example, evidenced by studies on viral cfDNA in non-invasive prenatal testing samples[91].

Parallel to conventional approaches, several specialized library preparation strategies have been developed to optimize sequencing readouts for specific applications. For instance, **Jag-seq**, a specialized wetlab and bioinformatics workflow established specifically to identify jagged ends and determine their lengths in diverse liquid biopsy types, including urine and plasma[92,93]. The latest version of Jag-seq uses methylated dCTP (5mC) during DNA end-repair step preceding dsLP adapter ligation, which later enables the detection of jaggedness by identifying CH sites (where C is followed by A, C, or T) near the ends of DNA fragments[93]. The readout is hereby similar to other methylation-aware techniques using bisulfite conversion, where unmethylated cytosines are converted to uracils, which later become thymines, while methylated cytosines remain unchanged. Another methylation-aware method is **SIFT-seq**, which tags freshly isolated DNA through bisulfite conversion[94]. Contaminating bacterial DNA introduced after tagging can be identified by the absence of cytosine conversion. This method is particularly useful in low-microbial biomass samples, where distinguishing contaminants from true biological signals is critical to avoid interference.

Identifying cfDNA origin: selecting computational tools for taxonomic labeling

The analytical sensitivity, specificity and overall accuracy of cfDNA-based diagnostics hinge on accurately identifying the taxonomic origin of each sequenced cfDNA molecule, defined as the microbial taxa or cells from which the cfDNA derives. This identification is typically achieved through the analysis of distinct (epi)genomic sequences that differentiate cfDNA from the normal background (or host) DNA. However, determining the origin of cfDNA is challenging due to the short length of these molecules. In cancer diagnostics, tumor-derived cfDNA may contain specific genomic alterations absent in normal DNA, such as single-nucleotide variants (SNVs), fusions and copy number variations (CNVs), or unique methylation patterns, which may aid in the identification of their origin. Furthermore, genomic variations can distinguish fetal cfDNA from maternal cfDNA and graft-derived cfDNA from host-derived cfDNA. Unlike human DNA, microbial DNA exhibits considerable sequence variability across cfDNA fragments due to its diverse origins, necessitating comparisons with established microbial genomes or taxonomically annotated databases for accurate identification. This process, known as assembly-free metagenomic profiling, is crucial due to the short length and scarcity of pathogenic microbial cfDNA, making assembly (Box 2) infeasible.

Nowadays, most microbial identification workflows begin with **identifying host-derived cfDNA through conventional mapping** before taxonomically labeling the remaining reads (see also Figure 2). To achieve accurate taxonomic assignments for the remaining reads, various computational tools have been developed. One significant challenge in taxonomic labeling is that standard alignment methods are often inadequate. While mapping to the human genome typically takes only hours to days, mapping against comprehensive databases, such as NCBI RefSeq — which includes a vast array of microbial genomes — demands (prohibitively) excessive computational resources. Additionally, ambiguity arises when a single cfDNA read aligns with multiple genomes, complicating origin determination and hindering accurate taxonomic assignment. Tools like PathoScope (2.0) effectively reassign ambiguously mapped reads, but this approach comes with considerable computational demands and limits on the size of the reference genome set[95,96].

There are other, non-mapping based, tools designed to address the challenges of high computational demands, as well as the ambiguity that arises from reads aligning with multiple genomes after host-read subtraction via host genome alignment. These taxonomic assignment tools can be categorized into two types: **taxonomic labelers**, which use a classifier to assign taxonomic identities to all reads, and **taxonomic profilers**, which provide relative abundance profiles without detailed labeling. Taxonomic labelers focus on identifying specific microorganisms in a sample by labeling all input cfDNA reads, while taxonomic profilers offer a broader overview of microbial composition, including the relative proportions of different taxa, without aiming to label every input read. For exploring microbial community composition, profiling is essential, whereas microbial labeling is more effective for detecting pathogens in clinical samples (see also Lu *et al.*, 2022[99]). Given that taxonomic mislabeling of sequencing reads can have diagnostic consequences, it is crucial to select a reliable taxonomic labeling tool that is sensitive enough to identify microbes at the genus, species, or even strain level.

Taxonomic classification algorithms (i.e. profilers and labellers) fall into three main categories: DNA-to-DNA, DNA-to-protein and DNA-to-marker classifiers. DNA-to-marker methods, like MetaPhlAn and mOTUs, use reference databases containing specific genes known to differentiate species. The 16S rRNA sequence is most classically used in bacterial metagenomics (see also Box 2). For more complex microbial communities that span multiple domains or kingdoms, combining multiple marker genes can be beneficial, with tools such as MetaPhlAn2 facilitating this approach[100]. **DNA-to-DNA classifiers**, including Kraken[101], Kraken2[102], Bracken[103] and PathSeq[104] and **DNA-to-protein tools** like Kaiju[105] and DIAMOND[106], compare sequences against larger genomic or protein databases. Protein-based classifiers do still require more computational power due to amino acid translation but are less sensitive to high mutational rates. In a benchmarking study by Ye *et al.* 2019[107], the performance of 20 metagenomic classifiers was assessed using simulated and experimental datasets. Previous studies have revealed significant variability in classifier performance, even when evaluated with identical benchmark datasets[108,109]. A unique feature of the benchmark from Ye *et al.* 2019[107] is the utility of a standardized reference database, which led them to conclude that DNA-to-DNA tools were the most effective, demonstrating high precision, recall, and a substantial proportion of species-level classifications. However, it is important to consider that this benchmarking primarily focused on bacterial species. Viral species, which exhibit higher mutation rates, may require DNA-to-protein tools for taxonomic labeling. On the other hand, research on fungal identification in shotgun metagenomics datasets demonstrated an advantage for Kraken (over DIAMOND and Kaiju) due to its speed, high sensitivity, specificity and ability to assign a Lowest Common Ancestor (LCA) prediction to each read[110] (see also Figure 2).

When working with a DNA-to-DNA tool, it is crucial to consider **reference database composition**. Databases such as NCBI RefSeq struggle with a **lack of comprehensive genomic reference data** with only 16,000 fungal and approximately 1.7 million bacterial assemblies, still representing only a small fraction of the estimated 10^{14} microbial taxa[111]. As a result, the origins of some reads remain unidentifiable, obscuring valuable information from clinically relevant microorganisms, as also demonstrated in the analyses of cumulative sequence data from multiple patient samples revealing novel taxa[30]. This is linked to the bias in reference genome databases toward easy-to-culture, well-studied organisms, ultimately compromising the taxonomic accuracy of DNA-to-DNA tools. It is also known that the accuracy of microbial sequence labeling improves with a comprehensive reference database that includes diverse taxonomic groups — bacteria, archaea, fungi and viruses. Larger databases enhance the likelihood of close matches, reducing misclassifications. However, while expanding the set of reference genomes can improve overall classification rates, it may also decrease the proportion of reads classified at the species level and increase computational demands[112,113]. Nevertheless, including all domains of life in classification databases is recommended[114], particularly the inclusion of host reference sequences (even after host reads identification by mapping) and vector sequences to identify potential contaminants[97]. Furthermore, it is advised to clean genomic sequences by removing low-complexity or incorrectly incorporated sequences from reference genomes[114,115]. Despite ongoing efforts, database optimization has not yet reached its full potential, in part due to the continual addition of new reference sequences and its dependence on specific clinical applications.

Identification of microbial contaminant DNA

Contaminant DNA refers to DNA that is not part of the clinical sample (i.e., the source) but may be introduced during the research process. To minimize the risk of contamination, it is recommended to use ultraclean reagents and work in a controlled environment (see Eisenhofer *et al.* 2019[116], for a review on eliminating contamination in low microbial biomass microbiome studies). Previous studies have identified multiple sources of contamination, including nucleic acid extraction kits[117], other laboratory consumables[82], human DNA from researchers and cross-contamination between samples (both in the wet lab and dry lab)[82,116,117]. Thus, contaminants can and will occur even with careful and clean laboratory practices, making their complete elimination challenging.

For computational identification of microbial contaminants, the SIFT-seq protocol is developed which can identify environmental contaminants from DNA isolation onwards, by an initial DNA tagging step[94]. This ssLP-based workflow has proven effective, reducing the prevalence of the commonly observed skin microbes (such as bacterium *Cutibacterium acnes*) by up to 35 times and eliminating (up to 76% of) frequently detected DNA contaminant genera in a cohort of plasma and urine samples. However, it is important to note that SIFT-seq utilizes Kaiju, a DNA-to-protein tool for taxonomic classification and could only label cfDNA as either a contaminant or a true biological signal when sufficient CpG sites are present within the molecule, thereby reducing the number of origin-identified reads and thus potentially sensitivity. SIFT-seq was initially used in the context of ssLP, but could theoretically also be applied in combination with dsLP and/or in the context of more targeted strategies (e.g. amplicon sequencing).

In standard library preparation workflows, *in silico* decontamination methods often ‘blacklist’ contaminating microbes from downstream analyses. To facilitate the computational identification of contaminants, it is recommended to record batch numbers and include ‘blank’ controls[116] — samples devoid of target host or microbial DNA — processed alongside experimental samples. These controls, prepared with the same reagents and protocols, help ensure that detected contaminants are a result of laboratory processes rather than biological sources. Although blank controls are valuable for decontamination efforts, it is generally inadvisable to remove all microbial taxa identified in these controls due to the potential for cross-contamination during sample handling[118] or sequencing[119]. Additionally, batch tracking can assist in retrospectively identifying significant differential abundances attributable to technical variables, as demonstrated in various studies[120].

More sophisticated computational decontamination techniques have been developed, often utilizing taxonomically labeled data as input. One of the most widely used statistical tools for this purpose is the *R* ***decontam*** package, which identifies contaminants through either a frequency-based or prevalence-based approach[121]. The frequency-based method uses DNA quantitation data from sample preparation, based on the assumption that contaminant DNA is typically present at consistently low concentrations across samples, while true sample DNA concentrations can vary significantly. Consequently, the relative frequency of contaminant DNA is inversely correlated with total DNA concentration, enabling statistical detection of contaminants. The prevalence-based approach of *decontam* relies on sequencing negative 'blank' controls. Contaminants are more likely to be detected in these blanks due to the absence of competing DNA, resulting in a higher prevalence of contaminants in these controls compared to true samples.

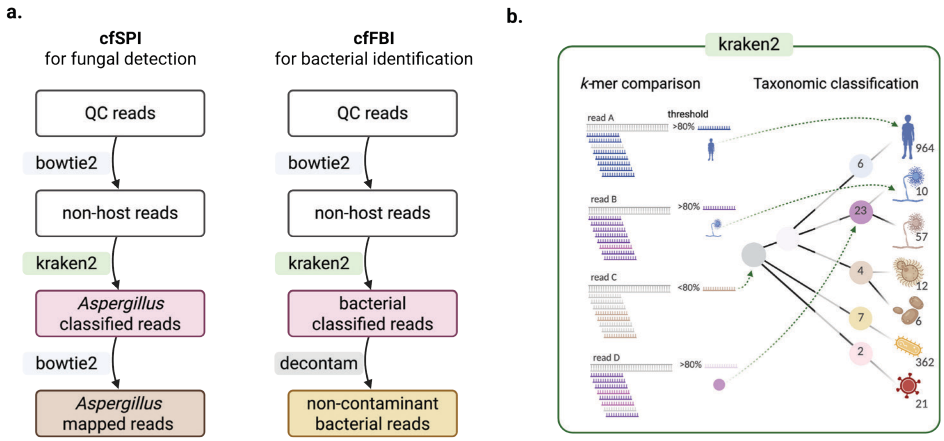


Figure 2. Schematic of a computational workflow for pathogen identification, featuring key concepts of the taxonomic labeler Kraken2. a. The left workflow, cfSPI, is designed for detecting pathogenic fungi, while the right workflow, cfFBI, targets sepsis-causing bacterial pathogens. After quality control (QC), host DNA is removed by excluding reads aligned with the host genome using Bowtie2. The remaining reads are classified using Kraken2 against a reference database. For cfSPI, fungal species reads are extracted and aligned with the genomic sequences of potential pathogen candidate, *Aspergillus*. In cfFBI, contaminant bacterial species are filtered out using the frequency-based decontamination method. **b.** Kraken examines the *k*-mers in the query sequence (i.e. read), searching the database (a compact hash table) for their positions in the taxonomy tree. By determining the most probable classification, it maps *k*-mers to the lowest common ancestor of all relevant taxa. Thereby it makes use of a classification threshold; if the percentage of *k*-mers matching a taxa is lower than the threshold, then the reads will not be classified. Figure created with BioRender.com (accessed on 22-10-2024).

A promising new approach to identifying microbial contaminants may furthermore be to analyze their fragmentomic characteristics. Recently, Dennis Lo's group in their study (Wang *et al.* 2023) demonstrated that analyzing end motifs, specifically CC- and GG-end motifs, can effectively differentiate pathogen-derived cfDNA from DNA contamination[18]. This differentiation is based on the premise that contaminant DNA undergoes different degradation processes than cfDNA derived from patient samples, leading to distinct fragmentomic characteristics.

Thesis scope and outline

In this thesis, we focus on utilizing shotgun second-generation sequencing to develop innovative liquid biopsy diagnostics for infectious diseases in mammals. In **Chapter 2** we evaluate the analytical performance of NGS methods for detecting *Aspergillus* in liquid biopsies from pediatric patients suspected of invasive pulmonary Aspergillosis (IPA). We assessed performance through *in silico* read simulations and fungal taxonomic labeling with various Kraken2 databases, highlighting the computational detection limits. Additionally, we examined the impact of sample workup on fungal yield and analytical sensitivity and evaluated diagnostic performance in a small cohort of potential IPA cases using our computational workflow, cfSPI (see Figure 2, for schematic of the workflow). In **Chapter 3** we searched for alternatives to improve fungal detection guided by fragmentomics characterisation, mediated by both ssLP and dsLP, and concluded that size selection could substantially increase *Aspergillus* fungal cfDNA abundance. **In Chapter 4**, we established an integrated wet lab and computational workflow, named cfFBI, to identify bacterial pathogens responsible for sepsis in newborn foals with systemic inflammatory response syndrome (SIRS). This workflow utilized *decontam* for the rigorous removal of contaminants (see also Figure 2). Given that SIRS is a dysregulated immune response often triggered by infections, we investigated host-related biomarkers in foals and found that mitochondrial DNA fraction and end-motifs of host cfDNA may be promising. Finally, in **Chapter 5**, we examine the contributions and limitations of our work in the context of cfDNA NGS pathogen identification, highlighting challenges such as establishing causality. We conclude by providing insights into potential directions for future research.

References

1. Flores-Mireles, A. L., Walker, J. N., Caparon, M. & Hultgren, S. J. Urinary tract infections: epidemiology, mechanisms of infection and treatment options. *Nat. Rev. Microbiol.* **13,** 269–284 (2015).
2. den Heijer, C. D. J., van Bijnen, E. M. E., Paget, W. J., Pringle, M., Goossens, H., Bruggeman, C. A., Schellevis, F. G., Stobberingh, E. E. & APRES Study Team. Prevalence and resistance of commensal Staphylococcus aureus, including meticillin-resistant S aureus, in nine European countries: a cross-sectional study. *Lancet Infect. Dis.* **13,** 409–415 (2013).
3. Alsayed, S. S. R. & Gunosewoyo, H. Tuberculosis: Pathogenesis, current treatment regimens and new drug targets. *Int. J. Mol. Sci.* **24,** (2023).
4. Rozenbaum, M. H., Pechlivanoglou, P., van der Werf, T. S., Lo-Ten-Foe, J. R., Postma, M. J. & Hak, E. The role of Streptococcus pneumoniae in community-acquired pneumonia among adults in Europe: a meta-analysis. *Eur. J. Clin. Microbiol. Infect. Dis.* **32,** 305–316 (2013).
5. Denning, D. W. Global incidence and mortality of severe fungal disease. *Lancet Infect. Dis.* **24,** e428–e438 (2024).
6. Pana, Z. D., Roilides, E., Warris, A., Groll, A. H. & Zaoutis, T. Epidemiology of invasive fungal disease in children. *J. Pediatric Infect. Dis. Soc.* **6,** S3–S11 (2017).
7. Lehrnbecher, T., Schöning, S., Poyer, F., Georg, J., Becker, A., Gordon, K., Attarbaschi, A. & Groll, A. H. Incidence and outcome of invasive fungal diseases in children with hematological malignancies and/or allogeneic hematopoietic stem cell transplantation: Results of a prospective multicenter study. *Front. Microbiol.* **10,** 681 (2019).
8. Poespoprodjo, J. R., Douglas, N. M., Ansong, D., Kho, S. & Anstey, N. M. Malaria. *Lancet* **402,** (2023).
9. Singer, M., Deutschman, C. S., Seymour, C. W., Shankar-Hari, M., Annane, D., Bauer, M., Bellomo, R., Bernard, G. R., Chiche, J.-D., Coopersmith, C. M., Hotchkiss, R. S., Levy, M. M., Marshall, J. C., Martin, G. S., Opal, S. M., Rubenfeld, G. D., van der Poll, T., Vincent, J.-L. & Angus, D. C. The third international consensus definitions for sepsis and septic shock (sepsis-3). *JAMA* **315,** 801–810 (2016).
10. Rudd, K. E., Johnson, S. C., Agesa, K. M., Shackelford, K. A., Tsoi, D., Kievlan, D. R., Colombara, D. V., Ikuta, K. S., Kissoon, N., Finfer, S., Fleischmann-Struzek, C., Machado, F. R., Reinhart, K. K., Rowan, K., Seymour, C. W., Watson, R. S., West, T. E., Marinho, F., Hay, S. I., Lozano, R., Lopez, A. D., Angus, D. C., Murray, C. J. L. & Naghavi, M. Global, regional, and national sepsis incidence and mortality, 1990-2017: analysis for the Global Burden of Disease Study. *Lancet* **395,** 200–211 (2020).
11. GBD 2016 Causes of Death Collaborators. Global, regional, and national age-sex specific mortality for 264 causes of death, 1980-2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet* **390,** 1151–1210 (2017).
12. Riedel, S. & Carroll, K. C. Early identification and treatment of pathogens in sepsis: Molecular diagnostics and antibiotic choice. *Clin. Chest Med.* **37,** 191–207 (2016).
13. Perez, K. K., Olsen, R. J., Musick, W. L., Cernoch, P. L., Davis, J. R., Peterson, L. E. & Musser, J. M. Integrating rapid diagnostics and antimicrobial stewardship improves outcomes in patients with antibiotic-resistant Gram-negative bacteremia. *J. Infect.* **69,** 216–225 (2014).
14. Whiles, B. B., Deis, A. S. & Simpson, S. Q. Increased time to initial antimicrobial administration is associated with progression to septic shock in severe sepsis patients. *Crit. Care Med.* **45,** 623–629 (2017).
15. Perez, K. K., Olsen, R. J., Musick, W. L., Cernoch, P. L., Davis, J. R., Land, G. A., Peterson, L. E. & Musser, J. M. Integrating rapid pathogen identification and antimicrobial stewardship significantly decreases hospital costs. *Arch. Pathol. Lab. Med.* **137,** 1247–1254 (2013).
16. Pfefferle, S., Reucher, S., Nörz, D. & Lütgehetmann, M. Evaluation of a quantitative RT-PCR assay for the detection of the emerging coronavirus SARS-CoV-2 using a high throughput system. *Euro Surveill.* **25,** (2020).
17. Peeling, R. W., Heymann, D. L., Teo, Y.-Y. & Garcia, P. J. Diagnostics for COVID-19: moving from pandemic response to control. *Lancet* **399,** 757–768 (2022).
18. Wang, G., Lam, W. K. J., Ling, L., Ma, M.-J. L., Ramakrishnan, S., Chan, D. C. T., Lee, W.-S., Cheng, S. H., Chan, R. W. Y., Yu, S. C. Y., Tse, I. O. L., Wong, W. T., Jiang, P., Chiu, R. W. K., Allen Chan, K. C. & Lo, Y. M. D. Fragment ends of circulating microbial DNA as signatures for pathogen detection in sepsis. *Clin. Chem.* **69,** 189–201 (2023).
19. Grumaz, S., Stevens, P., Grumaz, C., Decker, S. O., Weigand, M. A., Hofer, S., Brenner, T., von Haeseler, A. & Sohn, K. Next-generation sequencing diagnostics of bacteremia in septic patients. *Genome Med.* **8,** 73 (2016).
20. Blauwkamp, T. A., Thair, S., Rosen, M. J., Blair, L., Lindner, M. S., Vilfan, I. D., Kawli, T., Christians, F. C., Venkatasubrahmanyam, S., Wall, G. D., Cheung, A., Rogers, Z. N., Meshulam-Simon, G., Huijse, L., Balakrishnan, S., Quinn, J. V., Hollemon, D., Hong, D. K., Vaughn, M. L., Kertesz, M., Bercovici, S., Wilber, J. C. & Yang, S. Analytical and clinical validation of a microbial cell-free DNA sequencing test for infectious disease. *Nat. Microbiol.* **4,** 663–674 (2019).
21. Camargo, J. F., Ahmed, A. A., Lindner, M. S., Morris, M. I., Anjan, S., Anderson, A. D., Prado, C. E., Dalai, S. C., Martinez, O. V. & Komanduri, K. V. Next-generation sequencing of microbial cell-free DNA for rapid noninvasive diagnosis of infectious diseases in immunocompromised hosts. *F1000Res.* **8,** 1194 (2020).

22. Grumaz, C., Hoffmann, A., Vainshtein, Y., Kopp, M., Grumaz, S., Stevens, P., Decker, S. O., Weigand, M. A., Hofer, S., Brenner, T. & Sohn, K. Rapid next-generation sequencing-based diagnostics of bacteremia in septic patients. *J. Mol. Diagn.* **22,** 405–418 (2020).
23. Hong, D. K., Blauwkamp, T. A., Kertesz, M., Bercovici, S., Truong, C. & Banaei, N. Liquid biopsy for infectious diseases: sequencing of cell-free plasma to detect pathogen DNA in patients with invasive fungal disease. *Diagn. Microbiol. Infect. Dis.* **92,** 210–213 (2018).
24. Armstrong, A. E., Rossoff, J., Hollemon, D., Hong, D. K., Muller, W. J. & Chaudhury, S. Cell-free DNA next-generation sequencing successfully detects infectious pathogens in pediatric oncology and hematopoietic stem cell transplant patients at risk for invasive fungal disease. *Pediatr. Blood Cancer* **66,** e27734 (2019).
25. Hill, J. A., Dalai, S. C., Hong, D. K., Ahmed, A. A., Ho, C., Hollemon, D., Blair, L., Maalouf, J., Keane-Candib, J., Stevens-Ayers, T., Boeckh, M., Blauwkamp, T. A. & Fisher, C. E. Liquid biopsy for invasive mold infections in hematopoietic cell transplant recipients with pneumonia through next-generation sequencing of microbial cell-free DNA in plasma. *Clin. Infect. Dis.* **73,** e3876–e3883 (2021).
26. Huygens, S., Schauwvlieghe, A., Wlazlo, N., Moors, I., Boelens, J., Reynders, M., Chong, G.-L., Klaassen, C. H. W. & Rijnders, B. J. A. Diagnostic value of microbial cell-free DNA sequencing for suspected invasive fungal infections: A retrospective multicenter cohort study. *Open Forum Infect. Dis.* **11,** ofae252 (2024).
27. Lee, K. H., Won, D., Kim, J., Lee, J. A., Kim, C. H., Kim, J. H., Jeong, S. J., Ku, N. S., Choi, J. Y., Yeom, J.-S., Cho, H., Chung, H., Cheong, J.-W., Lee, S.-T., Jang, J. E., Shin, S. & Ahn, J. Y. Utility of plasma microbial cell-free DNA whole-genome sequencing for diagnosis of invasive aspergillosis in patients with hematologic malignancy or COVID-19. *J. Infect. Dis.* **228,** 444–452 (2023).
28. Burnham, P., Dadhania, D., Heyang, M., Chen, F., Westblade, L. F., Suthanthiran, M., Lee, J. R. & De Vlaminck, I. Urinary cell-free DNA is a versatile analyte for monitoring infections of the urinary tract. *Nat. Commun.* **9,** 2412 (2018).
29. Cheng, A. P., Cheng, M. P., Loy, C. J., Lenz, J. S., Chen, K., Smalling, S., Burnham, P., Timblin, K. M., Orejas, J. L., Silverman, E., Polak, P., Marty, F. M., Ritz, J. & De Vlaminck, I. Cell-free DNA profiling informs all major complications of hematopoietic cell transplantation. *Proc. Natl. Acad. Sci. U. S. A.* **119,** e2113476118 (2022).
30. Kowarsky, M., Camunas-Soler, J., Kertesz, M., De Vlaminck, I., Koh, W., Pan, W., Martin, L., Neff, N. F., Okamoto, J., Wong, R. J., Kharbanda, S., El-Sayed, Y., Blumenfeld, Y., Stevenson, D. K., Shaw, G. M., Wolfe, N. D. & Quake, S. R. Numerous uncharacterized and highly divergent microbes which colonize humans are revealed by circulating cell-free DNA. *Proc. Natl. Acad. Sci. U. S. A.* **114,** 9623–9628 (2017).
31. Kowarsky, M., Vlaminck, I. D., Okamoto, J., Neff, N. F., LeBreton, M., Nwobegabay, J., Tamoufe, U., Ledoux, J. D., Tafon, B., Kiyang, J., Saylors, K., Wolfe, N. D. & Quake, S. R. Cell-free DNA reveals potential zoonotic reservoirs in non-human primates. *bioRxiv* 481093 (2018). doi:10.1101/481093
32. Thongsripong, P., Chandler, J. A., Kittayapong, P., Wilcox, B. A., Kapan, D. D. & Bennett, S. N. Metagenomic shotgun sequencing reveals host species as an important driver of virome composition in mosquitoes. *Sci. Rep.* **11,** 8448 (2021).
33. Weerakoon, K. G. & McManus, D. P. Cell-free DNA as a diagnostic tool for human parasitic infections. *Trends Parasitol.* **32,** 378–391 (2016).
34. Lo, Y. M. D., Tsui, N. B. Y., Chiu, R. W. K., Lau, T. K., Leung, T. N., Heung, M. M. S., Gerovassili, A., Jin, Y., Nicolaides, K. H., Cantor, C. R. & Ding, C. Plasma placental RNA allelic ratio permits noninvasive prenatal chromosomal aneuploidy detection. *Nat. Med.* **13,** 218–223 (2007).
35. Ravitsky, V., Roy, M.-C., Haidar, H., Henneman, L., Marshall, J., Newson, A. J., Ngan, O. M. Y. & Nov-Klaiman, T. The emergence and global spread of noninvasive prenatal testing. *Annu. Rev. Genomics Hum. Genet.* **22,** 309–338 (2021).
36. Norton, M. E., Jacobsson, B., Swamy, G. K., Laurent, L. C., Ranzini, A. C., Brar, H., Tomlinson, M. W., Pereira, L., Spitz, J. L., Hollemon, D., Cuckle, H., Musci, T. J. & Wapner, R. J. Cell-free DNA analysis for noninvasive examination of trisomy. *N. Engl. J. Med.* **372,** 1589–1597 (2015).
37. Murtaza, M., Dawson, S.-J., Tsui, D. W. Y., Gale, D., Forshew, T., Piskorz, A. M., Parkinson, C., Chin, S.-F., Kingsbury, Z., Wong, A. S. C., Marass, F., Humphray, S., Hadfield, J., Bentley, D., Chin, T. M., Brenton, J. D., Caldas, C. & Rosenfeld, N. Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA. *Nature* **497,** 108–112 (2013).
38. Zhang, Q., Luo, J., Wu, S., Si, H., Gao, C., Xu, W., Abdullah, S. E., Higgs, B. W., Dennis, P. A., van der Heijden, M. S., Segal, N. H., Chaft, J. E., Hembrough, T., Barrett, J. C. & Hellmann, M. D. Prognostic and predictive impact of circulating tumor DNA in patients with advanced cancers treated with immune checkpoint blockade. *Cancer Discov.* **10,** 1842–1853 (2020).
39. Dawson, S.-J., Tsui, D. W. Y., Murtaza, M., Biggs, H., Rueda, O. M., Chin, S.-F., Dunning, M. J., Gale, D., Forshew, T., Mahler-Araujo, B., Rajan, S., Humphray, S., Becq, J., Halsall, D., Wallis, M., Bentley, D., Caldas, C. & Rosenfeld, N. Analysis of circulating tumor DNA to monitor metastatic breast cancer. *N. Engl. J. Med.* **368,** 1199–1209 (2013).
40. Burnham, P., Kim, M. S., Agbor-Enoh, S., Luikart, H., Valantine, H. A., Khush, K. K. & De Vlaminck, I. Single-stranded DNA library preparation uncovers the origin and diversity of ultrashort cell-free DNA in plasma. *Sci. Rep.* **6,** 27859 (2016).
41. Oellerich, M., Budde, K., Osmanodja, B., Bornemann-Kolatzki, K., Beck, J., Schütz, E. & Walson, P. D. Donor-derived cell-free DNA as a diagnostic tool in transplantation. *Front. Genet.* **13,** 1031894 (2022).

42. Burnham, P., Khush, K. & De Vlaminck, I. Myriad applications of circulating cell-free DNA in precision organ transplant monitoring. *Ann. Am. Thorac. Soc.* **14,** S237–S241 (2017).
43. Heitzer, E., Auinger, L. & Speicher, M. R. Cell-free DNA and apoptosis: How dead cells inform about the living. *Trends Mol. Med.* **26,** 519–528 (2020).
44. Jahr, S., Hentze, H., Englisch, S., Hardt, D., Fackelmayer, F. O., Hesch, R. D. & Knippers, R. DNA fragments in the blood plasma of cancer patients: quantitations and evidence for their origin from apoptotic and necrotic cells. *Cancer Res.* **61,** 1659–1665 (2001).
45. Rostami, A., Lambie, M., Yu, C. W., Stambolic, V., Waldron, J. N. & Bratman, S. V. Senescence, necrosis, and apoptosis govern circulating cell-free DNA release kinetics. *Cell Rep.* **31,** 107830 (2020).
46. Snyder, M. W., Kircher, M., Hill, A. J., Daza, R. M. & Shendure, J. Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues-of-origin. *Cell* **164,** 57–68 (2016).
47. Moss, J., Magenheim, J., Neiman, D., Zemmour, H., Loyfer, N., Korach, A., Samet, Y., Maoz, M., Druid, H., Arner, P., Fu, K.-Y., Kiss, E., Spalding, K. L., Landesberg, G., Zick, A., Grinshpun, A., Shapiro, A. M. J., Grompe, M., Wittenberg, A. D., Glaser, B., Shemer, R., Kaplan, T. & Dor, Y. Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. *Nat. Commun.* **9,** 5068 (2018).
48. Sun, K., Jiang, P., Chan, K. C. A., Wong, J., Cheng, Y. K. Y., Liang, R. H. S., Chan, W.-K., Ma, E. S. K., Chan, S. L., Cheng, S. H., Chan, R. W. Y., Tong, Y. K., Ng, S. S. M., Wong, R. S. M., Hui, D. S. C., Leung, T. N., Leung, T. Y., Lai, P. B. S., Chiu, R. W. K. & Lo, Y. M. D. Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. *Proc. Natl. Acad. Sci. U. S. A.* **112,** E5503–12 (2015).
49. Mattox, A. K., Douville, C., Wang, Y., Popoli, M., Ptak, J., Silliman, N., Dobbyn, L., Schaefer, J., Lu, S., Pearlman, A. H., Cohen, J. D., Tie, J., Gibbs, P., Lahouel, K., Bettegowda, C., Hruban, R. H., Tomasetti, C., Jiang, P., Chan, K. C. A., Lo, Y. M. D., Papadopoulos, N., Kinzler, K. W. & Vogelstein, B. The origin of highly elevated cell-free DNA in healthy individuals and patients with pancreatic, colorectal, lung, or ovarian cancer. *Cancer Discov.* **13,** 2166–2179 (2023).
50. Wang, E., Batey, A., Struble, C., Musci, T., Song, K. & Oliphant, A. Gestational age and maternal weight effects on fetal cell-free DNA in maternal plasma: Maternal factors affecting fetal cell-free DNA in maternal plasma. *Prenat. Diagn.* **33,** 662–666 (2013).
51. de Miranda, F. S., Claudio, L. M. A. M., de Almeida, D. S. M., Nunes, J. B., Barauna, V. G., Luiz, W. B., Vassallo, P. F. & Campos, L. C. G. Cell-free nuclear and mitochondrial DNA as potential biomarkers for assessing sepsis severity. *Biomedicines* **12,** (2024).
52. Yan, H. P., Li, M., Lu, X. L., Zhu, Y. M., Ou-Yang, W.-X., Xiao, Z. H., Qiu, J. & Li, S. J. Use of plasma mitochondrial DNA levels for determining disease severity and prognosis in pediatric sepsis: a case control study. *BMC Pediatr.* **18,** 267 (2018).
53. Kung, C.-T., Hsiao, S.-Y., Tsai, T.-C., Su, C.-M., Chang, W.-N., Huang, C.-R., Wang, H.-C., Lin, W.-C., Chang, H.-W., Lin, Y.-J., Cheng, B.-C., Su, B. Y.-J., Tsai, N.-W. & Lu, C.-H. Plasma nuclear and mitochondrial DNA levels as predictors of outcome in severe sepsis patients in the emergency room. *J. Transl. Med.* **10,** 130 (2012).
54. Markus, H., Zhao, J., Contente-Cuomo, T., Stephens, M. D., Raupach, E., Odenheimer-Bergman, A., Connor, S., McDonald, B. R., Moore, B., Hutchins, E., McGilvrey, M., de la Maza, M. C., Van Keuren-Jensen, K., Pirrotte, P., Goel, A., Becerra, C., Von Hoff, D. D., Celinski, S. A., Hingorani, P. & Murtaza, M. Analysis of recurrently protected genomic regions in cell-free DNA found in urine. *Sci. Transl. Med.* **13,** eaaz3088 (2021).
55. Gu, W., Deng, X., Lee, M., Sucu, Y. D., Arevalo, S., Stryke, D., Federman, S., Gopez, A., Reyes, K., Zorn, K., Sample, H., Yu, G., Ishpuniani, G., Briggs, B., Chow, E. D., Berger, A., Wilson, M. R., Wang, C., Hsu, E., Miller, S., DeRisi, J. L. & Chiu, C. Y. Rapid pathogen detection by metagenomic next-generation sequencing of infected body fluids. *Nat. Med.* **27,** 115–124 (2021).
56. Van den Ackerveken, P., Lobbens, A., Turatsinze, J.-V., Solis-Mezarino, V., Völker-Albert, M., Imhof, A. & Herzog, M. A novel proteomics approach to epigenetic profiling of circulating nucleosomes. *Sci. Rep.* **11,** 7256 (2021).
57. Han, D. S. C., Ni, M., Chan, R. W. Y., Chan, V. W. H., Lui, K. O., Chiu, R. W. K. & Lo, Y. M. D. The biology of cell-free DNA fragmentation and the roles of DNASE1, DNASE1L3, and DFFB. *Am. J. Hum. Genet.* **106,** 202–214 (2020).
58. Sanchez, C., Roch, B., Mazard, T., Blache, P., Dache, Z. A. A., Pastor, B., Pisareva, E., Tanos, R. & Thierry, A. R. Circulating nuclear DNA structural features, origins, and complete size profile revealed by fragmentomics. *JCI Insight* **6,** (2021).
59. Han, D., Li, R., Shi, J., Tan, P., Zhang, R. & Li, J. Liquid biopsy for infectious diseases: a focus on microbial cell-free DNA sequencing. *Theranostics* **10,** 5501–5513 (2020).
60. Thierry, A. R. Circulating DNA fragmentomics and cancer screening. *Cell Genom.* **3,** 100242 (2023).
61. Maxam, A. M. & Gilbert, W. A new method for sequencing DNA. *Proc. Natl. Acad. Sci. U. S. A.* **74,** 560–564 (1977).
62. Sanger, F. & Coulson, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.* **94,** 441–448 (1975).
63. Hyman, E. D. A new method of sequencing DNA. *Anal. Biochem.* **174,** 423–436 (1988).
64. Prober, J. M., Trainor, G. L., Dam, R. J., Hobbs, F. W., Robertson, C. W., Zagursky, R. J., Cocuzza, A. J., Jensen, M. A. & Baumeister, K. A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science* **238,** 336–341 (1987).
65. Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A. & Merrick, J. M. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **269,** 496–512 (1995).

66. Voelkerding, K. V., Dames, S. A. & Durtschi, J. D. Next-generation sequencing: from basic research to diagnostics. *Clin. Chem.* **55,** 641–658 (2009).
67. Barzilay, E. J., Schaad, N., Magloire, R., Mung, K. S., Boncy, J., Dahourou, G. A., Mintz, E. D., Steenland, M. W., Vertefeuille, J. F. & Tappero, J. W. Cholera surveillance during the Haiti epidemic--the first 2 years. *N. Engl. J. Med.* **368,** 599–609 (2013).
68. Vermeulen, C., Pagès-Gallego, M., Kester, L., Kranendonk, M. E. G., Wesseling, P., Verburg, N., de Witt Hamer, P., Kooi, E. J., Dankmeijer, L., van der Lugt, J., van Baarsen, K., Hoving, E. W., Tops, B. B. J. & de Ridder, J. Ultra-fast deep-learned CNS tumour classification during surgery. *Nature* **622,** 842–849 (2023).
69. Miga, K. H., Koren, S., Rhie, A., Vollger, M. R., Gershman, A., Bzikadze, A., Brooks, S., Howe, E., Porubsky, D., Logsdon, G. A., Schneider, V. A., Potapova, T., Wood, J., Chow, W., Armstrong, J., Fredrickson, J., Pak, E., Tigyi, K., Kremitzki, M., Markovic, C., Maduro, V., Dutra, A., Bouffard, G. G., Chang, A. M., Hansen, N. F., Wilfert, A. B., Thibaud-Nissen, F., Schmitt, A. D., Belton, J.-M., Selvaraj, S., Dennis, M. Y., Soto, D. C., Sahasrabudhe, R., Kaya, G., Quick, J., Loman, N. J., Holmes, N., Loose, M., Surti, U., Risques, R. A., Graves Lindsay, T. A., Fulton, R., Hall, I., Paten, B., Howe, K., Timp, W., Young, A., Mullikin, J. C., Pevzner, P. A., Gerton, J. L., Sullivan, B. A., Eichler, E. E. & Phillippy, A. M. Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585,** 79–84 (2020).
70. Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S. J., Diekhans, M., Logsdon, G. A., Alonge, M., Antonarakis, S. E., Borchers, M., Bouffard, G. G., Brooks, S. Y., Caldas, G. V., Chen, N.-C., Cheng, H., Chin, C.-S., Chow, W., de Lima, L. G., Dishuck, P. C., Durbin, R., Dvorkina, T., Fiddes, I. T., Formenti, G., Fulton, R. S., Fungtammasan, A., Garrison, E., Grady, P. G. S., Graves-Lindsay, T. A., Hall, I. M., Hansen, N. F., Hartley, G. A., Haukness, M., Howe, K., Hunkapiller, M. W., Jain, C., Jain, M., Jarvis, E. D., Kerpedjiev, P., Kirsche, M., Kolmogorov, M., Korlach, J., Kremitzki, M., Li, H., Maduro, V. V., Marschall, T., McCartney, A. M., McDaniel, J., Miller, D. E., Mullikin, J. C., Myers, E. W., Olson, N. D., Paten, B., Peluso, P., Pevzner, P. A., Porubsky, D., Potapova, T., Rogaev, E. I., Rosenfeld, J. A., Salzberg, S. L., Schneider, V. A., Sedlazeck, F. J., Shafin, K., Shew, C. J., Shumate, A., Sims, Y., Smit, A. F. A., Soto, D. C., Sović, I., Storer, J. M., Streets, A., Sullivan, B. A., Thibaud-Nissen, F., Torrance, J., Wagner, J., Walenz, B. P., Wenger, A., Wood, J. M. D., Xiao, C., Yan, S. M., Young, A. C., Zarate, S., Surti, U., McCoy, R. C., Dennis, M. Y., Alexandrov, I. A., Gerton, J. L., O'Neill, R. J., Timp, W., Zook, J. M., Schatz, M. C., Eichler, E. E., Miga, K. H. & Phillippy, A. M. The complete sequence of a human genome. *Science* **376,** 44–53 (2022).
71. Rhie, A., Nurk, S., Cechova, M., Hoyt, S. J., Taylor, D. J., Altemose, N., Hook, P. W., Koren, S., Rautiainen, M., Alexandrov, I. A., Allen, J., Asri, M., Bzikadze, A. V., Chen, N.-C., Chin, C.-S., Diekhans, M., Flicek, P., Formenti, G., Fungtammasan, A., Garcia Giron, C., Garrison, E., Gershman, A., Gerton, J. L., Grady, P. G. S., Guarracino, A., Haggerty, L., Halabian, R., Hansen, N. F., Harris, R., Hartley, G. A., Harvey, W. T., Haukness, M., Heinz, J., Hourlier, T., Hubley, R. M., Hunt, S. E., Hwang, S., Jain, M., Kesharwani, R. K., Lewis, A. P., Li, H., Logsdon, G. A., Lucas, J. K., Makalowski, W., Markovic, C., Martin, F. J., Mc Cartney, A. M., McCoy, R. C., McDaniel, J., McNulty, B. M., Medvedev, P., Mikheenko, A., Munson, K. M., Murphy, T. D., Olsen, H. E., Olson, N. D., Paulin, L. F., Porubsky, D., Potapova, T., Ryabov, F., Salzberg, S. L., Sauria, M. E. G., Sedlazeck, F. J., Shafin, K., Shepelev, V. A., Shumate, A., Storer, J. M., Surapaneni, L., Taravella Oill, A. M., Thibaud-Nissen, F., Timp, W., Tomaszkiewicz, M., Vollger, M. R., Walenz, B. P., Watwood, A. C., Weissensteiner, M. H., Wenger, A. M., Wilson, M. A., Zarate, S., Zhu, Y., Zook, J. M., Eichler, E. E., O'Neill, R. J., Schatz, M. C., Miga, K. H., Makova, K. D. & Phillippy, A. M. The complete sequence of a human Y chromosome. *Nature* **621,** 344–354 (2023).
72. Ashton, P. M., Nair, S., Dallman, T., Rubino, S., Rabsch, W., Mwaigwisya, S., Wain, J. & O'Grady, J. MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. *Nat. Biotechnol.* **33,** 296–300 (2015).
73. Lau, B. T., Almeda, A., Schauer, M., McNamara, M., Bai, X., Meng, Q., Partha, M., Grimes, S. M., Lee, H., Heestand, G. M. & Ji, H. P. Single-molecule methylation profiles of cell-free DNA in cancer with nanopore sequencing. *Genome Med.* **15,** 33 (2023).
74. Afflerbach, A.-K., Rohrandt, C., Brändl, B., Sönksen, M., Hench, J., Frank, S., Börnigen, D., Alawi, M., Mynarek, M., Winkler, B., Ricklefs, F., Synowitz, M., Dührsen, L., Rutkowski, S., Wefers, A. K., Müller, F.-J., Schoof, M. & Schüller, U. Classification of brain tumors by Nanopore sequencing of cell-free DNA from cerebrospinal fluid. *Clin. Chem.* **70,** 250–260 (2024).
75. Lo, Y. M., Zhang, J., Leung, T. N., Lau, T. K., Chang, A. M. & Hjelm, N. M. Rapid clearance of fetal DNA from maternal plasma. *Am. J. Hum. Genet.* **64,** 218–224 (1999).
76. Butler, T. M., Spellman, P. T. & Gray, J. Circulating-tumor DNA as an early detection and diagnostic tool. *Curr. Opin. Genet. Dev.* **42,** 14–21 (2017).
77. Hilt, E. E. & Ferrieri, P. Next generation and other sequencing technologies in diagnostic microbiology and infectious diseases. *Genes (Basel)* **13,** 1566 (2022).
78. Johnson, J. S., Spakowicz, D. J., Hong, B.-Y., Petersen, L. M., Demkowicz, P., Chen, L., Leopold, S. R., Hanson, B. M., Agresta, H. O., Gerstein, M., Sodergren, E. & Weinstock, G. M. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nat. Commun.* **10,** 5029 (2019).

79. Nilsson, R. H., Kristiansson, E., Ryberg, M., Hallenberg, N. & Larsson, K.-H. Intraspecific ITS variability in the kingdom fungi as expressed in the international sequence databases and its implications for molecular species identification. *Evol. Bioinform. Online* **4,** 193–201 (2008).
80. Schoch, C. L., Seifert, K. A., Huhndorf, S., Robert, V., Spouge, J. L., Levesque, C. A., Chen, W., Fungal Barcoding Consortium & Fungal Barcoding Consortium Author List. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 6241–6246 (2012).
81. Rhee, S.-Y., Kassaye, S. G., Jordan, M. R., Kouamou, V., Katzenstein, D. & Shafer, R. W. Public availability of HIV-1 drug resistance sequence and treatment data: a systematic review. *Lancet Microbe* **3,** e392–e398 (2022).
82. Laurence, M., Hatzis, C. & Brash, D. E. Common contaminants in next-generation sequencing that hinder discovery of low-abundance microbes. *PLoS One* **9,** e97876 (2014).
83. Bukin, Y. S., Galachyants, Y. P., Morozov, I. V., Bukin, S. V., Zakharenko, A. S. & Zemskaya, T. I. The effect of 16S rRNA region choice on bacterial community metabarcoding results. *Sci. Data* **6,** 190007 (2019).
84. Yang, B., Wang, Y. & Qian, P.-Y. Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. *BMC Bioinformatics* **17,** 135 (2016).
85. Zhou, Y., Wylie, K. M., El Feghaly, R. E., Mihindukulasuriya, K. A., Elward, A., Haslam, D. B., Storch, G. A. & Weinstock, G. M. Metagenomic approach for identification of the pathogens associated with diarrhea in stool specimens. *J. Clin. Microbiol.* **54,** 368–375 (2016).
86. Pendleton, K. M., Erb-Downward, J. R., Bao, Y., Branton, W. R., Falkowski, N. R., Newton, D. W., Huffnagle, G. B. & Dickson, R. P. Rapid pathogen identification in bacterial pneumonia using real-time metagenomics. *Am. J. Respir. Crit. Care Med.* **196,** 1610–1612 (2017).
87. Schlaberg, R., Queen, K., Simmon, K., Tardif, K., Stockmann, C., Flygare, S., Kennedy, B., Voelkerding, K., Bramley, A., Zhang, J., Eilbeck, K., Yandell, M., Jain, S., Pavia, A. T., Tong, S. & Ampofo, K. Viral pathogen detection by metagenomics and pan-viral group polymerase chain reaction in children with pneumonia lacking identifiable etiology. *J. Infect. Dis.* **215,** 1407–1415 (2017).
88. Kapp, J. D., Green, R. E. & Shapiro, B. A fast and efficient single-stranded genomic library preparation method optimized for ancient DNA. *J. Hered.* **112,** 241–249 (2021).
89. Gansauge, M.-T., Gerber, T., Glocke, I., Korlevic, P., Lippik, L., Nagel, S., Riehl, L. M., Schmidt, A. & Meyer, M. Single-stranded DNA library preparation from highly degraded DNA using T4 DNA ligase. *Nucleic Acids Res.* **45,** e79 (2017).
90. Cheng, J. C., Swarup, N., Wong, D. T. W. & Chia, D. A review on the impact of single-stranded library preparation on plasma cell-free diversity for cancer detection. *Front. Oncol.* **14,** 1332004 (2024).
91. Linthorst, J., Welkers, M. R. A. & Sistermans, E. A. Distinct fragmentation patterns of circulating viral cell-free DNA in 83,552 non-invasive prenatal testing samples. *Extracellular Vesicles and Circulating Nucleic Acids* **2,** 228–237 (2021).
92. Jiang, P., Xie, T., Ding, S. C., Zhou, Z., Cheng, S. H., Chan, R. W. Y., Lee, W.-S., Peng, W., Wong, J., Wong, V. W. S., Chan, H. L. Y., Chan, S. L., Poon, L. C. Y., Leung, T. Y., Chan, K. C. A., Chiu, R. W. K. & Lo, Y. M. D. Detection and characterization of jagged ends of double-stranded DNA in plasma. *Genome Res.* **30,** 1144–1153 (2020).
93. Xie, T., Wang, G., Ding, S. C., Lee, W.-S., Cheng, S. H., Chan, R. W. Y., Zhou, Z., Ma, M.-J. L., Han, D. S. C., Teoh, J. Y. C., Lam, W. K. J., Jiang, P., Chiu, R. W. K., Chan, K. C. A. & Lo, Y. M. D. High-resolution analysis for urinary DNA jagged ends. *NPJ Genom. Med.* **7,** 14 (2022).
94. Mzava, O., Cheng, A. P., Chang, A., Smalling, S., Djomnang, L.-A. K., Lenz, J. S., Longman, R., Steadman, A., Gómez-Escobar, L. G., Schenck, E. J., Salvatore, M., Satlin, M. J., Suthanthiran, M., Lee, J. R., Mason, C. E., Dadhania, D. & De Vlaminck, I. A metagenomic DNA sequencing assay that is robust against environmental DNA contamination. *Nat. Commun.* **13,** 4197 (2022).
95. Francis, O. E., Bendall, M., Manimaran, S., Hong, C., Clement, N. L., Castro-Nallar, E., Snell, Q., Schaalje, G. B., Clement, M. J., Crandall, K. A. & Johnson, W. E. Pathoscope: species identification and strain attribution with unassembled sequencing data. *Genome Res.* **23,** 1721–1729 (2013).
96. Hong, C., Manimaran, S., Shen, Y., Perez-Rogers, J. F., Byrd, A. L., Castro-Nallar, E., Crandall, K. A. & Johnson, W. E. PathoScope 2.0: a complete computational framework for strain identification in environmental or clinical sequencing samples. *Microbiome* **2,** 33 (2014).
97. Gihawi, A., Ge, Y., Lu, J., Puiu, D., Xu, A., Cooper, C. S., Brewer, D. S., Pertea, M. & Salzberg, S. L. Major data analysis errors invalidate cancer microbiome findings. *MBio* **14,** e0160723 (2023).
98. Sepich-Poore, G. D., McDonald, D., Kopylova, E., Guccione, C., Zhu, Q., Austin, G., Carpenter, C., Fraraccio, S., Wandro, S., Kosciolek, T., Janssen, S., Metcalf, J. L., Song, S. J., Kanbar, J., Miller-Montgomery, S., Heaton, R., Mckay, R., Patel, S. P., Swafford, A. D., Korem, T. & Knight, R. Robustness of cancer microbiome signals over a broad range of methodological variation. *Oncogene* **43,** 1127–1148 (2024).
99. Lu, J., Rincon, N., Wood, D. E., Breitwieser, F. P., Pockrandt, C., Langmead, B., Salzberg, S. L. & Steinegger, M. Metagenome analysis using the Kraken software suite. *Nat. Protoc.* **17,** 2815–2839 (2022).

100. Truong, D. T., Franzosa, E. A., Tickle, T. L., Scholz, M., Weingart, G., Pasolli, E., Tett, A., Huttenhower, C. & Segata, N. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* **12,** 902–903 (2015).
101. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15,** R46 (2014).
102. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20,** 257 (2019).
103. Lu, J., Breitwieser, F. P., Thielen, P. & Salzberg, S. L. Bracken: estimating species abundance in metagenomics data. *PeerJ Comput. Sci.* **3,** e104 (2017).
104. Kostic, A. D., Ojesina, A. I., Pedamallu, C. S., Jung, J., Verhaak, R. G. W., Getz, G. & Meyerson, M. PathSeq: software to identify or discover microbes by deep sequencing of human tissue. *Nat. Biotechnol.* **29,** 393–396 (2011).
105. Menzel, P., Ng, K. L. & Krogh, A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat. Commun.* **7,** 11257 (2016).
106. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12,** 59–60 (2015).
107. Ye, S. H., Siddle, K. J., Park, D. J. & Sabeti, P. C. Benchmarking metagenomics tools for taxonomic classification. *Cell* **178,** 779–794 (2019).
108. McIntyre, A. B. R., Ounit, R., Afshinnekoo, E., Prill, R. J., Hénaff, E., Alexander, N., Minot, S. S., Danko, D., Foox, J., Ahsanuddin, S., Tighe, S., Hasan, N. A., Subramanian, P., Moffat, K., Levy, S., Lonardi, S., Greenfield, N., Colwell, R. R., Rosen, G. L. & Mason, C. E. Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. *Genome Biol.* **18,** 182 (2017).
109. Lindgreen, S., Adair, K. L. & Gardner, P. P. An evaluation of the accuracy and speed of metagenome analysis tools. *Sci. Rep.* **6,** 19233 (2016).
110. Donovan, P. D., Gonzalez, G., Higgins, D. G., Butler, G. & Ito, K. Identification of fungi in shotgun metagenomics datasets. *PLoS One* **13,** e0192898 (2018).
111. Lennon, J. T. & Locey, K. J. More support for Earth's massive microbiome. *Biol. Direct* **15,** 5 (2020).
112. Nasko, D. J., Koren, S., Phillippy, A. M. & Treangen, T. J. RefSeq database growth influences the accuracy of k-mer-based lowest common ancestor species identification. *Genome Biol.* **19,** 165 (2018).
113. Méric, G., Wick, R. R., Watts, S. C., Holt, K. E. & Inouye, M. Correcting index databases improves metagenomic studies. *bioRxiv* 712166 (2019). doi:10.1101/712166
114. R Marcelino, V., Holmes, E. C. & Sorrell, T. C. The use of taxon-specific reference databases compromises metagenomic classification. *BMC Genomics* **21,** 184 (2020).
115. Lu, J. & Salzberg, S. L. Removing contaminants from databases of draft genomes. *PLoS Comput. Biol.* **14,** e1006277 (2018).
116. Eisenhofer, R., Minich, J. J., Marotz, C., Cooper, A., Knight, R. & Weyrich, L. S. Contamination in low microbial biomass microbiome studies: Issues and recommendations. *Trends Microbiol.* **27,** 105–117 (2019).
117. Salter, S. J., Cox, M. J., Turek, E. M., Calus, S. T., Cookson, W. O., Moffatt, M. F., Turner, P., Parkhill, J., Loman, N. J. & Walker, A. W. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* **12,** 87 (2014).
118. Jousselin, E., Clamens, A.-L., Galan, M., Bernard, M., Maman, S., Gschloessl, B., Duport, G., Meseguer, A. S., Calevro, F. & Coeur d'acier, A. Assessment of a 16S rRNA amplicon Illumina sequencing procedure for studying the microbiome of a symbiont-rich aphid genus. *Mol. Ecol. Resour.* **16,** 628–640 (2016).
119. Larsson, A. J. M., Stanley, G., Sinha, R., Weissman, I. L. & Sandberg, R. Computational correction of index switching in multiplexed sequencing libraries. *Nat. Methods* **15,** 305–307 (2018).
120. Zozaya-Valdés, E., Wong, S. Q., Raleigh, J., Hatzimihalis, A., Ftouni, S., Papenfuss, A. T., Sandhu, S., Dawson, M. A. & Dawson, S.-J. Detection of cell-free microbial DNA using a contaminant-controlled analysis framework. *Genome Biol.* **22,** 187 (2021).
121. Davis, N. M., Proctor, D. M., Holmes, S. P., Relman, D. A. & Callahan, B. J. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome* **6,** 226 (2018).

CHAPTER 2

NGS-based *Aspergillus* detection in plasma and lung lavage of children with invasive pulmonary aspergillosis

Emmy Wesdorp [1,2], Laura Rotte [3,*], Li-Ting Chen [1,2,*], Myrthe Jager [1,2,*], Nicolle Besselink [1,2], Carlo Vermeulen [1,2], Ferry Hagen [4,5,6], Tjomme van der Bruggen [7], Caroline Lindemans [3,8], Tom Wolfs [8], Louis Bont [8,#], Jeroen de Ridder [1,2,#]

[1] Center for Molecular Medicine, University Medical Center Utrecht, Utrecht, The Netherlands.

[2] Oncode Institute, Utrecht, The Netherlands.

[3] Hematopoietic stem cell transplantation, Princess Máxima Center for Pediatric Oncology, Utrecht, The Netherlands.

[4] Westerdijk Fungal Biodiversity Institute, Utrecht, The Netherlands.

[5] Department of Medical Microbiology, University Medical Center Utrecht, Utrecht, The Netherlands.

[6] Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam, The Netherlands.

[7] Department of Medical Microbiology, University Medical Centre Utrecht, The Netherlands.

[8] Department of Pediatric Infectious Diseases and Immunology, Wilhelmina Children's hospital, UMC Utrecht, Utrecht, The Netherlands.

*These authors contributed equally to this work

[#] Corresponding authors.

Published in NPJ Genomic Medicine; DOI: 10.1038/s41525-025-00482-8

Abstract

In immunocompromised pediatric patients, diagnosing invasive pulmonary aspergillosis (IPA) poses a significant challenge. Next-Generation Sequencing (NGS) shows promise for detecting fungal DNA but lacks standardization. This study aims to advance towards clinical evaluation of liquid biopsy NGS for *Aspergillus* detection, through an evaluation of wet-lab procedures and computational analysis. Our findings support using both CHM13v2.0 and GRCh38.p14 in host-read mapping to reduce fungal false-positives. We demonstrate the sensitivity of our custom kraken2 database, cRE.21, in detecting *Aspergillus* species. Additionally, cell-free DNA sequencing shows superior performance to whole-cell DNA sequencing by recovering higher fractions of fungal DNA in lung fluid (bronchoalveolar lavage [BAL] fluid) and plasma samples from pediatric patients with probable IPA. In a proof-of-principle, *A. fumigatus* was identified in 5 out of 7 BAL fluid samples and 3 out of 5 plasma samples. This optimized workflow can advance fungal-NGS research and represents a step towards enhancing diagnostic certainty by enabling more sensitive and accurate species-level diagnosis of IPA in immunocompromised patients.

Introduction

Invasive mold disease (IMD) is a threat to immunocompromised children, especially those with hematological malignancies or undergoing hematopoietic stem cell transplantation (HSCT)[1]. Despite receiving antifungal prophylaxis, breakthrough IMD can still occur, with an incidence of up to 20%[2–6], with *Aspergillus* being the most common cause of IMD[7]. Early and accurate identification of fungal pathogens is crucial for tailoring antifungal treatment, especially as diverse fungal pathogenic species require different antifungal treatments. For aspergillosis, an azole is recommended as first-line treatment, whereas for mucormycosis amphotericin B is the first-choice treatment[8].

The current diagnostic toolbox for IMD includes radiologic imaging, microbiological bronchoalveolar lavage (BAL) fluid analysis (i.e. culture, antigen testing and PCR) and antigen testing on serum. While valuable[9], these microbiological tests have limited sensitivity, particularly as prior antifungal treatment — common in pediatric patients — compromises their performance. For instance, BAL PCR shows a sensitivity of 0.17% (95% CI, 0.05–0.45), while BAL galactomannan has a sensitivity of 60%[10,11]. Additionally, the galactomannan antigen test lacks species level identification, which is crucial in the differentiation between invasive aspergillosis from invasive non-*Aspergillus* species. Therefore, there is an urgent clinical need to expand the diagnostic toolbox for early, sensitive and accurate species level diagnosis, preventing disease progression, improving patient outcomes and avoiding unnecessary exposure to prolonged toxic antifungal therapy.

Microbial next-generation sequencing (NGS) detects microbial DNA of pathogens in patients with infectious diseases and holds promise for IMD diagnosis[12–16]. Microbial NGS enables species-level identification of pathogens and is regarded as sensitive when applied to BAL samples[16], while its sensitivity in plasma appears to depend on the specific pathogen responsible for the IMD[15,16]. Yet, the preferred microbial NGS workflow for pediatric IMD diagnosis is unclear and multiple technical gaps exist. Computationally, there is a lack of standardization on taxonomy classification for fungal identification in samples from patients with IMD. Specifically, the impacts of reference database composition, genomic sequence processing (i.e. masking low-complexity and contaminated regions) and threshold settings for fungal diagnosis remain unclear. Given that only a small fraction of short-read sequences corresponds to potential pathogens, any taxonomic misassignment due to suboptimal computational methods or parameters can greatly affect pathogen identification and subsequent therapeutic decisions. Such misclassification may occur when reads coincidentally match to multiple genomes in the reference database or persistently match to an incorrect genome or no genome at all leading to false positive or false negative outcomes[17].

Additionally, it remains unclear what the optimal wet-lab strategy is to maximize fungal DNA yield. Specifically, little is known about the impact of sample source, DNA-type, and DNA isolation method on the success of fungal diagnosis. Although traditionally whole-cell DNA (wcDNA) sequencing was applied to samples collected near the infection sources, like BAL fluid, recent research indicates that BAL fluid cell-free DNA (cfDNA) sequencing may outperform wcDNA sequencing in pulmonary infections[18]. At the same time, plasma microbial cfDNA sequencing has also shown potential in mostly small cohort studies involving adult IMD patients[12,13,15,16], and one pediatric study[14], highlighting its potential but also emphasizing the need for further investigation. Methods like DNA isolation and adapter ligation can also affect fungal DNA abundance. Although single-stranded (ss) sequencing libraries are preferred for e.g. bacterial and viral cfDNA over double-stranded (ds) ligation-based DNA library preparation[19], further exploration through comparative evaluation is needed to determine the optimal approach for recovering fungal mold DNA from liquid biopsies.

In this study, we focus on the detection of *Aspergillus*, the predominant pathogen associated with pulmonary IMD. We aim to close above-mentioned technical knowledge gaps by optimizing six key steps through comparative experimental testing (Q#1-6; detailed in Fig. 1). We introduce a refined wet-lab strategy together with an open-source cfDNA pathogen identification workflow, referred to as cfDNA Single-strand Pathogen Identification pipeline (cfSPI). cfSPI is optimized for the detection of *Aspergillus* species with minimal false positives taxonomic mislabeling and maximum accuracy of true positive detections. The cfSPI pipeline uses paired-end Illumina sequencing data and incorporates host genome mapping. Unmapped reads are subsequently classified using *kraken2*[20], with enhancements such as an improved reference database through *dustmasking* and *cleanup*[20,21], and an optimal confidence threshold for classification accuracy. We show that these factors, all refined in this study, can impact *Aspergillus* classification accuracy[17,22–25].

Finally, as a proof-of-principle, we applied cfDNA sequencing with cfSPI to seven pediatric patients (7 BAL fluid; 5 plasma) with invasive pulmonary aspergillosis (IPA) and eighteen external controls (9 BAL fluid; 9 plasma), successfully detecting *Aspergillus* species in the majority (6/7) of these IPA cases. This work establishes the groundwork for large cohort evaluations of the accuracy and sensitivity of cfDNA NGS for *Aspergillus* diagnosis in suspected patients, ultimately contributing to the potential implementation of NGS in the IMD diagnostic work-up.

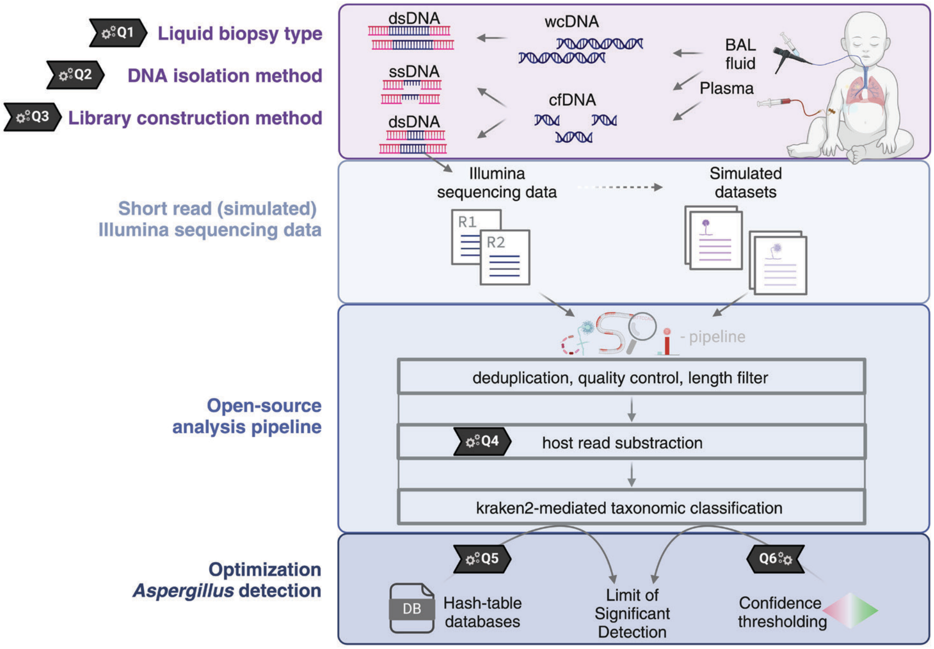

Q1 Liquid biopsy type
Q2 DNA isolation method
Q3 Library construction method
dsDNA
ssDNA
dsDNA
wcDNA
cfDNA
BAL fluid
Plasma
Short read (simulated) Illumina sequencing data
Illumina sequencing data
Simulated datasets
R1
R2
Open-source analysis pipeline
cfSPi - pipeline
deduplication, quality control, length filter
Q4 host read substraction
kraken2-mediated taxonomic classification
Optimization *Aspergillus* detection
Q5
Q6
DB
Hash-table databases
Limit of Significant Detection
Confidence thresholding

Fig. 1. NGS library strategies and the cfSPI open-source workflow for *Aspergillus* detection

Our comparative study employs Illumina shotgun sequencing to enhance *Aspergillus* DNA detection in liquid biopsies from pediatric immunocompromised patients. We set to optimize six key steps (Q#1-6) for microbial NGS-based *Aspergillus* diagnostics in IMD patients. We compared two NGS library preparation strategies — single-stranded (ss) and double-stranded (ds) ligation — across plasma and bronchoalveolar lavage (BAL) fluid samples (Q#1), alongside a comparison of BAL cell-free DNA (cfDNA) to whole-cell DNA (wcDNA) NGS (Q#2-3). After sequencing, we employ the open-source cell-free DNA pathogen identification workflow known as the cell-free DNA Single-strand Pathogen Identification pipeline (cfSPI), which is tailored for detecting *Aspergillus* species. Within the cfSPI pipeline, we conduct quality control of sequencing data followed by host-read subtraction (Q#4) and taxonomic classification using kraken2 with various hash-table genome reference databases (Q#5) and confidence thresholds (Q#6). To assess the accuracy of (*Aspergillus*) read classification, we further simulate short-read Illumina sequencing data. In our Limit of Significant Detection analysis, we investigate how hash-table database complexity and confidence thresholds impact the theoretical minimum number of molecules per million needed to detect significantly elevated *Aspergillus* taxon above the control background noise in control patient samples. All six key steps underwent optimization through comparative experimental testing.

Results

cfSPI is an open-source pipeline optimized for accurate *Aspergillus* detection. The pipeline processes paired-end Illumina sequencing data through quality filtering, host genome mapping, and classification using kraken2, a high-performance DNA-to-DNA tool leveraging 31-mer matches against a reference database. Each step is carefully fine-tuned, as described in the next section. To validate the results produced by cfSPI, we generated 87 simulated Illumina sequencing-like cfDNA datasets (55 *Aspergillus*, 7 *Penicillium*, 25 other fungi). Moreover, we Illumina sequenced samples of 7 probable invasive pulmonary aspergillosis and 18 external controls to further showcase utility of cfSPI on real patient data.

Optimizing host read subtraction and kraken2 database composition for the cfSPI pipeline

Detecting *Aspergillus*-derived DNA fragments from cfDNA sequencing data is a 'needle-in-a-haystack' challenge, where the vast majority of DNA reads will be derived from the host. For this reason, host-read subtraction through mapping to the host genome is a critical initial step in cfSPI to minimize the risk of overestimating microbial counts. Previous work already highlighted that improper subtraction can inflate bacterial counts[23]. There are a number of frequently used human genome versions, such as GRCh38.p14 and CHM13v2, which vary in completeness. The impact of mapping to these different genome versions on detecting fungal-derived reads in liquid biopsies remains unclear. To evaluate this, we mapped the sequencing reads from our external control samples (9 BAL; 9 plasma) to reference genomes GRCh38.p14 or CHM13v2. In addition, we performed *dual-mapping,* in which mapping was performed to a combined reference containing both GRCh38.p14 and CHM13v2. Our results revealed that both mapping to CHM13v2 and dual-mapping strategies significantly reduced the fraction of unmapped reads compared to mapping to GRCh38.p14, with rates dropping from 2.27% to 1.77% and 1.71%, respectively (Fig. 2a).

Unmapped reads were taxonomically classified using kraken2 (see *Methods*), using the default confidence threshold (CT=0). We found that fungal-classified reads dropped from 81.60 read per million (RPM) with GRCh38.p14 to 60.23 RPM and 58.86 RPM with CHM13v2 and dual-mapping[24]). These findings indicate that including CHM13v2 is essential to prevent inflated fungal counts. Overall, dual-mapping emerges as the preferred strategy for reducing misclassifications to the fungal kingdom.

Next, we focused on optimizing the kraken2-mediated taxonomic classification of non-human reads. To address the possibility of human reads remaining unidentified during host mapping, we evaluated whether incorporating CHM13v2 into the standard NCBI RefSeq database used by kraken2, which traditionally includes only GRCh38.p14, could help reduce misclassifications (see Supplementary Fig. 2 for database details). Indeed, inclusion of CHM13v2 (which we refer to as database 'uR.7') increased reads labeled as human (Fig. 2c, CT=0, the default setting in kraken2; Supplementary Fig. 3a CT=0.8), even after dual-mapping, while reducing fungal (Fig. 2d), but not other microbial counts (Supplementary Fig. 3b-c). Additionally, omission of CHM13v2 from the hash-table database led to misclassification of reads as microbial (Supplementary Fig. 4a), including as *Aspergillus* (Supplementary Fig. 4b). Therefore, in our cfSPI pipeline, we utilize kraken2 databases that include both CHM13v2 and GRCh38.p14 to further mitigate the risk of misclassifying human-derived reads as fungal or microbial taxa.

The default kraken2 NCBI RefSeq database ('uR.7 w/o CHM13v2') traditionally contains only seven of over 300 known *Aspergillus* species (Supplementary Fig. 2a-b), leading to inadequate *Aspergillus* classification. Specifically, 88.1% of reads from 55 simulated *Aspergillus* datasets remained unclassified (i.e. not classified to any taxa; Fig. 3b) when using a CT of 0.8, due to the absence of these species in the database (Fig. 3b). With the aim to enhance (species-level) classification of *Aspergillus*, we replaced the fungal genomes in the uR.7 database with **c**leaned, **d**ustmasked or **u**naltered fungal sequences from EuPathDB[26] and MycoCosm[27] (Fig. 3a; database details in Supplementary Fig. 2, Methods, and *Aspergillus* species in Suppl. Table 3). This resulted in new kraken2 databases: **u**RE.**21**, **d**RE.**21,** which include **21** *Aspergillus* species), and **u**RE.**31**, **d**RE.**31,** which include **31** *Aspergillus* species. As well as: **d**REM.258 and **d**REM.260, which include 258 and 260 *Aspergillus* species, respectively.

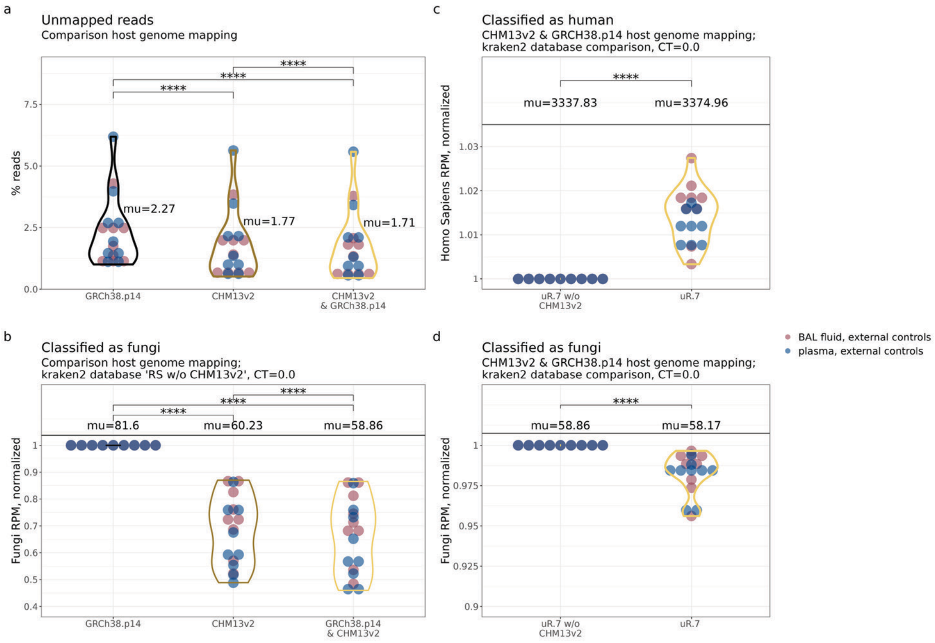
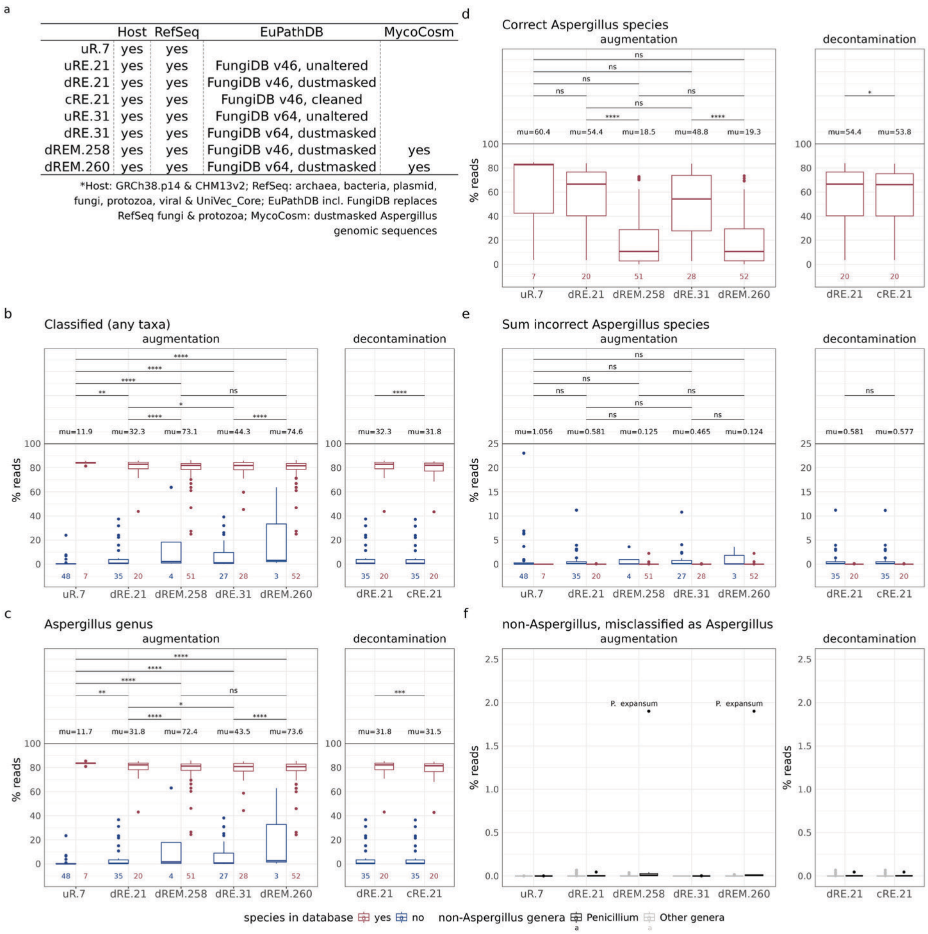


Fig. 2. Unraveling the impact of CHM13v2 host genome mapping for optimizing microbial read quantification in control samples

a. Percentage of reads remaining after mapping to the human reference genome (i.e., % unmapped reads) using the cfSPI workflow. Reads were mapped either to human reference genome GRCH38p14, CHM13v2, or a combination of these two. **b.** Fractional abundance of fungi (kingdom) kraken2 classified reads (CT=0, kraken2's default), after subtracting host reads via reference genome mapping, normalized to the old version of the human genome assembly (GRCh38.p14). Fractional abundance of **c.** human (species) and **d.** fungi (kingdom) classified reads when utilizing the 'CHM13v2-containing uR.7' or 'uR.7' database for kraken2 taxonomic classification (CT=0, kraken2's default) after dual-mapping to the host genome, normalized to 'uR.7'. In **a-d.**, each data point represents one control sample (9 BAL; 9 plasma), with colors indicating the sample type. Mean RPM values are denoted as 'mu'. Statistical analysis included one-tailed paired t-tests with Bonferroni correction (****, p ≤ 0.0001), utilizing the fractional abundance in RPM.

When evaluating the performance of these six new kraken2 databases, we first assessed the effects of dustmasking and cleanup, followed by the impact of database augmentation on the classification of simulated Illumina-like cfDNA *Aspergillus* datasets. Using a cleaned or dustmasked database proved crucial for preventing taxonomic misassignments of *Aspergillus*. We found a slight but significant reduction in overall classification rates (7.0–8.2%) and true positives at the genus (6.7–7.6%) and species levels (8.1–9.1%) with a CT of 0.8 (Supplementary Fig. 5a-c, e-g). However, classification accuracy improved significantly at CT values <0.2 (Supplementary Fig. 5h), reaching a 0.30–0.38% reduction in false positives (Supplementary Fig. 5d). These findings suggest that masking unreliable sequences in reference genomes is highly recommended. Second, we evaluated the impact of database augmentation on classification accuracy across different taxonomic levels. By incorporating fungal sequences from EuPathDB[26] and MycoCosm[27] into the kraken2 database, we identified a trade-off between database size and classification accuracy. Specifically, we observed the followings trends linked the extended databases: improved overall classification rates (Fig. 3b), increased true positives at the genus level (Fig. 3c), a broader range of detectable *Aspergillus* species

(Supplementary Table 4), and a reduced false positive rate during species classification (Fig. 2e; Supplementary Fig. 6). However, we also observed a lower percentage of reads classified at the species level (Fig. 3d). This while, misclassification of *Penicillium* as *Aspergillus* generally remained low but increased when expanding the database with MycoCosm genomes, particularly in dREM.260, where we noted 1.97% misclassification at the genus level and 1.90% at the species level (Fig. 2f).



a

	Host	RefSeq	EuPathDB	MycoCosm
uR.7	yes	yes		
uRE.21	yes	yes	FungiDB v46, unaltered	
dRE.21	yes	yes	FungiDB v46, dustmasked	
cRE.21	yes	yes	FungiDB v46, cleaned	
uRE.31	yes	yes	FungiDB v64, unaltered	
dRE.31	yes	yes	FungiDB v64, dustmasked	
dREM.258	yes	yes	FungiDB v46, dustmasked	yes
dREM.260	yes	yes	FungiDB v64, dustmasked	yes

*Host: GRCh38.p14 & CHM13v2; RefSeq: archaea, bacteria, plasmid, fungi, protozoa, viral & UniVec_Core; EuPathDB incl. FungiDB replaces RefSeq fungi & protozoa; MycoCosm: dustmasked Aspergillus genomic sequences

Fig. 3. Database-dependent taxon detection in simulated *Aspergillus* samples

a. Hash-table database composition overview (for comprehensive details on database composition, see Supplementary Fig. 2 and for details on database construction see *Methods* and Suppl. Table 5). Box plots displaying the results of the classification of simulated *Aspergillus* samples, including **b.** the overall classification rate (e.g. percentage reads classified to any taxon) as well the percentage of reads classified at the **c.** *Aspergillus* genus level, **d.** to correct *Aspergillus* species, and **e.** cumulative percentage to incorrect *Aspergillus* species. Read classification percentages are shown across databases (x-axis). **f.** Displaying the results of the classification of *Penicillium* (n=7) and other fungal genera (n=25). The boxplots show the percentage of non-*Aspergillus* simulated reads that were erroneously classified as *Aspergillus*. **b-f.** The CT was set at 0.8 for kraken2-based classification, to increase precision [20]. In **b-e.** the effects of hash-table database augmentation (on the left) and decontamination (on the right) on accuracy are tested via a one-tailed t-test with Bonferroni correction (*, p ≤ 0.05; **, p ≤ 0.01; ***, p ≤ 0.001; ****, p ≤ 0.0001; ns, p > 0.05) and the mean percentages are denoted as 'mu'.

Table 1. Clinical details of pediatric patients with probable IPA

	Patient information	Host factors	Radiology	Mycological evidence	Antifungal prophylaxis at time of diagnosis	EORTC/ MSG definition
A01	Girl, 11, Post-HSCT for inborn error of metabolism. Persistent fever in the setting of severe GVHD. Antifungal therapy was started based on repetitive positive serum GM. Additional diagnostic testing was performed after acute dyspnea.	HSCT r ecipient Severe GVHD Prolonged corticosteroid use	Negative HRCT thorax	positive serum GM 3.58 Hyphae: yes Positive BAL GM 5.6 BAL culture: *A. fumigatus* Negative phenotypic/ genotypic resistance test Functional endoscopic sinus surgery: negative culture, no hyphae Bronchoscopy: tracheobronchial pseudomembrane	Therapeutic L-AMB (daily 3 mg/kg)	Probable aspergillosis laryngo-tracheo-bronchitis
A02	Boy, 13, Post-HSCT for lymphoma Bronchiolitis obliterans treated with MP pulses. No clinical symptoms.	HSCT recipient Prolonged corticosteroid use	Suggestive HRCT thorax	Negative serum GM Hyphae: no Positive BAL GM 5.2 BAL culture: *A. fumigatus* Negative phenotypic/ genotypic resistance test	No antifungal prophylaxis or antifungal therapy	Probable pulmonary aspergillosis
A03	Boy, 18 months, Post HSCT for HLH Persistent fever, pre-engraftment.	HSCT recipient Neutropenia	Suggestive HRCT thorax	Positive serum GM 6.9 Hyphae: no Negative BAL GM BAL culture: negative	Isavuconazole prophylaxis Adequate trough level	Probable pulmonary aspergillosis
A04	Boy, 5, with high-risk ALL. Induction phase Persistent fever.	Hematologic malignancy	Suggestive HRCT thorax	Negative serum GM Hyphae: no Negative BAL GM BAL culture: *A. fumigatus* Negative phenotypic/ genotypic resistance test	Micafungin prophylaxis	Probable pulmonary aspergillosis
A05	Boy, 7, Post-HSCT for bone marrow failure Subfebrile temperature with coughing and tachypnea in the setting of graft failure.	HSCT recipient Neutropenia	Suggestive HRCT thorax	Negative serum GM Hyphae: no Negative BAL GM BAL culture: *A. fumigatus* TR34/L98H genotype resistance	Itraconazol prophylaxis No trough level measured	Probable pulmonary aspergillosis
A14	Girl, 1, with high-risk AML, induction therapy Persistent neutropenic fever	Hematologic malignancy Neutropenia	Suggestive HRCT thorax	Positive serum GM 2.5 Hyphae: no Positive BAL GM 7.2 BAL culture: negative	Micafungin prophylaxis	Probable pulmonary aspergillosis
A15	Boy, 16, Post-HSCT for NK cell malignancy Fever pre-engraftment	HSCT recipient Neutropenia	Suggestive HRCT thorax	Negative serum GM Hyphae: yes Positive BAL GM 9.7 BAL culture: *A. fumigatus* Negative phenotypic/ genotypic resistance test	Micafungin prophylaxis	Probable pulmonary aspergillosis

ALL; acute lymphoblastic leukemia, HSCT; hematopoietic stem cell transplantation, GVHD; graft versus host disease, GM; galactomannan, MP; methylprednisolone, NK; natural killer.

Overall, we conclude that medium-sized databases with 21 to 31 *Aspergillus* genomes are optimal for species level detection, while larger databases (258 to 260 genomes) excel in broader genus level detection. The curated cRE.21 database thereby demonstrated the highest sensitivity for species level detection (mean 53.8%; Fig. 3b), while dREM.260 excelled in genus level detection (mean 73.6%; Fig. 3c). Consequently, the cfSPI pipeline uses cRE.21 for species identification and dREM.260 for genus identification.

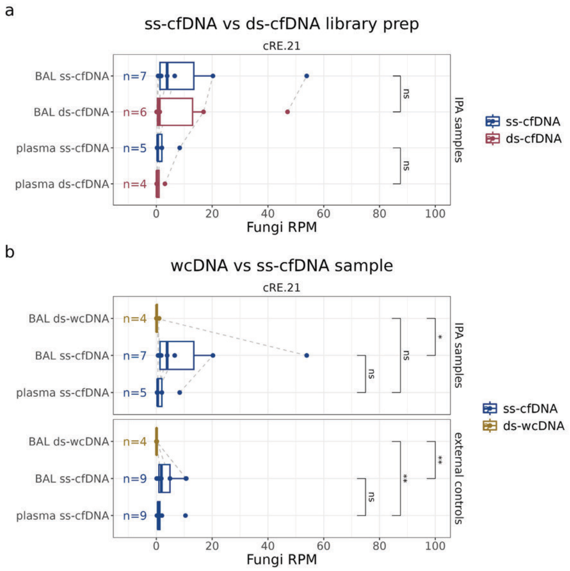


Fig. 4. ss-ligation of cfDNA most effective in retrieving fungal DNA

a-b. Boxplots showing the fungal (kingdom) fractional abundance in RPM, determined by cRE.21-mediated taxonomic classification (CT=0.8). This analysis was conducted, where possible, for both IPA patients and external control samples. **a.** Comparison fungal fractional abundance in sequencing libraries, emphasizing the impact of ss- versus ds-ligation based library preparation. **b.** Comparison of pellet wcDNA to supernatant cfDNA sequencing. **a-b.** Statistical significance is evaluated through a one-tailed Wilcoxon rank test with Bonferroni correction (*, $p \leq 0.05$; **, $p \leq 0.01$; ns, $p > 0.05$). Dotted lines connect sequenced libraries derived from samples collected from the same patient.

Optimizing sample workup for fungal NGS

Sample preparation can impact the sensitivity of shotgun microbial NGS for *Aspergillus* detection. We isolated cfDNA and wcDNA from BAL fluid and cfDNA from plasma, constructing sequencing libraries using either single-stranded (ss) or double-stranded (ds) ligation methods (for schematic overview see Fig. 1; for details see *Methods*). Illumina sequencing was performed on 71 libraries (see Suppl. Table 2). With the optimized computational workflow, elevated fungal counts are interpreted as indicative of improved fungal DNA retrieval rather than false positive observations.

Previous work, focusing on the cfDNA in plasma, demonstrated that ss-ligation produced a higher yield of short (<100 bp) microbial cfDNA compared to dsDNA ligation[19]. We assessed if the same holds true for fungal and *Aspergillus* cfDNA in our sample set of IPA patient- and external control samples, while making use of the cRE.21 database and a CT of 0.8. A noticeable trend in all samples suggests that ss-ligation generally resulted in elevated fungal (Fig. 4a) and *Aspergillus* (Supplementary Fig. 7a) relative abundance in both plasma and BAL samples (n.s., Wilcoxon rank-sum test; $p > 0.05$). We observed *Aspergillus* reads in almost all of these liquid biopsy samples (Supplementary Fig. 7c). In addition, ss-ligation resulted in a narrower library-size range compared to ds-ligation (Supplementary Fig. 8), circumventing DNA yield-reducing bead-based size selection (i.e. elimination of DNA molecules >700 bp; see *Methods* and Suppl. Table 2). Together, these results confirm that ss-cfDNA NGS is more effective than ds-DNA NGS for the recovery of fungal DNA from liquid biopsy samples[19].

Further analysis of ss-cfDNA workup revealed a significantly higher fungal (Fig. 4b) and *Aspergillus* (Supplementary Fig. 7b) abundance in the cfDNA when compared to wcDNA BAL sequencing ($p \leq 0.05$, one-tailed Wilcoxon's rank with Bonferroni correction). This difference could not be attributed to a difference in sequencing-depth as there is no correlation between the total read count and the relative number of fungal reads in our external control samples (Supplementary Fig. 7e; $p > 0.05$, Pearson correlation). These observations thus confirm earlier reports[18,28] that cfDNA contains a relatively higher fraction of fungal DNA molecules than wcDNA. Moreover, our study reveals that fungal counts in IPA patient BAL fluid ss-cfDNA samples are, on average, 5.4x higher than the relative fungal abundance in IPA patient plasma samples (n.s., one-tailed Wilcoxon's rank with Bonferroni correction). This may be attributed to the direct sampling of BAL fluid at the presumed site of the *Aspergillus* infection. Taken together, the ss-cfDNA library preparation method results in higher fungal DNA relative abundance from both BAL and plasma samples, thereby exhibiting slightly higher abundances in BAL fluid.

Optimizing confidence thresholding: analyzing theoretical minimum for *Aspergillus* detection

After application of the cfSPI pipeline some *Aspergillus* counts were above zero in control samples (see Supplementary Fig. 7b,d for *Aspergillus* prevalence and Supplementary Fig. 7b for relative abundance in external control samples). These *Aspergillus* background levels must thus be considered when conducting diagnostic testing. Overall, background levels in control samples were higher at the genus (dREM.260-mediated) than at the species level (cRE.21-mediated), and higher in plasma samples compared to BAL samples (mean 1.5x and 1.2x using the cRE.21 and dREM.260, respectively) (Fig. 5a,c). Furthermore, these background levels were influenced by classification confidence thresholding (Fig. 5a,c). Acknowledging that both computational choices thus directly affect our ability to detect elevated *Aspergillus* DNA levels in patients suspected of a pulmonary fungal infection, we developed a methodology to explore the relationship between fungal background levels, classification rates in simulated datasets, and theoretical Limits of Significant Detection (LoSD).

In this LoSD analysis, we computed the theoretical minimum number of molecules per million (MPM) necessary for the detection of significantly elevated *Aspergillus* taxa above the control background noise, i.e. the background levels observed in immunocompromised pediatric patients without suspicion of a fungal infection (see *Methods* for details). Based on the previously observed species level classification rates (Fig. 3), our hypothesis was that the cRE.21 database would outperform the dREM.260 for detecting *Aspergillus* in clinical samples. And indeed, our LoSD analysis showed that species detection with the dREM.260 necessitated a substantially higher cfDNA load than species level detection with the cRE.21 (Fig. 5b; for details see Supplementary Fig. 9a).

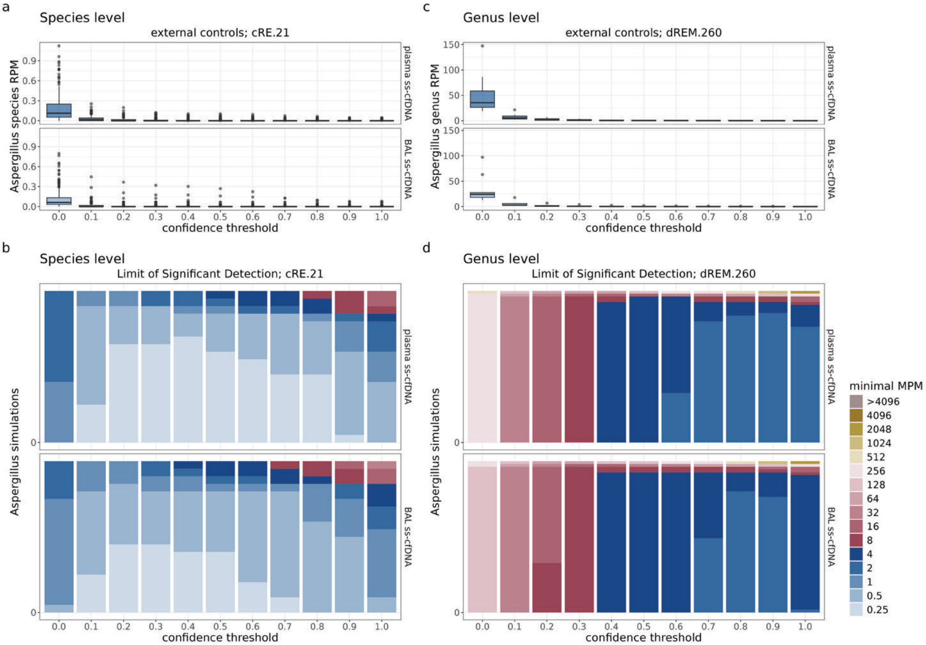


Fig. 5. Computational analysis theoretical minimum fraction for enhanced *Aspergillus* detection to optimize database and parameter selection

The fractional **a.** species level (cRE.21-mediated classification) and **c.** genus level (dREM.260-mediated classification) abundance of *Aspergillus* (in RPM) delineated for plasma (n=9) and for BAL fluid (n=11) external control samples. Leveraging the background levels from our external control samples and classification rates derived from simulations (not shown), we calculated the theoretical minimum fraction of *Aspergillus* molecules (in MPM) necessary for the detection of significantly elevated *Aspergillus* levels above the control background noise, as visualized in **b.,d.** The minimum of *Aspergillus* MPM was computed both at the **b.** species level (cRE.21-mediated) and **d.** genus level (dREM.260-mediated), at a theoretical sequencing depth of 70M reads/sample (for details, see *Methods*). The subsequent CT parameter optimization was based on the highest fraction of observation at <4 molecules per million.

Recognizing the interplay between database and CT choices in our LoSD analysis, we established the most optimal CT for each database. Species level cRE.21-mediated detection was most sensitive when employing a kraken2 CT of 0.4, while genus level dREM.260-mediated detection required a CT of 0.9 (Fig. 5b,d). These findings are based on plasma, where fungal load is generally lower, making optimal sensitivity critical for reliable detection. Employing the cRE.21 for species level detection (CT=0.4), we could detect *Aspergillus* species up to 4 MPM across all 20 simulated *Aspergillus* datasets of species included in the cRE.21 (Fig. 5b). When utilizing the dREM.260

(CT=0.9) for genus level detection, detection of ≤4 MPM was achieved in 93% of the plasma and BAL ss-cfDNA simulations (n=52), respectively (Fig. 5d). Surprisingly, the LoSD appeared relatively stable across different library sizes (Supplementary Fig. 10). Nevertheless, our findings discourage sequencing fewer than 40 million reads due to the adverse impact (i.e. substantial increase in the minimal required MPM) on the theoretical LoSD (Supplementary Fig. 10).

While the cRE.21 is thus established as the most sensitive for species level identification, we emphasize that the impact of database selection on the minimum required MPM can vary among different *Aspergillus* species. For example, the theoretical minimal *A. oryzae* MPM was 2 for cRE.21 and dREM.260, while for *A. niger* our analysis indicated a minimum of 0.25-0.5 MPM required with cRE.21 compared to 2 MPM with dREM.260 (Supplementary Fig. 9b-c). Exceptions notwithstanding, we confirmed our prior hypothesis that detection of *Aspergillus* species should be performed using cRE.21 supplemented by cREM.260 for genus level detection if species-specific results are negative.

Diagnostic performance assessment: a proof-of-principle with seven IPA patients

As a proof-of-principle, we applied ss-cfDNA cfSPI to seven IPA cases which met the inclusion criteria (see *Methods*; Supplementary Fig 11; Table 1) and were classified as probable according to the EORTC/MSG criteria [9] (A01-A05, A14-A15). Patient A03 had been classified as a probable IPA due to host factors (Table 1), repetitive high positive serum galactomannan, and suspected lesions on imaging (i.e. HRCT), but the initial BAL procedure showed negative results during diagnostic work-up. Subsequent repeat BAL procedure confirmed aspergillosis through positive BAL galactomannan at a later time point (this subsequent BAL sample was not included in our study). In our retrospective study, we subjected the initial BAL sample to ss-cfDNA NGS analysis. In total, we sequenced 7 BAL fluid samples (for A01-A05, A14-A15), paired with corresponding plasma (for A01-A05) and internal control plasma samples (for A01-A05) where possible.

Comparing IPA liquid biopsy samples (plasma and BAL of A01-A05 and A14-A15, obtained at diagnosis) to 18 external control liquid biopsy samples from immunocompromised pediatric patients without suspected IPA infection (9 BAL; 9 plasma), *A. fumigatus* was significantly elevated in 5/7 of IPA BAL samples and 3/5 IPA plasma samples (Fig. 6a; mean pairwise Fisher's exact test, $p \leq 0.001$; for details see *Methods*). Importantly, none of the 18 external control samples tested positive with either of the databases (Supplementary 13b,d), indicating high specificity for ss-cfDNA NGS in immunocompromised pediatric patients. Similar results were observed when comparing IPA plasma to the internal control plasma as shown in Supplementary Fig. 12a (pairwise Fisher's exact test, $p \leq 0.001$). The positive samples exhibited fractional abundance of *A. fumigatus* ranging between 1.26 and 40.90 RPM in BAL and 0.31 and 7.71 RPM in plasma (Fig. 6b). In total, 6/7 patients (all except patient A05) had a positive result in at least one liquid biopsy using cfSPI ss-cfDNA NGS. Notably, cfSPI did not detect *A. fumigatus* in the BAL of patient A05, aligning with negative GM test results at the time of collection. Furthermore, cfSPI was the only molecular test that could detect *Aspergillus* in patient A03 compared to standard fungal molecular diagnostics (Table 1), confirming the potential added value of our workflow.

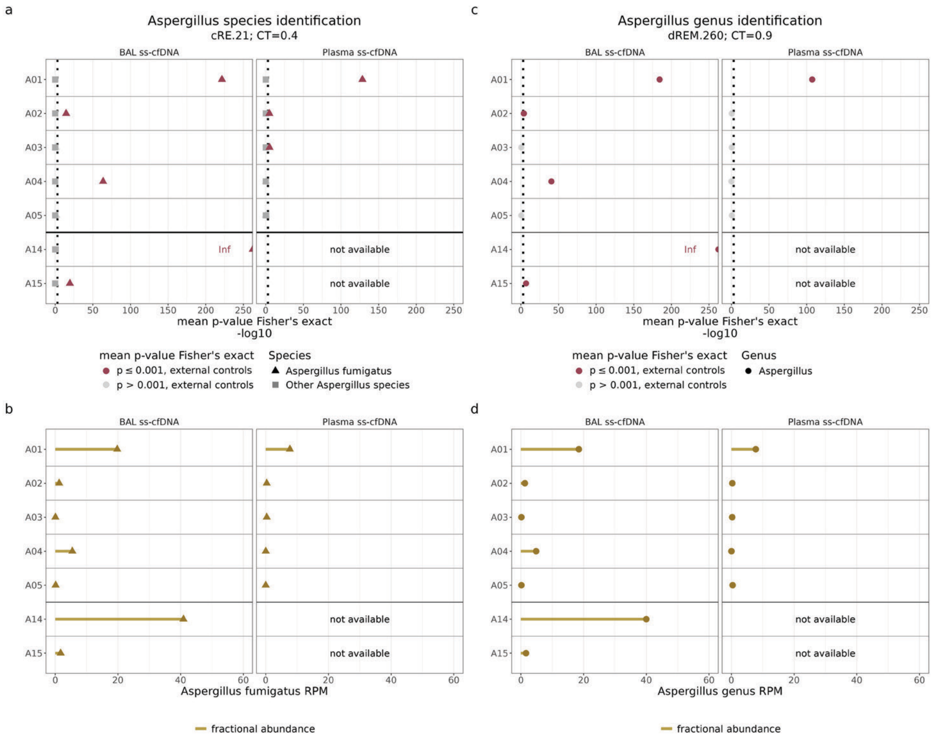


Fig. 6. Elevated *Aspergillus* levels in a subset of IPA patient samples processed via ss-cfDNA NGS

To compare the fractional abundance of *Aspergillus* taxon in patient samples with invasive pulmonary aspergillosis (IPA) to external control pediatric cancer patient samples, the one-tailed Fisher's exact test was utilized. **a.,d.** This analysis was performed both at the species level, using the cRE.21 database (CT=0.4), and at the genus level, using the dREM.260 (CT=0.9) (see *Methods* for details). Dot plots display the mean -log10-transformed computed p-values, with the significance threshold set at p = 0.001 indicated by a vertical dotted line. Instances exceeding the significance threshold are highlighted in dark red. Lollipop plot displaying the fractional abundance of **b.** *A. fumigatus,* determined using the cRE.21 (CT=0.4) and **d.** the fractional abundance of *Aspergillus* at the genus level, determined using the dREM.260 (CT=0.9), in BAL and plasma IPA samples. Abbreviations: Inf, infinite value; not available, samples not available.

The LoSD experiments demonstrated that if cRE.21-mediated species level diagnostics yield negative results, then the simultaneously conducted dREM.260-mediated genus level detection results should be interpreted. Among our IPA samples, the majority (8 out of 12) tested positive for *A. fumigatus* when using cRE.21, which led to no further interpretation of their dREM.260 results. The remaining 4 samples-(BAL samples from patients A03 and A05, as well as plasma samples from patients A04 and A05), also tested negative with the dREM.260 database (Fig. 6c-d; and Supplementary Fig. 12b).

Notably, the internal control sample of patient A05 — the only patient in whom we could not detect *Aspergillus* with cfSPI in samples collected at diagnosis (Fig. 6; Supplementary Fig. 12) — showed elevated *A. fumigatus* cfDNA-levels 24 days before diagnostic work-up (Supplementary Fig. 13a; see Supplementary Fig. 11 for timeline). To gain insight in the time course of *Aspergillus* ss-cfDNA levels within this patient, we subjected two additional plasma samples (plasma A05-2 and A05-3; collected respectively 31 and 38 days prior to diagnosis) for sequencing. Both samples showed no elevated *Aspergillus* cfDNA levels (Supplementary Fig. 13a,c; employing the cRE.21 and dREM.260, resp.), suggesting a potential false positive *Aspergillus* observation in patient A05 obtained by cfSPI 24 days before diagnosis. However, the positive internal control sample for patient A05 could also represent the detection of a true biological signal detected by cfSPI, even though the GM test was negative. In conclusion, elevated cfDNA levels detected by cfSPI in the absence of clinical or radiological symptoms, as observed in patient A05, should be interpreted with considerable caution at this time.

As a quality control, we systematically aligned the reads classified as *A. fumigatus* from each sample to all *Aspergillus* genomes incorporated in constructing the cRE.21 database. Across positive cases, a median of 97.9% and 97.6% of *A. fumigatus* classified reads aligned with a high mapping quality to *A. fumigatus* strains A1163 (=CBS121325) and Af293 (=CBS126847), respectively (Supplementary Fig. 14). Mapping to the closely related species *A. fischeri* and *A. novofumigatus*, on the other hand, only resulted in 54.0% and 36.2% mapped reads and low mapping quality (Supplementary Fig. 14). Minimal alignment was observed for all other *Aspergillus* species (Supplementary Fig. 14a). Reassuringly, a uniform distribution of cRE.21 classified reads was observed when mapping to the consensus genome *A. fumigatus* Af293 (Supplementary Fig. 15). Taken together, these findings confirm the presence of true-positive *A. fumigatus* classified cfDNA reads.

Discussion

Our objective was to enhance the standardization of liquid-biopsy shotgun microbial NGS and to translate this knowledge into the cfSPI open-source analysis workflow that can be readily accessed and utilized by the scientific community. Our analysis involved both real-world (75 sequence libraries, generated from 55 different liquid biopsy samples; Suppl. Table 2) and simulated Illumina sequencing data (of 87 simulated datasets). We demonstrated that dual mapping to the human CHM13v2.0 and GRCh38.p14 references, along with the incorporation of both genomes into the classification database, is essential to prevent incorrect classification of human reads as microbial reads, potentially reducing the chances of false positive diagnoses. Additionally, we demonstrated that the NCBI RefSeq kraken2 standard database is unable to detect most *Aspergillus* species, thus highlighting the necessity of database extension, particularly accomplished by incorporation of (pathogenic) fungal species and closely related taxa to enable species level *Aspergillus* detection. From a wet lab perspective, our results indicate that ss-cfDNA sequencing is a superior method for detecting fungal cfDNA compared to wcDNA sequencing. This is supported by the fact that our wcDNA isolation protocol yielded a maximum fungal count of only 2.55 RPM while the cfDNA yielded up to 12.55 RPM. The insights gained from our comparative research are poised to enhance the optimization of microbial NGS procedures and contribute to exploration of more precise and reliable liquid biopsy-based diagnostics for IMD.

In our retrospective proof-of-principle, we demonstrated that cfSPI enables the detection of elevated levels of *A. fumigatus* in 71% (5/7) of BAL samples and 60% (3/5) of plasma samples compared to controls. This suggests that ss-cfDNA sequencing of BAL fluid supernatant — a relatively understudied liquid biopsy type in pediatrics — may offer diagnostic potential, given that BAL fluid fungal fractions typically exceed those in patient plasma samples. This finding aligns with previous research on detecting bacterial and viral pathogens in local body fluids[29,30], as well as prior studies on *Aspergillus* detection in BAL samples from adults[31]. Nevertheless, due to the minimally invasive nature of plasma sampling, there remains a strong interest in exploring the diagnostic potential of plasma microbial NGS for IMD. The observed sensitivity of plasma sequencing (60%) is consistent with previous sequencing findings in adult cohorts, where sensitivity ranging between 38.5%[16] and 61%[15], suggesting that BAL testing might be unnecessary in up to 60% of IPA cases. Although this is a small proof-of-principle cohort, sequencing of BAL samples resulted in a better sensitivity than the standard diagnostics in BAL[10,11]. Importantly, we noted minimal false positive rates, indicating that cfSPI offers compelling evidence for the presence of *Aspergillus*.

While the exclusive detection of *A. fumigatus* in our IPA patient set is in line with findings by Hill *et al.*[15], the size and composition of our IPA patient set limits our ability to evaluate the effectiveness of cfSPI ss-cfDNA NGS in identifying other *Aspergillus* species. Nonetheless, our simulations suggest the capability of detecting diverse *Aspergillus* species and distinguishing clinically relevant ones such as *A. fumigatus* and *A. terreus*. The limited availability of retrospective samples also restricts our ability to definitively demonstrate (or rule out) potential additional benefits of dREM.260 as a supplementary test for pan-*Aspergillus* detection at the genus level. Therefore, our hypothesis that in cases where the pathogen in a patient is not covered by the compact curated cRE.21 database, resulting in negative species-specific findings, the supplementary dREM.260 results could be crucial for comprehensive pan-*Aspergillus* detection, requires further investigation and validation. Moreover, a comparison between our optimized ss-cfDNA NGS workflow for *Aspergillus* detection and conventional diagnostics remains impossible due to the highly limited size of our proof-of-principle and exclusive application of ss-cfDNA to probable IPA cases, thus precluding insight into its diagnostic performance in fungal infections.

To advance this technical research towards potential clinical implementation, our workflow still requires an external validation cohort comprising pediatric patients suspected of IMD, including those classified as possible, probable, and proven cases. A critical step in demonstrating clinical applicability is testing whether our cfSPI workflow maintains its sensitivity and specificity outside of a retrospective, controlled research environment. By making our workflow open-source and designed for in-house use, we aim to facilitate external validation efforts while minimizing turnaround times by eliminating the need for sample shipping. This approach also supports the development of local expertise and ensures adaptability to specific populations through the seamless integration of customized classification databases. These features provide a significant advantage over commercial tests like the Karius Test, which notably shares substantial overlap with the genomes integrated in our cRE.21 database (see Supplementary Fig. 16).

Collectively, our study provides valuable insights in the use of ss-cfDNA NGS for *Aspergillus* detection in pediatric immunocompromised hosts. The cfSPI framework introduced here has the potential to expedite future fungal-NGS investigations through its open-source analysis workflow. Our findings, complemented with new simulations and LoSD analysis and coupled to benchmarking using short-read sequencing data from fungal culture isolates, can solidify the groundwork for enhancing these open-source detection tests, not only for *Aspergillus* but also for other IMD.

Methods

Aim of this study

The aim of this study is optimization of the microbial NGS workflow tailored to *Aspergillus* detection in liquid biopsy samples. This involves shotgun sequencing of liquid biopsy samples, such as blood or BAL fluid, with the objective of sequencing intact microbial whole-cell DNA wcDNA or small DNA molecules from degraded microbes (i.e. cell-free DNA, cfDNA). Following microbial NGS, taxonomic identification of the DNA source is performed by tracing the origin of sequenced DNA molecules. This process includes human read subtraction, taxonomic classification, and subsequent statistical analysis to determine if there is supporting evidence for a pathogenic microbe. In this study, we refined both the wet-lab and computational workflow (cfSPI) by optimizing six key steps (Q#1-6) in microbial NGS-based fungal diagnostics, as detailed in Fig. 1.

Diagnostic work-up of IMD suspected patients includes clinical microbiology

At the Princess Máxima Center for Pediatric Oncology (PMC) in Utrecht, The Netherlands, patients suspected of IMD are evaluated by a chest high-resolution computerized tomography (HRCT) scan. If suspected lesions for IMD are seen on HRCT, a BAL is performed. The BAL fluid provides material for microscopy, fungal culture, galactomannan (GM) assay (Platelia Aspergillus Ag, Bio-Rad), and molecular diagnostics. A BAL GM >1.0 is considered to be positive according to the criteria of the European Organization for Research and Treatment of Cancer and Mycoses Study Group (EORTC/MSG) criteria[9].

Retrospective proof-of-principle: inclusion of IPA patient samples

We included plasma and BAL fluid from immunocompromised pediatric patients who were diagnosed with an IPA at the PMC between 2020 and 2023. We searched manually for cases with probable or proven IPA according to the EORTC/MSG criteria[9] (Table. 1).

The date of diagnosis was based on the date of BAL retrieval with a positive microbiological finding. One BAL fluid sample and one blood plasma sample (time-matching the BAL fluid sample or at least one or two days around the date of diagnosis) were used from each case, as well as a plasma sample collected approximately fourteen days prior to the date of diagnosis (Supplementary Fig. 11). These plasma samples collected earlier in time served as internal control samples. We also included BAL fluid and plasma samples from pediatric immunocompromised patients without IPA, so called external control samples. BAL fluid control samples were used from patients pre-HSCT, during an anesthetic procedure for line insertion[32].

All materials used for this study were clinical samples routinely stored at -70°C. Before freezing, the plasma samples were prepared by centrifuging EDTA plasma to remove the cells. BAL fluids were stored directly after use without prior centrifugation. Our study sets a high standard by incorporating both internal and external control samples, surpassing other studies that solely reported sequencing results from suspected IPA cases.

In total, this study encompassed 25 pediatric patients, including 7 diagnosed with probable IPA and 18 external immunocompromised controls (9 plasma samples and 9 BAL fluids). All included patients provided written informed consent for participation in the biobank for storage and use of their rest materials (International Clinical Trials Registry Platform: NL7744; https://onderzoekmetmensen.nl/en/trial/21619). For use of samples and data in this study we refer to local Biobank and Data Access Committee approval. All patients have also provided informed consent for sequencing and use of these data for publication.

Diagnostic sensitivity and specificity

One of the aims of this pilot study was to examine the diagnostic sensitivity of microbial NGS-based detection of *Aspergillus* DNA in immunocompromised pediatric patients. The diagnostic sensitivity was gauged by the proportion of probable IPA patients displaying significantly heightened levels of at least one *Aspergillus* species, or at the *Aspergillus* genus level, in at least one liquid biopsy sample (blood or BAL fluid) in comparison to external controls.

The specificity of microbial NGS was determined by the proportion of external control patient samples that displayed significantly elevated levels of at least one *Aspergillus* species or at the *Aspergillus* genus level, in at least one liquid biopsy sample (blood or BAL fluid) in comparison to external controls.

Sample preparation and DNA isolation

Plasma was obtained from EDTA blood samples by centrifugation, 10 min at 1,500 g, followed by an additional centrifugation step for 5 min at 12,000 g, to remove all cells (both centrifugation steps at room temperature). Plasma was subsequently stored at -80°C. Fresh BAL fluid samples were stored at -80°C until further processing. To separate the (whole) cells from the BAL fluid the samples were centrifuged for 5 min at 13,000 rpm at 4°C. Subsequently, the pellet and supernatant were processed in separate DNA isolation methods for wcDNA sequencing and cfDNA sequencing, respectively.

cfDNA nucleic acids were isolated from BAL supernatant, EDTA plasma and sterile Dulbecco’s Phosphate Buffered Saline (DPBS), using the Circulating Nucleic Acid Kit (Qiagen, 55114) with the following modifications to the manufacturer’s protocol. A select set of our BAL fluid and plasma samples were complemented with DPBS, up to a total volume of 2 mL. Furthermore, the lysis time was increased from 30 to 60 minutes and the final elution of cfDNA was done with Nuclease Free water (Invitrogen, 10977-035).

DNA from 5 (out of 13) BAL pellets was extracted from whole cells by mechanical bead beating (see Suppl. Table 2, "internal" DNA isolation) where 50 µL DPBS and a 5 mm stainless steel-bead (Qiagen, 69989) were added to the BAL pellet. Samples were beaten for 2 min (with a frequency of 2 5 1/s) on the TissueLyser II (Qiagen) and subsequently diluted in ATL buffer (Qiagen, 939011) and transferred to a fresh tube. Additional ATL buffer was added to a total volume of 300 µL, with an overnight incubation at 56°C after addition of Proteinase K (Qiagen, 19131). wcDNA isolation was performed using the DNeasy Blood & Tissue Kit (Qiagen, 69506), following manufacturers protocol adjusting subsequent volumes to accommodate the larger lysis volume.

From the remaining 8 of our 13 BAL pellet samples (see Suppl. Table 2, *external* DNA isolation) wcDNA was isolated according to a standard protocol for fungal DNA isolation prior to RT PCR (UMC Utrecht, Dept. Medical Microbiology, The Netherlands). In short, BAL pellets were bead beaten and snap lysed (by freezing them at -80°C and heating them to 96°C). After addition of a lysis buffer, nucleic acids were isolated using the MagNA Pure system (Roche).

All DNA samples were quantified using the Qubit dsDNA High Sensitivity assay kit or Broad Range assay kit (Thermofisher Scientific, Q32854 and Q32853, respectively). The DNA fragment length distribution and concentration of the cfDNA was evaluated using the TapeStation 2200, D1000HS kit (Agilent, 5067-5585).

Preparation of next-generation sequencing libraries

For ss-ligation based DNA-capture the SRSLY PicoPlus NGS Library Prep Kit for Illumina (Claret Bioscience, CBS-K250B-96) was used, according to the manufacturer's Moderate Fragment Retention version of the protocol, using max 5ng cfDNA as input and 11 PCR cycles. For ds-ligation-based cf/wcDNA-capture the KAPA Library Preparation Kit (Roche) was used, with a maximum of 50ng of input DNA and 8-12 PCR cycles. For ds-capture of the BAL pellet DNA, the manufacturer's protocol was followed. For ds-capture of cfDNA the following changes to the manufacturer's protocol have been made: no fragmentation step and following similar bead clean up steps as the moderate fragment retention protocol to improve the yield of small fragments. All library preparations were quantified using the Qubit dsDNA High Sensitivity Assay Kit (Thermofisher Scientific, 32854) and size distribution was analyzed using the TapeStation 2200, either the D1000 and/or D5000 kits (Agilent, 5067-5583/5067-5589). Some samples showed an aberrant size distribution (substantial fraction of >700 bp fragments); these samples were subjected to a bead-based size selection protocol (see Suppl. Table 2), to remove long fragments as Illumina sequencing is optimized for fragments <700 bp. Concentration (see Suppl. Table 2, total DNA yield in ng) and size were re-evaluated after size selection. Samples were pooled equimolarly and submitted for sequencing.

Next-generation sequencing

Sequencing of all 75 Illumina libraries was conducted on the NovaSeq 6000, 2x 150 bp reads, resulting in 42 to 218 million reads per ss-cfDNA sample, 23 to 126 million reads per ds-cfDNA patient sample, and 34 to 62 million reads per wcDNA patient sample.

Read simulations

In order to evaluate the impact of kraken2's reference hash-table database composition and confidence threshold on the *Aspergillus* classification accuracy, we simulated cfDNA-like high-throughput datasets from 87 complete, scaffold, or draft genomes derived from the NCBI RefSeq, among which 55 *Aspergillus*, 7 *Penicillium*, and 25 other pathogenic fungal genomes with a Illumina read simulator tool named ReSeq[33]. To create a realistic error profile of sequencing reads, the unprocessed ss-cfDNA sequencing reads from plasma of patient A01 (i.e. A01Pasp, for details on sample workup see Suppl. Table 2) sequenced with Illumina Novaseq 6000 2x 150 bp were mapped to the human GRCh38.p14 reference genome using Bowtie2 (run with option "-X 2000") alignment software[34]. Subsequently, the reference sequence statistics were determined using ReSeq (option "--statsOnly"), and 151 bp paired-end reads were simulated *in-silico* using the illuminaPE ReSeq command (option "--noBias")[33]. For each genome we simulated between 99,050 and 101,023 reads; for details on simulated datasets and the number of reads per dataset see Suppl. Table 1.

Database construction

For the purpose of fungal nucleic acid detection, we utilized 9 kraken2 classification databases, namely uR.7, uR.7 w/o CHM13v2, cRE.21, dRE.21, uRE.21, uRE.31, dRE.31, dREM.258 and dREM.260 of which details on the construction and composition are reported in Suppl. Table 5.

Prior to hash-table construction, specific genomic regions of the reference genomes were masked to prevent spurious misclassifications. This masking procedure involved *dustmasking*, where low-complexity regions were masked using the DUST algorithm[21] as advised by the developers of kraken2 , more rigorous *decontamination* efforts thereby masking contaminant organism sequences, or a combination of dustmasking and decontamination (i.e. full *cleanup*), such as in the work of Lu and Salzberg[35]. Contaminant organisms' sequences are sequences within genome assemblies that do not accurately represent the organism's genetic information.

Database names indicate the masking procedure used (**u**naltered, **d**ustmasked or **c**leaned), the database sources (**R**efSeq, **E**uPathDB and/or **M**ycoCosm) and the number of *Aspergillus* species included (Supplementary Fig. 2). For example, **cRE.21** refers to the **c**leaned version of a combination of **R**efSeq and **E**uPathDB with a total of **21** *Aspergillus* species. Each database was built using the NCBI taxonomic information, which was downloaded on 05-05-2023.

The NCBI RefSeq[22] is a comprehensive and curated collection of nucleotide sequences, encompassing a wide range of species, which — in the context of kraken2 classification — are often used as a standard reference hash-table database construction. Using the "kraken2-build --download-taxonomy" command, genomes of the NCBI RefSeq were downloaded on 15-05-2023 (RefSeq release 218, file creation date 05-05-2023), including 1,495 archaea, 285,827 bacteria, 498 fungi, 98 protozoa, 14,979 viral and 1 human sequence plus 3,137 contigs which were part of the UniVec_Core. UniVec_Core comprises oligonucleotide and vector sequences sourced from bacteria, phage, yeast, and synthetic constructs, excluding vector sequences of mammalian origin.

The EuPathDB encompasses a curated set of genomic sequences of 386 pathogenic fungi, protists, oomycetes as well as evolutionarily related non-pathogenic species[26]. EuPathDB genomic sequences were downloaded on 16-05-2023 (file creation date was 28-10-2020), consisting of the following subsets: AmoebaDB (n=30), CryptoDB (n=18), FungiDB (n=164), GiardiaDB (n=10), MicrosporidiaDB (n=35), PiroplasmaDB (n=10), PlasmoDB (n=45), ToxoDB (n=33), TrichDB (n=1) and TriTrypDB (n=42). Upon inspection post-download, we observed that two fungal genomes were omitted from the purified edition of FungiDB-46. Jennifer Lu, one of the authors of the contaminant removal paper[35], graciously supplied us with the latest version of these two genomes:

FungiDB-54_PgraminisCRL75-36-700-3_Genome_cleaned_v_final.fna
FungiDB-54_Ptriticina1-1BBBDRace1_Genome_cleaned_v_final.fna.

An updated version of the *seqid2taxid.map* used for database construction was also provided. In total, EuPathDB version 46 includes 27 *Aspergillus* genomes, representing 21 *Aspergillus* species, while version 64 contains 38 genomes, representing 31 *Aspergillus* species.

The MycoCosm[27] is a web-based resource and information portal developed by the Joint Genome Institute (JGI) for fungal genomics. It provides access to a comprehensive collection of fungal genomes, associated functional annotations, and tools for comparative analysis. We acquired all 763 genomic sequences of *Aspergillus* assembly scaffolds from this resource, along with their corresponding taxonomic annotations.

Host read subtraction

Our objective was to eliminate potential false positives microbial reads originating from incomplete host read subtraction. To achieve this, we employed mapping to either the GRCh38.p14, CHM13v2, or a dual-mapping utilizing a consolidated reference index for Bowtie2 alignment, encompassing both GRCh38.p14 and CHM13v2, thereby avoiding a two-step mapping process.

Sequencing and synthetic data processing: the cfSPI-pipeline

Illumina sequencing and synthetic data were processed using the Snakemake[36] cfSPI-pipeline available in our Github repository (https://github.com/AEWesdorp/cfSPI/cfspi/). In short, duplicates were removed (using nubeam[37]), after which high-quality sequencing data was generated (using fastp[38]) by default removal of low quality reads and usage of a low complexity filter as well as by adapter removal and removal of short (<35 bp) reads (using AdapterRemoval[39]). After subtraction of host sequences by mapping to the human reference genome using Bowtie2[34], the remaining paired-end reads were taxonomically classified using kraken2[20] with the 9 kraken2 databases specified above, employing a confidence threshold (CT) ranging from 0.0 (no filter on the fraction of *k-mers* matching, default setting) to 1.0 (100% of *k*-mers within the read match a taxa, very stringent setting), increments of 0.1.

Remapping of classified reads

Reads classified as *Aspergillus* at the species level or below, with the cRE.21 database, were aligned to the respective *Aspergillus* species genomic sequences used for the cRE.21 database construction, through Bowtie2.

Relative abundance per taxon

The relative abundance per taxon within each sample was quantified as Reads Per Million (RPM), a normalized measure that accounts for i.e. differences in sequencing depth. RPM was calculated according to the following formula:

$$\mathit{RPM} = \mathit{totalSumTaxonReads} / \mathit{totalNumQCReads} * 1{,}000{,}000$$

totalSumTaxonReads corresponds to the number of reads classified at each taxon (e.g. genus- or species level) and all lower-ranking taxons belonging to the same clade. *totalNumQCReads* is the count of reads that passed quality control (i.e. number of reads remained after duplicate removal, low-quality and low-complexity reads filtering, and adapter removal; see Supplementary Fig. 1a).

Quantification and analytical analysis

Quantification and analytical examination were conducted using R version 4.2.0. After the analysis, downstream outcomes and visual representations were generated to enhance the interpretability of the results.

Limits of Significant Detection

In order to explore the relationship between fungal background levels in immunocompromised individuals, the observed classification rate in simulated datasets, and the resulting theoretical 'Limits of Significant Detection', we formulated the following methodology. First, we defined our taxon, database, and CT of interest. Second, we set the *theoretical sequencing depth* (*tSD*), ranging from 10 to 100 million reads (with intervals of 15 million) as well as a theoretical number of *molecules per million* (*mpm*) ranging from 0.25 to 4096. Third, we obtained the median *background abundance* (*ba*) observed in our external immunocompromised samples for the specified taxon, database, and CT of interest (normalized, RPM). Thirdly, we determined the *classification rate* (*cr*) observed in our simulated datasets for the specified taxon, database, and CT of interest. Subsequently, we applied the following calculation to determine the total taxon read count in our artificial sample, rounded to the nearest integer:

$$\mathit{total\ taxon\ count} = \lfloor (\mathit{tSD} * \mathit{ba}) + (\mathit{tSD} * \mathit{mpm} * \mathit{cr}) + 0.5 \rfloor$$

Following this, we applied a one-tailed Fisher's exact test to assess whether the observed total taxon count in our theoretical samples significantly differed from that in our external control samples. A mean $p \leq 0.001$ was deemed statistically significant. The lowest mpm value with a p-value ≤ 0.001 was reported as the MPM.

Identification of elevated *Aspergillus* levels

Following taxonomic classification of all clinical samples, we conducted a one-tailed Fisher's exact test to assess statistical differences in the read count of a specified taxon between our patient samples and internal/external control samples. This test was based on comparing the following two counts in the contingency tables:

1. The number of reads classified at the taxon of interest, including reads at the specified taxonomic level (e.g., genus- or species level) and all lower-ranking taxa within the same clade.
2. The number of reads remaining after duplicate removal, low-quality, and low-complexity read filtering, excluding those classified at the taxon of interest.

The significance level was set at $p \leq 0.001$, calculated by deriving the mean of all Fisher's exact tests conducted across the samples. This analysis aimed to identify meaningful differences in taxon-specific read counts between patient and control groups, considering the overall composition and diversity of microbial taxa in the studied samples.

Statistical analysis

To evaluate the influence of computational choices, including host-read subtraction and database composition, we utilized the one-tailed paired t-test. For comparisons involving sample types and library preparation, we applied the one-tailed Wilcoxon rank-test. To account for multiple testing, we employed Bonferroni correction in these analyses.

Declarations

Data availability

Data will be made available on reasonable request, through the European Genome-phenome Archive (EGA) under accession number EGAS00001008021. Additionally, taxonomic abundance matrixes are provided via the GitHub repository upon publication.

Code availability

Details regarding the cfSPI pipeline can be found here:

https://github.com/AEWesdorp/cfSPI/cfspi/ .

The code, encompassing data simulation scripts and the analysis of data presented in this manuscript, are available here: https://github.com/AEWesdorp/cfSPI/

Acknowledgements

We acknowledge the Utrecht Sequencing Facility (USEQ) for providing sequencing service and data (USEQ is subsidized by the University Medical Center Utrecht and The Netherlands X-omics Initiative; NWO project 184.034.019). The authors also would like to thank C. Beudeker and M. van der Flier for obtaining blood plasma samples through the DIAMONDS study. Furthermore, we thank J. Lu for generously supplying an updated version of the EuPathDB *seqid2taxid.map* and the inclusion of two supplementary genomic FungiDB sequences. The authors would also like to thank A.K.M. Rosendahl Huber for assistance with the preparation of the manuscript as well as W.A.A. Steenhuijsen Piters, A.C. Fluit and J.D.F. Groot-Mijnes for their critical feedback on the setup of the study.

Competing interests

EW is supported by a Vidi Fellowship (639.072.715) to **JdR** from the Dutch Organization for Scientific Research (Nederlandse Organisatie voor Wetenschappelijk Onderzoek, NWO). **FH** has received products/financial compensation from Pathonostics, OLM Diagnostics, Altona Diagnostics, EWC Diagnostics, CHROMagar, and IMMY in the context of product validation (and has published or will publish about the outcome of these studies). **CL** has received financial support from Pfizer for educational purposes. **LB** has regular interaction with pharmaceutical and other industrial partners. **LB** has not received personal fees or other personal benefits, but UMCU has received significant funding (>€100,000 per industrial partner) for investigator-initiated studies from AstraZeneca, Sanofi, Janssen, Pfizer, MSD, and MeMed Diagnostics, as well as major funding from the Bill and Melinda Gates Foundation and through public-private partnerships such as the IMI-funded RESCEU and PROMISE projects, involving partners GSK, Novavax, Janssen, AstraZeneca, Pfizer, and Sanofi, along with substantial funding from Julius Clinical for participation in clinical studies sponsored by AstraZeneca, Merck, and Pfizer, and minor funding (€1,000-25,000 per industrial partner) for consultation, DSMB membership, or invited lectures by Ablynx, Bavaria Nordic, GSK, Novavax, Pfizer, Moderna, AstraZeneca, MSD, Sanofi, and Janssen. **LB** is the founding chairman of the ReSViNET Foundation. **JdR** is cofounder and CTO of Cyclomics, a genomics company. **LR, LC, MJ, NB, CV** and **TvdB** declare no competing interests.

Authors contributions

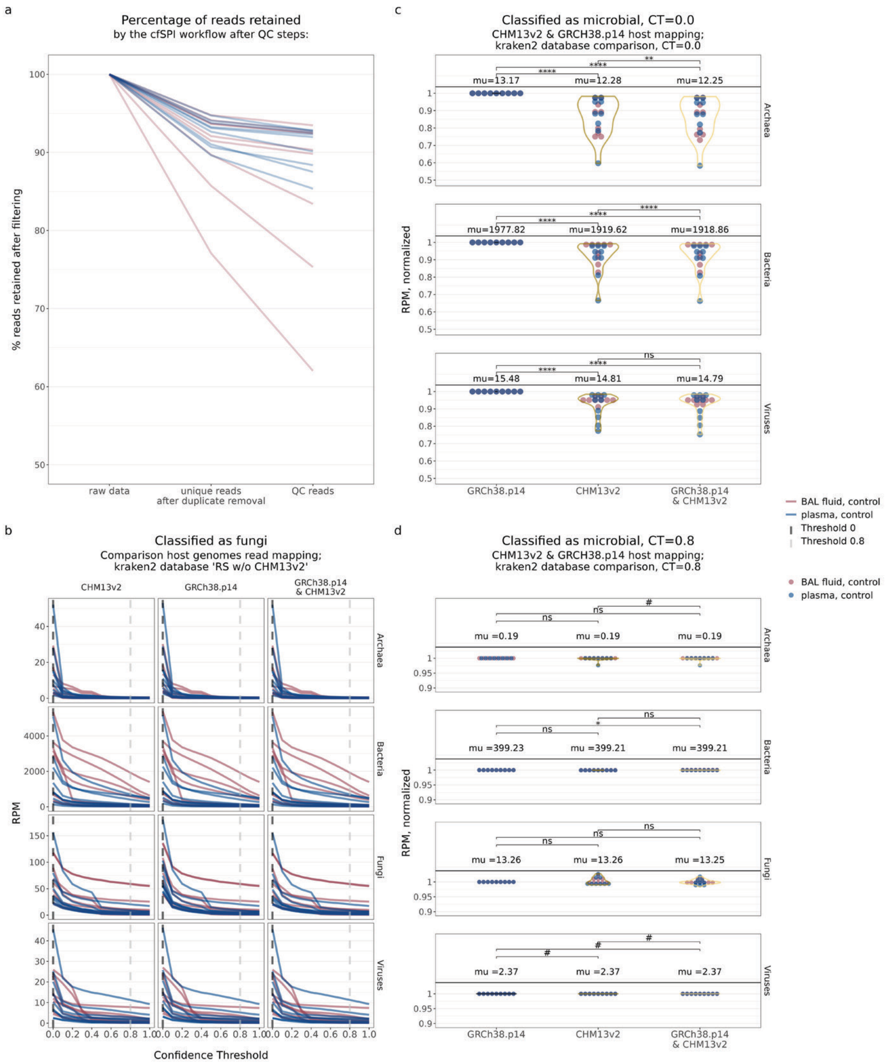
EW, **LR**, **LC** and **MJ** conceived experiments and wrote the article. **LR** searched for and collected samples, managed patients, provided clinical information and sample data and interpreted with **EW** sequencing data to clinical data. **EW**, **LC** and **NB** performed experiments. **EW** and **LC** contributed to constructing the data analysis pipeline and conducting the data analysis. **EW** curated tables and created figures. **MJ**, **CV**, **FH**, **TvdB**, **CL**, **TW**, **LB** and **JdR** designed experiments, contributed to the writing of the article and/or provided (clinical) information.

References

1. Loeffen, E. A. H. *et al.* Treatment-related mortality in children with cancer: Prevalence and risk factors. *Eur. J. Cancer* **121**, 113–122 (2019).
2. Pana, Z. D., Roilides, E., Warris, A., Groll, A. H. & Zaoutis, T. Epidemiology of invasive fungal disease in children. *J. Pediatric Infect. Dis. Soc.* **6**, S3–S11 (2017).
3. Lehrnbecher, T. *et al.* Incidence and outcome of invasive fungal diseases in children with hematological malignancies and/or allogeneic hematopoietic stem cell transplantation: Results of a prospective multicenter study. *Front. Microbiol.* **10**, 681 (2019).
4. Cesaro, S. *et al.* Retrospective study on the incidence and outcome of proven and probable invasive fungal infections in high-risk pediatric onco-hematological patients. *Eur. J. Haematol.* **99**, 240–248 (2017).
5. Bartlett, A. W. *et al.* Epidemiology of invasive fungal infections in immunocompromised children; an Australian national 10-year review. *Pediatr. Blood Cancer* **66**, e27564 (2019).
6. Kazakou, N. *et al.* Invasive fungal infections in a pediatric hematology-oncology department: A 16-year retrospective study. *Curr. Med. Mycol.* **6**, 37–42 (2020).
7. Bury, D. *et al.* Clinical presentation and outcome of invasive mould disease in paediatric patients with acute lymphoblastic leukaemia. *EJC Paediatric Oncology* **3**, 100143 (2024).
8. Groll, A. H. *et al.* 8th European Conference on Infections in Leukaemia: 2020 guidelines for the diagnosis, prevention, and treatment of invasive fungal diseases in paediatric patients with cancer or post-haematopoietic cell transplantation. *Lancet Oncol.* **22**, e254–e269 (2021).
9. Donnelly, J. P. *et al.* Revision and update of the consensus definitions of invasive fungal disease from the European Organization for research and Treatment of cancer and the Mycoses Study Group education and research consortium. *Clin. Infect. Dis.* **71**, 1367–1376 (2020).
10. Reinwald, M. *et al.* Therapy with antifungals decreases the diagnostic performance of PCR for diagnosing invasive aspergillosis in bronchoalveolar lavage samples of patients with haematological malignancies. *J. Antimicrob. Chemother.* **67**, 2260–2267 (2012).
11. de Mol, M. *et al.* Diagnosis of invasive pulmonary aspergillosis in children with bronchoalveolar lavage galactomannan: BAL Galactomannan Aspergillosis Children. *Pediatr. Pulmonol.* **48**, 789–796 (2013).
12. Hong, D. K. *et al.* Liquid biopsy for infectious diseases: sequencing of cell-free plasma to detect pathogen DNA in patients with invasive fungal disease. *Diagn. Microbiol. Infect. Dis.* **92**, 210–213 (2018).
13. Ma, X. *et al.* Invasive pulmonary aspergillosis diagnosis via peripheral blood metagenomic next-generation sequencing. *Front. Med. (Lausanne)* **9**, 751617 (2022).
14. Armstrong, A. E. *et al.* Cell-free DNA next-generation sequencing successfully detects infectious pathogens in pediatric oncology and hematopoietic stem cell transplant patients at risk for invasive fungal disease. *Pediatr. Blood Cancer* **66**, e27734 (2019).
15. Hill, J. A. *et al.* Liquid biopsy for invasive mold infections in hematopoietic cell transplant recipients with pneumonia through next-generation sequencing of microbial cell-free DNA in plasma. *Clin. Infect. Dis.* **73**, e3876–e3883 (2021).
16. Huygens, S. *et al.* Diagnostic value of microbial cell-free DNA sequencing for suspected invasive fungal infections: A retrospective multicenter cohort study. *Open Forum Infect. Dis.* **11**, ofae252 (2024).
17. R Marcelino, V., Holmes, E. C. & Sorrell, T. C. The use of taxon-specific reference databases compromises metagenomic classification. *BMC Genomics* **21**, 184 (2020).
18. He, P. *et al.* Comparison of metagenomic next-generation sequencing using cell-free DNA and whole-cell DNA for the diagnoses of pulmonary infections. *Front. Cell. Infect. Microbiol.* **12**, 1042945 (2022).
19. Burnham, P. *et al.* Single-stranded DNA library preparation uncovers the origin and diversity of ultrashort cell-free DNA in plasma. *Sci. Rep.* **6**, 27859 (2016).
20. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).
21. Morgulis, A., Gertz, E. M., Schäffer, A. A. & Agarwala, R. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. *J. Comput. Biol.* **13**, 1028–1040 (2006).
22. Nasko, D. J., Koren, S., Phillippy, A. M. & Treangen, T. J. RefSeq database growth influences the accuracy of k-mer-based lowest common ancestor species identification. *Genome Biol.* **19**, 165 (2018).
23. Gihawi, A. *et al.* Major data analysis errors invalidate cancer microbiome findings. *MBio* **14**, e0160723 (2023).
24. Wright, R. J., Comeau, A. M. & Langille, M. G. I. From defaults to databases: parameter and database choice dramatically impact the performance of metagenomic taxonomic classification tools. *Microb. Genom.* **9**, (2023).
25. Méric, G., Wick, R. R., Watts, S. C., Holt, K. E. & Inouye, M. Correcting index databases improves metagenomic studies. *bioRxiv* 712166 (2019) doi:10.1101/712166.
26. Warrenfeltz, S. *et al.* EuPathDB: The eukaryotic pathogen genomics database resource. *Methods Mol. Biol.* **1757**, 69–113 (2018).

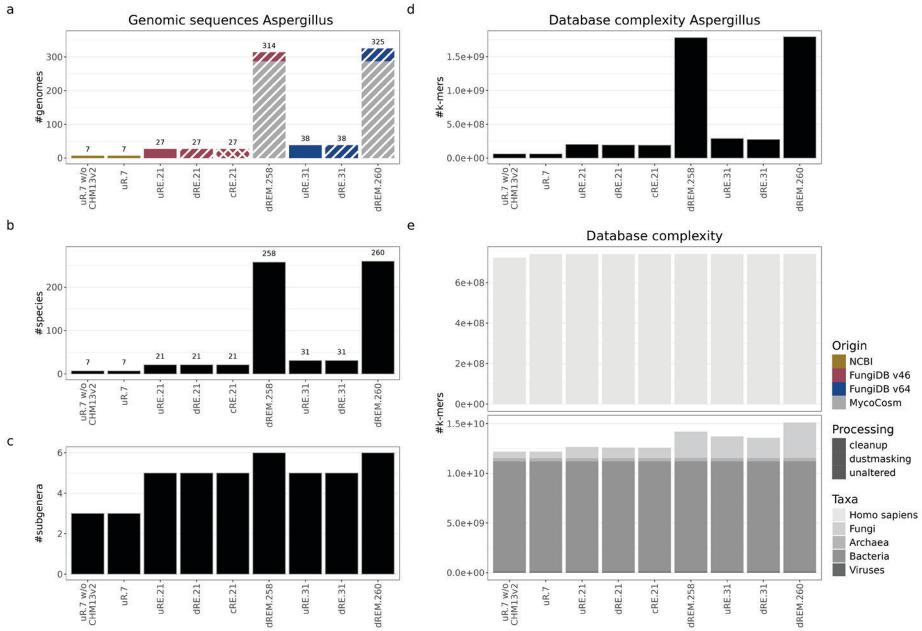
27. Ahrendt, S. R., Mondo, S. J., Haridas, S. & Grigoriev, I. V. MycoCosm, the JGI's fungal genome portal for comparative genomic and multiomics data analyses. *Methods Mol. Biol.* **2605**, 271–291 (2023).
28. Yu, L. *et al.* Metagenomic next-generation sequencing of cell-free and whole-cell DNA in diagnosing central nervous system infections. *Front. Cell. Infect. Microbiol.* **12**, 951703 (2022).
29. Gu, W. *et al.* Detection of cryptogenic malignancies from metagenomic whole genome sequencing of body fluids. *Genome Med.* **13**, 98 (2021).
30. Sun, T. *et al.* A paired comparison of plasma and bronchoalveolar lavage fluid for metagenomic next-generation sequencing in critically ill patients with suspected severe pneumonia. *Infect. Drug Resist.* **15**, 4369–4379 (2022).
31. Chen, S., Kang, Y., Li, D. & Li, Z. Diagnostic performance of metagenomic next-generation sequencing for the detection of pathogens in bronchoalveolar lavage fluid in patients with pulmonary infections: Systematic review and meta-analysis. *Int. J. Infect. Dis.* **122**, 867–873 (2022).
32. Versluys, A. B. *et al.* High diagnostic yield of dedicated pulmonary screening before hematopoietic cell transplantation in children. *Biol. Blood Marrow Transplant.* **21**, 1622–1626 (2015).
33. Schmeing, S. & Robinson, M. D. ReSeq simulates realistic Illumina high-throughput sequencing data. *Genome Biol.* **22**, 67 (2021).
34. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
35. Lu, J. & Salzberg, S. L. Removing contaminants from databases of draft genomes. *PLoS Comput. Biol.* **14**, e1006277 (2018).
36. Mölder, F. *et al.* Sustainable data analysis with Snakemake. *F1000Res.* **10**, 33 (2021).
37. Dai, H. & Guan, Y. Nubeam-dedup: a fast and RAM-efficient tool to de-duplicate sequencing reads without mapping. *Bioinformatics* **36**, 3254–3256 (2020).
38. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
39. Schubert, M., Lindgreen, S. & Orlando, L. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Res. Notes* **9**, 88 (2016).

Supplementary Figures



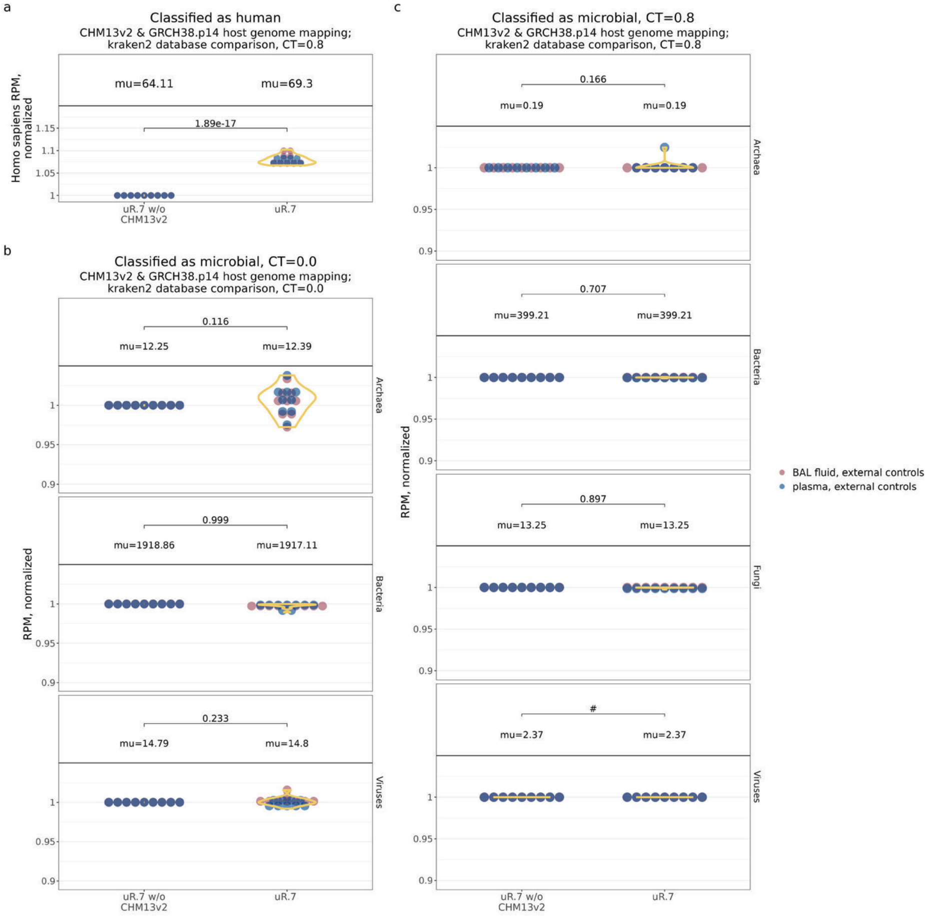
Supplementary Fig. 1. Impact of CHM13v2 host genome mapping on archeal and viral read quantification

a. Percentage of quality controlled (QC) reads retained by the cfSPI workflow after 1) removal of duplicates and 2) removal of low-quality and short reads. **b.** CT dependent fraction of archaeal, bacterial, fungal and viral classified reads, after host read subtraction by reference genome mapping. Black and grey lines represent CT of 0.0 and 0.8, respectively. **c-d.** Fractional abundance of archaeal, bacterial, fungal (only in **d.**) and viral kraken2 classified reads when utilizing a **c.** CT of 0.0 (kraken2's default) or a **d.** CT of 0.8 (for a high precision [20]), after subtracting host reads via reference genome mapping, normalized to the old version of the human genome assembly (GRCh38.p14). In **a-d.**, each data point/line represents one control sample (9 BAL; 9 plasma), with colors indicating the sample type. In **c-d.**, Mean PRM values are denoted as 'mu'. Statistical analysis included one-tailed paired t-tests with Bonferroni correction (*, p ≤ 0.05; **, p ≤ 0.01; ****, p ≤ 0.0001; ns, p > 0.05; #, t-test not performed due to all paired values being identical), utilizing the fractional abundance in RPM.



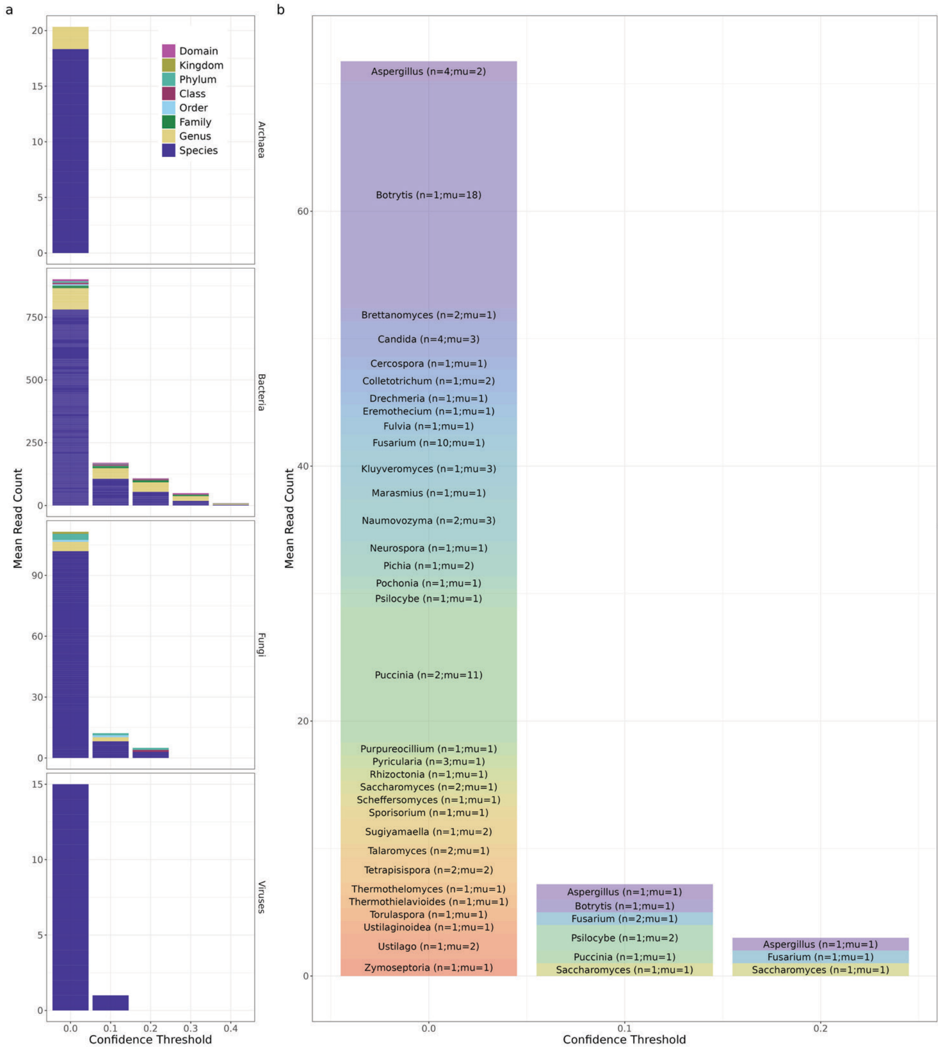
Supplementary Fig. 2. Comparison between kraken2 classification database complexity

a. Number of *Aspergillus* genomes, sourced from NCBI RefSeq, FungiDB (a subset of the EuPathDB), and/or MycoCosm, within diverse kraken2 classification databases (x-axis). Genomic sequences underwent various processing steps, including cleanup (crossing lines), dustmasking (diagonal lines), or no processing (denoted as 'u' for unaltered), before database construction. The number of **b.** *Aspergillus* species and **c.** *Aspergillus* subgenera in the diverse databases. **d.** Count of *Aspergillus* *k*-mers in diverse hash-table databases. **e.** Count of human and microbial *k*-mers in diverse hash-table databases. Database characteristics are summarized on the x-axis using colors to reflect the origin of the *Aspergillus* genomic sequences, patterns to denote the processing status of the *Aspergillus* genomes in the database.



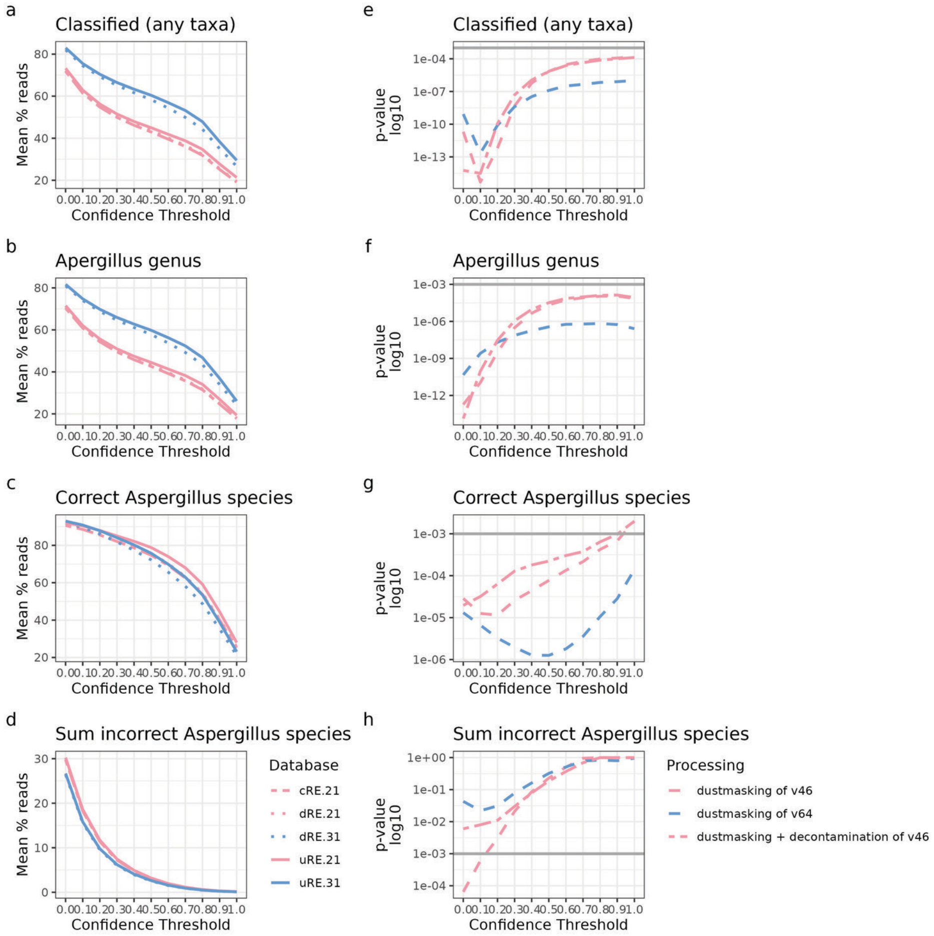
Supplementary Fig. 3. Integration CHM13v2 into the kraken2 database: exploring implications for host and microbial read counts

a. Fractional abundance of human classified reads when utilizing the 'CHM13v2-containing uR.7' or 'uR.7' database for kraken2 taxonomic classification (CT=0.8, kraken2's default) after dual-mapping to the host genome, normalized to 'uR.7'. **b-c.** Fractional abundance of archaeal, bacterial, fungal (only in **c.**) and viral kraken2 classified reads when utilizing the 'CHM13v2-containing uR.7' or 'uR.7' database for kraken2 taxonomic classification with a CT of 0.0 and 0.8, resp., after dual-mapping to the host genome, normalized to 'uR.7'. Each data point represents one control sample (9 BAL; 9 plasma), with colors indicating the sample type. Mean RPM values are denoted as 'mu'. Statistical analysis included one-tailed paired t-tests with Bonferroni correction (#, t-test not performed due to all paired values being identical), utilizing the fractional abundance in RPM.



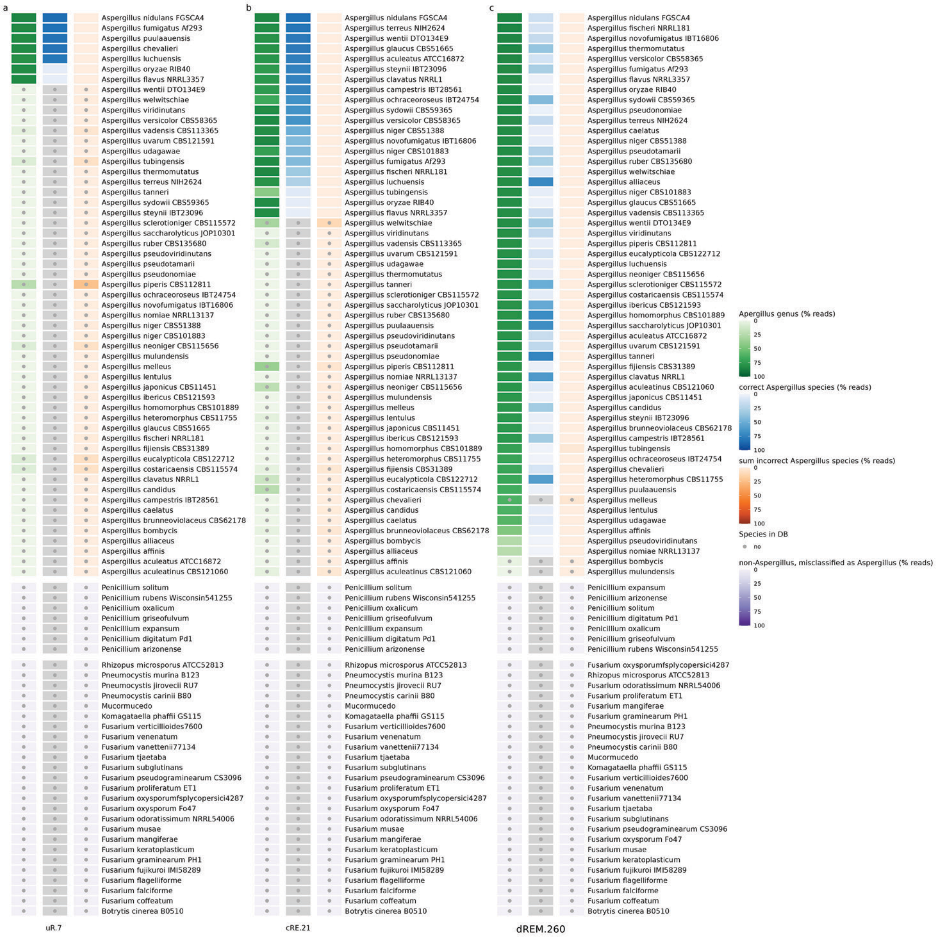
Supplementary Fig. 4. Exclusion CHM13v2 in the kraken2 database leads to misclassification of human reads as fungal species

We traced individual reads classified as human by the 'CHM13v2-containing uR.7' and assess their classification in the 'uR.7 w/o CHM13v2' database to identify misclassification. **a.** Mean classified microbial (i.e. archaeal, bacterial, fungal and viral) read counts in control samples (n=18) at various CTs. Colors indicate the microbial taxonomic level at which reads were classified. **b.** Mean counts of fungal species level classifications in control samples (n=18). Colors represent fungal genera, with numerical annotations denoting the count of fungal species within each genus and the mean number of reads ('mu').



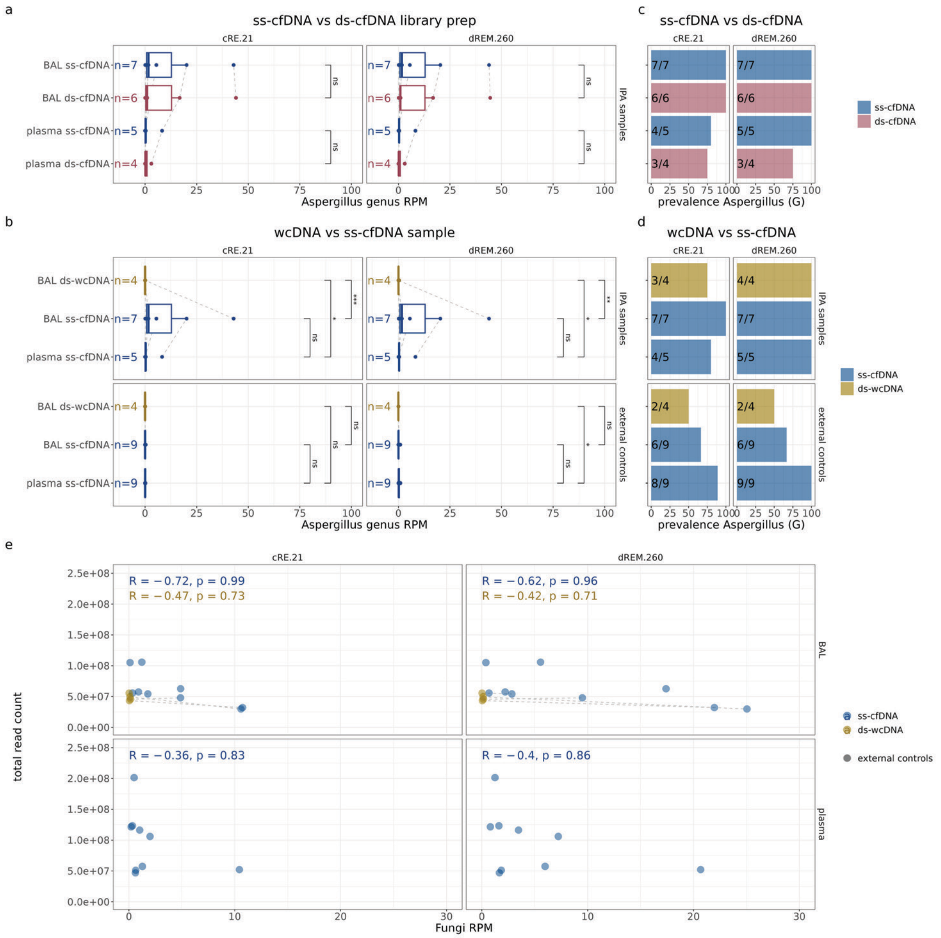
Supplementary Fig. 5. Assessing the impact of genomic dustmasking and decontamination on taxonomic classification in immunocompromised control samples

The line plots in **a-d.** depict the average across our 54 separate *Aspergillus* simulated Illumina datasets remained unclassified, observed at various CTs. The metrics include **a.** overall classification, **b.** *Aspergillus* genus level, **c.** correct *Aspergillus* species, and **d.** incorrect species. **e-h.** To assess the impact of dustmasking alone and dustmasking plus decontamination (i.e. cleanup) on these four metrics, we conducted a paired one-tailed t-test with Bonferroni correction between the percentage of reads classified in a database that underwent dustmasking and/or decontamination and their unaltered counterpart (uRE.21 and uRE31). Each data point in **e-h.** shows the outcome of this t-test, with the significance threshold at p = 0.001 indicated by a gray horizontal line.



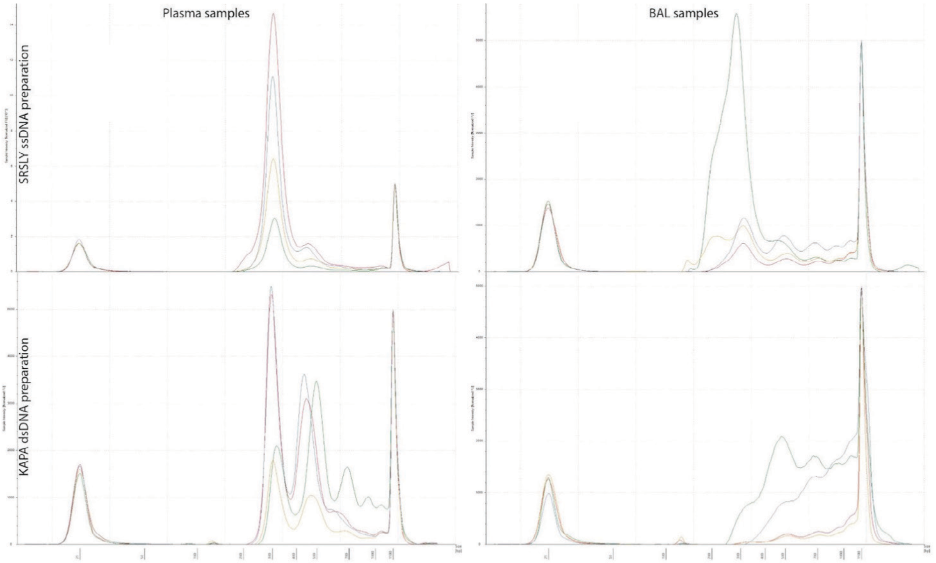
Supplementary Fig. 6. Heatmap taxon detection simulated fungal datasets using uR.7, cRE.21, and dREM.260 databases

Classification results of simulated fungal datasets, including *Aspergillus* (n=54), *Penicillium* (n=7), and other clinically relevant species (n=25), using three distinct databases: **a.** uR.7, **b.** cRE.21, and **c.** dREM.260, all with a CT at 0.8. For the *Aspergillus* simulated set, percentages of reads assigned to the *Aspergillus* genus (green), correct *Aspergillus* species (blue), or incorrect *Aspergillus* assignments (purple) are presented. In non-*Aspergillus* simulations, the percentage of misclassified reads to either the *Aspergillus* genus or species is indicated in orange. The datasets include simulated species present in the classification database, or absent (marked by gray dots). The y-axis lists *Aspergillus* species sorted based on their percentage classified reads at the species level (uR.7, cRE.21) or genus level (dREM.260).



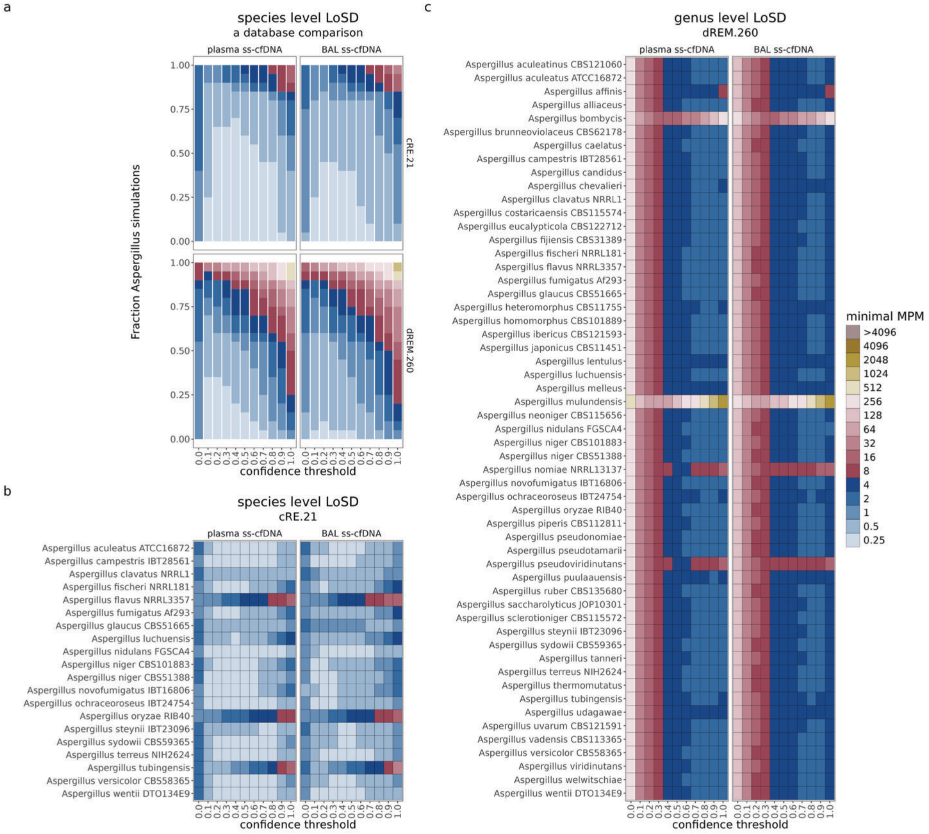
Supplementary Fig. 7. Influence of library preparation on *Aspergillus* genus detection and sequence depth on fungal counts

a-b. *Aspergillus* fractional abundance in RPM, emphasizing **a.** ss- versus ds-ligation and b. ds-wcDNA versus ss-cfDNA sequencing in IPA (top row) and external control samples (bottom row) when employing the cRE.21 (left) and d.REM.260 (right) database. Statistical significances between library preparation and/or sample types are evaluated by one-tailed Wilcoxon rank test with Bonferroni correction (*, $p \leq 0.05$; **, $p \leq 0.01$; ***, $p \leq 0.001$; ns, $p > 0.05$). **c-d.** Barplot of *Aspergillus* genus detection prevalence using cRE.21 (left), or dREM.260 (right) databases, emphasizing **c.** ss- versus ds-ligation and **d.** ds-wcDNA versus ss-cfDNA sequencing. **e.** Scatter plot with Pearson correlation between the total read count and fractional fungal abundance (RPM) in external control samples. Pearson correlation test (*R*, one-tailed) with a $p > 0.05$ indicated there was no positive correlation between the two variables for cRE.21 (left) and dREM.260 (right) databases. BAL samples that underwent both ss-cfDNA and ds-wcDNA preparation are connected by a gray dashed line.



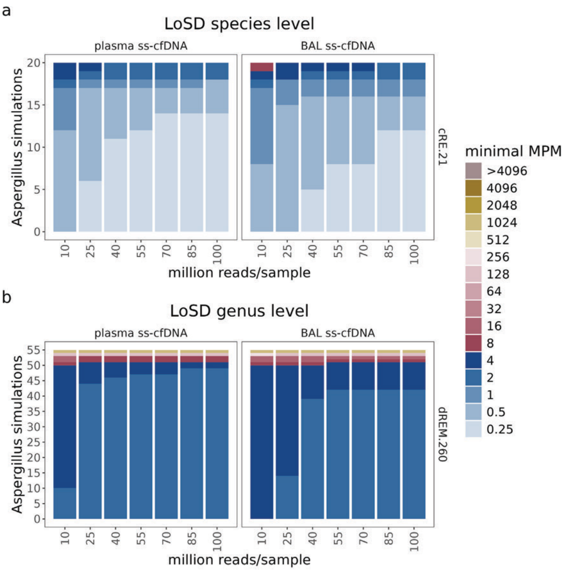
Supplementary Fig. 8. Length profile analysis of ssDNA-ligated sequencing libraries in BAL and plasma samples using TapeStation

This figure illustrates the electropherogram (TapeStation 2200, D1000HS kit) capturing the length profiles of sequencing libraries from plasma and BAL fluid samples. The libraries were generated using ss-cfDNA (SRLY ssDNA preparation, top row) or ds-cfDNA (KAPA dsDNA preparation, bottom row), providing insights into the diverse profiles associated with each preparation method. For each electropherogram profile, x-axis represents the estimated fragment lengths (in bp), y-axis represents relative abundance, with reference peak at 25 bp and 1500 bp.



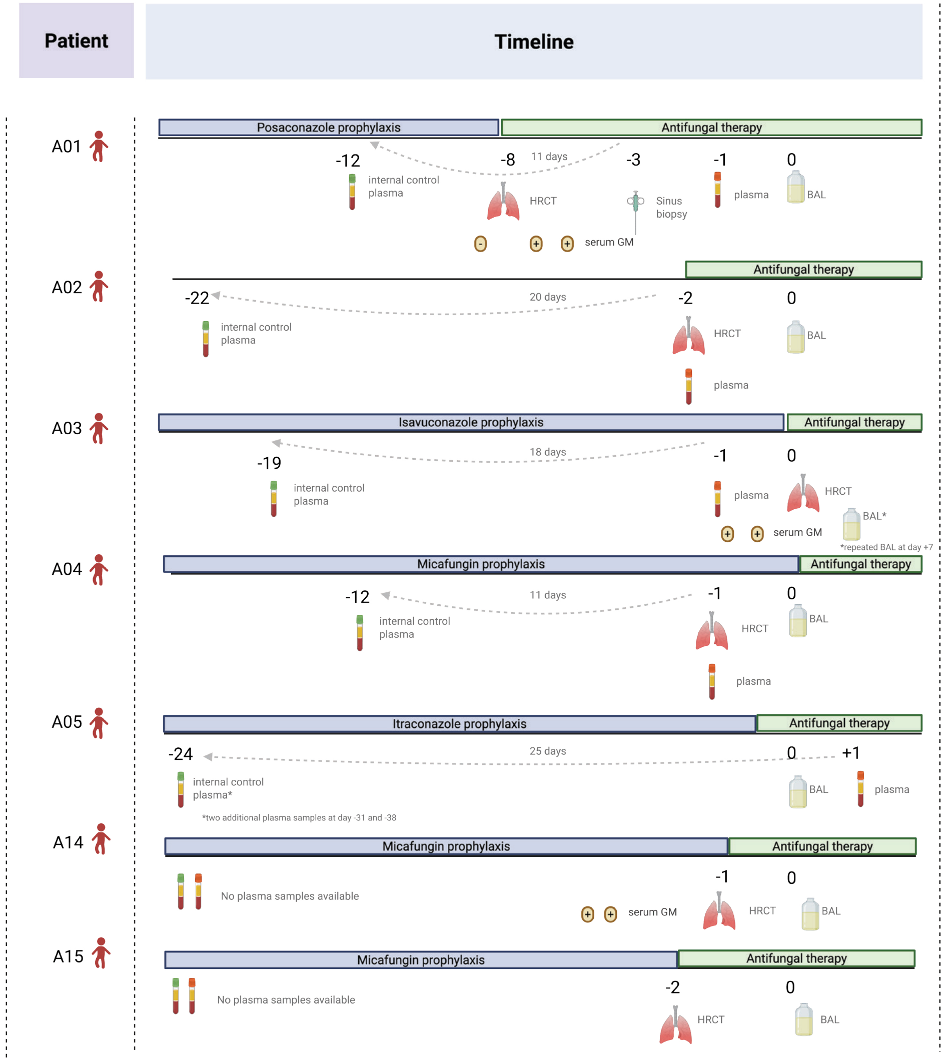
Supplementary Fig. 9. LoSD per simulated *Aspergillus* dataset

Results of LoSD analysis aiming to compute the theoretical minimum fraction of *Aspergillus* molecules necessary for the detection of significantly elevated *Aspergillus* taxon above the control background noise. Fisher's exact test was employed to determine significant differences in *Aspergillus* **a-b.** species and **c.** genera count compared to our external control samples, with a mean p ≤ 0.001 considered statistically significant. **b-c.** The categorical heatmap depicts the computed limits of significance in MPM at a sequencing depth of 70 million for each *Aspergillus* simulation (y-axis; n=54) for plasma ss-cfDNA (left) or BAL ss-cfDNA (right), either **b.** at the species level (mediated by the cRE.21 database) or **c.** at the genus level (mediated by the dREM.260 database) with variable CTs from 0.0-1.0 (x-axis).



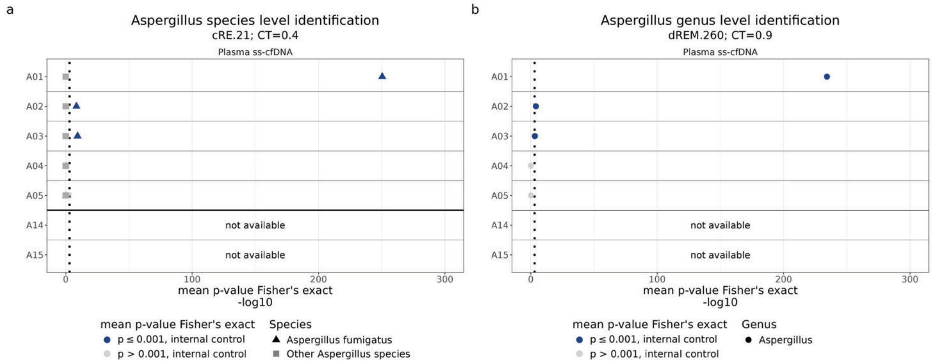
Supplementary Fig. 10. Impact of sequencing depth on LoSD in simulated *Aspergillus* datasets

Variable sequencing depth results in different minimal MPM required for *Aspergillus* detection above the control background noise, i.e. LoSD. Barplots showing cumulative counts of simulated *Aspergillus* species. Fisher's exact test assessed significant differences at *Aspergillus* **a.** species level in 20 simulated *Aspergillus* species (mediated by the cRE.21 database) and **b.** genus level in 54 simulated *Aspergillus* species (mediated by the dREM.260 database), compared to external control samples, with a mean p ≤ 0.001 considered statistically significant. The categorical heatmap illustrates computed limits of significance between 0.25 and >4096 molecules per million (MPM) at sequencing depths ranging between 10 and 100 million reads per sample (x-axis) for each Aspergillus simulation (n=54).



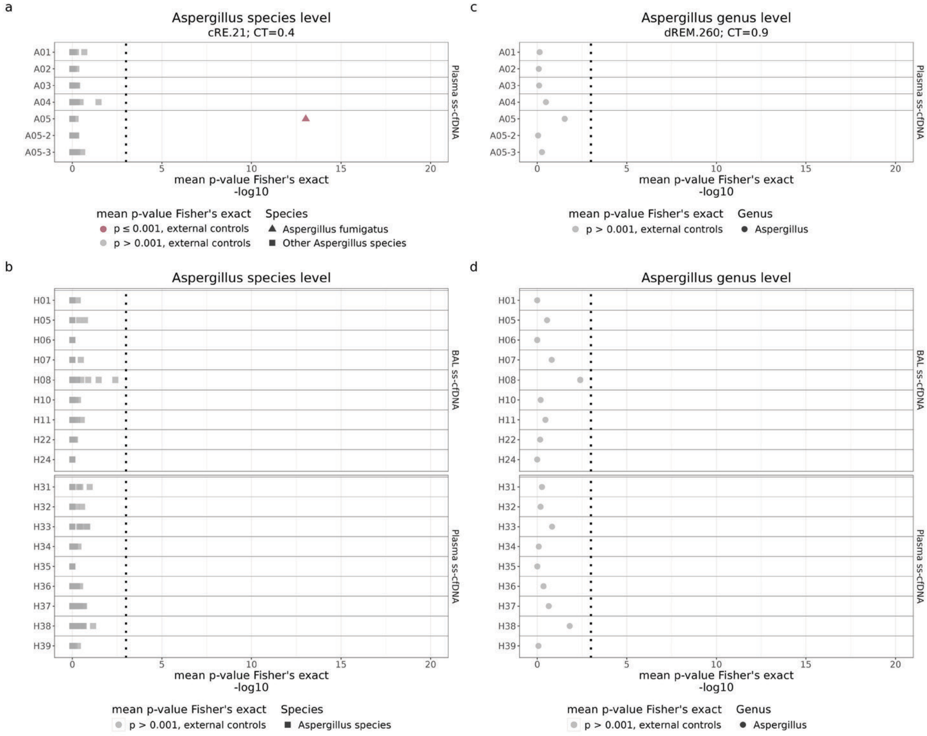
Supplementary Fig. 11. Timeline clinical work-up of probable aspergillosis cases including antifungal treatment in relation to sample retrievement

Illustration of clinical timeline of all seven probable IPA patients A01-A05, A14-A15, detailing antifungal treatment, the timing of various diagnostic workups and retrieval of blood plasma. In our proof-of-principle study, both a plasma and a BAL fluid sample were included for each patient. An additional internal control plasma sample was included, acquired between 11-25 days preceding the diagnostic workup.



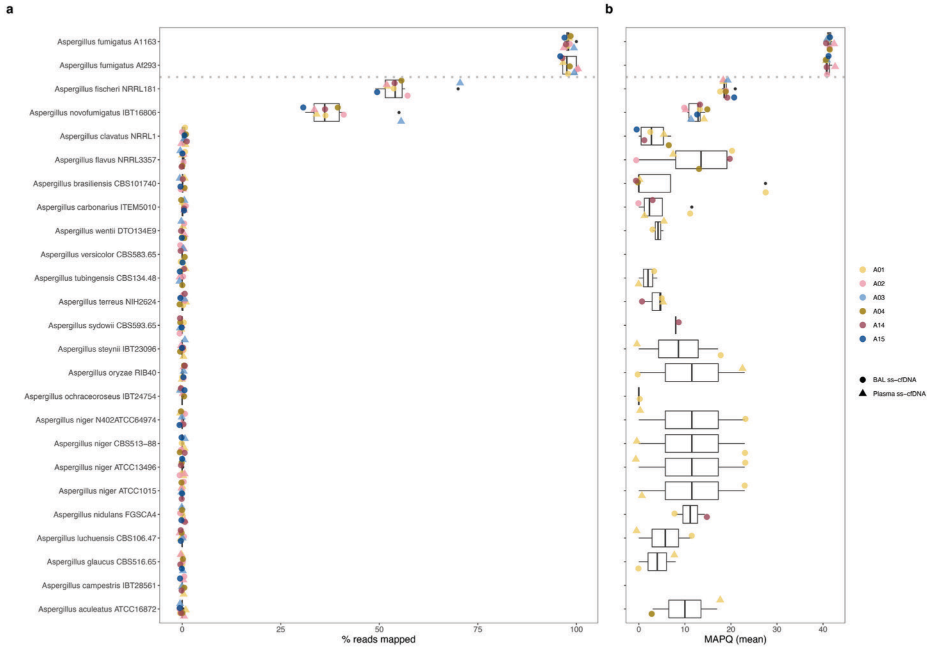
Supplementary Fig. 12. Elevated *Aspergillus* levels in plasma samples with probable IPA compared to internal controls

The one-tailed Fisher's exact test was employed to assess the fractional abundance of *Aspergillus* in ss-cfDNA NGS plasma samples from patients with probable IPA. In **a-b** this comparison was made against an internal control plasma sample from the same patients, which had been collected at an earlier time point (for schematic see Supplementary Fig 11; for details see *Methods*). Comparison was made both at the **a.** species level, using the cRE.21 (CT=0.4), and **b.** at the genus level, using the dREM.260 (CT=0.9) (see *Methods* for details).



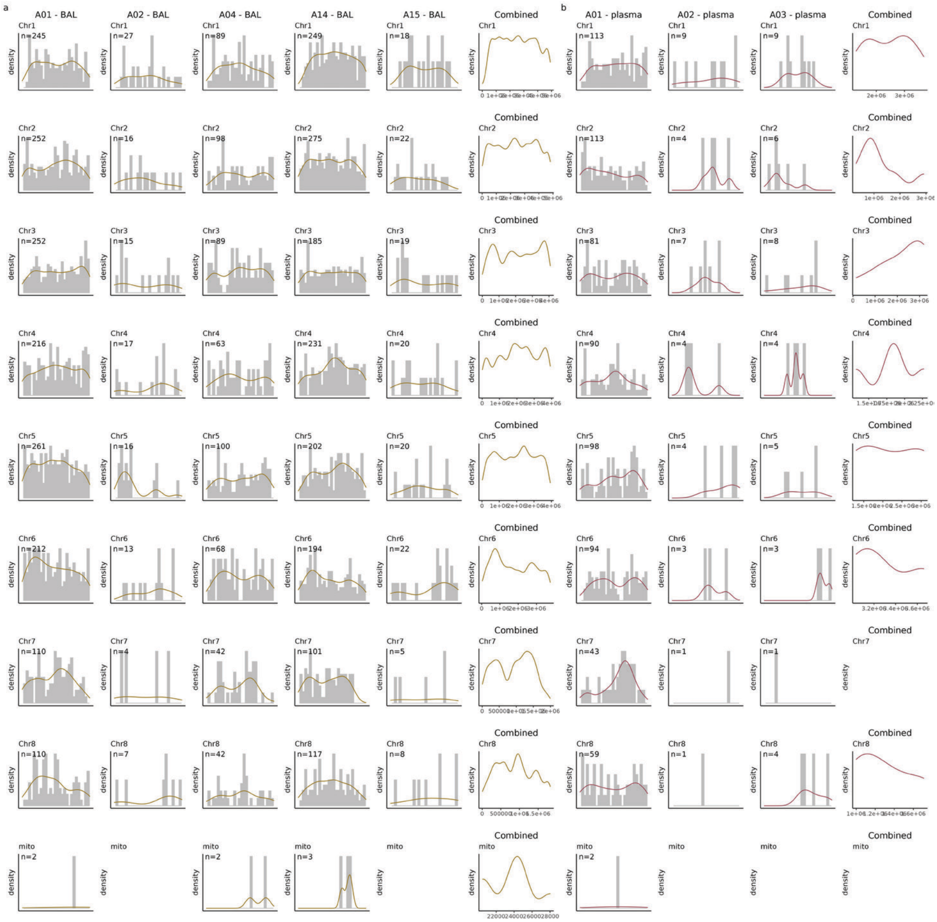
Supplementary Fig. 13. *Aspergillus* levels in control ss-cfDNA NGS samples

The one-tailed Fisher's exact test was utilized to compare the fractional abundance of *Aspergillus* in a given sample to a control set of samples. Comparison entails ss-cfDNA NGS read counts of **a-b.** *Aspergillus* at the species level while employing the cRE.21 (CT=0.4) plus **c-d.** *Aspergillus* at the genus level while employing the dREM.260 (CT=0.9). We employed the one-tailed Fisher's exact test to compare **a,c.** each internal control sample (derived from a IPA patient) against external control samples (derived from immunocompromised patients), as well as to compare **b,d.** each external control sample against all other external control samples (leave-one-out principle). Dot plots display the mean -log10-transformed computed p-values, with the significance threshold set at p=0.001 indicated by a vertical dotted line. Instances exceeding the significance threshold are highlighted in red.



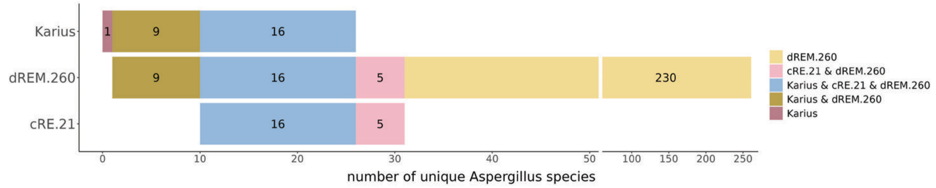
Supplementary Fig. 14. Alignment *A. fumigatus* classified reads to *Aspergillus* genomes integrated in the cRE.21 database

Reads classified as *A. fumigatus* in IPA patient samples with a significantly heightened *A. fumigatus* ss-cfDNA abundance in their plasma (n=3) and/or BAL fluid (n=5) sample were realigned to all *Aspergillus* genomes integrated in the cRE.21 database. Boxplots showing **a.** the percentage of ss-cfDNA NGS reads aligned and **b.** the average mapping quality (MAPQ) of the aligned reads. Each datapoint represents a sample.



Supplementary Fig. 15. Read distribution of *A. fumigatus* classified reads to the genome of *A. fumigatus* strain Af293

Histogram and density analysis of ss-cfDNA NGS reads, categorized using the cRE.21 at CT=0.4, and aligned to the *A. fumigatus* reference genome. Sequence reads obtained from a patient exhibiting notably elevated *A. fumigatus* ss-cfDNA levels in their **a.** BAL fluid and/or **b.** plasma were re-mapped to the genome of *A. fumigatus* strain Af293. The alignment was systematically displayed across varied chromosomes (rows), with each column symbolizing a unique IPA patient sample. The count of aligned reads is provided for each subplot in the upper-left corner (n). The alignment of all **a.** BAL and **b.** plasma samples are summarized in yellow and dark-red respectively, and demonstrates uniformity across different chromosomes.



Supplementary Fig. 16. Comparative analysis of pathogen list concordance with cRE.21 and dREM.260 databases

The bar plot visually depicts the (dis)concordance in *Aspergillus* species quantity between the pathogen list from Karius and the ones in the cRE.21 and dREM.260 databases.

Supplementary table

Suppl. Tables 1-5 can be found here: https://doi.org/10.1038/s41525-025-00482-8

CHAPTER 3

Exploring cell-free DNA fragmentomics to improve *Aspergillus* detection in invasive mold infections

Emmy Wesdorp [1,2], Myrthe Jager [1,2], Jeroen de Ridder [1,2]

[1] Center for Molecular Medicine, University Medical Center Utrecht, Utrecht, The Netherlands.
[2] Oncode Institute, Utrecht, The Netherlands.

Unpublished

Abstract

Plasma cell-free DNA (cfDNA) next-generation sequencing (NGS) shows great potential for diagnosing infectious diseases, including life-threatening invasive mold infections. However, its diagnostic utility is limited by extremely low counts of fungal cfDNA, requiring excessive sequencing depth, which drives up costs. This study examines how library preparation affects the detection of fungal species, particularly *Aspergillus fumigatus*, a leading cause of invasive mold infections in immunocompromised patients, and explores whether size selection of cfDNA or selection based on cfDNA end-motis can enhance *A. fumigatus* relative abundance. Our analysis of sequencing data from liquid biopsy samples of immunocompromised patients demonstrates that single-stranded library preparation (ssLP) increases the proportion of *A. fumigatus* fragments by effectively capturing shorter fragments that are typically underrepresented in double-stranded library preparation (dsLP). We found that approximately 50% of *A. fumigatus* cfDNA fragments in plasma ssLP were ≤100 bp, compared to only 5% for host cfDNA. These findings prompted us to optimize size selection *in silico* based on yield and enrichment, concluding that selecting fragments up to 100 bp significantly enhances the recovery of fungal cfDNA in both plasma and bronchoalveolar lavage (BAL) samples, achieving a 3- to 8-fold enrichment of the *A. fumigatus* signal, respectively. Additionally, *in silico* selection of fragments starting with a CG at the 5’ end can increase *Aspergillus* relative abundance by about 3-fold, regardless of library preparation method used. These computational results provide a promising pathway for optimizing wet lab protocols to enhance *Aspergillus* detection in low-fungal plasma samples from patients suspected of invasive pulmonary aspergillosis, aiding the development of more sensitive and cost-effective minimally invasive diagnostic methods.

Background

Cell-free DNA (cfDNA) molecules are short DNA fragments found in liquid biopsies. Next-generation sequencing (NGS) of these molecules have enabled a plethora of diagnostic applications, including early cancer detection, non-invasive prenatal testing and microbial pathogen detection. The majority of cfDNA (>99%) typically originates from (the hematopoietic lineage of) the host[1–6]. A much smaller fraction of cfDNA may come from other sources. The fraction of microbial cfDNA is particularly low (around 0.01-0.001%[5]), of which only a subset is linked to active pathogens. For every pathogen-derived cfDNA molecule sequenced, more than a thousand host-derived cfDNA molecules are also sequenced. Consequently, a significant challenge in cfDNA NGS-based pathogen detection is that the majority of the sequencing budget — often several hundred euros per sample — is consumed by sequencing host-derived reads that offer no diagnostic value for identifying pathogens. This inefficiency ultimately diminishes the cost-effectiveness and diagnostic efficiency of the test.

To address this challenge, it is crucial to maximize the capture and readout of relevant microbial cfDNA. One approach is to simply increase sequencing depth, but this also raises costs and may increase turnaround times. Alternatively, enrichment techniques that increase the proportion of microbial cfDNA in the sequencing library could improve efficiency by enhancing the detection of pathogen-derived sequences, reducing costs, and accelerating clinical diagnostics. These techniques could furthermore decrease the total read count, making the assay more compatible with third-generation sequencing technologies.

Previous studies have shown that single-stranded library preparation (ssLP) can increase the relative abundance of, for example, bacterial and viral cfDNA, over conventional double-stranded library preparation (dsLP)[5]. ssLP captures a wider range of cfDNA types, including jagged, nicked, and single-stranded molecules[7], and is more effective at detecting shorter cfDNA fragments (40–100 bp), whereas dsLP detects few microbial cfDNA molecules under 100 bp[5]. In Wesdorp *et al.* (under review), we further gathered evidence demonstrating that ssLP enhances the recovery of fungal DNA. However, despite this advantage, fungal DNA levels remained low — often limited to just a few molecules per sample — even in patients suspected of invasive pulmonary aspergillosis (IPA). These extremely low levels highlight the urgent need for additional enrichment strategies to significantly improve the recovery and detection of the "needle-in-a-haystack" microbial cfDNA. Such strategies would increase the likelihood of detecting fungal pathogen-derived cfDNA, thereby potentially enhancing diagnostic capabilities.

In this study, we reanalyzed the dataset from Wesdorp *et al.* (under review), comprising of 12 lung lavage and plasma liquid biopsy samples from patients suspected of IPA. In these samples, we previously identified elevated levels of *Aspergillus fumigatus,* a major cause of invasive fungal disease, in 5 out of 7 lung lavage samples and 3 out of 5 plasma samples. Our primary goal is to understand why ssLP only marginally enriches *Aspergillus*-derived cfDNA, and more importantly, to explore more effective methods for enrichment of cfDNA derived from *Aspergillus fumigatus*. To this end, we present an *in silico* analysis designed to support our long-term objective of developing an optimized wet lab approach for pathogen-derived signal enrichment, enhancing the detection of pathogens that might otherwise go undetected due to their low abundance. For the purpose of our *in silico* analysis, we make use of liquid biopsy samples from immunocompromised children, a population at high risk for invasive fungal diseases, with and without IPA, including those with hematological malignancies and those undergoing hematopoietic stem cell transplantation. *Aspergillus fumigatus* thereby serves as a proof-of-principle example, with the expectation that the strategies developed will be applicable to other *Aspergillus* species and fungal pathogens.

Fragmentomics, the study of cfDNA fragmentation patterns, is well-established for plasma-derived host cfDNA. However, our understanding of cfDNA fragmentation in immunocompromised patients with hematological disorders remains limited. To address this gap, we first mapped the chromosomal cfDNA fragmentation landscape in both plasma and bronchoalveolar lavage (BAL) supernatant from immunocompromised patients. This foundational work allowed us to compare host cfDNA findings to *Aspergillus*-derived fragmentomics, elucidating how ssLP enhances fungal abundance and enabling the development of tailored *Aspergillus* enrichment strategies. This insight facilitated the development of tailored *Aspergillus* enrichment strategies. We conducted *in silico* exploration of two enrichment strategies: size selection and end-motif selection, based on the hypothesis that *Aspergillus*-derived cfDNA differs from host cfDNA in key physical fragmentomic characteristics, such as fragment length and end-motif. Our findings demonstrated that selecting for these characteristics enhances the relative abundance of *Aspergillus* cfDNA. Our approach builds on previous research into optimized DNA isolation and library preparation[8], as well as studies demonstrating that size selection techniques, applied *in silico*, can effectively enrich short cfDNA fragments. This method has been shown to improve ratios such as viral-to-host[9] as well as increase the relative bacterial fraction[10], making it particularly effective for gathering sufficient evidence to identify or confirm the presence of a pathogen.

Results

ssLP achieves effective capture of sub-nucleosomal fragments of host chromosomal cfDNA

We characterized the host-derived cfDNA fraction in our liquid biopsy Illumina dataset (see *Methods*) to examine how library preparation strategies influence the fragmentomic characteristics of chromosomal cfDNA, which constitutes the majority of the cfDNA pool. Our dataset comprised of 24 plasma samples from immunocompromised pediatric patients, including 9 samples from immunocompromised controls (“External Controls”), 7 samples from patients with suspected fungal infections — 5 with invasive pulmonary aspergillosis (IPA) and 2 with other mold infections — and 8 samples collected from these same patients prior to the onset of fungal symptoms (“Internal Controls”). Additionally, we included 21 BAL samples (9 External controls, 11 IPA cases, and 1 Other mold). All samples were processed using SRSLY-mediated ssLP methods (Wesdorp *et al.*, under review), with some also undergoing KAPA-mediated double-stranded library preparation (dsLP) (Fig. 1a, schematic library preparations in Supplementary Fig. 1), resulting in 63 paired-end (2 × 150 bp) sequenced libraries. To account for sequencing depth variations (see Supplementary Table 1 for number of reads/library), we normalized the data within each library and compared the relative frequencies of cfDNA length profiles and end-motifs. We focussed our analysis on short cfDNA molecules (≤ 500 bp).

To examine the effect of library preparation on host-fragment length, we analyzed samples prepared using dsLP and ssLP, as these methods have previously been applied for *Aspergillus* detection in our immunocompromised cohort. In plasma samples processed with dsLP, we observed a prominent peak at 166 bp and a less prominent peak at 320 bp, corresponding to the typical peaks linked to mono- and di-nucleosomes, respectively. The 166 bp peak was also present in plasma samples processed using ssLP. Similarly, BAL samples had peaks at 340 bp and 166 bp in both library preparations, although the 340 bp peak was only prominent in dsLP. Notably, the 166 bp single chromatosome peak in ssLP was more symmetrical compared to the unilateral right shoulder seen in dsLP. We thereby confirmed that BAL host chromosomal cfDNA fragments are longer than plasma cfDNA fragments, which is in line with previous findings[11]. Most importantly, however, our comparison of library preparation methods reaffirmed that dsLP is more suitable for studying longer, host chromosomal cfDNA, while ssLP offers higher resolution for short, sub-nucleosomal cfDNA fragments. This difference is due to ssLP’s enhanced sensitivity to the full spectrum of cfDNA forms in liquid biopsy samples, including fragments with nicks, overhangs, and single-stranded DNA[12].

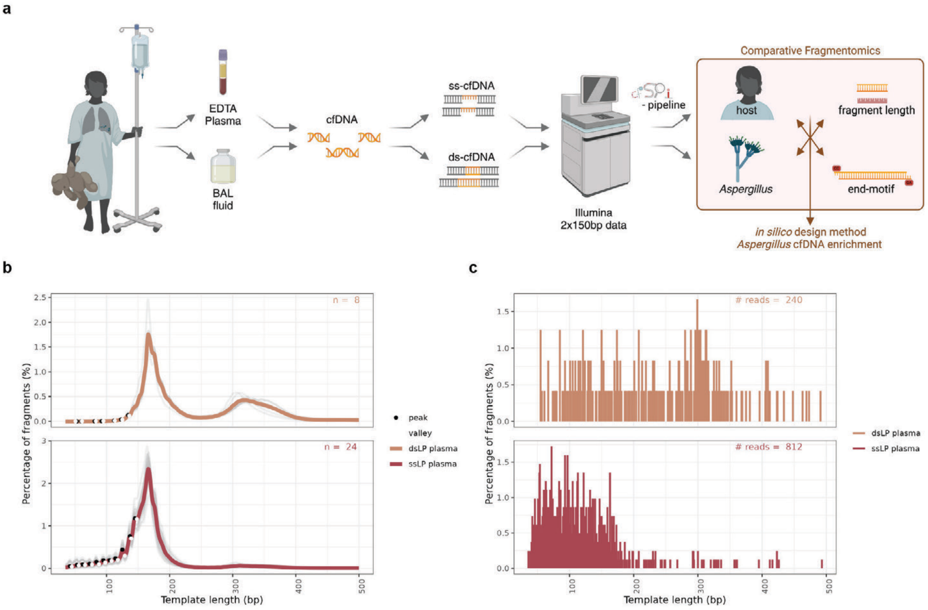


Fig. 1. Comparative fragmentomics analysis of host and *Aspergillus fumigatus* cfDNA using ssLP and dsLP library preparations

a. In previous work (Wesdorp *et al*., under review) on liquid biopsy sequencing in immunocompromised pediatric patients, cfDNA was isolated from EDTA plasma (n=24) and BAL fluid supernatants (n=21) from patients at risk of invasive mold infections. Liquid biopsy microbial DNA sequencing was performed both after double-stranded library preparation (dsLP) with the KAPA kit and single-stranded library preparation (ssLP) with the SRSLY method[23]. This approach yielded 45 ssLP and 18 dsLP sequencing libraries, allowing us to compare the effectiveness of these methods in detecting pathogenic *Aspergillus* species in liquid biopsies. The prepared libraries were sequenced using Illumina sequencing (2x 150 bp) and analyzed using the cfSPI-pipeline. Building on this (metagenomic) data, the current study focuses on the fragmentomic characteristics of host and *Aspergillus* cfDNA. This analysis aims to address the limited enrichment of *Aspergillus*-derived cfDNA observed with ssLP and to identify more effective methods for enhancing the recovery of cfDNA from *Aspergillus fumigatus*. To this end, we explore *in silico* enrichment strategies, based on size and end-motif selection, grounded in the hypothesis that the physical characteristics of *Aspergillus*-derived cfDNA differ significantly from those of host-derived cfDNA. **b.** Human chromosomal cfDNA fragment length distribution in plasma samples after ssLP and dsLP preparation, shown as an average density plot (red) over 24 ssLP and 8 dsLP plasma specimens (grey). Periodicity analysis revealed variable valleys (white points) and peaks (black points) within the sub-nucleosomal size range. **c.** Percentage of *Aspergillus fumigatus*-derived cfDNA fragments per 1-bp length bin in plasma samples, following ssLP and dsLP preparation. Data is pooled from all samples, including 24 ssLP and 8 dsLP BAL samples. The total number of reads classified as *A. fumigatus* across all dsLP or ssLP is displayed in the upper-right corner.

In addition to nucleosomal peaks, both plasma and BAL exhibited the previously described characteristic ~10 bp oscillating pattern in sub-chromatosome fragments in both library preparations (Fig. 1b, Supplementary Fig. 2a)[13]. However, the number and intensity of these oscillations varied between the two library preparations, even with the same cfDNA sample (Supplementary Fig. 3). ssLP generally showed more oscillations, with 11 valleys in plasma and 10 valleys in BAL, and a higher average periodicity index (PI). A higher PI indicates that there were stronger periodic patterns present, of 1.52 in plasma and 1.67 in BAL. In contrast, dsLP showed fewer detectable oscillations with 7 valleys in plasma and 10 valleys in BAL, and a lower PI of 1.52 in plasma and 1.54 in BAL. These differences in oscillation also confirm previous findings, and are attributed to nucleosome wrapping and dsLP end-polishing effects, with ssLP providing a more accurate representation of true fragment lengths and end-nucleotide(s)[1,14]. These observations reinforce previous findings that ssLP provides a more precise view of cfDNA fragment structure, particularly in capturing oscillations of short sub-nucleosomal fragments. To summarize, our fragmentomics findings align with existing literature, indicating no major deviations in host cfDNA fragmentomic characteristics, this despite alterations in our patient cohort's hematopoietic and immune systems.

Elevated relative abundance of *Aspergillus fumigatus* cfDNA achieved through short fragment capture with ssLP

In our recent work, we observed that ssLP achieved a slight (non-significant) increase in fungal and *Aspergillus*-derived cfDNA relative abundance. However, we did not previously analyze the length of these fungal-derived sequences. To compare the abundance and lengths of *Aspergillus fumigatus*-derived cfDNA fragments between library preparations, we reanalyzed our data using an optimized Kraken2 database and threshold. We observed a relatively higher abundance of *A. fumigatus* (reads per million (RPM)) (Supplementary Fig. 4a) in ssLP compared to dsLP, although this difference was not statistically significant (Wilcoxon's rank with Bonferroni correction). The median abundance of *A. fumigatus*-derived cfDNA was 1.26 RPM (IQR: 0.11–4.51) in plasma and 0.31 RPM (IQR: 0.01–0.32) in BAL samples from IPA patients, indicating that the level of *A. fumigatus* reads remains low even after ssLP.

We mapped the reads assigned to *A. fumigatus* by Kraken2 onto the *A. fumigatus* genome and analyzed their length profiles. Pooled reads—combining all samples—showed median fragment sizes of 236 bp (IQR: 150–307 bp) in plasma after dsLP and 105 bp (IQR: 72–142 bp) after ssLP (Fig. 1c). For BAL samples, the median lengths were 270 bp (IQR: 204–330 bp) for dsLP and 129 bp (IQR: 94–175 bp) for ssLP (Supplementary Fig. 2b). Compared to host cfDNA (both chromosomal and mitochondrial), *A. fumigatus*-derived cfDNA is generally shorter (Fig. 2a; Supplementary Fig. 4b), confirming findings from previous work, which primarily focused on bacteria[5,10]. Notably, when comparing abundances across cfDNA lengths, ssLP demonstrated a lower *A. fumigatus*-to-host read ratio across various lengths compared to dsLP (Supplementary Fig. 4c-d). This indicates that the higher RPMs for *A. fumigatus* in ssLP-prepared samples primarily result from the method's enhanced recovery of shorter fragments.

Increased abundance of *Aspergillus fumigatus* through *in silico* selection of short cfDNA fragments

Given the differences in length between *A. fumigatus* and host cfDNA (host: chromosomal and mitochondrial), we investigated whether *in silico* selection of short cfDNA fragments could increase fungal abundance as a proxy for potential *in vitro* size selection and how this would affect detection sensitivity. To this end, we first assessed the ratio of *A. fumigatus* to human reads across different fragment lengths, then performed *in silico* size selection of all reads and finally evaluate its impact on the number of liquid biopsy IPA samples showing increased fungal abundance compared to immunocompromised controls (i.e., detection sensitivity). The results of these analyses are presented here. This work represents a step towards the overarching goal of designing enrichment strategies to be applied *in vitro*, enabling improved detection of pathogens that might otherwise remain undetected due to their low abundance.

To start, we calculated a ratio between the *cumulative* fraction of *A. fumigatus* reads and the *cumulative* fraction of host (chromosomal and mitochondrial) reads. The cumulative fraction represents the proportion of total reads that are of a particular length or shorter, summed across all read lengths starting from 35bp up to a specified point. This allowed us to compare how the distribution of fragment lengths for *A. fumigatus* reads relates to that of the host cfDNA (Fig. 2a,b; Supplementary Fig. 4e). The *cumulative* fraction ratios peaked at different lengths, depending on the library preparation method and biopsy type. dsLP generally showed a higher enrichment peak at small fragment sizes than ssLP, but the number of absolute observations was extremely low in dsLP DNA fragments < 150bp. Regardless of library preparation type, the ratios peaked between 50-75 bp for plasma samples (Fig. 2b), whereas for BAL samples, the peak occurred around 100 bp (Supplementary Fig. 4e). This suggests that short-read size selection appears particularly advantageous for improving detection of *A. fumigatus*, especially in ssLP libraries.

To determine the optimal cutoff for enhancing microbial abundance, we performed *in silico* size selections across various fragment sizes (up to 50 bp, 75 bp, 100 bp, 125 bp, and 150 bp) for all cfDNA molecules in the pool, including those of microbial origin and other molecules of undefined origin. Our analysis focused on three key metrics: relative abundance (expressed as RPM), prevalence (defined as the presence of at least one *A. fumigatus* read), and the retained fraction of total reads after *in silico* size selection.

In liquid biopsy samples from IPA patients prepared with ssLP, selecting 35–50 bp fragments resulted in the greatest increase in *A. fumigatus* RPM (Fig. 2c, middle panel; Supplementary Fig. 5b, middle panel). However, this stringent selection criterion reduced the number of IPA samples with at least one detected *A. fumigatus* read (Fig. 2c, left panel; Supplementary Fig. 5b, left panel), highlighting a trade-off: while size selection can enhance fungal abundance, it may also reduce the likelihood of pathogen detection, defined here as the presence of at least one *A. fumigatus* read. Additionally, selecting 35–50 bp fragments decreased the average fraction of retained fragments per library to less than 0.8% in plasma (Fig. 2c, right panel), underscoring an additional trade-off between boosting fungal abundance and preserving sufficient DNA after size selection, reducing sequencing library complexity.

To balance fungal enrichment with DNA retention in our IPA samples following size selection, we recommend selecting fragments in the 35-100 bp range for ssLP plasma and ssLP BAL samples. This approach was estimated to retain, on average, 8% of plasma ssLP fragments and 15% of BAL ssLP fragments (Fig. 2c, right panel; Supplementary Fig. 5b, right panel). Furthermore, the prevalence of *Aspergillus* in IPA samples remained the same after size selection of ssLP libraries. In contrast, the percentage of reads retained from the dsLPs is notably lower after *in silico* size selection, with fewer than 0.5% under 100bp (Supplementary Fig. 5a,c; right panel), presenting challenges for *in vitro* application. Thus, stringent size selection is not advisable for dsLP, whereas selecting fragments <100 bp does enrich *A. fumigatus* reads in ssLP libraries.

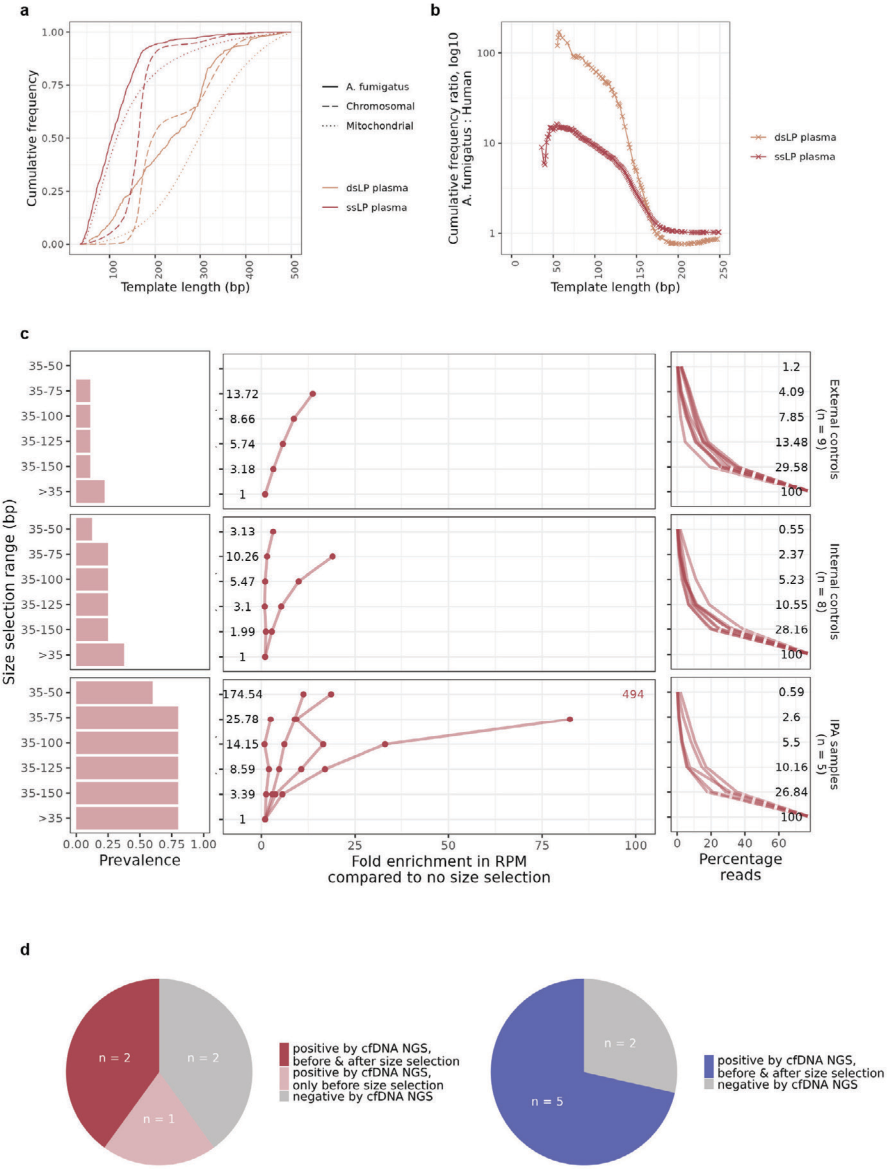


Fig. 2. *Aspergillus fumigatus* cfDNA size selection-based enrichment effective after ssLP

a. Cumulative distribution plot of *Aspergillus fumigatus*- and human-derived (chromosomal and mitochondrial) cfDNA fragment lengths by ddLP and ssLP in plasma. Lines are colored by sample type and library preparation. **b.** Cumulative frequency ratios of *A. fumigatus* to host chromosomal cfDNA in plasma after ssLP and dsLP. Lines are colored by sample type and library preparation. **c.** Effect of *in silico* size selection on plasma samples after ssLP, showing outcomes when selecting fragments up to 50, 75, 100, 125, or 150 bp, or with no size selection (>35). Left panel: prevalence of *Aspergillus fumigatus* detection, indicating the presence of at least one read post-ssLP with/without size selection. Middle panel: Fold enrichment in the relative abundance of *A. fumigatus*, expressed in reads per million (RPM) of quality-controlled reads, relative to its abundance without size selection. Right panel: percentage of quality-controlled reads retained following *in silico* size selection. A dotted line approximates the trend leading up to 100% when no size selection was applied (>35). **d.** Effect of size selection (range 35-100) on diagnostic yield using cfDNA NGS in plasma (left) and BAL (right).

Selecting ssLP reads up to 100 bp resulted in an average 8-fold enrichment of *A. fumigatus* reads across ssLP plasma samples *in silico*. Furthermore, we were able to detect *A. fumigatus* both before and after size selection, with a notable 3-fold enrichment specifically in BAL ssLP samples. Selecting fragments in the 35-100 bp range also increased *A. fumigatus* RPM in IPA ssLP plasma samples from 0.31 (IQR: 0.01-0.32) to 0.32 (IQR: 0.27-5.19) and raised the median *A. fumigatus* RPM in BAL IPA samples from 1.26 (IQR: 0.11-4.51) to 3.86 (IQR: 0.81-11.87). However, this *in silico* size selection also reduced the number of IPA plasma samples showing elevated *A. fumigatus* cfDNA abundance relative to external controls. Importantly, only one patient, patient A03, no longer exhibited elevated levels (mean pairwise Fisher's exact test, $p > 0.001$; Fig. 2d). It is worth noting that patient A03's BAL sample tested negative by cfDNA NGS (data not shown; see Wesdorp *et al*., under review (Chapter 2) for details), suggesting that evidence for invasive *A. fumigatus* infection in A03 based on cfDNA NGS was limited even before size selection. Overall, these findings suggest that selecting fragments up to 100 bp after ssLP library preparation can enhance fungal cfDNA abundance but may also increase the risk of false negatives in patient testing when sequencing is limited to only the enriched fragment set.

A noteworthy finding is that, despite retaining only about 8 to 15% of ssLP fragments on average after selection, *A. fumigatus* cfDNA levels remained elevated across the five BAL fluid samples and the remaining two plasma samples (from patients A01 and A02), even after a ~12-fold reduction in sequencing coverage. Notably, neither the maximum possible amount of cfDNA from collected samples nor the highest achievable cfDNA sequencing depth in multiplexed libraries was attained, as indicated by a median PCR duplicate rate of 7.1 (IQR: 6.2–10.3). This suggests that with additional sequencing and more input, even samples with currently undetectable or low fungal DNA could show increased abundance, potentially surpassing the background levels observed in the control patient set, provided that the same efforts in input and sequencing are applied to both. These observations underscore that further optimization — by increasing cfDNA input, refining size selection after library preparation in the wet lab, and enhancing sequencing depth — could improve the capture and sequencing of unique *Aspergillus*-derived fragments.

Effect of library preparation on host chromosomal cfDNA fragment ends

We next mapped host-derived cfDNA end-motifs — the nucleotide sequences at the terminal ends of fragments — and examined how e.g. library preparation methods influence the relative observed abundance of these motifs. This analysis is crucial for informing strategies to enrich fungal signals, such as through end-motif dependent host cfDNA depletion following library preparation. Typically, these motifs are defined by the final end-nucleotide, the last two nucleotides (2-mers), or the last four nucleotides (4-mers) at the fragment ends[15–18].

Our analysis revealed significant differences in end-nucleotide composition between ssLP and dsLP preparations in plasma samples (Supplementary Fig. 6a; Kruskal-Wallis test, $p < 0.05$), as well as for certain end-nucleotides in BAL samples (Supplementary Fig. 6b). Since we performed paired-end sequencing, we were able to assess both the 5′ end of double-stranded fragments (after blunting) generated by dsLP, as well as the 5′ and 3′ ends of single-stranded molecules captured by ssLP. Generally, however, the end-nucleotide composition of read 1 (R1) was more comparable between both library preparation methods, while read 2 (R2) exhibited more pronounced differences between dsLP and dsLP (Supplementary Fig. 6a-b). This trend was further supported in our analysis of 2-mer end-motifs (Supplementary Fig. 6b-c), where end-motif differences between ssLP and dsLP in R2 are likely caused by the end-repair processes in the dsLP method[7,16]. These findings highlight the substantial impact of library preparation on the observed fragment ends. Specifically, the dsLP method fills 3' single-stranded overhangs using 5'→3' polymerase activity and trims 5' overhangs with 3'→5' exonuclease activity, whereas the ssLP method, which omits these repair steps, preserves the native ends of cfDNA (see schematic of library preparations in Supplementary Fig. 1).

A closer examination revealed that in plasma, C-ends were the most abundant (Supplementary Fig. 6a-b), consistent with previous studies[19], except for R2 in cfDNA prepared using ssLP. The presence of a second cytosine, a CC-end, was even more enriched (Supplementary Fig. 6c-d). Conversely, other 2-mers, such as AG- and TG-ends were significantly more prevalent in R2 from ssLP-prepared plasma cfDNA. In BAL samples, the enriched 2-mers exhibited considerable variability among individual samples, which was more apparent after ssLP (Supplementary Fig. 6d). Furthermore, all samples showed notable reduced frequency of CG-ends (Supplementary Fig. 6c-d), in line with the low 1.1% abundance of CG dinucleotides in the human genome (as determined for the chm12v2.0, by our in-house analysis). These analyses indicate that also genomic content affects end-motif composition.

Comparison end-motif landscape of *Aspergillus*- versus host-derived cfDNA

We next investigated whether the observed differences in end-motifs can be leveraged for *Aspergillus* enrichment. Previous research has demonstrated that cfDNA fragments derived from microbes, particularly pathogenic bacteria, exhibit distinct end-motif preferences[15]. To investigate fungal enrichment potential through end-motifs, we analyzed the theoretical ratio of *A. fumigatus* to host-derived reads in R1 or R2 of our IPA patient samples, selecting reads based on their 1-mer end-motif — A, T, C, or G. Our analysis revealed significant differences in the *A. fumigatus* : host ratio for reads with specific 1-mer end-nucleotides when normalized and compared to the whole cfDNA pool with no end-nucleotide selection (Supplementary Fig. 8). Notably, a combination of fragments starting with either C or G — referred to as the "C&G" end-nucleotide selection — generally yielded the highest *A. fumigatus* : host ratio. A significantly increased *A. fumigatus* : host ratio was observed following ssLP and "C&G" selection, in both R1 and R2 reads (Supplementary Fig. 8b). This trend was also evident in some dsLP libraries, although less consistently over R1 and R2 (Supplementary Fig. 8b). These *in silico* findings suggest that *in vitro* end-motif selection might present a promising strategy for pathogen cfDNA enrichment, particularly in ssLP and certain dsLP samples.

Expanding the analysis to dinucleotide start motifs, we observed that fragments starting with CG exhibited a substantial increase in the *A. fumigatus* : host ratio, particularly in BAL samples. The enrichment potential of the CG motif seems to reflect intrinsic differences in nucleotide abundance between host genome and the *Aspergillus fumigatu*s genome, where CG dinucleotides account for 1.1% and 5.4%, respectively (based on our analysis of the Chm13v2.0 human reference genome and the *A. fumigatus Af293* consensus genome). This implies that refining start-motif selection to CG motifs could enhance *Aspergillus* abundance (Supplementary Fig. 9).

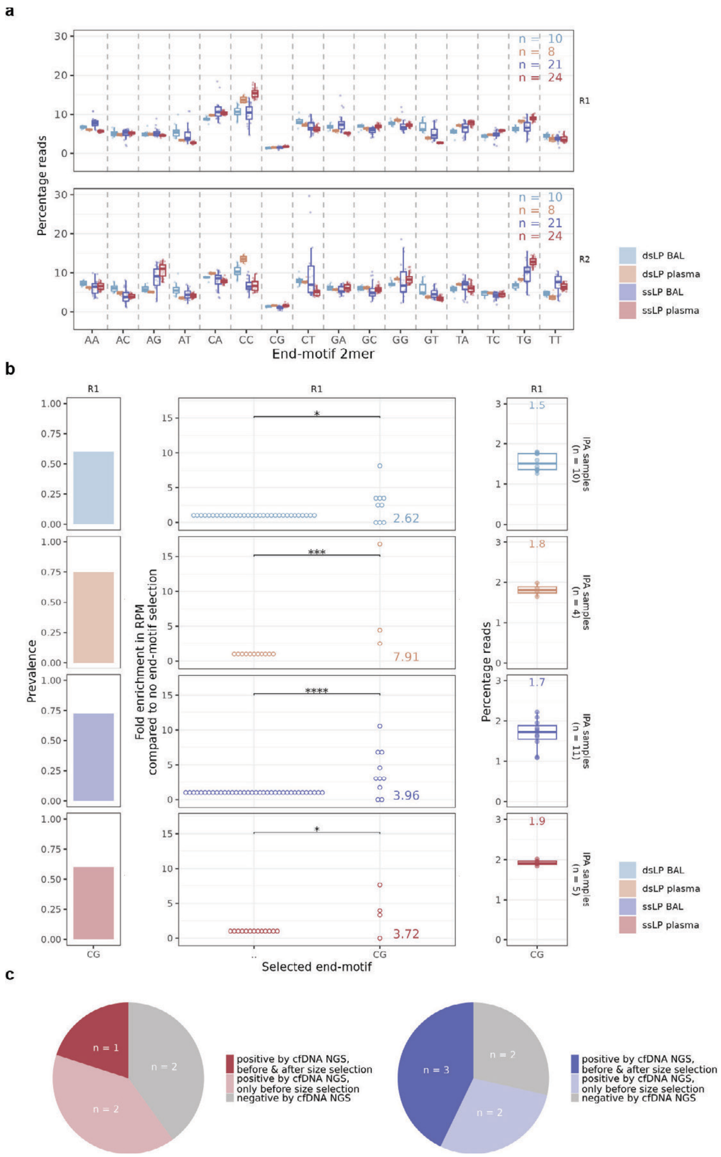


Fig. 3. Enrichment of *Aspergillus fumigatus* cfDNA using 'CG' end-motif selection

a. 2-mer end-motif composition in read 1 (R1, top) and read 2 (R2, bottom) of BAL and plasma following ssLP or dsLP. **b.** This figure illustrates the results of *in silico* end-motif selection for enriching cfDNA fragments from *Aspergillus fumigatus* by targeting fragments that begin with the nucleotide sequence \`CG\` at the p5 side of the adapter (read 1). Left panel: Panel displays the effect of \`CG\` motif selection on the prevalence of *Aspergillus fumigatus* detection. It shows the number of samples with at least one read detected post-ssLP, compared to the total number of samples showing an *A. fumigatus* signal in the absence of end-motif selection. Middle panel: Dotplot illustrates the fold enrichment of *Aspergillus fumigatus* in terms of reads per million (RPM) of quality-controlled reads. The data are presented relative to the abundance observed without \`CG\` end-motif selection. Right panel: Dotplot panel quantifies the percentage of reads retained following \`CG\` end-motif selection. Mean values are indicated in panels middle and right panel. **c.** Effect of 'CG' end-motif selection on diagnostic yield using cfDNA NGS in plasma (left) and BAL (right).

Considering the SRSLY-mediated ssLP design, start-motif selection could theoretically be implemented *in vitro* by simply substituting the random overhang in the split adapter sequence with an initial di-nucleotide complementary sequence in the bottom strand of the P5 adapter. Upon examining the *A. fumigatus* RPM after *in silico* CG end-motif selection, we indeed observed an average 3.9-fold increase across libraries (Fig. 3b). Approximately 1.5-2.0% of the cfDNA pool exhibited a CG end-motif (Read 1 in Fig. 3b; for other motifs and Read 2, see Supplementary Fig. 10), indicating that it was retained *in silico*. However, end-motif selection also resulted in a decrease in overall prevalence (i.e., the detection of at least one read; Fig. 3b). Consequently, the end-motif selection reduced the number of cases with increased *A. fumigatus* abundance by 50%, with only 1/5 plasma and 3/7 BAL samples from IPA patients (diagnosed according to the EORTC/MSG criteria; for details, see Wesdorp *et al.* (under review) showing significantly elevated levels (mean pairwise Fisher's exact test, $p > 0.001$; Fig. 3c) compared to immunocompromised controls. In summary, while CG-based end-motif selection may yield a modest (~3-fold) increase in *Aspergillus* relative abundance, it also substantially reduces detection sensitivity in our IPA patient samples. This suggests that end-motif fragment selection would require processing or sequencing larger volumes of material to make this approach viable for fungal diagnostic application.

End-motif host chromosomal cfDNA fragment ends might be related to patient status

We observed variability in chromosome end-motif frequencies across samples prepared with the same method, prompting us to investigate whether a patient's immunocompromised status might influence fragment ends. Hierarchical clustering of 2-mer frequencies showed that plasma samples from the same patient, especially A01 and A02, clustered more closely than those from different patients, like A05 (Supplementary Fig. 11a-b). This suggests that end-motif frequency similarities are stronger within individual patients, possibly influenced by conditions such as hematological or immunocompromised status.

Further assessment aimed at determining whether fragment ends could differentiate host statuses showed that end-motif clustering did, however, not distinguish mold-infected samples (IPA and other molds) from controls in plasma (Supplementary Fig. 11). In contrast, BAL ssLP samples from IPA patients — previously shown to be *Aspergillus*-positive through shotgun sequencing in our prior work (Wesdorp *et al.*, under review) — did exhibit clustering (Supplementary Fig. 11c-d). The clustering of three out of five IPA BAL samples with a positive cfDNA NGS test suggests that variability in BAL end-motifs, largely driven by library preparation, may also correlate with clinical status. Further research is needed to fully investigate this potential relationship.

Discussion

This study aimed to enhance the chance of detecting *Aspergillus fumigatus* through liquid biopsy cfDNA NGS, while minimizing the sequencing of host cfDNA to reduce associated costs. Although our long-term goal is to develop wet lab methods for enriching pathogen-derived cfDNA, we here report on initial tests with multiple library preparation techniques, combined with *in silico* size and end-motif selection, using plasma and BAL fluid samples from immunocompromised patients with invasive IPA. Our key findings are that *Aspergillus fumigatus*-derived cfDNA exhibits distinct fragment length distributions compared to human cfDNA, with a higher proportion of shorter fragments (≤ 100 bp) in both plasma and BAL samples. Our analysis further indicated that size selection after ssLP could enrich *A. fumigatus* cfDNA by up to 8-fold in plasma and 3-fold in BAL, suggesting potential for improving fungal detection in immunocompromised patients. It should be noted that these findings await *in vitro* validation. While size selection appears advantageous only after ssLP, 'CG'-based end-motif enrichment demonstrated promise after both ssLP and dsLP. This approach provided a modest, yet consistent, ~3-fold enrichment for *A. fumigatus*-derived cfDNA.

A major limitation of the selection strategies is that some IPA samples no longer demonstrated increased *A. fumigatus* cfDNA abundance above the background level of the external immunocompromised control patients. For example, following 'CG' end-motif selection, we observed a 50% reduction in IPA liquid biopsies with detectable elevated *A. fumigatus* levels. While this decrease is undesirable for diagnostics, it is important to note that these results were obtained after a >50-fold reduction in selected reads (as only 1.5–2.0% of reads contained a CG motif at the P5-adaptor side), *in silico*. This drastic reduction in reads also reduced the statistical power of our pairwise analyses between IPA and control samples. Hence, it is now crucial to test the presented approaches *in vitro* while simultaneously increasing sequencing depth (which will now be directed at the more interesting size ranges and end-motifs), to identify the full potential of size- and end-motif selection. Despite this, diagnosis remained possible for half of the patients, suggesting potential for its application, provided that increasing input material for size selection and sequencing can recover some of the diagnostic potential lost during this reduction, while still maintaining lower sequencing costs.

Interestingly, our findings also suggest that sequencing could be reduced by up to 85% in BAL samples following ssLP and size selection, without compromising diagnostic yield. If such reductions are maintained through *in vitro* size selection, platforms like Oxford Nanopore Technology may emerge as viable alternatives to current Illumina-based testing. While size selection may slightly increase wet lab costs and processing time, the substantial reduction in sequencing costs makes this approach attractive, improving accessibility in smaller clinics or low(er)-income countries. Additionally, ONT offers the advantage of faster time-to-results, enabling more rapid sequential testing for patients.

While our fragmentomics characterization work shows promising potential for enrichment through fragment selection, it is limited by factors such as small sample size and challenges in detecting low-abundance *Aspergillus* DNA in shallow-depth libraries. These limitations prevented an assessment of the combined effects of size- and 'CG' end-motif selection, which warrants further study. Limitations also emerged concerning the impact of sample storage on BAL cfDNA length — confirmed again to be longer than in plasma samples[11] — and end-motif signatures, as uncontrolled storage conditions (e.g., room temperature without DNase inhibition) likely contributed to greater variability in length profiles and end-motifs for BAL samples compared to plasma samples[20,21]. Future research should thus prioritize *in vitro* testing with controlled sample storage and increased sequencing depth to assess diagnostic yield. Additionally, broadening the focus beyond library preparation to include cfDNA isolation from biofluids[8], potential biases in the sequencing process, and computational factors — such as classification biases related to fragment size — could be advantageous. Finally, as mentioned, it is worthwhile to combine these approaches with platforms like ONT sequencing.

In summary, this study demonstrated the presence of highly fragmented *Aspergillus*-derived cfDNA in human plasma specimens. These findings underscore the value of size selection in increasing the relative abundance of fungal cfDNA, which could enable the detection of pathogens that might otherwise remain undetected due to their low abundance in shotgun next-generation sequencing for fungal detection.

Dataset

We utilized shotgun Illumina sequencing data, with written informed consent provided for participation in the biobank (International Clinical Trials Registry Platform: NL7744;https://onderzoekmetmensen.nl/en/trial/21619), which is unrelated to the current study presented here. Data will be made available on reasonable request, with the EGA number pending. Briefly, for 45 samples (24 plasma, 21 BAL) of 32 patients, 2x150 bp Illumina sequencing data was available, whereby sequencing libraries were generated using two different approaches. For ssLP the SRSLY PicoPlus NGS Library Prep Kit (Claret Bioscience, CBS-K250B-96) had been employed. In short, this method was applied to 45 samples (24 plasma; 21 BAL) according to the manufacturer's instruction for the Moderate Fragment Retention protocol. For dsLP the KAPA Library Preparation Kit (Roche) had been employed. For 18 samples (8 plasma; 10 BAL), with some small modifications to the manufacturer's protocol (Wesdorp *et al.*, under review).

For upstream analysis, we utilized the cfSPI-pipeline, as previously described (Wesdorp *et al.*, under review), to deduplicate and remove low-quality or low-complexity reads (incl. the removal of < 35 bp reads), followed by host-read identification through host genome mapping using *Bowtie2* (v2.5.1) and taxonomic classification of non-human mapped reads, using *kraken2* (v2.1.2). In more details: for host genome mapping, we used both the reference genome *CHM13v2* and a combined reference comprising of the *GRCh38.p14* and *CHM13v2* genomes. Mapping exclusively to *CHM13v2* was conducted for host cfDNA fragmentomics analysis. In contrast, dual mapping to *GRCh38.p14* and *CHM13v2* (the *default* approach) was implemented to minimize false-positive microbial identifications, as discussed in our previous work (Wesdorp *et al.*, under review). Following dual mapping, we employed the *EPRSc2* database and Kraken2 with a classification threshold of 0.4 (for more details, see https://github.com/AEWesdorp/ESCALA/tree/main/scripts/01_PreProcessing).

Fragmentomics analysis host versus *Aspergillus fumigatus*: length profiling and end-motif after mapping

***Host-derived reads**:*

After mapping the *CHM13v2* host-reference genome using the cfSPI pipeline, reads aligned to host chromosomal contigs or mitochondrial contig were extracted with *Samtools* (v1.3.1).

Following mapping, first and second reads were separated using *Samtools* (v1.3.1). Reads with a MAPQ score below 30 (-q 30) were filtered out, along with secondary (-F 256) and supplementary alignments (-F 2048). We extracted the cfDNA template length (TLEN) and end-motif information (3 bp) and quantified their abundance using a custom script, processing R1 and R2 reads separately.

For details, see:
https://github.com/AEWesdorp/ESCALA/blob/main/scripts/01_PreProcessing/workflow/Snakemake_PreProcessing.

***Aspergillus fumigatus*-derived reads:**

Reads classified by Kraken2 as *Aspergillus fumigatus* (NCBI:txid746128) or its daughter taxa (using the cfSPI pipeline, employing the *EPRSc2* database with a classification threshold of 0.4) were retrieved from non-host-mapped read files using a custom script. These reads were then mapped to the *A. fumigatus Af293* strain (CBS126847).

Following mapping, first and second reads were separated using *Samtools* (v1.3.1). Reads with a MAPQ score below 30 (-q 30) were filtered out, along with secondary (-F 256) and supplementary alignments (-F 2048). We extracted the cfDNA template length (TLEN) and end-motif information (3 bp) and quantified their abundance using a custom script, processing R1 and R2 reads separately.

For details see:
https://github.com/AEWesdorp/ESCALA/tree/main/scripts/02_mapAspergillus/.

For circulating cfDNA study, which typically targets short DNA molecules, we excluded longer cfDNA fragments (>500 bp) using a custom R script (v4.2.1). Reads shorter than 35 bp were filtered out during the quality control steps of our cfSPI pipeline. Additionally, reads containing 'N' in their end-motif were removed, and for reads with a negative TLEN, we obtained the reverse complement of their 3 bp end-motif. End-motifs were defined with the first base as the outermost for both R1 and R2 reads. For additional details and code on these post-pressing steps performed for both host and *Aspergillus fumigatus* mapped reads, see:

https://github.com/AEWesdorp/ESCALA/blob/main/figures/scripts/PP_host_MT_TLEN_EndMotif.ipynb and

https://github.com/AEWesdorp/ESCALA/blob/main/figures/scripts/PP_Aspergillus_TLEN_EndMotif.ipynb, respectively.

Length periodicity analysis

To quantify the periodicity as observed in the fragment length analysis, we used a measure similar to the periodicity index (*PI*) as described in work from the group of Dennis Lo[22]. For the identification of peaks and valleys in the oscillating length pattern, we thereby used the *pracma* package (with *minpeakdistance* = 7, *npeaks* = 25, *threshold* = 0). Of note, different valleys and peaks were identified in different liquid biopsy samples by different library preparations (for annotation see Fig. 1b and Supplementary Fig. 2a, peaks (black points) and valleys (white points)).

Subsequently, we calculated the relative difference between the valley and its surrounding peak, only for sub-150 bp valleys, and only for those valleys for which both peaks where within less than 9 bp distant from the valley, through the following formula:

$$RPM = totalNumAspergillusfumigatusReads / totalNumQCReads * 1x10^6$$

Where the *Vi* is the frequency of cfDNA fragments at a particular length valley *i*. And *Pil* is the frequency at the left valley relative to the peak *i*, whereby the distance between *Pil* and *Vi* is less than 9 bp. *Pir* is the frequency at the right peak relative to the valley *i*, whereby the distance between *Vi* and *Pil* is less than 9 bp. Finally, we computed the mean of all calculated PIs to summarize the periodicity across samples.

In silico size selection

For upstream analysis, we employed the cfSPI-pipeline to deduplicate and filter out low-quality or low-complexity reads, including the removal of reads < 35 bp reads. Following quality control, we identified the taxonomic origin of the reads using Kraken2, specifically targeting *Aspergillus fumigatus*. We calculated the abundance of *A. fumigatus* in reads per million (RPM) using the formula:

$$PI=(Pil + Pir) / (Vi * 2)$$

Next, we employed the *bbmerge.sh* script from the BBMap suite (v39.01) to merge paired-end reads and estimate fragment lengths (for success rates see Supplementary Fig. 6c). Merged reads were selected based on size ranges of 35 to 50, 75, 100, 125 or 150 bp for further analysis. To evaluate the impact of *in silico* size selection on the relative abundance of *Aspergillus fumigatus*, we quantified both the number of quality-controlled reads and the number classified as *A. fumigatus* by Kraken2 within the same size range. The relative abundance of *A. fumigatus* in each sample was then expressed as RPM, calculated using the formula mentioned above. We also calculated fold-enrichment by normalizing RPM values obtained prior to size selection.

In silico end-motif selection

For upstream analysis, we employed the cfSPI-pipeline, to deduplicate and filter out low-quality or low-complexity reads, including the removal of reads < 35 bp reads. Following quality control, we identified the taxonomic origin of the reads using Kraken2, specifically targeting *Aspergillus fumigatus*. As previously described we obtained the *Aspergillus fumigatus* end-motif information after mapping the Kraken2 classified reads to the *A. fumigatus Af293* strain (CBS126847). The end-motif data for the quality-controlled reads was extracted from the FASTQ files using a custom script, available at: https://github.com/AEWesdorp/ESCALA/blob/main/scripts/04_nucleotideFreq/.

After excluding all reads containing unknown nucleotides in their end-motifs (i.e., any nucleotide other than A, T, C, or G), we calculated the abundance of *A. fumigatus* in the same was as described in the methods section *"In silico* size selection".

This calculation was performed both before and after selection for the CG end-motif.

Effect of size or end-moti selection on diagnostic yield

To identify elevated levels of *Aspergillus fumigatus* after taxonomic classification of all clinical samples, we performed a one-tailed Fisher's exact test. This test assessed statistical differences in the read counts of *Aspergillus fumigatus* between our patient samples and external control samples, based on two counts in contingency tables:

The number of *Aspergillus fumigatus* reads (including all lower-ranking taxa within the same clade) after size selection or end-motif selection.

The number of reads remaining after filtering for duplicates, low quality, and low complexity, as well as after size selection or end-motif selection, excluding those classified as *Aspergillus fumigatus*.

The significance level was set at $p \leq 0.001$, calculated as the mean of all Fisher's exact tests conducted across the samples. This analysis aimed to identify meaningful differences in *Aspergillus fumigatus* read counts between patient samples and immunocompromised external controls.

From our previous work we know that, without size selection or end-motif selection, 5/7 IPA BAL samples tested positive, while 3/5 IPA plasma samples showed statistically increased *Aspergillus fumigatus* levels, when compared to controls (Wesdorp *et al.*, under review). In this study, we assessed the diagnostic yield in relation to these earlier findings.

2-mer similarity analysis of host chromosomal end-motifs

To assess the 2-mer end-motif composition of host chromosomal-derived reads across samples, we applied normalization by overall end-motif frequency before performing hierarchical clustering using Euclidean distance. The observed 2-mer frequencies were then compared to expected values, with the expected 2-mer frequency set at 6.25%. A log10 transformation was applied, followed by centering and scaling of each end-motif across all samples.

Statistics

To evaluate the increase in *Aspergillus fumigatus* cfDNA fractions, expressed in reads per million (RPM), after selecting for ssLP compared to dsLP, we conducted a one-tailed Wilcoxon rank-sum test with Bonferroni correction. Similarly, for assessing the increase in *Aspergillus fumigatus* cfDNA fractions after end-motif selection, we also performed a one-tailed Wilcoxon rank-sum test with Bonferroni correction. Additionally, we used Kruskal-Wallis tests for multiple comparisons to conduct a non-directional, non-parametric ANOVA to assess differences in cfDNA end-motifs between ssLP and dsLP.

Software

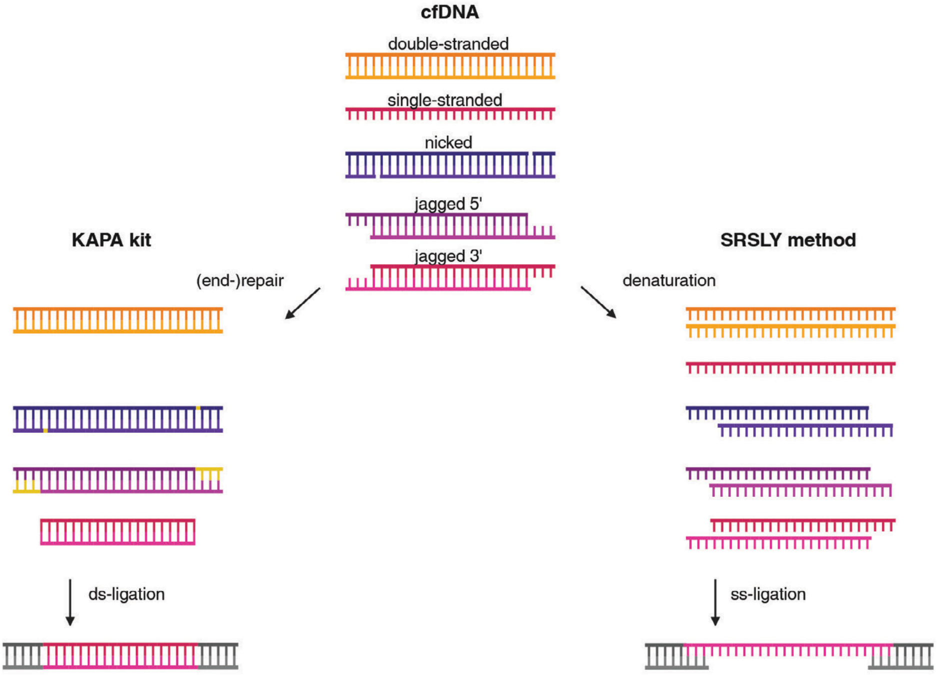
Data and statistical analyses were performed using R (v. 4.2.0). Figures were generated in R (v. 4.2.0), while illustrations were created using BioRender and Adobe Illustrator (2024, v. 28.6).

References

1. Snyder, M. W., Kircher, M., Hill, A. J., Daza, R. M. & Shendure, J. Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues-of-origin. *Cell* **164,** 57–68 (2016).
2. Sun, K., Jiang, P., Chan, K. C. A., Wong, J., Cheng, Y. K. Y., Liang, R. H. S., Chan, W.-K., Ma, E. S. K., Chan, S. L., Cheng, S. H., Chan, R. W. Y., Tong, Y. K., Ng, S. S. M., Wong, R. S. M., Hui, D. S. C., Leung, T. N., Leung, T. Y., Lai, P. B. S., Chiu, R. W. K. & Lo, Y. M. D. Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. *Proc. Natl. Acad. Sci. U. S. A.* **112,** E5503–12 (2015).
3. Moss, J., Magenheim, J., Neiman, D., Zemmour, H., Loyfer, N., Korach, A., Samet, Y., Maoz, M., Druid, H., Arner, P., Fu, K.-Y., Kiss, E., Spalding, K. L., Landesberg, G., Zick, A., Grinshpun, A., Shapiro, A. M. J., Grompe, M., Wittenberg, A. D., Glaser, B., Shemer, R., Kaplan, T. & Dor, Y. Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. *Nat. Commun.* **9,** 5068 (2018).
4. Mattox, A. K., Douville, C., Wang, Y., Popoli, M., Ptak, J., Silliman, N., Dobbyn, L., Schaefer, J., Lu, S., Pearlman, A. H., Cohen, J. D., Tie, J., Gibbs, P., Lahouel, K., Bettegowda, C., Hruban, R. H., Tomasetti, C., Jiang, P., Chan, K. C. A., Lo, Y. M. D., Papadopoulos, N., Kinzler, K. W. & Vogelstein, B. The origin of highly elevated cell-free DNA in healthy individuals and patients with pancreatic, colorectal, lung, or ovarian cancer. *Cancer Discov.* **13,** 2166–2179 (2023).
5. Burnham, P., Kim, M. S., Agbor-Enoh, S., Luikart, H., Valantine, H. A., Khush, K. K. & De Vlaminck, I. Single-stranded DNA library preparation uncovers the origin and diversity of ultrashort cell-free DNA in plasma. *Sci. Rep.* **6,** 27859 (2016).
6. Jiang, P., Chan, C. W. M., Chan, K. C. A., Cheng, S. H., Wong, J., Wong, V. W.-S., Wong, G. L. H., Chan, S. L., Mok, T. S. K., Chan, H. L. Y., Lai, P. B. S., Chiu, R. W. K. & Lo, Y. M. D. Lengthening and shortening of plasma DNA in hepatocellular carcinoma patients. *Proc. Natl. Acad. Sci. U. S. A.* **112,** E1317–25 (2015).
7. Cheng, J. C., Swarup, N., Wong, D. T. W. & Chia, D. A review on the impact of single-stranded library preparation on plasma cell-free diversity for cancer detection. *Front. Oncol.* **14,** 1332004 (2024).
8. Chang, A., Mzava, O., Lenz, J. S., Cheng, A. P., Burnham, P., Motley, S. T., Bennett, C., Connelly, J. T., Dadhania, D. M., Suthanthiran, M., Lee, J. R., Steadman, A. & De Vlaminck, I. Measurement biases distort cell-free DNA fragmentation profiles and define the sensitivity of metagenomic cell-free DNA sequencing assays. *Clin. Chem.* **68,** 163–171 (2021).
9. Phung, Q., Lin, M. J., Xie, H. & Greninger, A. L. Fragment size-based enrichment of viral sequences in plasma cell-free DNA. *J. Mol. Diagn.* **24,** 476–484 (2022).
10. Kisat, M. T., Odenheimer-Bergman, A., Markus, H., Joseph, B., Srivatsan, S. N., Contente-Cuomo, T., Khalpey, Z., Keim, P., O'Keeffe, T., Askari, R., Salim, A., Rhee, P. & Murtaza, M. Plasma metagenomic sequencing to detect and quantify bacterial DNA in ICU patients suspected of sepsis: A proof-of-principle study: a proof-of-principle study. *J. Trauma Acute Care Surg.* **91,** 988–994 (2021).
11. Nair, V. S., Hui, A. B.-Y., Chabon, J. J., Esfahani, M. S., Stehr, H., Nabet, B. Y., Zhou, L., Chaudhuri, A. A., Benson, J., Ayers, K., Bedi, H., Ramsey, M., Van Wert, R., Antic, S., Lui, N., Backhus, L., Berry, M., Sung, A. W., Massion, P. P., Shrager, J. B., Alizadeh, A. A. & Diehn, M. Genomic profiling of bronchoalveolar lavage fluid in lung cancer. *Cancer Res.* **82,** 2838–2847 (2022).
12. Bokelmann, L., Glocke, I. & Meyer, M. Reconstructing double-stranded DNA fragments on a single-molecule level reveals patterns of degradation in ancient samples. *Genome Res.* **30,** 1449–1457 (2020).
13. Kostyuk, S., Smirnova, T., Kameneva, L., Porokhovnik, L., Speranskij, A., Ershova, E., Stukalov, S., Izevskaya, V. & Veiko, N. GC-rich extracellular DNA induces oxidative stress, double-strand DNA breaks, and DNA damage response in human adipose-derived mesenchymal stem cells. *Oxid. Med. Cell. Longev.* **2015,** 782123 (2015).
14. Sanchez, C., Roch, B., Mazard, T., Blache, P., Dache, Z. A. A., Pastor, B., Pisareva, E., Tanos, R. & Thierry, A. R. Circulating nuclear DNA structural features, origins, and complete size profile revealed by fragmentomics. *JCI Insight* **6,** (2021).
15. Wang, G., Lam, W. K. J., Ling, L., Ma, M.-J. L., Ramakrishnan, S., Chan, D. C. T., Lee, W.-S., Cheng, S. H., Chan, R. W. Y., Yu, S. C. Y., Tse, I. O. L., Wong, W. T., Jiang, P., Chiu, R. W. K., Allen Chan, K. C. & Lo, Y. M. D. Fragment ends of circulating microbial DNA as signatures for pathogen detection in sepsis. *Clin. Chem.* **69,** 189–201 (2023).
16. Jiang, P., Sun, K., Peng, W., Cheng, S. H., Ni, M., Yeung, P. C., Heung, M. M. S., Xie, T., Shang, H., Zhou, Z., Chan, R. W. Y., Wong, J., Wong, V. W. S., Poon, L. C., Leung, T. Y., Lam, W. K. J., Chan, J. Y. K., Chan, H. L. Y., Chan, K. C. A., Chiu, R. W. K. & Lo, Y. M. D. Plasma DNA end-motif profiling as a fragmentomic marker in cancer, pregnancy, and transplantation. *Cancer Discov.* **10,** 664–673 (2020).
17. Jin, C., Liu, X., Zheng, W., Su, L., Liu, Y., Guo, X., Gu, X., Li, H., Xu, B., Wang, G., Yu, J., Zhang, Q., Bao, D., Wan, S., Xu, F., Lai, X., Liu, J. & Xing, J. Characterization of fragment sizes, copy number aberrations and 4-mer end motifs in cell-free DNA of hepatocellular carcinoma for enhanced liquid biopsy-based cancer detection. *Mol. Oncol.* **15,** 2377–2389 (2021).

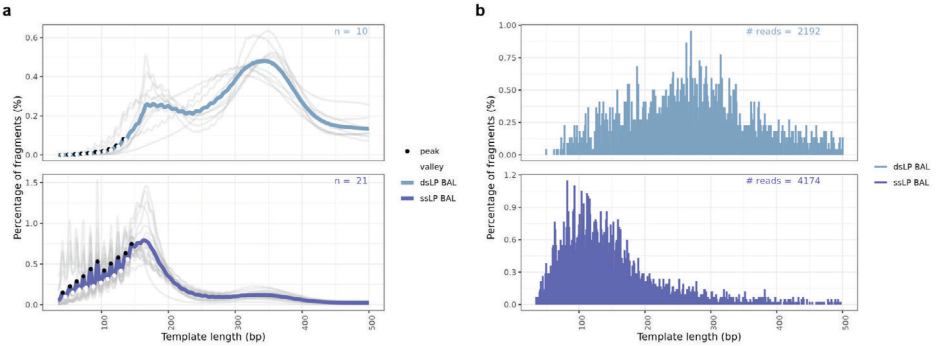
18. Serpas, L., Chan, R. W. Y., Jiang, P., Ni, M., Sun, K., Rashidfarrokhi, A., Soni, C., Sisirak, V., Lee, W.-S., Cheng, S. H., Peng, W., Chan, K. C. A., Chiu, R. W. K., Reizis, B. & Lo, Y. M. D. Dnase1l3 deletion causes aberrations in length and end-motif frequencies in plasma DNA. *Proc. Natl. Acad. Sci. U. S. A.* **116,** 641–649 (2019).
19. Han, D. S. C., Ni, M., Chan, R. W. Y., Chan, V. W. H., Lui, K. O., Chiu, R. W. K. & Lo, Y. M. D. The biology of cell-free DNA fragmentation and the roles of DNASE1, DNASE1L3, and DFFB. *Am. J. Hum. Genet.* **106,** 202–214 (2020).
20. Mouliere, F. A hitchhiker's guide to cell-free DNA biology. *Neurooncol. Adv.* **4,** ii6–ii14 (2022).
21. Malentacchi, F., Pizzamiglio, S., Verderio, P., Pazzagli, M., Orlando, C., Ciniselli, C. M., Günther, K. & Gelmini, S. Influence of storage conditions and extraction methods on the quantity and quality of circulating cell-free DNA (ccfDNA): the SPIDIA-DNAplas External Quality Assessment experience. *Clin. Chem. Lab. Med.* **53,** 1935–1942 (2015).
22. Xie, T., Wang, G., Ding, S. C., Lee, W.-S., Cheng, S. H., Chan, R. W. Y., Zhou, Z., Ma, M.-J. L., Han, D. S. C., Teoh, J. Y. C., Lam, W. K. J., Jiang, P., Chiu, R. W. K., Chan, K. C. A. & Lo, Y. M. D. High-resolution analysis for urinary DNA jagged ends. *NPJ Genom. Med.* **7,** 14 (2022).
23. Troll, C. J., Kapp, J., Rao, V., Harkins, K. M., Cole, C., Naughton, C., Morgan, J. M., Shapiro, B. & Green, R. E. A ligation-based single-stranded library preparation method to analyze cell-free DNA and synthetic oligos. *BMC Genomics* **20,** 1023 (2019).

Supplementary Figures



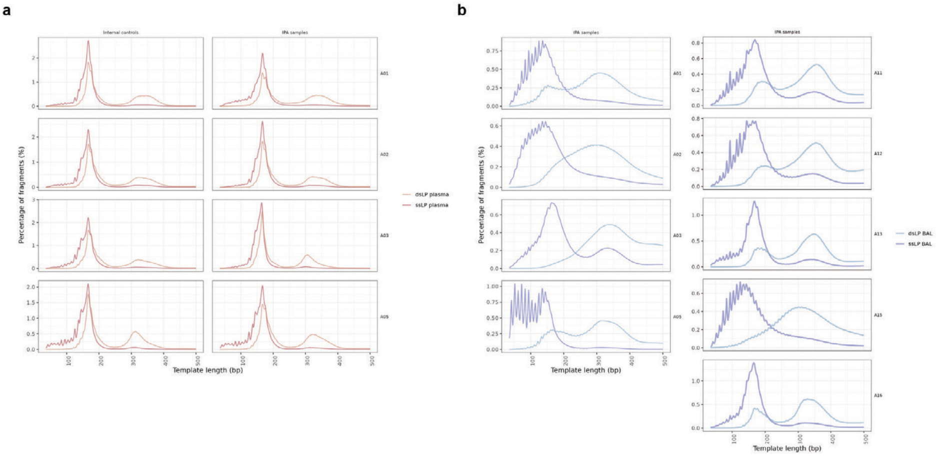
Supplementary Fig. 1. Schematic representation of library preparation methods: KAPA Kit vs. SRSLY method

Circulating free DNA (cfDNA) exists in various forms within liquid biopsies, including jagged, nicked, single-stranded, and double-stranded molecules. Several methods have been developed to capture cfDNA for sequencing library preparation, enabling analysis via Illumina sequencing. In this paper, we generated sequencing libraries using either the KAPA kit for double-stranded library preparation (dsLP) or the SRSLY method for single-stranded library preparation (ssLP). The KAPA kit-mediated dsLP repairs nicks and jagged ends to facilitate double-stranded adapter ligation (ds-ligation; fill-in repair depicted in yellow). The end-repair fills in 3’ single-stranded overhangs using 5’→3’ polymerase activity and trims 5’ single-stranded overhangs with 3’→5’ exonuclease activity. In contrast, the ssLP method starts with the denaturation of cfDNA, subsequently enabling all single-stranded molecules to ligate to a random-sequence splint adapter (ss-ligation). This approach preserves the native ends of cfDNA fragments by avoiding blunt-end repair. By bypassing these steps, the ssLP method retains both the length of the original cfDNA fragments and their native end nucleotides.



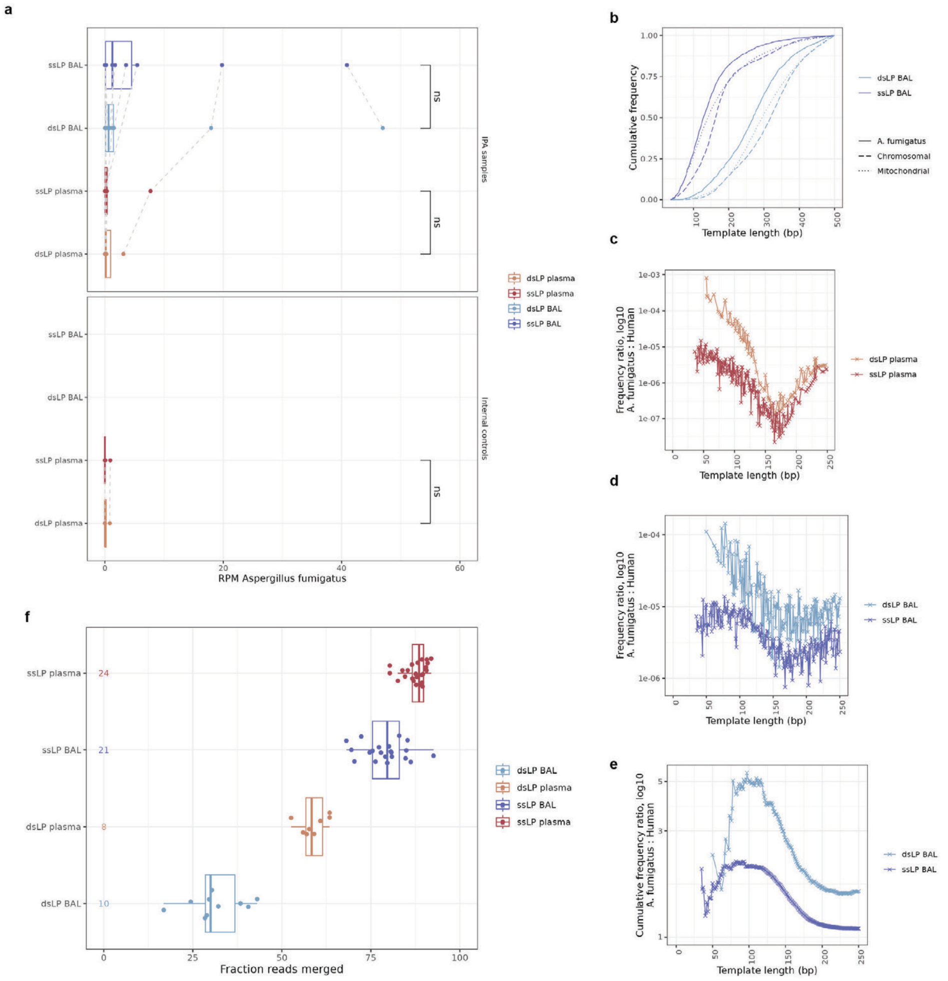
Supplementary Fig. 2. Assessment of cfDNA fragment lengths in BAL: human chromosomal and *Aspergillus fumigatus* after ssLP and dsLP

a. Distribution of human chromosomal cfDNA fragment lengths in BAL fluid samples after ssLP and dsLP preparation, shown as an average density plot (blue) over 21 ssLP and 10 dsLP specimens (grey). Our periodicity analysis revealed variable valleys (white points) and peaks (black points) within the sub-nucleosomal size range. **b.** Percentage of *Aspergillus fumigatus*-derived cfDNA fragments per 1-bp length bin in BAL fluid samples, following ssLP and dsLP preparation. Data is pooled from all samples, including 21 ssLP and 10 dsLP BAL samples. The total number of reads classified as *A. fumigatus* across all dsLP or ssLP is displayed in the upper-right corner.



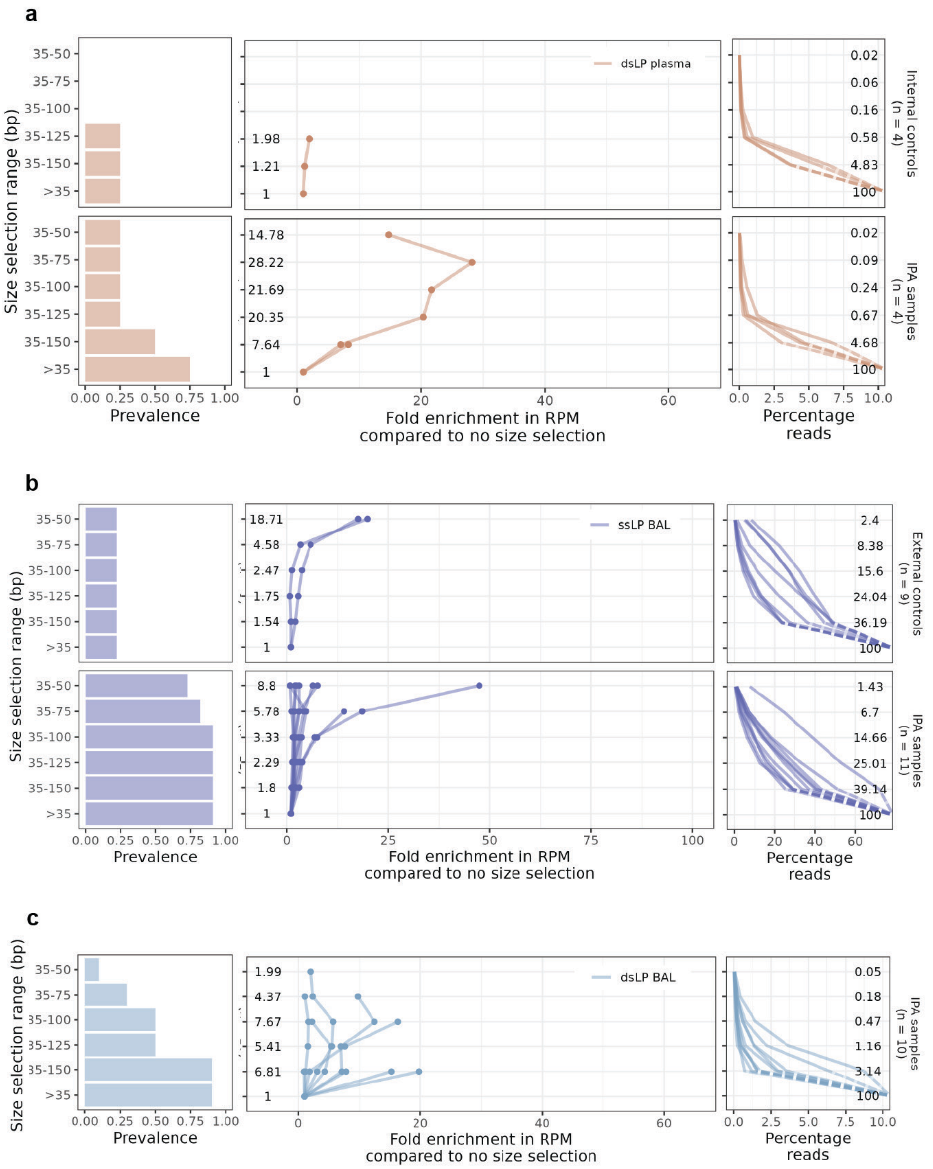
Supplementary Fig. 3. Comparative analysis of cfDNA fragment length distributions in plasma and BAL fluid samples from (IPA) patients after ssLP or dsLP

a. Human chromosomal cfDNA fragment length distribution in plasma samples from patients A01, A02, A03, and A05, after ssLP or dsLP preparation, shown as a density plot. Samples were derived prior to (left) or during (right) the IPA episode. **b.** Human chromosomal cfDNA fragment length distribution in BAL fluid samples, after ssLP or dsLP preparation, from several patients, presented as a density plot. All samples were derived during the IPA episode.



Supplementary Fig. 4. Abundance and length of *Aspergillus fumigatus* cfDNA fragments, in relation to host cfDNA

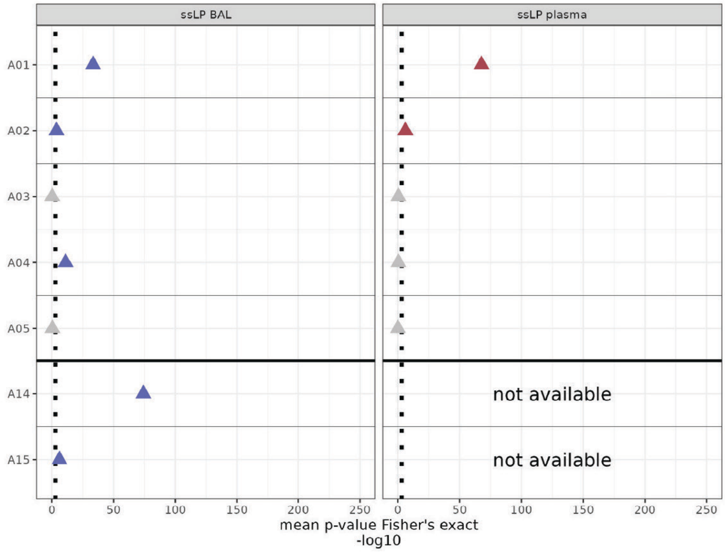
a. Boxplots showing the *Aspergillus fumigatus* fractional abundance in RPM, determined by Kraken2's database cRE.21, with a taxonomic classification threshold at 0.4 (Wesdorp *et al.*, under review). This analysis was conducted for both IPA patients and internal control samples whenever both conditions were present. Statistical significance between ssLP and dsLP is evaluated through a one-tailed Wilcoxon rank test with Bonferroni correction (ns, $p > 0.05$). **b.** Cumulative distribution plot of *Aspergillus fumigatus*- and human-derived (chromosomal and mitochondrial) cfDNA fragment lengths by ddLP and ssLP in BAL. **c-d.** Frequency ratios of *A. fumigatus* to host chromosomal cfDNA in **c.** plasma and **d.** BAL after ssLP and dsLP. **e.** Cumulative frequency ratios of *A. fumigatus* to host chromosomal cfDNA in BAL after ssLP and dsLP. In **b-e.** lines are colored by sample type and library preparation. **f.** Boxplot displaying the fraction of quality-controlled reads successfully merged by *bbmerge.sh* (see *Methods* for details), used to assess cfDNA fragment lengths and evaluate the impact of in silico size selection.



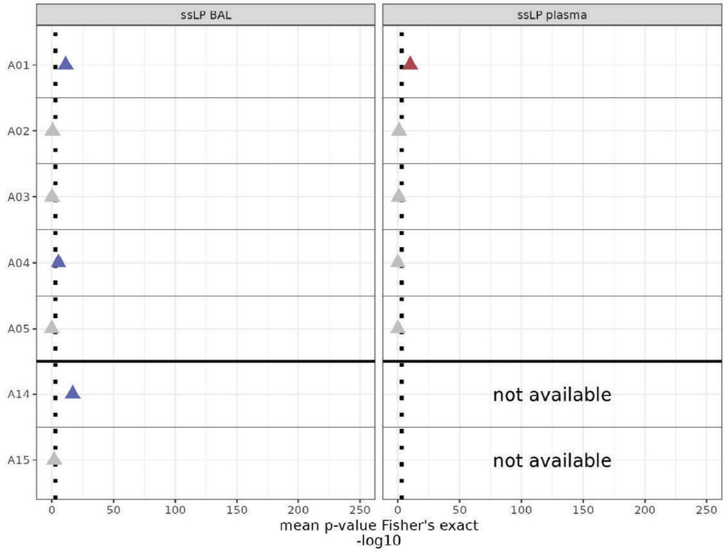
Supplementary Fig. 5. *Effect of* in silico *size selection* on dsLP plasma as well as on ssLP and dsLP BAL

Effect of *in silico* size selection on cfDNA in **a.** plasma samples after dsLP, **b.** BAL samples after ssLP, and **c.** BAL samples after dsLP, showing outcomes for selecting fragments up to 50, 75, 100, 125, or 150 bp, or with no size selection (>35). Left panel: prevalence of *Aspergillus fumigatus* detection, indicating the presence of at least one read post-ssLP with/without size selection. Middle panel: Fold enrichment in the relative abundance of *A. fumigatus*, expressed in reads per million (RPM) of quality-controlled reads, relative to its abundance without *in silico* size selection. Right panel: percentage of quality-controlled reads retained following *in silico* size selection. A dotted line approximates the trend leading up to 100% when no size selection was applied (>35).

a

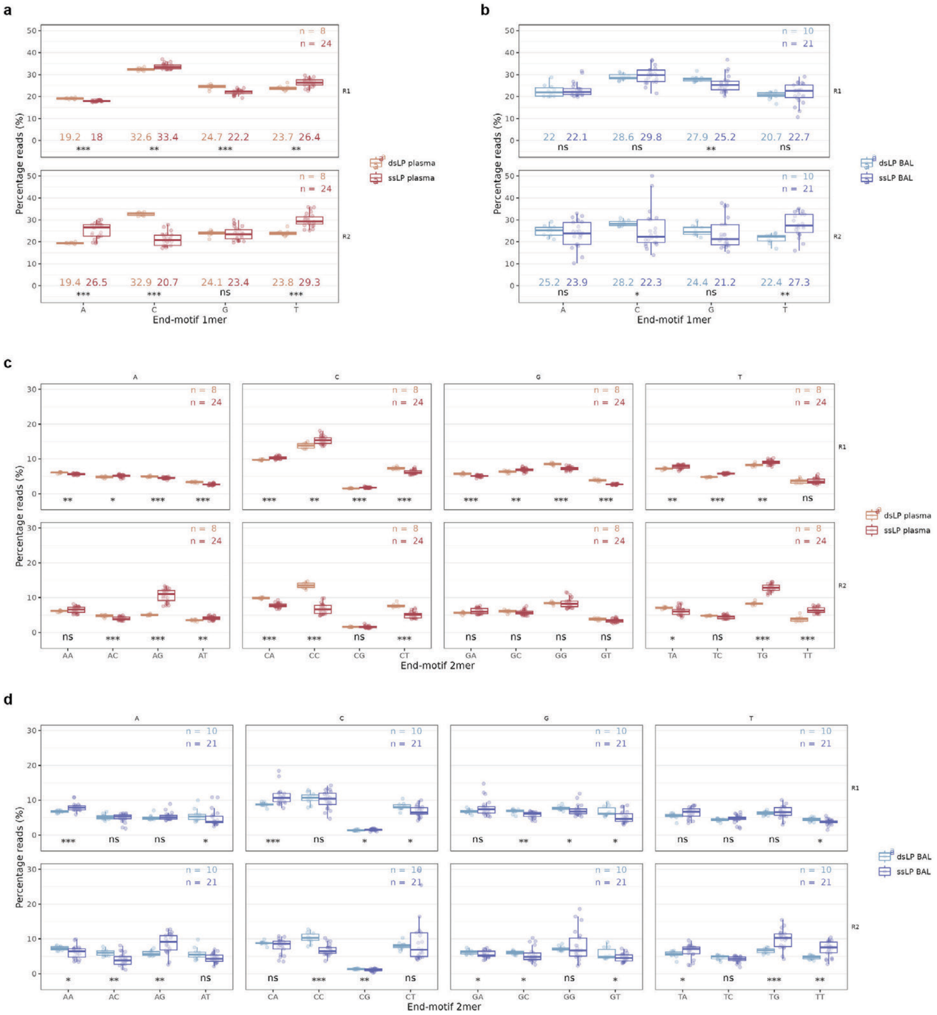


b



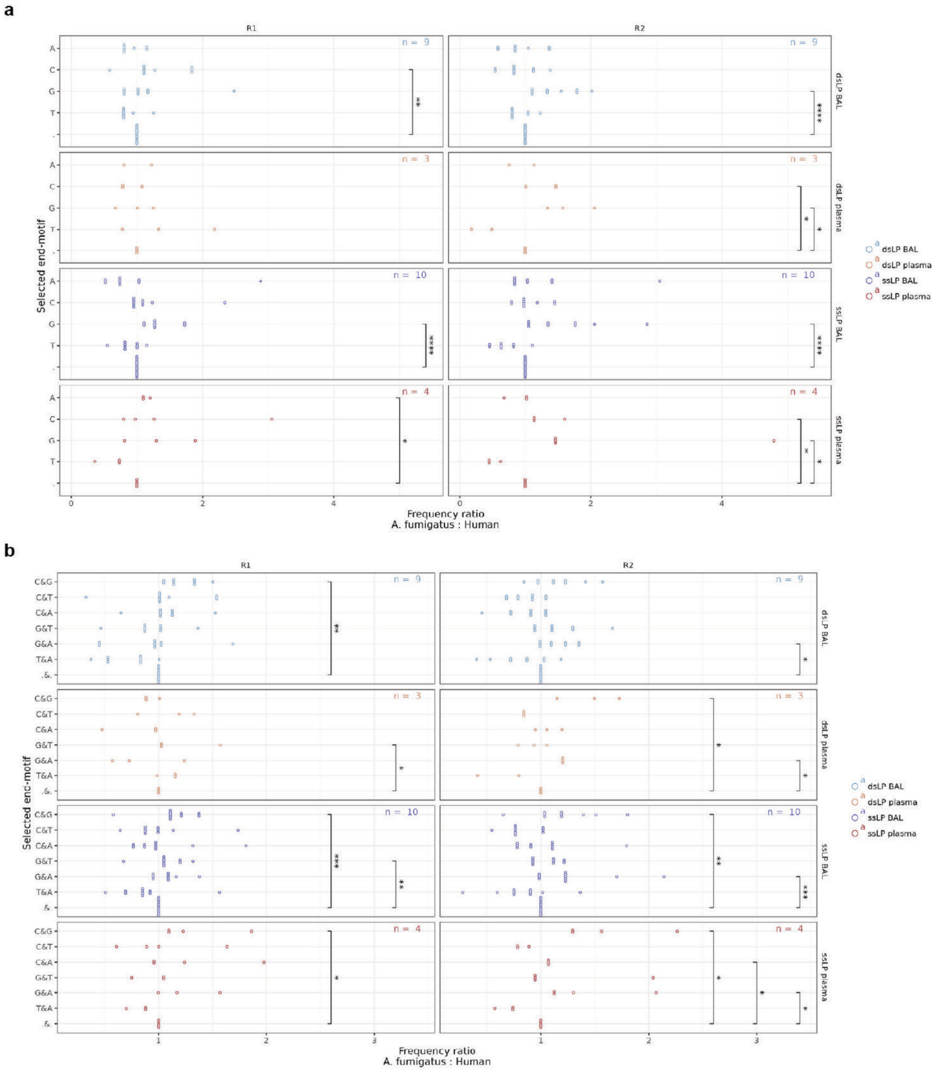
Supplementary Fig. 6. Elevated Aspergillus fumigatus levels in IPA patient samples processed with ss-cfDNA NGS and size- or 'CG' end-motif selection

To compare the fractional abundance of *Aspergillus fumigatus* in patient samples with invasive pulmonary aspergillosis (IPA) to those from external immunocompromised pediatric controls, a paired one-tailed Fisher's exact test was performed. The results are displayed as dot plots representing the mean -log10-transformed p-values, with a significance threshold set at p = 0.001 (vertical dotted line). Points surpassing this threshold are highlighted in color: blue for BAL samples in the (left panel) and red for plasma samples (right panel). **a.** Shows data after size selection (up to 100 bp), and **b.** displays data after read 1 'CG' end-motif selection.



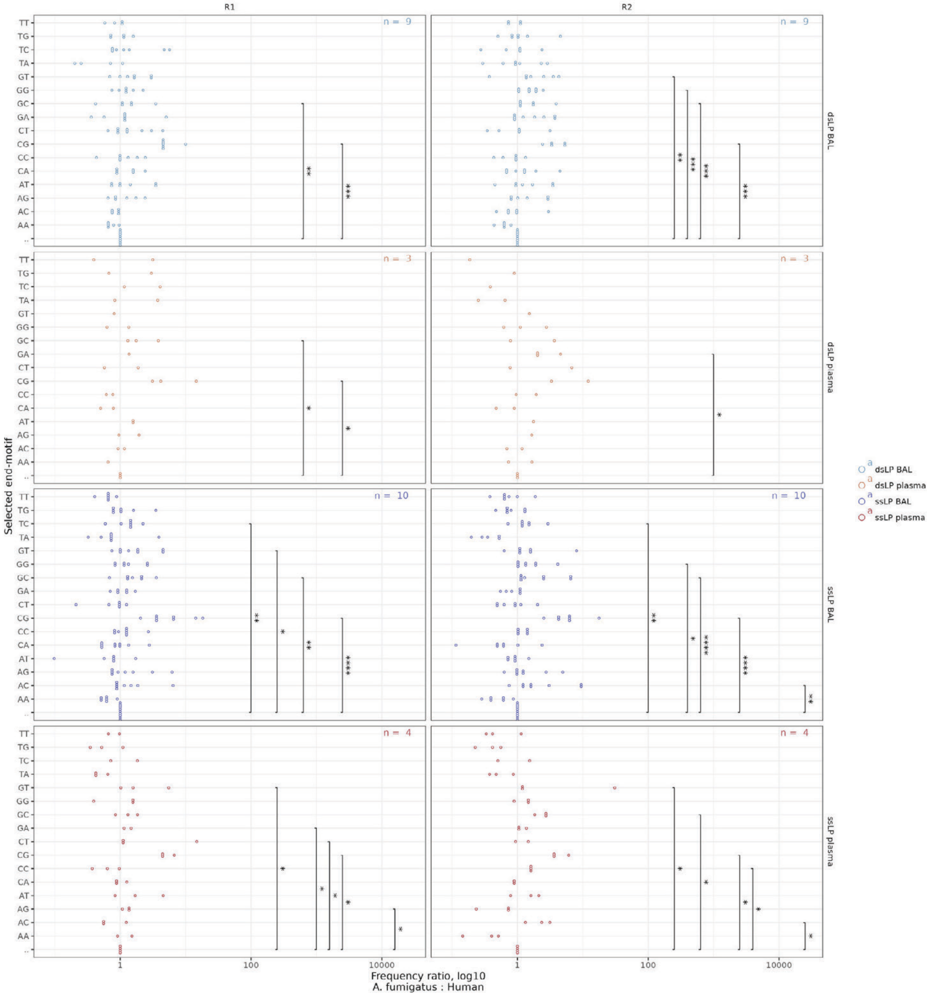
Supplementary Fig. 7. Effects of ssLP and dsLP on cfDNA end-motis in plasma and BAL

a. 1-mer end-nucleotide composition in read 1 (R1, top) and read 2 (R2, bottom) of plasma cfDNA fragments after ssLP (n=24) and dsLP (n=8), displayed as a boxplot. **b.** 1-mer end-nucleotide composition in read 1 (R1, top) and read 2 (R2, bottom) of BAL cfDNA fragments after ssLP (n=21) and dsLP (n=10), displayed as a boxplot. **c.** 2-mer end-motif composition in read 1 (R1, top) and read 2 (R2, bottom) of plasma cfDNA fragments after ssLP (n=24) and dsLP (n=8), displayed as a boxplot. **d.** 2-mer end-motif composition in read 1 (R1, top) and read 2 (R2, bottom) of BAL cfDNA fragments after ssLP (n=21) and dsLP (n=10) displayed as a boxplot. In **a-d,** each dot represents a single sample.



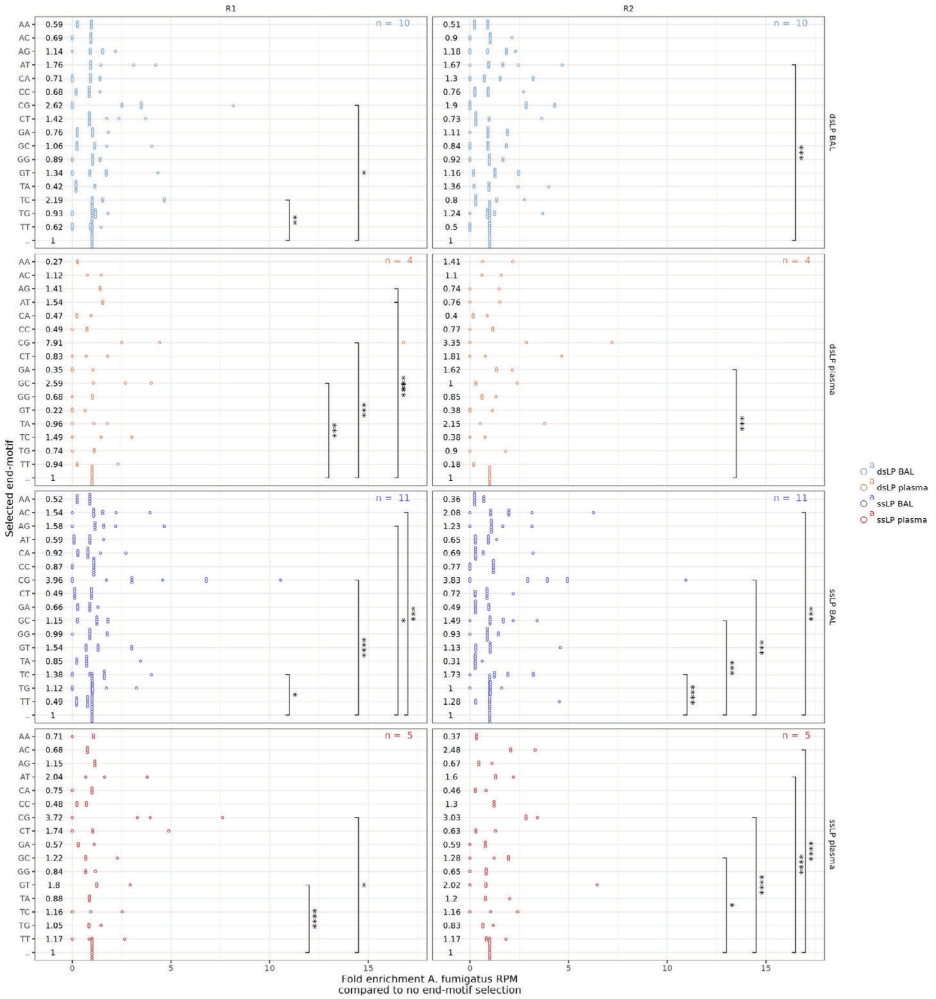
Supplementary Fig. 8. Impact of end-nucleotide selection on *A. fumigatus* to host read ratios

a. The ratio of *A. fumigatus* to host-derived reads in each sequenced library stratified by starting nucleotide, is shown. Values are normalized and compared to the ratio observed without end-motif selection (denoted by "." on the y-axis). **b.** The ratio of *A. fumigatus* to host-derived reads when selecting a combination of two starting nucleotides, such as for example C or G — referred to as the "C&G" end-nucleotide selection on the y-axis — is shown for each sequenced library. These ratio values are normalized and compared to the abundance without end-motif selection (denoted by ".&." on the y-axis). **a-b.** Each dot in the plot represents a library, illustrating results with and without end-motif selection after normalization. A one-sided Wilcoxon test was performed to evaluate enrichment due to end-motif selection compared to the abundance ratio without selection (*, $p \leq 0.05$; **, $p \leq 0.01$; ***, $p \leq 0.001$; ****, $p \leq 0.0001$); ns indicates $p > 0.05$ and is not shown in the figure.



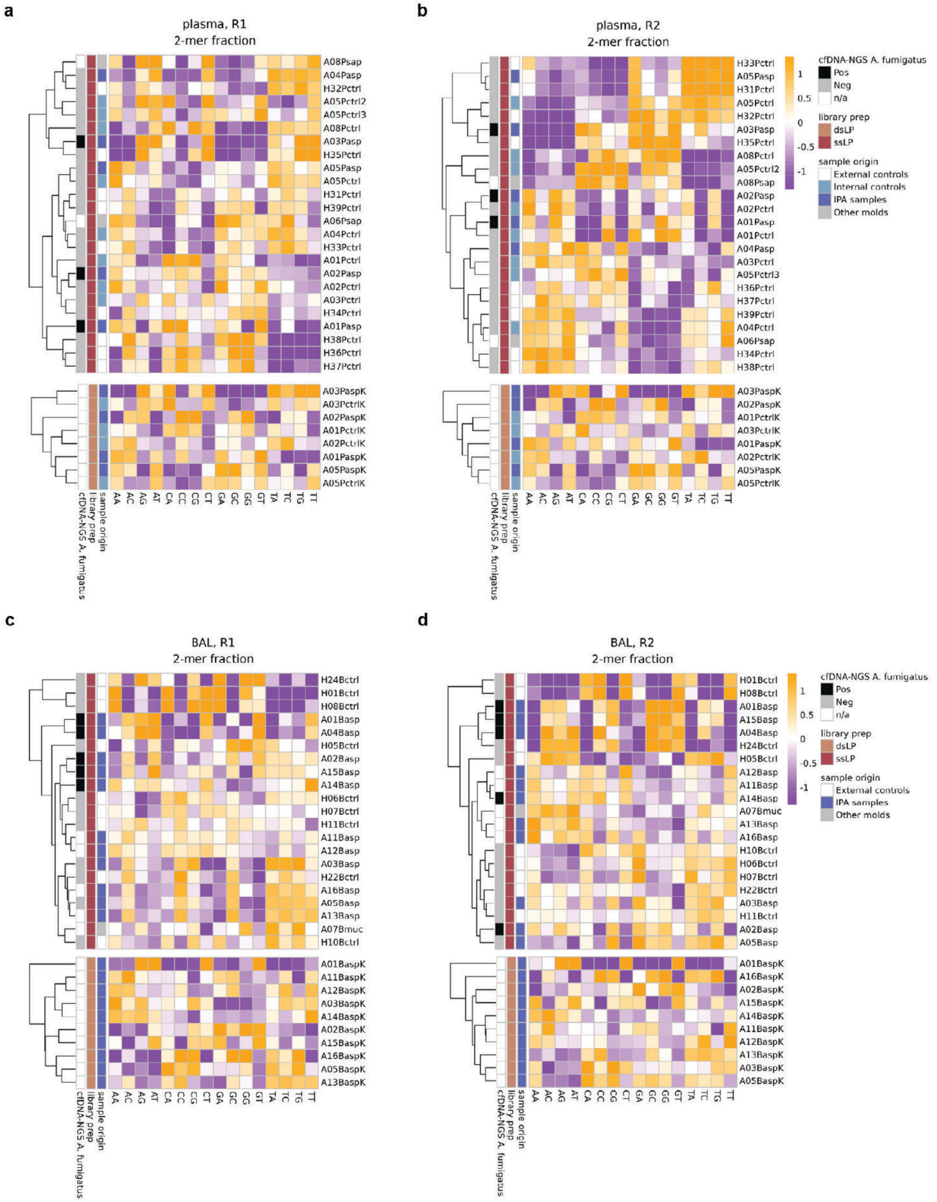
Supplementary Fig. 9. Impact of dinucleotide end-motif selection on *A. fumigatus* to host read ratios

Ratio of *A. fumigatus* to host-derived reads in each sequenced library, filtered by di-nucleotide end-motif, compared to abundance without end-motif selection (denoted by ".." on the y-axis). Each dot in the dot plot represents a library, illustrating results with and without end-motif selection after normalization. A one-sided Wilcoxon test was performed to evaluate enrichment due to end-motif selection compared to the abundance ratio without selection (*, p ≤ 0.05; **, p ≤ 0.01; ***, p ≤ 0.001; ****, p ≤ 0.0001); ns indicates p > 0.05 and is not shown in the figure.



Supplementary Fig. 10. Impact of dinucleotide end-motif selection on *A. fumigatus* relative abundance

This figure illustrates the relative abundance of *A. fumigatus* in each sequenced library, filtered by dinucleotide end-motif, with abundance expressed in reads per million (RPM) and normalized to the RPM of *A. fumigatus* without end-motif selection (denoted by ".." on the y-axis). The left panels presents results for read 1 (R1), while the right panels shows results for read 2 (R2). Each dot in the dot plot represents a library, displaying the results with and without end-motif selection after normalization. Mean values are indicated per end-motif, in both reads. A one-sided Wilcoxon test was used to assess the enrichment from end-motif selection relative to the abundance ratio without selection (*p ≤ 0.05; **p ≤ 0.01; ***p ≤ 0.001; ****p ≤ 0.0001); values with p > 0.05 are not shown (ns).



Supplementary Fig. 11. Hierarchical clustering of BAL fluid samples based on normalized 2-mer end-motif frequencies of human chromosomal fragments

Hierarchical clustering of **a-b.** plasma and **c-d.** BAL fluid samples based on normalized 2-mer end-motif frequencies of human chromosomal fragments at the start of either **a,c.** Read 1 (R1) or **b,d.** Read 2 (R2). The 2-mer composition was normalized to expected values assuming equal prevalence, log10-transformed, and then centered and scaled per end-motif before clustering using Euclidean distance.

Supplementary Table 1

sample id	number of reads
A01Basp	98,321,639
A01BaspK	55,609,143
A01Pasp	106,310,487
A01PaspK	65,829,490
A01Pctrl	132,958,863
A01PctrlK	76,168,618
A02Basp	103,778,807
A02BaspK	71,972,971
A02Pasp	122,536,249
A02PaspK	70,952,319
A02Pctrl	95,631,106
A02PctrlK	72,569,835
A03Basp	63,114,840
A03BaspK	94,522,010
A03Pasp	134,086,014
A03PaspK	95,250,616
A03Pctrl	94,998,548
A03PctrlK	84,416,488
A04Basp	122,592,500
A04Pasp	108,073,315
A04Pctrl	140,908,754
A05Basp	56,298,735
A05BaspK	96,255,822
A05Pasp	50,569,414
A05PaspK	57,636,593
A05Pctrl	64,649,290
A05Pctrl2	128,470,500
A05Pctrl3	92,668,605
A05PctrlK	71,219,334
A06Psap	118,453,883
A07Bmuc	128,885,353
A08Pctrl	90,757,643
A08Psap	144,080,679
A11Basp	57,733,250
A11BaspK	42,993,752
A12Basp	51,116,400
A12BaspK	34,449,639
A13Basp	67,638,855
A13BaspK	52,796,500
A14Basp	44,846,443
A14BaspK	37,388,970
A15Basp	87,102,780
A15BaspK	126,031,442
A16Basp	99,108,366
A16BaspK	23,036,239
H01Bctrl	60,335,786

Supplementary Table 1. *Continued*

sample id	number of reads
H05Bctrl	42,663,948
H06Bctrl	57,578,929
H07Bctrl	48,147,577
H08Bctrl	67,629,827
H10Bctrl	60,411,430
H11Bctrl	62,251,070
H22Bctrl	117,178,912
H24Bctrl	112,487,405
H31Pctrl	65,567,338
H32Pctrl	52,153,511
H33Pctrl	61,083,618
H34Pctrl	218,198,338
H35Pctrl	57,851,704
H36Pctrl	134,010,898
H37Pctrl	125,224,364
H38Pctrl	114,299,419
H39Pctrl	131,389,878

CHAPTER 4

Bacterial cell-free DNA profiling reveals co-elevation of multiple bacteria in newborn foals with suspected sepsis

Li-Ting Chen [1,2,*], Emmy Wesdorp [1,2,*], Myrthe Jager [1,2], Esther W. Siegers [3], Mathijs J.P. Theelen [3], Nicolle Besselink [1,2], Carlo Vermeulen [1,2], Aldert L. Zomer [4], Els M. Broens [4], Jaap A. Wagenaar [4,5], Jeroen de Ridder [1,2]

[1] Center for Molecular Medicine, University Medical Center Utrecht, Utrecht University, 3584 CX Utrecht, The Netherlands

[2] Oncode Institute, 3521 AL, Utrecht, the Netherlands

[3] Department of Clinical Sciences, Faculty of Veterinary Medicine, Utrecht University, 3584 CM Utrecht, The Netherlands

[4] Department of Biomolecular Health Sciences, Faculty of Veterinary Medicine, Utrecht University, 3584 CL Utrecht, the Netherlands

[5] Wageningen Bioveterinary Research, 8221 RA Lelystad, The Netherlands

*These authors contributed equally to this work

Provisionally accepted in iScience

Abstract

Background: Sepsis is the leading cause of death in newborn foals. This study investigates whether cell-free DNA (cfDNA) sequencing can enhance bacterial pathogen detection in foals with suspected sepsis and addresses existing knowledge gaps and diagnostic challenges.

Methods: We developed a **f**oal cfDNA sequencing for **b**acteria **i**dentification (cf**FBI**) workflow, integrating wetlab and computational protocols to detect increased bacterial cfDNA abundance in blood. Specifically, cfFBI focusses on enriching bacterial cfDNA molecules and preventing false positive bacterial identifications. cfFBI was applied to blood samples of 25 hospitalized foals categorized according to the neonatal Systemic Inflammatory Response Syndrome (nSIRS) criteria and 7 healthy foals.

Results: cfDNA levels of potential sepsis-causing bacterial genera were elevated in all 11 nSIRS-positive foals compared to healthy foals (n=7), and in 8/11 (72.7%) when compared to both nSIRS-negative (n=4; nSIRS=0) and healthy foals, with multiple genera elevated in 5/11 (45.5%). The total cfDNA concentration, bacterial cfDNA fraction and bacterial diversity were not different between the foal groups. However, nSIRS-positive foals showed significantly different end-motifs in host chromosomal cfDNA, and a decrease in host mitochondrial cfDNA fraction.

Conclusions: This study is the first to demonstrate that cfDNA sequencing in blood samples from newborn foals enables detection of pathogenic bacteria and can help identify novel host-related sepsis biomarkers. The elevated presence of multiple sepsis-causing genera in nSIRS-positive foals and the difference in end-motif, suggests that multibacterial elevation may be more common than previously thought. These findings indicate that cfDNA sequencing holds promise as a future diagnostic tool for identifying sepsis in newborn foals.

Introduction

Sepsis is defined as “a life-threatening organ dysfunction caused by a dysregulated host response to infection”, hallmarked by the systemic inflammatory response syndrome (SIRS) and often caused by a bacterial infection [1,2]. SIRS arises when pathogen- and damage-associated molecular patterns (PAMPs and DAMPs), as well as neutrophil extracellular traps (NETs), are recognized by the immune system [3,4]. Dysregulated innate and adaptive immune responses, in combination with overactivation of the coagulation system can result in multiple organ dysfunction, followed by multiple organ failure, and ultimately resulting in death [5–7]. In newborn foals, sepsis stemming from a bacterial infection is an important cause of morbidity and mortality during the first week of life [8–10]. Due to the rapid progression of sepsis, early recognition, prompt identification of the causative bacterial pathogen, and timely initiation of effective antimicrobial therapy are critical for improving survival rates[11].

Despite being a common cause of death in newborn foals, knowledge gaps persist regarding the pathogenesis, diagnosis, and treatment of sepsis. For instance, multiple bacteria are known to co-occur in human sepsis patients [12,13], and a similar phenomenon is likely in newborn foals [11,14]. However, traditional culture methods often lack the sensitivity to detect such co-infections. Additionally, when multiple bacteria are cultured from a single sample it is frequently dismissed as contamination in clinical settings. A further complication in understanding sepsis pathogenesis and diagnostics is that newborn foals absorb immunoglobulins from the colostrum over the gastrointestinal barrier, during which bacteria (including those that can cause sepsis) can also enter the bloodstream [15]. Although transient bacteremia is a normal physiological process, it remains unclear why in some foals this might lead to SIRS and sepsis, while in the majority it does not.

Early sepsis diagnosis and timely identification of the causative microbe(s) remains challenging in foals with current diagnostic tools. To aid prompt identification of foals at risk for sepsis, several scoring systems have been developed. These systems use either four (SIRS) or six (neonatal SIRS, nSIRS) objective clinical criteria (Supplementary Table 1) to identify foals that potentially have sepsis [2]. However, these scoring systems have limited sensitivity (SIRS 60%; nSIRS 42%) and specificity (SIRS 69%; nSIRS 76%) for detecting neonatal sepsis [2]. Bacterial infection, the other hallmark of sepsis, is typically identified through blood cultures, which enable bacteriological identification and subsequent antimicrobial susceptibility profiling. However, the sensitivity of bacterial detection through culture is only 25-45% in foals with sepsis [16–18]. Quantitative PCR (qPCR) systems have a higher sensitivity (87%) [19–21], but are only able to detect a finite set of pathogens, leading to false negative results for pathogens not included in the test. Additionally, false positive results can occur in both culture and qPCR in case of transient bacteremia or sample contamination [22]. As a result of the low sensitivity and specificity of current diagnostic tools, many newborn foals with sepsis remain undiagnosed or misdiagnosed. Given that foals can deteriorate rapidly within hours, there is an urgent need for improved diagnostic tools for earlier clinical intervention. Thus, expanding diagnostic capabilities and enhancing our understanding of sepsis-causing bacteria in foals is essential.

Cell-free DNA (cfDNA) are short DNA fragments found in body fluids, including plasma, which are released upon cell and microorganism death [23]. In human medicine, sequencing of plasma microbial cfDNA shows great promise for detecting bacterial pathogens in conditions including sepsis [12,24,25]. The advantages of cfDNA short-read sequencing include its culture-independent nature, a reasonable turnaround time of 2-3 days (with the potential to speed this to less than one day using alternative sequencing platform [26]), and the ability to facilitate the unbiased discovery of new pathogens that have not been previously cultured [24,25]. High-throughput sequencing of cfDNA may also reveal general differences in microbial composition in plasma associated with disease development [12,24]. The abundance and characteristics such as fragment lengths, fragment end-motifs, and mapping locations of cfDNA molecules, can reveal information on physiological and pathological processes such as the immune response [27–29]. In human plasma, approximately 99.5% of the cfDNA originates from the host [30], which are typically ~167 bp in length [31,32]. Microbial cfDNA is shorter, with a substantial fraction being smaller than 100bp in plasma [33–35]. Taxonomic classification and quantification of the microbial cfDNA provides a multi-pathogen, minimally invasive, accurate assay for diagnosing sepsis in humans [24,25], with cfDNA end-motifs potentially enhancing the overall diagnostic process [36,37].

In this study, our primary objective is to assess the potential of blood cfDNA sequencing for detecting elevated cfDNA levels of bacteria associated with sepsis in newborn foals with nSIRS. The secondary objectives focus on analyzing the overall cfDNA bacterial composition and investigating host cfDNA factors, including their origin and end-motif. While previous research has focused on cfDNA concentrations in septic foals [38,39], sequencing of cfDNA has not been previously conducted, marking this research as a pioneering effort in the field. This endeavor prompted us to create a specialized **f**oal cfDNA sequencing for **b**acterial **i**dentification (cf**FBI**) workflow, incorporating both wetlab and open-sourced computational workflows optimized for detecting pathogenic bacteria in foals with suspected sepsis. By applying the cfFBI pipeline to 32 newborn foals we aim to assess the viability of this approach not only as an alternative diagnostic tool, but also as a method to deepen the understanding of the pathophysiology associated with nSIRS.

Results

cfDNA sequencing in newborn foals with sepsis using cfFBI

To investigate the potential of cfDNA sequencing in the context of equine neonatal sepsis, we prospectively included 25 newborn sick foals admitted to Utrecht University Equine Hospital in The Netherlands as well as seven healthy newborn foals (H) (Fig. 1a; Supplementary Table 2, 3). All foals included in this study were between 0 and 6 days of age (Supplementary Tables 2-3). Based on the nSIRS criteria (Supplementary Table 1) [2], 11 of these foals were nSIRS-positive (S+; nSIRS≥3), four were nSIRS-negative with zero positive nSIRS parameters (nS-; nSIRS=0), and 10 were nSIRS-negative but had one or two positive nSIRS parameters (sS-; nSIRS=1-2) (Fig. 1a; Supplementary Table 3). Three of the eleven S+ (27%), two of the four nS- (50%) and three of the ten sS- (30%) foals had a positive bacterial blood culture (Supplementary Table 3). We focussed the analysis on comparing S+ against nS- and/or H foals, not sS- foals, as the presence of clinical sepsis-related signs in the sS-group (Supplementary Table 3) suggest that some of these foals could have a bacterial infection or even sepsis. nS- and H foals together represent a realistic clinically relevant background, especially as all nS- samples are derived from the same hospital setting as the S+ samples, ensuring that we account for potential biases related to sample handling and environmental factors.

To enable assessment of elevated levels of pathogenic bacteria in the S+ population, we created a cfFBI wetlab and computational workflow (Fig. 1a-c). Given that the microbial cfDNA fraction is known to be minute in plasma [33], cfFBI employs a specialized wetlab strategy to enrich for bacterial cfDNA molecules while remaining untargeted in multi-pathogen detection. The cfFBI wetlab cfDNA workflow therefore consists of a ligation-based single-stranded cfDNA library preparation method followed by a bead-based size selection step which effectively enriches short (<100bp) fragments (Fig. 1b-e), which is the known size range for bacterial cfDNA [40].

For all foal plasma cfDNA sequencing libraries, between 10 and 50 million paired-end reads were obtained (Supplementary Fig. 1a; Supplementary Table 4). Since bacterial fractions account for less than 0.5% of the total cfDNA even after enrichment [30], we deemed it crucial to prevent misclassification, especially false positives. cfFBI's computational pipeline tackles this challenge through a multi-step process designed to minimize such errors. Previous findings demonstrate a reduction of false positive microbial counts by mapping to a more comprehensive host reference genome [41]. Therefore, cfFBI maps to the latest horse reference genome, EquCab3, along with all other horse genomes available on NCBI, totaling 11 genomes (Supplementary Fig. 2a,b), increasing the average host fraction of total cfDNA by 0.5%. Second, cfFBI taxonomically classifies the remaining unmapped reads using Kraken2 [42], with a custom database that includes all 11 horse genomes, as well as human genomes and all NCBI complete microbial genomes which improves species assignment. Third, suspected microbial contaminants are excluded from downstream analyses by testing whether the levels of microbial species correlated with the volume of reagents used in cfFBI [43].

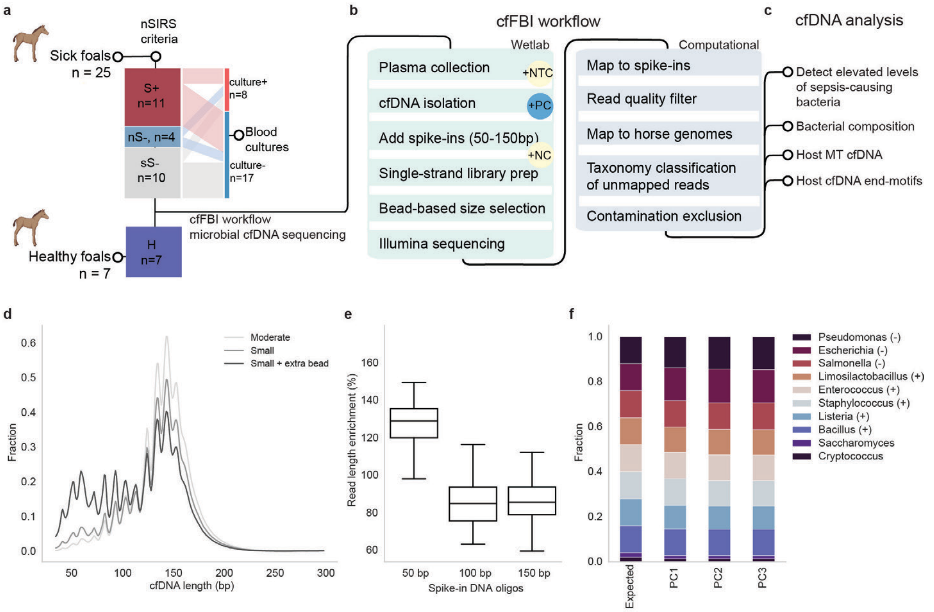


Figure 1. cfFBI pipeline, a cell-free DNA sequencing workflow designed to enhance bacterial identification in foals suspected of sepsis.

a. Foal cohort and SIRS categorization: The cohort includes sepsis-suspected foals and healthy (H) controls. Foals were categorized based on nSIRS criteria as SIRS-positive (S+; nSIRS ≥ 3), SIRS-negative with no symptoms (nS-; nSIRS=0), or SIRS-negative with symptoms (sS-; nSIRS=1-2). The alluvial plot shows the number of ill foals with positive blood cultures at hospital admission. **b.** Schematic of cfFBI workflow: cfDNA is isolated from foal blood plasma and mixed with synthetic DNA oligos (50, 100, 150 bp). A ligation-based library preparation and bead-based size selection enrich for short microbial fragments (<100 bp). After paired-end Illumina sequencing, spike-in sequences and low-quality reads are filtered out. Remaining reads are mapped to host genomes, and unmapped reads are classified taxonomically using Kraken2 with a customized database. Suspected contaminants are finally excluded. The workflow includes diverse controls: positive controls (PC), no-template controls (NTC), and negative controls (NC). **c.** Comparative analyses were performed within this study, comparing S+ versus H and nS-. Specifically, we focused on variations in host mitochondrial (MT) cfDNA, chromosomal host cfDNA end-motifs, bacterial load and diversity, and the abundance of potential pathogenic bacteria in septic foals **d.** Comparison of three ligation-based single-strand library preparation methods for enriching short cfDNA fragments: the 'moderate small' and 'extreme small' protocols from the SRSLY NGS Library Prep Kit, plus an additional bead-based size selection after the 'extreme small' protocol. The plot displays the template length size distribution of host cfDNA reads for each method. **e.** Boxplot showing the enrichment or depletion of synthetic DNA oligos (50, 100, and 150 bp) in foal plasma samples (n = 32). Synthetic oligos of 50, 100, and 150 bp were spiked in at equimolar ratios. **f.** Stacked bar plot showing the taxonomic classification results for a sonicated mock community with 10 microbial species. The symbols (-) and (+) after each bacteria genus name indicate whether the species is Gram-negative or Gram-positive. It compares expected versus observed species ratios from three technical replicates using the cfFBI workflow and Bracken abundance re-estimation.

To ascertain the consistency and effectiveness of the cfFBI workflow across library preparations, we first tested the classification accuracy using positive control samples (PCs) from sonicated mock microbial community DNA containing eight bacterial species and two yeasts. In all three technical replicate PCs, the eight bacterial genera were detected at levels consistent with the known microbial composition [44], with relative abundance being highly similar across PCs (Fig. 1f; Supplementary Table 5). The consistent detection of the correct microbes in the appropriate ratios across the three different PCs, shows robustness in both the wetlab and the bioinformatics workflow.

Decontamination and bacterial species composition assessment

We first analyzed the bacterial species composition in the samples, concentrating on both the bacterial fraction and its diversity. Typically 4.2% of the non-mapped reads were confidently classified using Kraken2 (Confidence threshold of 0.8; Supplementary Fig. 2b), of which 1.1% were classified as a bacteria at species level. To ensure accurate analysis of true biological signals, we first removed potential contaminants (see *Methods*), excluding 0.00319% of bacterial reads classified at the species level that were identified as contaminants (Supplementary Table 5). Most of these contaminant species were detected in the negative controls as well (i.e. NTCs and NCs; for details on sample collection see *Methods* and Fig. 1b; for results see Supplementary Table 6 and Supplementary Figs. 3-4), indicating that they are likely contaminants from cfDNA isolation or library preparation. The remaining cfDNA reads classified as bacterial species and not identified as contaminants were aggregated for all downstream bacterial composition analyses.

After decontamination, the median bacterial species-classified cfDNA fraction was 0.0083% (range 0.0010-5.5%) (Fig 2a). This total bacterial fraction moderately correlated with age (r=0.43, Supplementary Fig. 5a,b) and was more variable in the S+ group compared to the other two groups, with some samples showing notably high levels, including one outlier at 5.5% (Fig. 2a,d). Although most of the foals with a high bacterial fraction (above 0.0002) were S+ foals (4/6; 66%), the difference between groups was not statistically significant (Kruskal-Wallis test with Dunn's multiple comparison tests) (Fig. 2d,e). Between 75 and 1126 (median 327) species were found across 50 to 525 (median 192) genera in each foal. Interestingly, different foals exhibited distinct top abundant species (Fig 2b,c). In total, 4,284 bacterial species across 1,250 different genera were detected, with *Actinobacillus*, *Acinetobacter*, *Streptococcus*, and *Flavobacterium* being the most prevalent, contributing a median of 12.3% per foal.

Samples exhibited high variability in species composition (Fig. 2b,c), with the Shannon index revealing greater microbial diversity in sick foals (nS- and S+) compared to healthy foals (Fig. 2f,g), with age having no significant impact (Supplementary Fig. 5c,d). However, neither species richness nor diversity metrics (Supplementary Fig. 6) effectively distinguished between healthy, nS-, and S+ foals (Fig. 2b,c).

Co-elevation of multiple sepsis-causing genera observed in foals with sepsis

The primary objective of this study is to evaluate the potential of blood cfDNA sequencing for detecting elevated levels of sepsis-causing bacteria in newborn foals with nSIRS. To pinpoint bacteria associated with sepsis in the S+ foals, we compared the aggregated counts of species from the 16 most frequently cultured pathogenic genera found in culture-positive foals with sepsis (Fig. 3; Supplementary Fig. 7-9) [45]. One or multiple pathogenic genera were higher in 8/11 of the S+ foals compared to both nS- and H foals while the other 3/11 were higher compared to H foals alone (Supplementary Fig. 8), meaning that all S+ foals showed elevated levels of at least one pathogenic genus. In the follow-up analysis, we combined these results for species that showed an increase when compared to either H foals or both nS- and H foals.

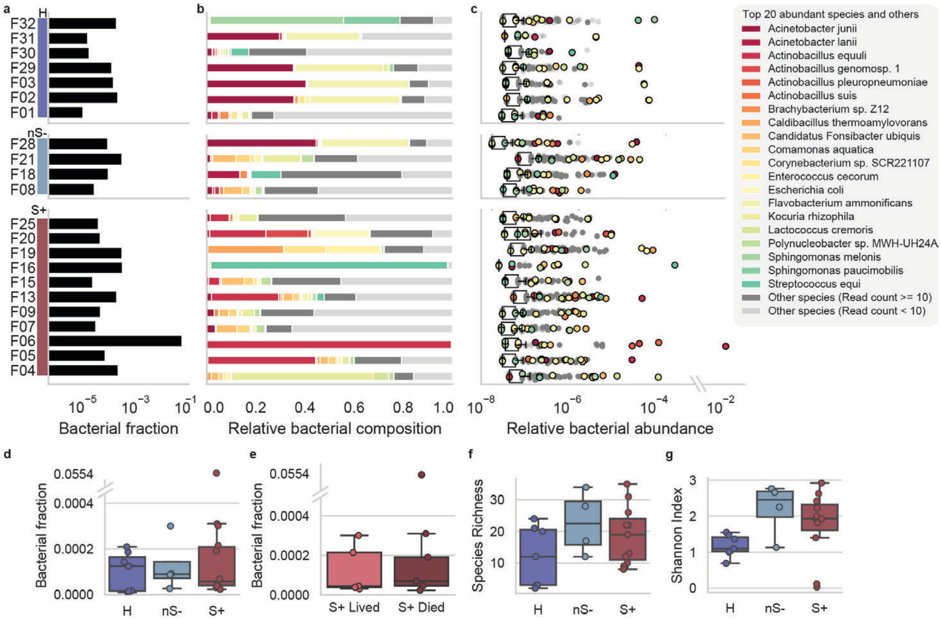


Figure 2. cfDNA bacterial load and diversity in foal plasma samples.

a. Bacterial fraction in each sample in three categories H, nS- and S+, represented in a log scale. **b.** Relative bacterial composition (normalized to total bacterial reads) in each sample. The top 20 abundant species in all samples were colored (see colors in the legend), the other species with more than 10 exact counts were colored with dark grey color and other species with less than 10 exact counts were colored with light grey color. **c.** Relative abundance (normalized to total cfDNA reads) in each bacterial species, represented in a log scale. The top 20 abundant species in all samples were colored (see colors in the legend), the other species with more than 10 exact counts were colored with dark grey color and other species with less than 10 exact counts were colored with light grey color. **d.** Fraction of cfDNA fragments taxonomically classified as bacterial origin (after removal of contaminant species) and its association with disease status. No significant difference is observed between groups. An outlier at 0.0554 is represented with a broken y-axis. Foals in S+ group showed the largest variation compared to the other two groups. (Standard deviation: H: 0.00008, nS-: 0.00011, S+, 0.0167). **e.** Fraction of cfDNA fragments taxonomically classified as bacterial origin (after removal of contaminant species) and its association with severity of disease in the S+ group. No significant difference is observed between groups. An outlier at 0.0554 is represented with a broken y-axis. **f.** Species richness, representing the number of bacterial species identified (after removal of contaminant species) in each plasma sample, and its association with disease status. **g.** Shannon index, indicating the evenness of classified bacterial species distribution (after removal of contaminant species) within each foal plasma sample, and the association between Shannon index with disease status. **d-g.** Disease status groups are H (n = 7), nS- (n = 4), and S+ (n = 11). For severity of disease, focus is on S+ cases with either survival (n = 5) or death (n = 6). Boxes represent the 25th percentile (bottom), median, and 75th percentile (top), with whiskers extending to the rest of the distribution within 1.5 times the inter-quartile range.

Overall, *Actinobacillus* and *Pantoea* were most frequently increased in 6/11 (54.5%) and in 4/11 (36.4%) of the S+ foals, respectively (Fig. 3; Supplementary Fig. 7; Supplementary Table 7). Co-elevation of multiple genera occurred in 5/11 (45.5%) foals (Fig. 3), with co-elevation of *Actinobacillus* and *Escherichia* being most common (3/11; 27.2%; Fig. 3). On the contrary, *Acinetobacter* was the only genus with higher frequencies in multiple H and nS- foals compared to S+ foals (Supplementary Fig. 9), suggesting that the elevated relative abundances of the 16 genera tested represent a genuine biological signal specific to the S+ foals. We did not observe clear relationships between bacterial elevation and the survival outcome of S+ foals (Fig. 3), suggesting that survival may be influenced by factors beyond the bacterial elevation, including the foal's immune response and the reaction to treatment. Taken together, these microbial cfDNA sequencing results show that sepsis in foals may have a multi-bacterial nature. Furthermore, the results emphasize that microbial cfDNA sequencing may hold potential for newborn foal sepsis diagnosis, although larger studies are required to establish the sensitivity and specificity of the technique.

Given this promise, we further investigated the sS- foals, which were excluded from previous analyses due to the ambiguous disease state. Based on the low sensitivity of the nSIRS criteria (42%) [2] and clinical symptoms observed in sS- foals, we expect some foals with sepsis in the sS- group as well. *Acinetobacter* levels were elevated in 4/10 sS- foals compared to S+ (Supplementary Fig. 9), resembling the H foals. Conversely however, the majority of sS- foals displayed trends similar to S+ foals, including increased levels of *Actinobacillus* (6/10) and *Escherichia* (4/10), when compared to H alone or H and nS-. Additionally, 70% of sS- foals exhibited co-elevation of multiple genera (Supplementary Fig. 8). The similarities in bacterial co-elevation between sS- and S+ foals, coupled with the low sensitivity of the nSIRS criteria [2], suggest that some foals with sepsis may have been overlooked. Alternatively, it could mean that the foals with a low nSIRS score are in an earlier stage of sepsis-development or suffer from other bacterial infections, both leading to an increase in microbial levels without many clinical nSIRS-symptoms.

Species-level bacterial identification can be used for clinical decision making, including guidance on selection of antimicrobial treatment. Therefore, we evaluated bacterial species-level elevations in the 16 most common genera associated with foal sepsis. Across the 16 genera, 22 pathogenic species were elevated in one or multiple S+ foals (Supplementary Table 8, Supplementary Fig. 7). *Actinobacillus equuli*, *Actinobacillus pleuropneumoniae* and *Escherichia coli* were most frequently elevated in 6/11, 3/11 and 3/11 S+ foals, respectively. Most elevated species corresponded with their respective elevated genera (31/35 observations). However, *Staphylococcus equinus* was higher in F05, *Acinetobacter haemolyticus* was elevated in F04, and *Acinetobacter lanii* as well as *Acinetobacter wanghuae* were elevated in F20, suggesting that species-level information can provide some additional leads compared to genus-level analyses (Supplementary Table 8). However, our analysis also urges caution interpreting sequencing results, particularly about potential misassignment of reads to closely related species, such as the identification of *Actinobacillus pleuropneumoniae*, which is not typically listed as a sepsis-causing species for foals.

Bacterial culture is the golden standard for identifying bacteria [18,22], but suffers from low sensitivity and false positive observations [16–18]. To evaluate the concordance between traditional culture and bacteria identified (as elevated) by microbial cfDNA profiling through sequencing, we compared the results of genus-level blood cultures to cfDNA sequencing after excluding potential contaminants. Notably, elevated levels of sepsis-causing bacterial genera were found by cfDNA sequencing in all three S+ and three sS- foals with positive bacterial blood cultures, indicating that cfDNA sequencing effectively detects bacterial abnormalities associated with sepsis (Supplementary Fig. 9). The overlap between detected genera, however, was limited, with only 43% (3/7) of the culture-identified bacterial genera showing elevation in the cfDNA (Supplementary Table 9). In an additional 29% (2/7) of cases, reads had been assigned to the cultured genera, but cfDNA levels did not surpass those as detected in H and nS- foals, suggesting that the bacteria can be simply present rather than elevated (Supplementary Table 9). Taken together, we observe low concordance between culture and cfDNA-based pathogen identification in newborn foals with SIRS, potentially due to the fact that cfDNA sequencing detects presence and elevated levels of DNA of sepsis-causing bacterial taxa, while culture detects viable bacteria.

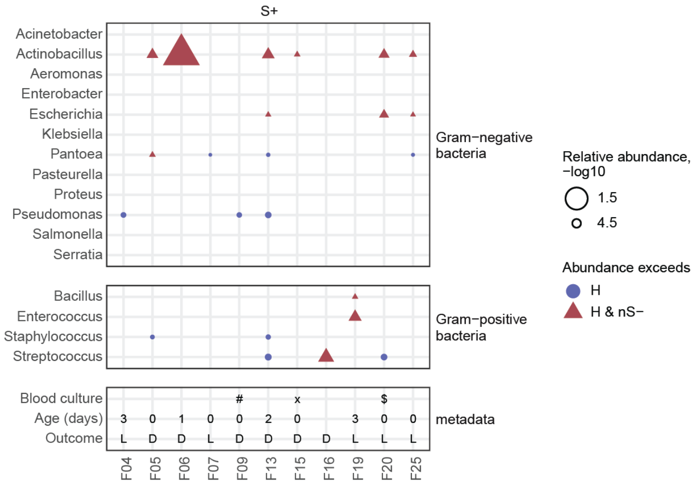


Figure 3. Species abundance and bacterial co-elevation detection in foal cohort samples

Dotplot displaying the detection of the 16 most frequently cultured pathogenic genera. Gram-negative species are shown at the top, and Gram-positive species are shown in the middle. Dots represent genera detected with at least 10 reads. Blue circles indicate genera with a relative abundance higher than in H foals, while red triangles denote genera with a relative abundance higher than in both H and nS- foals. Metadata is represented at the bottom, including blood culture if positive, age at hospital presentation if known, as well as survival outcome (L, Lived; D, Died). Special symbols in the blood culture section: '#' Gram-negative rod (non-fermenter), 'x' *Actinobacillus equuli* & *Streptococcus pneumoniae* and '$' *Staphylococcus* coagulase-negative.

Associations of host cfDNA with nSIRS status

Since most of the sequenced reads are mapped to the host reference genome and it is recognized that these host-derived reads can offer insights related to infection related tissue damage [46], host response to infection [23] and sepsis [47,48], we next investigated differences in host cfDNA between S+ and H and/or nS- foals, and between S+ foals that lived to S+ foals that died. Confirming previous results in foals [38], but differing from observations in humans [47–49], total cfDNA levels in plasma were not significantly elevated in S+ foals compared to H and nS- foals

(Kruskal-Wallis with Dunn's multiple comparison; S+ vs. H: p=0.92, Z=0.73; S+ vs. nS- p>0,99, Z=0.31), nor in S+ foals that lived compared to S+ foals that died (Mann-Whitney U Test, p=0.32, U=9, Fig. 4a,b). This suggests that total cfDNA levels cannot be used to diagnose sepsis in foals, as previously reported [38]. Strikingly, opposite to MT cfDNA levels in human sepsis patients [49,50], the MT cfDNA fraction of foal host origin was significantly lower in S+ foals compared to H foals (Fig. 4c). Similarly, a significant decrease in MT cfDNA was observed in S+ foals that died compared to those that survived (Fig. 4d). Of note, none of these variables were significantly different between isolation batches, library preparation batches, and operators (Mann-Whitney U Test with Bonferroni Correction, Supplementary Fig. 10).

As host end-motifs can give insight into the activities of nucleases and the interplay with innate immune response such as NETs [28,51], we proceeded to investigate host chromosomal cfDNA end-motifs. An enrichment in 5′ C-end and 3′ G-end cfDNA reads was present in all samples (Fig. 4e-h, Supplementary Table 10). Specifically 3′ C-end cfDNA reads were significantly decreased in S+ foals compared to H and nS- foals, while 5′ A-end cfDNA reads were significantly decreased in S+ foals compared to H foals (Kruskal-Wallis test with Dunn's multiple comparison tests, S+ vs. H: p=0.004, Z=3.09; Fig. 4e,g, Supplementary Table 10). Collectively, these results indicate that mitochondrial cfDNA levels and end-motifs could serve as potential biomarkers for SIRS and its prognosis in foals.

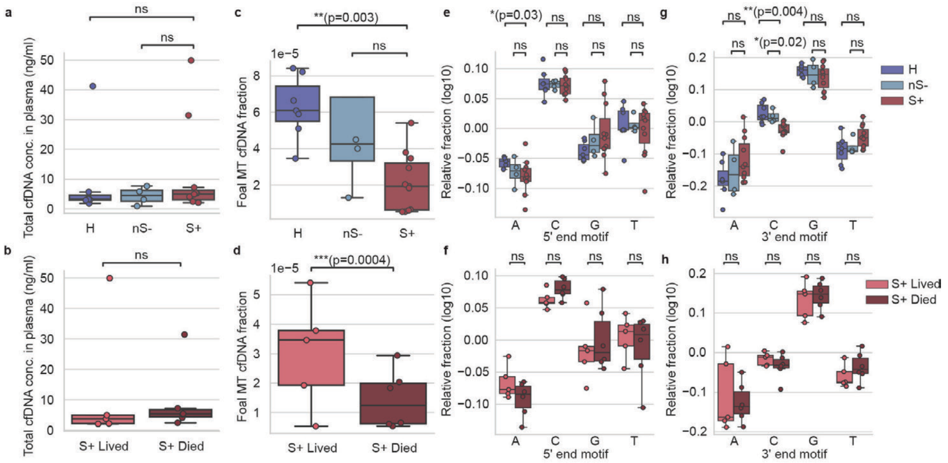


Figure 4. Host cfDNA abundance and its association with disease status and severity of disease.

a. Total cfDNA concentration in plasma samples and its association with disease status. **b.** Total cfDNA concentration in plasma samples and its association with its severity of disease. **c.** Fraction of cfDNA fragments of host mitochondrial origin, and their association with disease status. **d.** Fraction of cfDNA fragments of host mitochondrial origin, and association with severity of disease. **e.** Normalized base content fraction at the 5′ end of host chromosomal cfDNA fragments and its correlation with disease status. **f.** Normalized base content fraction at the 5′ end of host chromosomal cfDNA fragments and its association with severity of disease. **g.** Normalized base content fraction at the 3′ end of host chromosomal cfDNA fragments and its association with disease status. **h.** Normalized base content fraction at the 3′ end of host chromosomal cfDNA fragments and its association with severity of disease. Disease status groups are H (n = 7), nS- (n = 4), and S+ (n = 11). For severity of disease, focus is on S+ cases with either survival (n = 5) or death (n = 6). Boxes represent the 25th percentile (bottom), median, and 75th percentile (top), with whiskers extending to the rest of the distribution within 1.5 times the interquartile range.

Discussion

We introduce cfFBI, a cell-free DNA sequencing workflow designed to enhance bacterial identification in foals through a combination of optimized wetlab and computational procedures. cfFBI specifically enriches small, microbial cfDNA molecules which are present at minute levels within the cfDNA pool derived from a single tube of blood. cfFBI's computational steps, including host mapping and decontamination, are optimized to minimize false bacterial identifications. Using cfFBI, we applied cfDNA sequencing to newborn foals for the first time and detected elevated bacterial levels in 8/11 nSIRS-positive foals compared to levels in (sick) nSIRS-negative foals, while the other 3/11 nSIRS-positive foals showed at least one bacterial genus elevated compared to healthy foals alone. Interestingly, we find co-elevation of multiple pathogenic bacteria in 5/11 (45.5%) of nSIRS-positive foals. Although it was already known that bacterial culture can provide positive results for multiple species [11–14,52–54], it remained unclear if these were true co-occurrences, similar to the co-occurrence observed in human sepsis patients [12,13] or a result of contamination [17,18,55]. The frequent observation of co-elevation in cfDNA in this study suggests that multiple genera may actually jointly contribute to sepsis in newborn foals. Further validation and follow-up research is needed to determine the potential implications of these findings.

Until now, most knowledge about the bacteria causing sepsis in foals has been based on culture-dependent techniques, while culture is known to have only 25-45% sensitivity in foals with sepsis [16–18]. When comparing bacteria detected by cfDNA sequencing to those identified through culture in the cohort, we observed limited concordance (3/7 (43%)). This discrepancy may arise because bacterial culture detects only viable bacteria, while cfDNA sequencing reveals both the presence and increased abundance of bacterial cfDNA resulting from recent cell death. The two techniques thereby capture a different aspect of the complex underlying pathophysiology. In large-scale human studies, microbial cfDNA showed higher sensitivity and specificity than blood cultures for detecting clinically relevant pathogens, resulting in an enhanced patient survival, and a reduction of overall antimicrobial use in patients with sepsis [56,57]. Ultimately, the two techniques may turn out to be complementary, with cfDNA sequencing providing reliable multi-pathogen results, while culture provides valuable insights into antimicrobial resistance of the identified bacterial species.

Diagnosis of sepsis in newborn foals is challenging, with most tools suffering from a low sensitivity and specificity. This issue also applies to the nSIRS scoring system used here, where foals with sepsis can have a nSIRS score of less than 3 [2]. To minimize the impact of potential false negative septic foals in the nSIRS-negative group, we excluded the nSIRS- foals with a nSIRS score of 1-2 when setting background level for bacterial elevation analysis in nSIRS-positive foals, as some of these cases may have an unindicated sepsis or a bacterial infection. Simultaneously, the nSIRS-positive group may still include foals without sepsis, so without a bacterial infection, as sepsis is currently defined as a combination of SIRS with a bacterial infection. This challenge clearly shows the need for additional diagnostic tools for improved sepsis diagnosis in foals.

Although this study clearly indicates the potential of cfDNA sequencing in newborn foals for the first time, it should be noted that the current cohort has limited statistical power to detect significantly elevated microbes in nSIRS-positive foals above background. We opted for a conservative approach of identifying elevated bacterial levels by testing if the value is above the highest observation in the control background samples without aiming to test for significance. A larger cohort study involving more newborn foals is essential to fully assess the potential of cfDNA sequencing for diagnosing sepsis. This would also allow a comparison to healthy and nSIRS-negative foals separately, where the first comparison is informative to gain more insight into biology, whereas the latter comparison can provide tools useful in the clinic. Increased cohort size would also be beneficial to study abundance of bacterial species or genera in a more unbiased manner, without focusing exclusively on the 16 most frequently observed bacteria in culture. This could provide new insights into biologically relevant, yet hard-to-culture, bacterial taxa associated with sepsis.

In addition to validating the microbial cfDNA observations of the current study, a larger study could aim to further explore the trends we found in the host cfDNA. This includes confirming the decreased MT cfDNA fraction in nSIRS-positive foals which aligns with previous studies in foals [38,39], but contrary to what is observed in humans [49,50]. Moreover, such a study would enable the validation of the absence of the expected elevation in total cfDNA in nSIRS-positive cases [38,39], which would typically be indicative of increased tissue damage and cell death as is seen in humans [58]. It could also shed light on findings on the (complementary or concordant) relationship between host cfDNA and microbial cfDNA which we did not observe in the current cohort (Supplementary Table 10). Finally, the differences in end-motifs in host cfDNA in nSIRS-positive versus nSIRS-negative foals could be validated in a larger cohort, potentially revealing additional biomarkers. By combining host and pathogen information from plasma cfDNA, more insights of host transcription profiles such as innate immune response activities can be obtained, as was already shown in human sepsis patients [23,59].

Currently, foals suspected of having sepsis are treated with broad spectrum antimicrobials until culture and susceptibility testing results become available after approximately 72 hours. In cases that fail to improve within this time period, antimicrobial treatment regimens are adjusted based on historical information on prevalence and susceptibility of bacteria causing sepsis in foals in that specific geographic area. As sepsis and organ dysfunction can develop rapidly, the antimicrobial susceptibility tests are rarely timely to aid the treatment regimen. By utilizing a faster and more sensitive technique [26] for identifying potential pathogens, coupled with the continually decreasing costs of sequencing combined with more targeted approaches (e.g. for antimicrobial resistance genes [60]), future adjustments to antimicrobial therapy can be made earlier, potentially increasing the survival chances of foals with sepsis.

Methods

Foal cohort

We prospectively included 25 sick foals admitted to the Utrecht University Equine Hospital (Utrecht, The Netherlands), between March 1st, 2021 and July 1st, 2022. For diagnostic purposes, two blood samples (up to 20 mL each) and one blood sample of 10 mL were collected aseptically from the jugular or cephalic vein, either by venipuncture or through a newly placed intravenous catheter immediately upon hospitalization. The two 20 mL samples were placed into 70 mL brain heart infusion broth + SPS (Biotrading, Mijdrecht, the Netherlands) and transported to the Veterinary Microbiological Diagnostic Center where the bottles were incubated at 37°C for 18-24h. After incubation, Gram-staining was performed followed by inoculation on two sheep blood agars (SBA), chocolate agar (CHOC) and MacConkey agar (MAC; Biotrading, Mijdrecht, the Netherlands). One SBA and MAC agar were incubated aerobically, while the other SBA was incubated anaerobically and the CHOC agar was incubated microaerobically; all at 37°C for 5-7 days. Agars and broths were checked daily for bacterial growth. If bacterial growth was detected, identification took place using Maldi-TOF (Bruker, Bremen, Germany). The 10 mL sample of blood was collected directly into a Streck tube (see "Sample preparation and nucleic acid isolation"). Diagnostic and clinical data of sick foals were recorded and later extracted from the medical information system (Supplementary Table 2).

In addition to the sick foals, seven healthy (H) newborn foals were enrolled in this study, four from Utrecht University Equine Hospital (Utrecht, The Netherlands) and three from Dierenkliniek Emmeloord (Emmeloord, The Netherlands). In the healthy foals, at the moment of blood collection for the routine check for passive transfer of immunity, 10 mL of blood was collected aseptically for cfDNA sequencing. Blood cultures were not performed on the healthy foal samples.

nSIRS criteria were used to classify foals into groups (Supplementary Table 1) [2], whereby an nSIRS score ≥ 3 was considered nSIRS-positive (S+), a nSIRS score of 0 was considered nSIRS-negative (nS-) and an nSIRS score of 1-2 was considered symptomatic nSIRS-negative (sS-). The foal cohort thus comprised 11 S+ foals, four nS- foals, seven healthy foals, and 10 sS- foals. The demographics, including foal age at presentation, breed, sex, as well as dam age, gestation length, and parity, are detailed in Supplementary Tables 2-3.

Sample preparation and nucleic acid isolation

Blood for cfDNA sequencing was collected aseptically in Streck Cell-Free BCT (Streck #230257). Plasma extraction involved centrifugation for 10 minutes at 1600 g (at room temperature), followed by an additional centrifugation step for 10 minutes at 16,000 g (at 4 °C) to eliminate all cells and debris. The resulting plasma samples were then stored at -80 °C.

For nucleic acid isolation from plasma, the Circulating Nucleic Acid Kit (Qiagen, 55114) was employed with specific modifications to the manufacturer's protocol. First, a subset of the samples was supplemented to 5mL using PBS (Supplementary Table 4), prior to isolation. Second, the lysis time was extended from 30 to 60 minutes. Finally, cfDNA was eluted in 28 or 35 µL of Nuclease Free water (Invitrogen, 10977-035), and measured by the Qubit dsDNA High Sensitivity Assay Kit or Broad Range Assay Kit (Thermofisher Scientific, Q32854 and Q32853, respectively).

Sequencing library preparation using single-strand ligation based DNA-capture

For library preparation quality control purposes, plasma DNA was supplemented with synthetic spike-ins, equaling 0.2% of the total DNA input. The synthetic spike-ins consisted of an equimolar mix of three single-stranded DNA sequences that were 50, 100, and 150 bp in length (sequences of these spike-ins listed in Supplementary Table 11). Since the SRSLY splint adapter (refer to the next section for more details about SRSLY) contains a 7-base random overhang, the spike-ins were designed to include a random overhang sequence of the same length.

Then, the SRSLY PicoPlus NGS Library Prep Kit was used to prepare sequencing libraries (Claret BioScience, CBS-K250B-96). Briefly, DNA input molecules were denatured and kept as single-stranded molecules using a thermostable single-stranded DNA binding protein. The single-stranded DNA was then ligated to SRSLY splint adapters, followed by an indexing PCR [61]. To enrich short fragments in all foal samples, we used the small fragment retention version of the SRSLY PicoPlus NGS Library Prep Kit protocol along with an additional bead-based size selection step (Ampure XP, A63882). In a separate experiment (Fig. 1d) we compared a short fragment retention and moderate fragment retention protocol combined with/without a customized extra step of bead-based selection in a separate experiment, by which we established that we would use the small fragment retention version with a customized extra step of bead-based selection in all other samples in this study. The complete description of this experiment can be found in the method section “Short fragmentation length enrichment analysis”.

All libraries were quantified using the Qubit dsDNA High Sensitivity Assay Kit (Thermofisher Scientific, 32854) and size distribution was analyzed using the Tapestation 2200 and the D1000 kits (Agilent, 5067-5583). Foal sample sequencing libraries were pooled equimolar, with positive (see “Preparing positive control samples imitating microbial cfDNA fragments”) and negative controls (see “Negative controls for identifying contaminants in low microbial load samples”), albeit at a threefold lower molar ratio than the foal cfDNA libraries. This pool was subsequently enriched for sub-100 bp cfDNA molecules by a bead-based size selection step (Ampure XP, A63882). After this bead-based size selection, the concentration and size of the library pool were measured using the TapeStation 2200 and the D1000 kit.

Preparing positive control samples imitating microbial cfDNA fragments

As a positive control, we made use of a sonicated mock community DNA (ZymoBIOMICS Microbial Community DNA Standard, D6305) containing a mixture of genomic DNA of 10 microbial strains: *Listeria monocytogenes, Pseudomonas aeruginosa, Escherichia coli, Salmonella enterica, Lactobacillus fermentum, Enterococcus faecalis, Staphylococcus aureus, Bacillus subtilis, Saccharomyces cerevisiae* and *Cryptococcus neoformans*. In short, 2 µL ZymoBIOMICS standard was supplemented with 88 µl of LowTE (10mM Tris, 0.1mM EDTA), before shearing using the Covaris S2 at 6-8°C, with continuous degassing, a duty cycle of 10%, intensity set to 5, and 200 cycles per burst for 14 minutes. Bead-based size selection was then performed to enrich for DNA fragments shorter than 200 bp (Ampure XP, A63882), using an initial 1.1x volume of beads followed by adding a 3x volume of beads to the supernatant, to mimic cfDNA. Three ng of sheared, size-selected mock community DNA supplemented with 6 pg of synthetic spike-in DNA were used as input for the SRSLY library preparation. Since the next-generation sequencing libraries were prepared in three separate batches, we included one positive control sample for each batch, resulting in a total of three positive controls (PC1, PC2, and PC3).

After sequencing, the computational cfFBI workflow was applied to the positive control (PC) samples, including Bracken abundance re-estimation to refine the relative fractions of each species and genus (see “Sequencing Data Processing Using the cfFBI Pipeline” for details). The observations for these 10 species and their respective genera are provided in this study, including their relative fractions at both the species and genus levels, as well as the variance among the PC samples.

Of note: the positive controls (PC1-PC3) served three purposes. First, to validate the effectiveness of the protocol in each experimental batch. Second, to ensure the wet lab and computational workflow can accurately produce representative species and genera of interest. Third, to confirm that data generated from independent library preparations are comparable.

Negative controls for identifying contaminants in low microbial load samples

Due to the risk of contamination in low microbial load samples, we incorporated a set of four negative controls (NTC1, NTC2, NCMQiso, NCMQlib). Among these, two (NTC1, NTC2) consisted of 5 ml PBS that underwent the entire process of cfDNA and SRSLY-mediated NGS sequencing library preparation. Another two control samples contained Nuclease-Free water that was utilized for the elution (NC1MQiso) of cfDNA after cfDNA isolation and the supplemention of up to 18 ul that was added to samples before library preparation (NC1MQlib). These two samples underwent the process of SRSLY library preparation. No spike-in DNA sequences were added to the negative control samples.

Next-generation sequencing

Library sequencing was executed on the NovaSeq 6000 platform with 2 x 150 bp reads. This process yielded a range of 20 to 66 million reads per cfDNA library, between 8.6 and 10.3 million reads for each positive control (PC), and between 6.0 and 9.1 million reads for negative controls (NTC1, NTC2, NC1MQiso, and NC1MQlib).

Sequencing data processing using the cfFBI-pipeline

Illumina sequencing and synthetic data underwent processing via the cfFBI-pipeline, available on our Github repository (https://github.com/AEWesdorp/cfFBI/tree/main/pipeline). In a nutshell, bbduk.sh from tool BBmap [62] was employed to detect and eliminate reads containing 50mer, 100mer, or 150mer synthetic spike-in sequences. Subsequently, duplicate removal was carried out using nubeam-dedup [63], followed by default read quality filtering using fastp[64] to generate high-quality sequencing data. The quality filtering included removing low-quality reads, implementing a low complexity filter, adapter removal, and discarding short reads (< 35bp) using AdapterRemoval [65].

For horse read sequence identification, we tested two strategies via host genome mapping (bowtie2 [66]). The first strategy utilized the reference genome *Equus caballus* EquCab3.0 from NCBI RefSeq (accessed on Nov 8th, 2022). The second strategy incorporated all 10 additional genomic sequences available for *Equus caballus* within NCBI RefSeq (accessed on Feb 5th, 2024), bringing the total to 11 genomes: EquCab3.0, 57H, 25H, 16H, 7H, 2H, 9H, 30H, Ajinai1.0, LipY764, and EquCab2.0.

The latter strategy is adopted in the cfFBI pipeline. After host sequence subtraction, remaining paired-end reads underwent taxonomic classification using Kraken2 [42], a highly regarded metagenomic tool that performs exact *k*-mer alignment to a reference database for rapid per-read taxonomic classification (for details about the adapted database, see: “Taxonomic database construction and taxonomy classification”). Sequencing data were processed with a confidence threshold (CT) of 0.8 for all described databases, the selection of the CT was based on previous work [67] which demonstrated that a CT of 0.8 results in the highest average precision when using the NCBI database. After Kraken2 classification, Bracken [44,67] was employed to re-estimate the abundance of species within the metagenomic PCs (PC1-PC3; as specified in the cfFBI’s config file). Of note, Bracken abundance re-estimation was applied for PCs but not for foal cfDNA samples according to the Kraken software suite authors' recommendation [68].Resulting host-mapping reads, classified reads, and bacterial-classified reads were normalized to QC-passed reads in each sample unless otherwise specified.

Taxonomic database construction and taxonomy classification

For this study, we constructed a custom Kraken2 hash-table database that includes 11 horse genomes, two human genomes, and all complete microbial genomes from NCBI (downloaded as of May 15th, 2023). The microbial component comprises 285,825 bacterial, 14,977 viral, 496 fungal, 1,493 archaeal, and 96 protozoal genome assemblies. To build this database, genomic sequences from the NCBI RefSeq database were downloaded using the *kraken2-build --download-taxonomy* command for archaea, bacteria, fungi, human, plasmid, protozoa, UniVec_Core (contaminant sequences), and viral genomes. Additionally, all 11 genomic sequences for Equus caballus (EquCab3.0, 57H, 25H, 16H, 7H, 2H, 9H, 30H, Ajinai1.0, LipY764, and EquCab2.0) from NCBI RefSeq were downloaded (as of 05-02-2024). The database also incorporated the human genome GRCh38.p14 (obtained directly from NCBI RefSeq) and CHM13v2.0 (added manually). Kraken2 databases were built using the default settings (*kraken2-build*).

Short fragmentation length enrichment analysis

To evaluate the efficiency of short (<100 bp) microbial cfDNA fragment enrichment across various protocols, we tested different versions of the SRSLY PicoPlus NGS Library Prep Kit (Claret BioScience, CBS-K250B-96), including short fragment retention and moderate fragment retention protocol combined with or without a customized extra step of bead-based selection. After preparing these four different libraries, the libraries were pooled and sequenced with NextSeq 2000. On average 29 million paired-end (2 x 150 bp) reads were obtained. The cfFBI computational workflow was applied to all four samples, using only the EquCab3.0 as a reference genome. Reads mapped to the host chromosomal contigs were extracted using samtools (v1.3.1) with a minimal mapping quality score of >= 30. Subsequent processing and length analysis were performed by a customized processing script using R (v4.2.1) (for details, see: https://github.com/AEWesdorp/cfFBI/tree/main/fragmentomics).

Host mitochondrial and host read end-motifs analysis

Sequences mapped to mitochondrial contig (RefSeq contig name NC_001640.1) and chromosomal contigs of *Equus caballus* EquCab3.0 were extracted for host read mitochondrial and host read end-motif analyses using samtools (v1.19, v1.3.1). Amount of reads mapped to mitochondria with a quality score of >= 40 was divided by the number of reads mapped to all chromosomal and mitochondrial DNA with a quality score of >= 40 in EquCab3.0 to calculate MT cfDNA fraction. To investigate read end-motifs, we analyzed the most terminal base of R1 and R2 in all reads that were mapped with a quality score of >= 30, using a custom R script. Counts of each terminal base were tallied, then normalized to the expected fraction (equally distributed). As an example, for Motif A in 1-mer end-motif:

$$\text{Motif A Relative fraction (log10)} = \log_{10}\left(\text{motifA} / \left(\sum(\text{motifA}+\text{motifB}+\text{motifC}+\text{motifD}) /4\right) \right)$$

Contamination identification in sequencing runs with concentration-based methods

Low biomass microbial sequencing is sensitive to any DNA sequence present in samples, including contaminants. Decontamination, which refers to the process of removing contaminants from findings, is crucial to exclude uninformative findings. Previously, "*decontam*", a statistical method which identifies and removes reagent-related contaminant sequences has been proposed for metagenomics data. This method, implemented in an R package, detects contaminants by analyzing their correlation with DNA concentration (frequency-based method) as well as their presence in negative controls (prevalence-based method)[43] (illustration adapted in Supplementary Fig. 4b). We adapted the frequency-based method to identify contaminants resulting from the cfDNA isolation step and/or the SRSLY library preparation step (Supplementary Fig. 4a).

This frequency-based method assumes that contaminants exist at the same concentration in input reagents in different samples. Therefore, contaminants are more abundant in samples with low DNA input and consequently samples with lower DNA yield. We measured the cfDNA yield after isolation and used (whenever possible) 5 ng of cfDNA as input for the library preparation, to standardize the input DNA for this step. As a result, more isolation-related contaminants were expected in lower input DNA concentration samples. After SRSLY sequencing library preparation we again measured the total DNA yield, and pooled samples in equimolar amounts for sequencing, thus, more library preparation-related contaminants were expected in samples with low total DNA yield.

Therefore, both cfDNA isolation yield and SRSLY sequencing library preparation yield were used for contaminant identification. Normalized species classified counts of all species were correlated with (1) the inverse input DNA volume used for the library preparation, against correlation to a constant value; and correlated with (2) final DNA yield after library preparation, against correlation to a constant value. Testing the null hypothesis of whether each species was not a contaminant from step (1) and/or step (2) with the R package *decontam*. This derived a p-value (significance) of the likelihood of being a contaminant for each species. The authors of the decontam package suggested identifying the trough in the distribution of p-values to set as a cutoff to identify contaminants. We set a p-value cutoff at 0.25 for both steps and checked whether the species occurred in at least six samples, as a lenient approach to identify as many suspected contaminants as possible to prevent false positive findings as suspected pathogens (Supplementary Fig. 4f,g). We performed this across all isolation and library preparation samples without batch-specific testing due to small batch sizes in this experiment. The identified contaminant species were removed from the further analysis that involves classification results. To validate that the identified species were true contaminants, we checked their presence in four negative controls (NTC1, NTC2, NC1MQiso, and NC1MQlib, see Supplementary Fig. 5).

Bacterial load calculation

After excluding contaminant species, read counts of all other bacterial species were aggregated and normalized to the total number of quality-filtered reads, to calculate what we call the "total fraction of bacterial cfDNA".

Bacterial diversity measurements

For bacterial diversity measurement, it is important to avoid analytical noise arising from potential false positive taxonomy classification. Apart from removing species deemed as contaminants from the above mentioned method, we also exclude all observations that were fewer than ten classified reads. Species richness was calculated as the number of species present in each sample. This includes only species that have more than ten classified reads in any sample and are not identified as contaminant species. The Shannon index, or Shannon entropy, measures biodiversity by considering both the abundance and evenness of species present in a sample [69]. We also measured Bray-Curtis dissimilarity to assess compositional differences between sample pairs. In order to calculate this Bray-Curtis dissimilarity for all pairs, counts were log10-transformed. Results were visualized using hierarchical clustering with single linkage (nearest neighbor) to illustrate the relationships between samples.

Identification elevated bacterial taxa

Using the cfFBI workflow, we aimed to detect elevated levels of frequently observed bacterial pathogens in newborn foals with sepsis [45]. These bacterial pathogens include 12 Gram-negative genera (*Serratia*, *Salmonella*, *Pseudomonas*, *Proteus*, *Pasteurella*, *Pantoea*, *Klebsiella*, *Escherichia*, *Enterobacter*, *Aeromonas*, *Actinobacillus*, and *Acinetobacter*) and four Gram-positive genera (*Streptococcus*, *Staphylococcus*, *Enterococcus*, and *Bacillus*). We visualized the taxonomy-classified normalized read counts for all species within these 16 genera (Supplementary Fig. 7) and aggregated these counts to form a total count for each genus. Comparisons were made by contrasting the foals' (e.g., S+) aggregated genus-level counts against the maximum aggregated genus-level counts observed in H and/or nS- foals to identify abundances that exceeded those in the control group (i.e., H and/or nS-). Conversely, we also compared the aggregated genus-level counts of the non-S+ foals against those observed in the S+ group. In all comparisons, genera detected at low abundance (fewer than 10 reads) were excluded.

Statistics

We utilized Kruskal-Wallis tests followed by Dunn's multiple comparison tests to conduct a directional non-parametric ANOVA for comparing total cfDNA levels, mitochondrial cfDNA fraction, host cfDNA end-motifs, bacterial cfDNA fraction, species richness, and Shannon indices among groups composed of H, nS- or S+ foals. For the comparison between two groups (S+ Lived and S+ Died) we used Mann-Whitney U tests. Additionally, Mann-Whitney U tests with Bonferroni correction were employed to assess whether variables of interest were confounded by factors such as the isolation batch, library preparation batch, and the location of the sample during library preparation.

Software

Data and statistical analyses were carried out using Python (v3.10) with the packages numpy (v1.26.0) and statannotations (v0.6.0), and GraphPad Prism (v10.3.0). Figures were created using R (v.4.2.0) and Python (v3.10) with the packages seaborn (v0.11.2), pandas (v1.5.3), and compiled with Adobe illustrator (2024, v28.6). Illustrations were created using BioRender and Adobe illustrator (2024, v28.6).

Declarations

Code availability

The code related to analysis and visualization of content in this manuscript are deposited at a Github repository: https://github.com/AEWesdorp/cfFBI. The repository is open access with a GNU general public license version 3.

Data availability

Metagenomic sequencing data (FASTQ files) have been deposited in the European Nucleotide Archive (ENA) Browser under accession PRJEB77374.

Acknowledgments

This study is supported by the funding of Stichting Vrienden Diergeneeskunde, the Faculty of Veterinary Medicine, Utrecht University, and a Vidi Fellowship (639.072.715) to JdR from the Dutch Organization for Scientific Research (Nederlandse Organisatie voor Wetenschappelijk Onderzoek, NWO). We thank Rhoxane Korthals from Dierenkliniek Emmeloord for her assistance in sample collection. We thank Joost van Rosmalen for his valuable input on statistical analysis. We acknowledge the Utrecht Sequencing Facility (USEQ) for providing sequencing services and data. USEQ is subsidized by the University Medical Center Utrecht and The Netherlands X-omics Initiative (NWO project 184.034.019). We thank Marc Pagès-Gallego and Dieter Stoker for proofreading the manuscript.

Author Contributions

LC, EW, MJ, ES, MT, CV, JdR conceptually designed the study. ES and MT collected samples, gathered clinical information, and gave clinical input. EW, ES, NB, CV, MT prepared samples. EW, NB and CV optimized protocols and generated the sequencing libraries. LC, EW, MJ contributed to data analysis. CV, AZ, EB, JW provided input on the experiments and analyses. LC, EW, MJ wrote the manuscript. JdR coordinated the study. All authors read and approved the final manuscript.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used chatGPT in order to rephrase sentences. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Declaration of interests

JdR is founder and shareholder of Cyclomics BV, a genomics company. The other authors declare no competing interests.

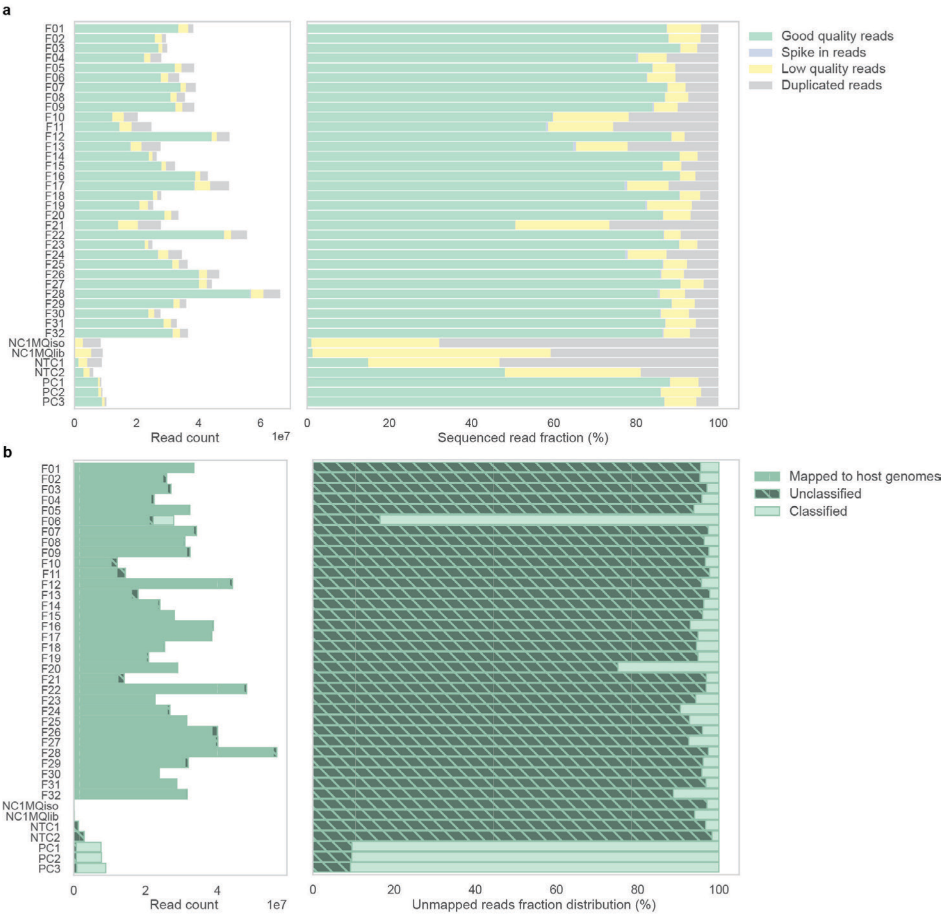
References

1. Singer, M., Deutschman, C.S., Seymour, C.W., Shankar-Hari, M., Annane, D., Bauer, M., Bellomo, R., Bernard, G.R., Chiche, J.-D., Coopersmith, C.M., et al. (2016). The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA *315*, 801–810.
2. Wong, D.M., Ruby, R.E., Dembek, K.A., Barr, B.S., Reuss, S.M., Magdesian, K.G., Olsen, E., Burns, T., Slovis, N.M., and Wilkins, P.A. (2018). Evaluation of updated sepsis scoring systems and systemic inflammatory response syndrome criteria and their association with sepsis in equine neonates. J. Vet. Intern. Med. *32*, 1185–1193.
3. Cicchinelli, S., Pignataro, G., Gemma, S., Piccioni, A., Picozzi, D., Ojetti, V., Franceschi, F., and Candelli, M. (2024). PAMPs and DAMPs in Sepsis: A Review of Their Molecular Features and Potential Clinical Implications. Int. J. Mol. Sci. *25*. https://doi.org/10.3390/ijms25020962.
4. Denning, N.-L., Aziz, M., Gurien, S.D., and Wang, P. (2019). DAMPs and NETs in Sepsis. Front. Immunol. *10*, 2536.
5. Dunkel, B., and Corley, K.T.T. (2015). Pathophysiology, diagnosis and treatment of neonatal sepsis. Equine Vet. Educ. *27*, 92–98.
6. Eaton, S. (2023). Neonatal sepsis – Pathology and clinical signs. Equine Vet. Educ. https://doi.org/10.1111/eve.13796.
7. Sheats, M.K. (2019). A Comparative Review of Equine SIRS, Sepsis, and Neutrophils. Front Vet Sci *6*, 69.
8. Cohen, N.D. (1994). Causes of and farm management factors associated with disease and death in foals. J. Am. Vet. Med. Assoc. *204*, 1644–1651.
9. Wohlfender, F.D., Barrelet, F.E., Doherr, M.G., Straub, R., and Meier, H.P. (2009). Diseases in neonatal foals. Part 1: the 30 day incidence of disease and the effect of prophylactic antimicrobial drug treatment during the first three days post partum. Equine Vet. J. *41*, 179–185.
10. Galvin, N., and Corley, K. (2010). Causes of disease and death from birth to 12 months of age in the Thoroughbred horse in Ireland. Ir. Vet. J. *63*, 37–43.
11. Theelen, M.J.P., Wilson, W.D., Byrne, B.A., Edman, J.M., Kass, P.H., and Magdesian, K.G (2019). Initial antimicrobial treatment of foals with sepsis: Do our choices make a difference? Vet. J. *243*. https://doi.org/10.1016/j.tvjl.2018.11.012.
12. Chen, P., Li, S., Li, W., Ren, J., Sun, F., Liu, R., and Zhou, X.J. (2020). Rapid diagnosis and comprehensive bacteria profiling of sepsis based on cell-free DNA. J. Transl. Med. *18*, 5.
13. Wang, Y., and Huang, X. (2018). Sepsis after uterine artery embolization-assisted termination of pregnancy with complete placenta previa: A case report. J. Int. Med. Res. *46*, 546–550.
14. Gayle, J.M., Cohen, N.D., and Keith Chaffin, M. (1998). Factors Associated with Survival in Septicemic Foals: 65 Cases (1988–1995). J. Vet. Intern. Med. *12*, 140–146.
15. Perkins, G.A., and Wagner, B. (2015). The development of equine immunity: Current knowledge on immunology in the young horse. Equine Vet. J. *47*, 267–274.
16. Hytychová, T. ’ana, and Bezděková, B. (2015). Retrospective evaluation of blood culture isolates and sepsis survival rate in foals in the Czech Republic: 50 cases (2011-2013). J. Vet. Emerg. Crit. Care *25*, 660–666.
17. Russell, C.M., Axon, J.E., Blishen, A., and Begg, A.P. (2008). Blood culture isolates and antimicrobial sensitivities from 427 critically ill neonatal foals. Aust. Vet. J. *86*, 266–271.
18. Giancola, S., and Hart, K.A. (2023). Equine blood cultures: Can we do better? Equine Vet. J. *55*, 584–592.
19. Poltavchenko, G.M. (1990). [Effects of diazepam and N(6)-cyclohexyladenosine on the level of diazepam-binding inhibitor in structures of the hippocampus during immobilization stress]. Biull. Eksp. Biol. Med. *110*, 166–167.
20. Elmas, C.R., Koenig, J.B., Bienzle, D., Cribb, N.C., Cernicchiaro, N., Coté, N.M., and Weese, J.S. (2013). Evaluation of a broad range real-time polymerase chain reaction (RT-PCR) assay for the diagnosis of septic synovitis in horses. Can. J. Vet. Res. *77*, 211–217.
21. Oeser, C., Pond, M., Butcher, P., Bedford Russell, A., Henneke, P., Laing, K., Planche, T., Heath, P.T., and Harris, K. (2020). PCR for the detection of pathogens in neonatal early onset sepsis. PLoS One *15*, e0226817.
22. Hackett, E.S., Lunn, D.P., Ferris, R.A., Horohov, D.W., Lappin, M.R., and McCue, P.M. (2015). Detection of bacteraemia and host response in healthy neonatal foals. Equine Vet. J. *47*, 405–409.
23. Cheng, A.P., Burnham, P., Lee, J.R., Cheng, M.P., Suthanthiran, M., Dadhania, D., and De Vlaminck, I. (2019). A cell-free DNA metagenomic sequencing assay that integrates the host injury response to infection. Proc. Natl. Acad. Sci. U. S. A. *116*, 18738–18744.
24. Pietrzak, B., Kawacka, I., Olejnik-Schmidt, A., and Schmidt, M. (2023). Circulating Microbial Cell-Free DNA in Health and Disease. Int. J. Mol. Sci. *24*. https://doi.org/10.3390/ijms24033051.
25. Blauwkamp, T.A., Thair, S., Rosen, M.J., Blair, L., Lindner, M.S., Vilfan, I.D., Kawli, T., Christians, F.C., Venkatasubrahmanyam, S., Wall, G.D., et al. (2019). Analytical and clinical validation of a microbial cell-free DNA sequencing test for infectious disease. Nat Microbiol *4*, 663–674.
26. Nielsen, M.E., Søgaard, K.K., Karst, S.M., Krarup, A.L., Nielsen, H.L., and Albertsen, M. (2024). A new method using rapid Nanopore metagenomic cell-free DNA sequencing to diagnose bloodstream infections: a prospective observational study. medRxiv, 2024.05.09.24307053. https://doi.org/10.1101/2024.05.09.24307053.

27. Serpas, L., Chan, R.W.Y., Jiang, P., Ni, M., Sun, K., Rashidfarrokhi, A., Soni, C., Sisirak, V., Lee, W.-S., Cheng, S.H., et al. (2019). Dnase1l3 deletion causes aberrations in length and end-motif frequencies in plasma DNA. Proceedings of the National Academy of Sciences *116*, 641–649.
28. Zhou, Z., Ma, M.-J.L., Chan, R.W.Y., Lam, W.K.J., Peng, W., Gai, W., Hu, X., Ding, S.C., Ji, L., Zhou, Q., et al. (2023). Fragmentation landscape of cell-free DNA revealed by deconvolutional analysis of end motifs. Proc. Natl. Acad. Sci. U. S. A. *120*, e2220982120.
29. Thierry, A.R. (2023). Circulating DNA fragmentomics and cancer screening. Cell Genom *3*, 100242.
30. Kowarsky, M., Camunas-Soler, J., Kertesz, M., De Vlaminck, I., Koh, W., Pan, W., Martin, L., Neff, N.F., Okamoto, J., Wong, R.J., et al. (2017). Numerous uncharacterized and highly divergent microbes which colonize humans are revealed by circulating cell-free DNA. Proc. Natl. Acad. Sci. U. S. A. *114*, 9623–9628.
31. Sanchez, C., Roch, B., Mazard, T., Blache, P., Dache, Z.A.A., Pastor, B., Pisareva, E., Tanos, R., and Thierry, A.R. (2021). Circulating nuclear DNA structural features, origins, and complete size profile revealed by fragmentomics. JCI Insight *6*. https://doi.org/10.1172/jci.insight.144561.
32. Yu, S.C.Y., Deng, J., Qiao, R., Cheng, S.H., Peng, W., Lau, S.L., Choy, L.Y.L., Leung, T.Y., Wong, J., Wong, V.W.-S., et al. (2023). Comparison of Single Molecule, Real-Time Sequencing and Nanopore Sequencing for Analysis of the Size, End-Motif, and Tissue-of-Origin of Long Cell-Free DNA in Plasma. Clin. Chem. *69*, 168–179.
33. Chang, A., Mzava, O., Lenz, J.S., Cheng, A.P., Burnham, P., Motley, S.T., Bennett, C., Connelly, J.T., Dadhania, D.M., Suthanthiran, M., et al. (2021). Measurement Biases Distort Cell-Free DNA Fragmentation Profiles and Define the Sensitivity of Metagenomic Cell-Free DNA Sequencing Assays. Clin. Chem. *68*, 163–171.
34. Huang, Y.-F., Chen, Y.-J., Fan, T.-C., Chang, N.-C., Chen, Y.-J., Midha, M.K., Chen, T.-H., Yang, H.-H., Wang, Y.-T., Yu, A.L., et al. (2018). Analysis of microbial sequences in plasma cell-free DNA for early-onset breast cancer patients and healthy females. BMC Med. Genomics *11*, 16.
35. Chen, H., Zheng, Y., Zhang, X., Liu, S., Yin, Y., Guo, Y., Wang, X., Zhang, Y., Zhao, C., Gai, W., et al. (2024). Clinical evaluation of cell-free and cellular metagenomic next-generation sequencing of infected body fluids. J. Advert. Res. *55*, 119–129.
36. Wang, G., Lam, W.K.J., Ling, L., Ma, M.-J.L., Ramakrishnan, S., Chan, D.C.T., Lee, W.-S., Cheng, S.H., Chan, R.W.Y., Yu, S.C.Y., et al. (2023). Fragment Ends of Circulating Microbial DNA as Signatures for Pathogen Detection in Sepsis. Clin. Chem. *69*, 189–201.
37. Jiang, P., Sun, K., Peng, W., Cheng, S.H., Ni, M., Yeung, P.C., Heung, M.M.S., Xie, T., Shang, H., Zhou, Z., et al. (2020). Plasma DNA End-Motif Profiling as a Fragmentomic Marker in Cancer, Pregnancy, and Transplantation. Cancer Discov. *10*, 664–673.
38. Colmer, S.F., Luethy, D., Abraham, M., Stefanovski, D., and Hurcombe, S.D. (2021). Utility of cell-free DNA concentrations and illness severity scores to predict survival in critically ill neonatal foals. PLoS One *16*, e0242635.
39. Hobbs, K.J., Cooper, B.L., Dembek, K., and Sheats, M.K. (2024). Investigation of Extracted Plasma Cell-Free DNA as a Biomarker in Foals with Sepsis. Vet. Sci. China *11*. https://doi.org/10.3390/vetsci11080346.
40. Burnham, P., Kim, M.S., Agbor-Enoh, S., Luikart, H., Valantine, H.A., Khush, K.K., and De Vlaminck, I. (2016). Single-stranded DNA library preparation uncovers the origin and diversity of ultrashort cell-free DNA in plasma. Sci. Rep. *6*, 27859.
41. Gihawi, A., Ge, Y., Lu, J., Puiu, D., Xu, A., Cooper, C.S., Brewer, D.S., Pertea, M., and Salzberg, S.L. (2023). Major data analysis errors invalidate cancer microbiome findings. MBio *14*, e0160723.
42. Wood, D.E., Lu, J., and Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. Genome Biol. *20*, 257.
43. Davis, N.M., Proctor, D.M., Holmes, S.P., Relman, D.A., and Callahan, B.J. (2018). Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. Microbiome *6*, 1–14.
44. Lu, J., Breitwieser, F.P., Thielen, P., and Salzberg, S.L. (2017). Bracken: estimating species abundance in metagenomics data. PeerJ Comput. Sci. *3*, e104.
45. Theelen, M.J.P., Wilson, W.D., Byrne, B.A., Edman, J.M., Kass, P.H., Mughini-Gras, L., and Magdesian, K.G. (2020). Differences in isolation rate and antimicrobial susceptibility of bacteria isolated from foals with sepsis at admission and after ≥48 hours of hospitalization. J. Vet. Intern. Med. *34*, 955–963.
46. Jin, X., Wang, Y., Xu, J., Li, Y., Cheng, F., Luo, Y., Zhou, H., Lin, S., Xiao, F., Zhang, L., et al. (2023). Plasma cell-free DNA promise monitoring and tissue injury assessment of COVID-19. Mol. Genet. Genomics *298*, 823–836.
47. Jing, Q., Leung, C.H.C., and Wu, A.R. (2022). Cell-Free DNA as Biomarker for Sepsis by Integration of Microbial and Host Information. Clin. Chem. *68*, 1184–1195.
48. Charoensappakit, A., Sae-Khow, K., Rattanaliam, P., Vutthikraivit, N., Pecheenbuvan, M., Udomkarnjananun, S., and Leelahavanichkul, A. (2023). Cell-free DNA as diagnostic and prognostic biomarkers for adult sepsis: a systematic review and meta-analysis. Sci. Rep. *13*, 19624.

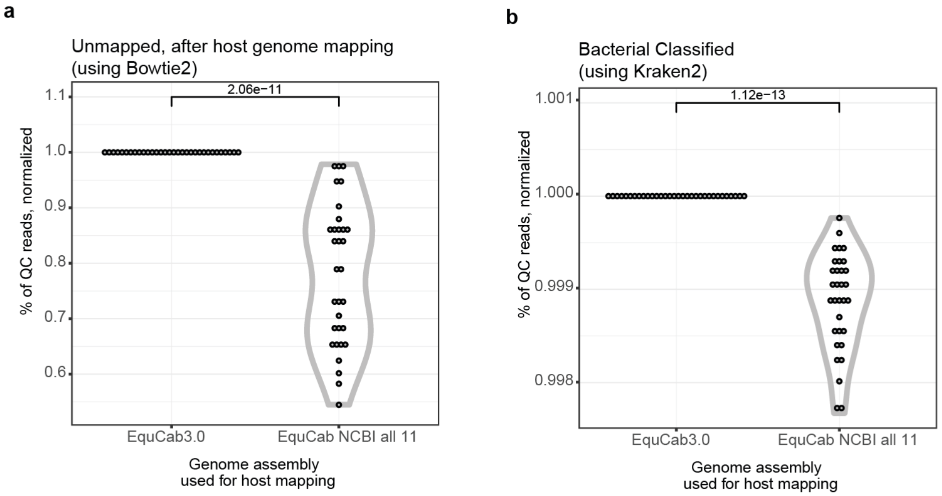
49. Kung, C.-T., Hsiao, S.-Y., Tsai, T.-C., Su, C.-M., Chang, W.-N., Huang, C.-R., Wang, H.-C., Lin, W.-C., Chang, H.-W., Lin, Y.-J., et al. (2012). Plasma nuclear and mitochondrial DNA levels as predictors of outcome in severe sepsis patients in the emergency room. J. Transl. Med. *10*, 130.
50. Yan, H.P., Li, M., Lu, X.L., Zhu, Y.M., Ou-Yang, W.-X., Xiao, Z.H., Qiu, J., and Li, S.J. (2018). Use of plasma mitochondrial DNA levels for determining disease severity and prognosis in pediatric sepsis: a case control study. BMC Pediatr. *18*, 267.
51. Henry, B.M., de Oliveira, M.H.S., Cheruiyot, I., Benoit, J., Rose, J., Favaloro, E.J., Lippi, G., Benoit, S., and Pode Shakked, N. (2022). Cell-Free DNA, Neutrophil extracellular traps (NETs), and Endothelial Injury in Coronavirus Disease 2019-(COVID-19-) Associated Acute Kidney Injury. Mediators Inflamm. *2022*, 9339411.
52. Murni, I.K., Duke, T., Daley, A.J., Kinney, S., and Soenarto, Y. (2018). True Pathogen or Contamination: Validation of Blood Cultures for the Diagnosis of Nosocomial Infections in a Developing Country. J. Trop. Pediatr. *64*, 389–394.
53. Weinstein, M.P. (2003). Blood culture contamination: persisting problems and partial progress. J. Clin. Microbiol. *41*, 2275–2278.
54. Hall, K.K., and Lyman, J.A. (2006). Updated review of blood culture contamination. Clin. Microbiol. Rev. *19*, 788–802.
55. Corley, K.T.T., Pearce, G., Magdesian, K.G., and Wilson, W.D. (2007). Bacteraemia in neonatal foals: clinicopathological differences between Gram-positive and Gram-negative infections, and single organism and mixed infections. Equine Vet. J. *39*, 84–89.
56. Park, S.Y., Chang, E.J., Ledeboer, N., Messacar, K., Lindner, M.S., Venkatasubrahmanyam, S., Wilber, J.C., Vaughn, M.L., Bercovici, S., Perkins, B.A., et al. (2023). Plasma Microbial Cell-Free DNA Sequencing from over 15,000 Patients Identified a Broad Spectrum of Pathogens. J. Clin. Microbiol. *61*. https://doi.org/10.1128/jcm.01855-22.
57. Grumaz, S., Grumaz, C., Vainshtein, Y., Stevens, P., Glanz, K., Decker, S.O., Hofer, S., Weigand, M.A., Brenner, T., and Sohn, K. (2019). Enhanced Performance of Next-Generation Sequencing Diagnostics Compared With Standard of Care Microbiological Diagnostics in Patients Suffering From Septic Shock. Crit. Care Med. *47*. https://doi.org/10.1097/CCM.0000000000003658.
58. Diaz, L.A., Jr, and Bardelli, A. (2014). Liquid Biopsies: Genotyping Circulating Tumor DNA. J. Clin. Oncol. https://doi.org/10.1200/JCO.2012.45.2011.
59. Kalantar, K.L., Neyton, L., Abdelghany, M., Mick, E., Jauregui, A., Caldera, S., Serpa, P.H., Ghale, R., Albright, J., Sarma, A., et al. (2022). Integrated host-microbe plasma metagenomics for sepsis diagnosis in a prospective cohort of critically ill adults. Nature Microbiology *7*, 1805.
60. Serpa, P.H., Deng, X., Abdelghany, M., Crawford, E., Malcolm, K., Caldera, S., Fung, M., McGeever, A., Kalantar, K.L., Lyden, A., et al. (2022). Metagenomic prediction of antimicrobial resistance in critically ill patients with lower respiratory tract infections. Genome Med. *14*, 74.
61. Troll, C.J., Kapp, J., Rao, V., Harkins, K.M., Cole, C., Naughton, C., Morgan, J.M., Shapiro, B., and Green, R.E. (2019). A ligation-based single-stranded library preparation method to analyze cell-free DNA and synthetic oligos. BMC Genomics *20*, 1023.
62. BBMap: A Fast, Accurate, Splice-Aware Aligner (2014).
63. Dai, H., and Guan, Y. (2020). Nubeam-dedup: a fast and RAM-efficient tool to de-duplicate sequencing reads without mapping. Bioinformatics *36*, 3254–3256.
64. Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics *34*, i884–i890.
65. Schubert, M., Lindgreen, S., and Orlando, L. (2016). AdapterRemoval v2: rapid adapter trimming, identification, and read merging. BMC Res. Notes *9*, 88.
66. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.
67. Wright, R.J., Comeau, A.M., and Langille, M.G.I. (2023). From defaults to databases: parameter and database choice dramatically impact the performance of metagenomic taxonomic classification tools. Microb Genom *9*. https://doi.org/10.1099/mgen.0.000949.
68. Lu, J., Rincon, N., Wood, D.E., Breitwieser, F.P., Pockrandt, C., Langmead, B., Salzberg, S.L., and Steinegger, M. (2022). Metagenome analysis using the Kraken software suite. Nat. Protoc. *17*, 2815–2839.
69. Shannon, C.E. (1948). A mathematical theory of communication. Bell Syst. Tech. J. *27*, 379–423.

Supplementary Figures



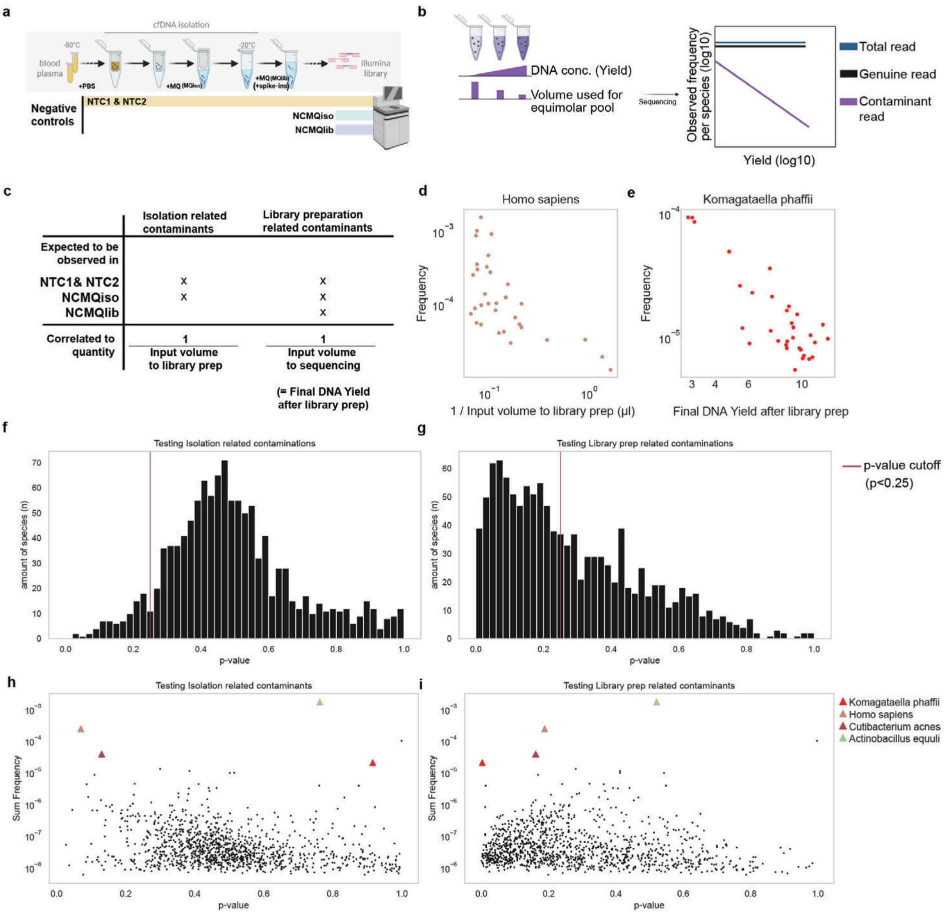
Supplementary Figure 1. Read counts and fractions per Illumina sequencing library, as determined by cfFBI

a. The stacked bar graphs display the number (left) and proportion (right) of paired-end Illumina sequenced reads for each sequencing library. Colors represent the fractions of spike-in reads, duplicated reads, low-quality reads, and high-quality reads, as subsequently determined by the cfFBI workflow pipeline. **b.** The stacked bar graphs illustrate the number of high-quality reads. for each sequencing library. Colors indicate the fractions of reads mapped to the host genome, taxonomically classified and unclassified reads, as subsequently determined by the cfFBI workflow pipeline. **c.** The relative proportion of reads not mapped to host for each sequencing library. Colors the fractions of taxonomically classified and unclassified reads, as subsequently determined by the cfFBI workflow pipeline.



Supplementary Figure 2. Comparison of read host mapping and bacterial classification using different horse genome reference genomes

a. Percentage of quality-controlled reads remaining after mapping to the human reference genome using the cfFBI workflow. Two conditions were compared: using the EquCab3 reference genome alone, and using a compendium that includes all available horse genomes on NCBI (a total of 11 genomes; EquCab NCBI all 11). The results are normalized to those obtained with EquCab3. Statistical analysis was conducted using one-tailed paired t-tests after normalization. **b.** Percentage of quality-controlled reads classified as bacterial by Kraken2, after subtracting host reads via reference genome mapping. Two conditions were compared: using the EquCab3 reference genome alone, and using a compendium that includes all available horse genomes on NCBI (a total of 11 genomes; EquCab NCBI all 11). The Kraken2 bacterial classification results are normalized to those obtained with the EquCab3 reference genome. Statistical analysis was conducted using one-tailed paired t-tests after normalization.

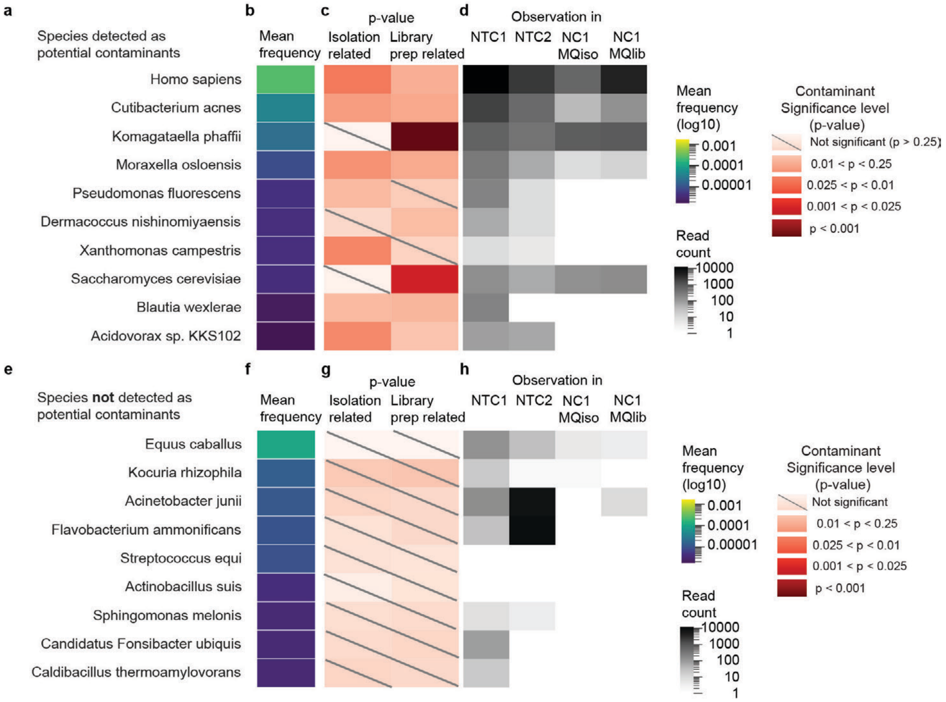


Supplementary Figure 3. Identification of contaminant taxonomic species

a. Schematic of four negative controls used in our study: Non-Template Controls (NTC1, NTC2), which underwent cfDNA DNA isolation and NGS library preparation, and Nuclease-Free water controls (NC1MQiso, NC1MQlib), which underwent only NGS library preparation. **b.** Contaminants in our samples may have been introduced during cfDNA isolation or sequencing library preparation. Contaminant DNA is expected to be present in low, relatively uniform concentrations across laboratory equipment and kits, leading to similar levels across samples. In contrast, cfDNA concentration — and thus the yield from cfDNA isolation or library preparation — can vary significantly between samples. Consequently, the expected frequency of contaminant DNA decreases as the total DNA (yield) in the sample increases (purple), while the frequency of non-contaminant DNA remains consistent (black). Figure adapted from (Davis et al. 2018) **c.** Table presents the expected cfDNA isolation and library preparation contaminants in each negative control sample (top), along with how input volumes serve as an inverse proxy for yield (bottom). **d.** Dotplot showing the relative frequency of *Homo sapiens* reads, an identified cfDNA isolation-related contaminant, versus the input volume used for library preparation (inverse proxy for cfDNA yield) across 32 samples. **e.** Dotplot showing the relative frequency of *Komagataella phaffi* reads, a proven library preparation-related contaminant, versus the input volume used for pooling (inverse proxy for library preparation yield) across 32 samples. **f.** Histogram of p-values derived from the decontam method for cfDNA isolation-related contaminants in all species tested (n = 1167). The p-value detection classification threshold is set at 0.25, indicated by the red vertical line; species falling below this threshold are identified as cfDNA isolation-related contaminants. **g.** Histogram of p-values derived from the decontam method for library preparation-related contaminants in all species tested (n = 1167). The p-value detection classification threshold is set at 0.25, indicated by the red vertical line; species falling below this threshold are identified as library preparation-related contaminants. **h.** Scatter plot displaying the summed frequency across

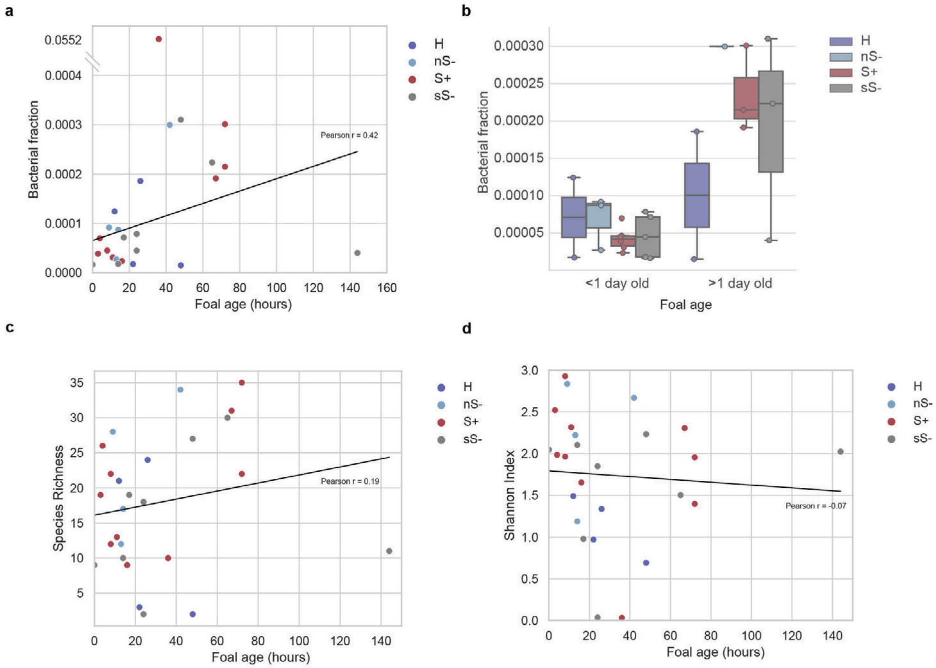
Supplementary Figure 3. *Continued*

all foals against p-values derived from the decontam method for cfDNA isolation-related contaminants. Contaminants *Homo sapiens* ($p < 0.25$) and *Cutibacterium acnes* ($p < 0.25$) are highlighted, indicated by red and brown triangles, respectively. The NGS library preparation-associated contaminant *Komagataella phaffi* (not significant) is also marked with a triangle. Additionally, *Actinobacillus equuli*, a highly abundant microbial species (frequently observed pathogen in previous research) that is not significantly detected as a contaminant, is represented by a green triangle. **i.** Scatter plot displaying the summed frequency across all foals against p-values derived from the decontam method for library preparation-related contaminants, highlighting contaminants. Contaminants *Homo sapiens* ($p < 0.25$) and *Cutibacterium acnes* ($p < 0.25$) are highlighted, indicated by red and brown triangles, respectively. The library preparation-associated contaminant *Komagataella phaffi* ($p < 0.25$) is also marked with a triangle. Additionally, *Actinobacillus equuli*, a highly abundant pathogen that is not significant, is represented by a green triangle.



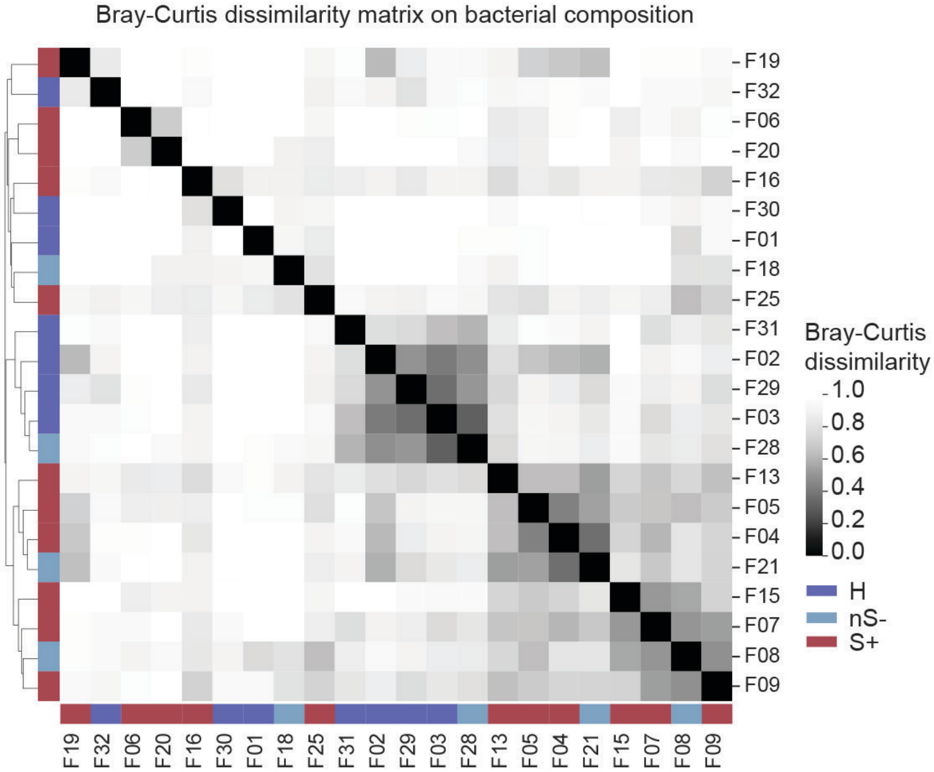
Supplementary Figure 4. Top species detected as contaminants and not detected as contaminants and their significance and observation in negative controls.

a. Top 10 species detected as potential contaminants sorted by mean relative abundance. **b.** Heatmap illustrates the mean frequency of potential contaminant species listed in **a.** The color reflects the log10 mean frequency of each species across all foal samples. **c.** Heatmap displays p-values indicating whether the contaminants are associated with the cfDNA isolation process or the library preparation process. Significance levels range from not significant (light pink, $p > 0.25$) to highly significant (dark red). **d.** Observations of contaminant species in various negative control conditions: NTC1, NTC2, NC1MQiso, and NC1MQlib. The color intensity reflects the log10 count frequency. **e.** Top 10 species non-contaminant species sorted by mean relative abundance. **f.** Heatmap illustrates the mean frequency of potential non-contaminant species listed in **e.** The color reflects the log10 mean frequency of each species across all foal samples. **g.** Heatmap of p-values indicating whether the species not detected as contaminants are related to the isolation process or the library preparation process. All were $p > 0.25$. **h.** Observations of non-contaminant species in various negative control conditions: NTC1, NTC2, NC1MQiso, and NC1MQlib. The color intensity reflects the log10 count frequency.

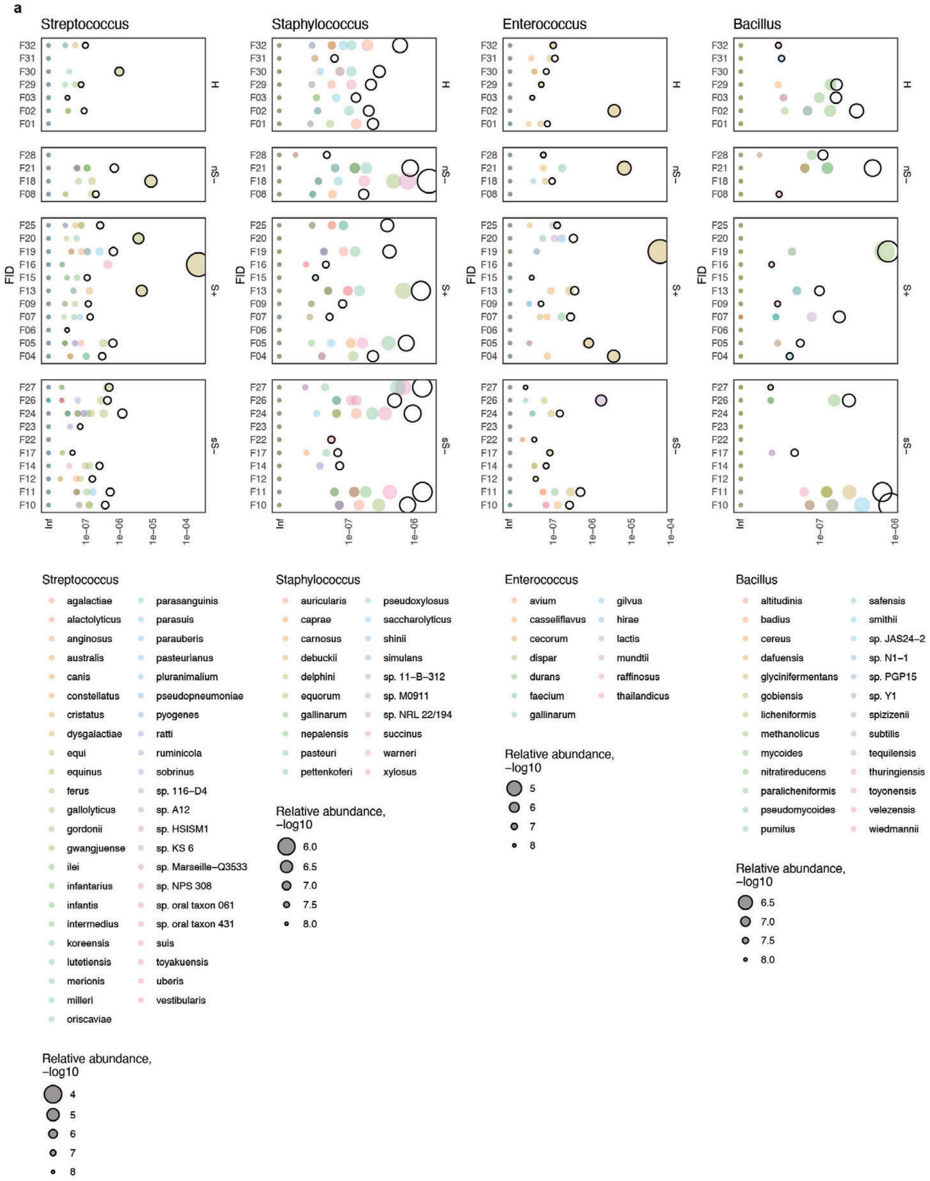


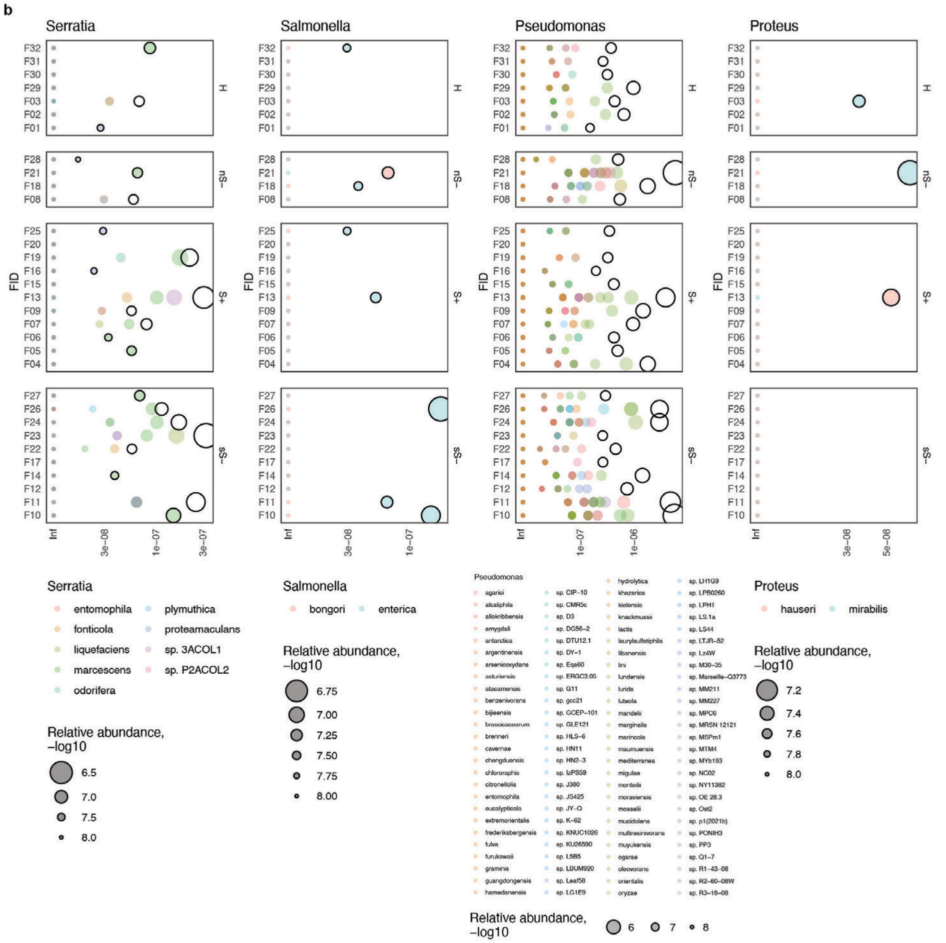
Supplementary Figure 5. Bacterial fraction, species richness, and Shannon index in foals at various ages, categorized by disease status.

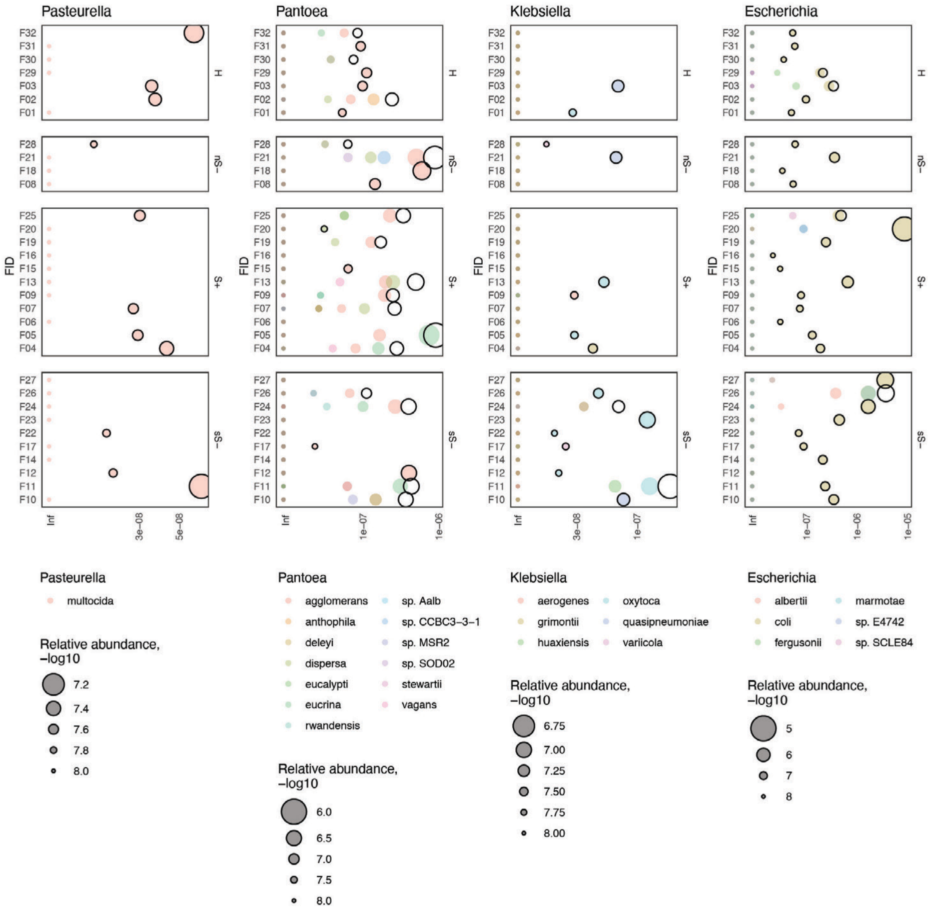
a. Scatter plot showing the bacterial fraction in relation to foal age at presentation (hours). Data points are color-coded based on disease status. A weak Pearson correlation (*r*=0.42) was calculated excluding outlier data point at bacterial fraction = 0.0552 to avoid strong influence on fitting caused by the outlier. If including this data point, the Pearson correlation is *r*=0.03. **b.** Box plot illustrates the bacterial fraction in foals less than one day old compared to those more than one day old. Data points are grouped by health status as shown in **a.**, with whiskers extending to the rest of the distribution within 1.5 times the interquartile range. **c.** Scatter plot depicting species richness in relation to foal age at presentation (hours). Data points are color-coded by health status as in **a.**. Weak correlation was observed (Pearson correlation *r*=0.19). **d.** Scatter plot showing the Shannon index in relation to foal age at presentation (hours). Data points are color-coded by health status as in **a.**. Weak negative correlation was observed (Pearson correlation *r*=0.19).

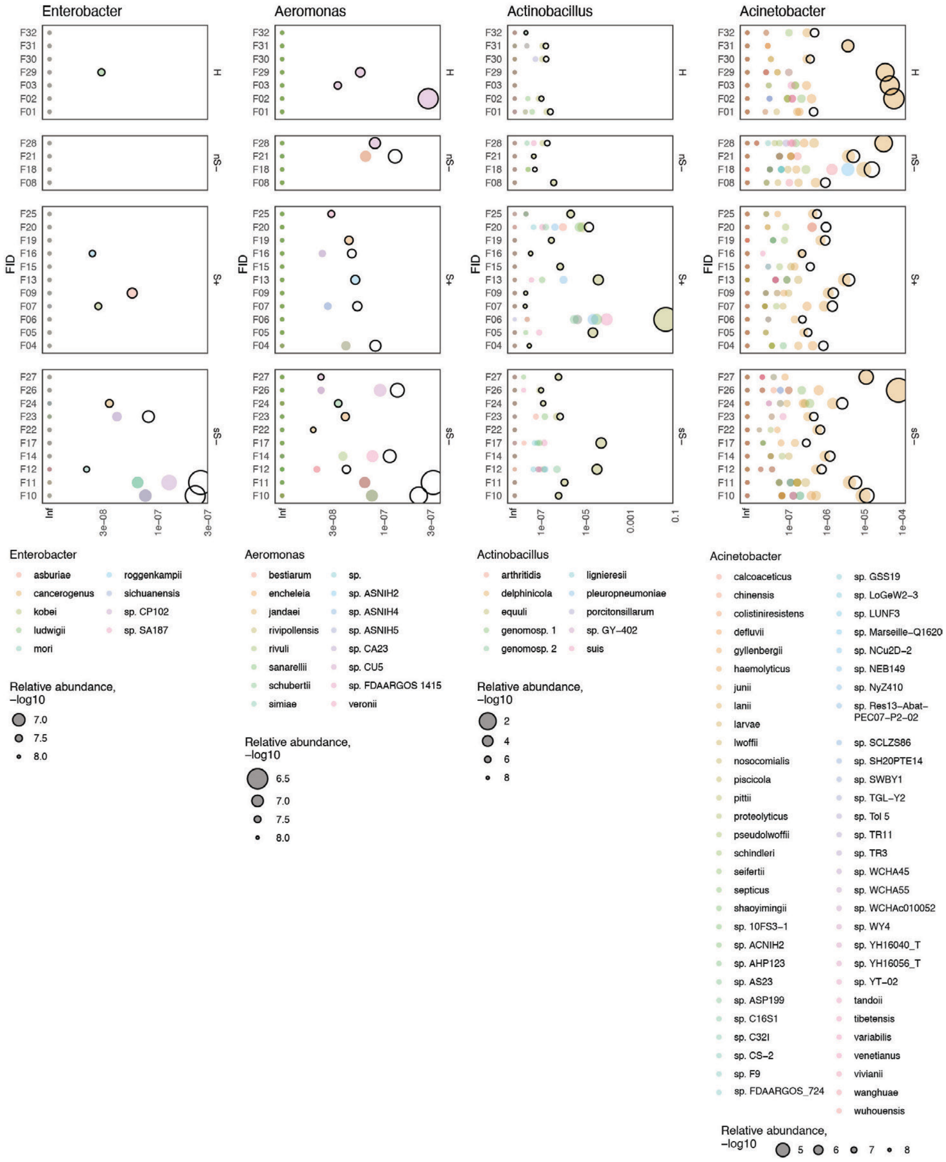


Supplementary Figure 6. Bray-Curtis dissimilarity matrix showing the distances in log transformed bacterial composition between samples. Distances are calculated for each pair of foal plasma samples, and the samples are clustered based on the minimum linkage of Euclidean distances of pairwise Bray-Curtis dissimilarity.


a
Streptococcus
Staphylococcus
Enterococcus
Bacillus
F32
F31
F30
F29
F03
F02
F01
H
F28
F21
F18
F08
nS−
F25
F20
F19
F16
F15
F13
F09
F07
F06
F05
F04
S+
F27
F26
F24
F23
F22
F17
F14
F12
F11
F10
sS−
FID
Inf
1e−07
1e−06
1e−05
1e−04
Streptococcus
agalactiae
alactolyticus
anginosus
australis
canis
constellatus
cristatus
dysgalactiae
equi
equinus
ferus
gallolyticus
gordonii
gwangjuense
ilei
infantarius
infantis
intermedius
koreensis
lutetiensis
merionis
milleri
oriscaviae
parasanguinis
parasuis
parauberis
pasteurianus
pluranimalium
pseudopneumoniae
pyogenes
ratti
ruminicola
sobrinus
sp. 116−D4
sp. A12
sp. HSISM1
sp. KS 6
sp. Marseille−Q3533
sp. NPS 308
sp. oral taxon 061
sp. oral taxon 431
suis
toyakuensis
uberis
vestibularis
Relative abundance, −log10
4
5
6
7
8
Staphylococcus
auricularis
caprae
carnosus
debuckii
delphini
equorum
gallinarum
nepalensis
pasteuri
pettenkoferi
pseudoxylosus
saccharolyticus
shinii
simulans
sp. 11−B−312
sp. M0911
sp. NRL 22/194
succinus
warneri
xylosus
Relative abundance, −log10
6.0
6.5
7.0
7.5
8.0
Enterococcus
avium
casseliflavus
cecorum
dispar
durans
faecium
gallinarum
gilvus
hirae
lactis
mundtii
raffinosus
thailandicus
Relative abundance, −log10
5
6
7
8
Bacillus
altitudinis
badius
cereus
dafuensis
glycinifermentans
gobiensis
licheniformis
methanolicus
mycoides
nitratireducens
paralicheniformis
pseudomycoides
pumilus
safensis
smithii
sp. JAS24−2
sp. N1−1
sp. PGP15
sp. Y1
spizizenii
subtilis
tequilensis
thuringiensis
toyonensis
velezensis
wiedmannii
Relative abundance, −log10
6.5
7.0
7.5
8.0

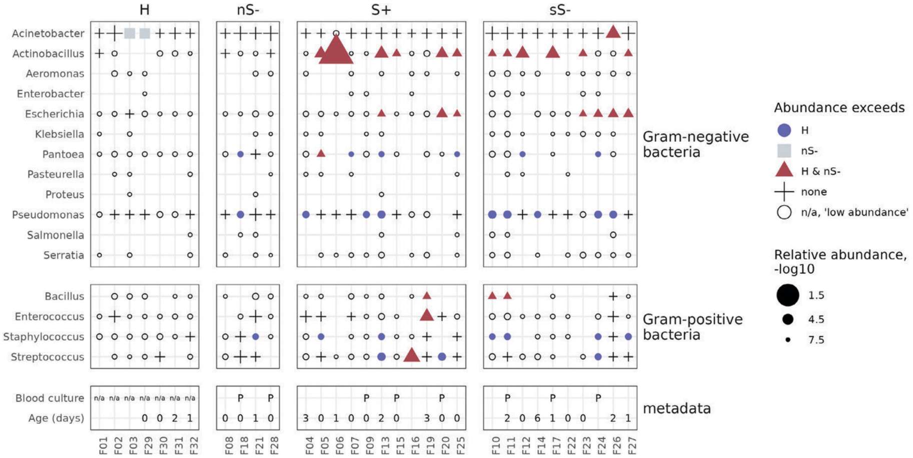

b
Serratia
Salmonella
Pseudomonas
Proteus
FID
F32
F31
F30
F29
F03
F02
F01
F28
F21
F18
F08
F25
F20
F19
F16
F15
F13
F09
F07
F06
F05
F04
F27
F26
F24
F23
F22
F17
F14
F12
F11
F10
H
nS−
S+
sS−
Inf
3e−08
1e−07
3e−07
1e−06
5e−08
Serratia
entomophila
fonticola
liquefaciens
marcescens
odorifera
plymuthica
proteamaculans
sp. 3ACOL1
sp. P2ACOL2
Relative abundance, −log10
6.5
7.0
7.5
8.0
Salmonella
bongori
enterica
Relative abundance, −log10
6.75
7.00
7.25
7.50
7.75
8.00
Pseudomonas
agarici
alcaliphila
allokribbensis
amygdali
antarctica
argentinensis
arsenicoxydans
asturiensis
atacamensis
benzenivorans
bijieensis
brassicacearum
brenneri
cavernae
chengduensis
chlororaphis
citronellolis
entomophila
eucalypticola
extremorientalis
frederiksbergensis
fulva
furukawaii
graminis
guangdongensis
hamedanensis
sp. CIP−10
sp. CMR5c
sp. D3
sp. DG56−2
sp. DTU12.1
sp. DY−1
sp. Eqa60
sp. ERGC3:05
sp. G11
sp. gcc21
sp. GCEP−101
sp. GLE121
sp. HLS−6
sp. HN11
sp. HN2−3
sp. IzPS59
sp. J380
sp. JS425
sp. JY−Q
sp. K−62
sp. KNUC1026
sp. KU26590
sp. L5B5
sp. LBUM920
sp. Leaf58
sp. LG1E9
hydrolytica
khazarica
kielensis
knackmussii
lactis
laurylsulfatiphila
libanensis
lini
lundensis
lurida
luteola
mandelii
marginalis
marincola
maumuensis
mediterranea
migulae
monteilii
moraviensis
mosselii
mucidolens
multiresinivorans
muyukensis
ogarae
oleovorans
orientalis
oryzae
sp. LH1G9
sp. LPB0260
sp. LPH1
sp. LS.1a
sp. LS44
sp. LTJR−52
sp. Lz4W
sp. M30−35
sp. Marseille−Q3773
sp. MM211
sp. MM227
sp. MPC6
sp. MRSN 12121
sp. MSPm1
sp. MTM4
sp. MYb193
sp. NC02
sp. NY11382
sp. OE 28.3
sp. Ost2
sp. p1(2021b)
sp. PONIH3
sp. PP3
sp. Q1−7
sp. R1−43−08
sp. R2−60−08W
sp. R3−18−08
Relative abundance, −log10
6
7
8
Proteus
hauseri
mirabilis
Relative abundance, −log10
7.2
7.4
7.6
7.8
8.0


Pasteurella
Pantoea
Klebsiella
Escherichia
FID
H
−Su
+S
−Ss
F32
F31
F30
F29
F03
F02
F01
F28
F21
F18
F08
F25
F20
F19
F16
F15
F13
F09
F07
F06
F05
F04
F27
F26
F24
F23
F22
F17
F14
F12
F11
F10
Inf
3e−08
5e−08
1e−07
1e−06
1e−05
Pasteurella
multocida
Relative abundance, −log10
7.2
7.4
7.6
7.8
8.0
Pantoea
agglomerans
anthophila
deleyi
dispersa
eucalypti
eucrina
rwandensis
sp. Aalb
sp. CCBC3−3−1
sp. MSR2
sp. SOD02
stewartii
vagans
Relative abundance, −log10
6.0
6.5
7.0
7.5
8.0
Klebsiella
aerogenes
grimontii
huaxiensis
oxytoca
quasipneumoniae
variicola
Relative abundance, −log10
6.75
7.00
7.25
7.50
7.75
8.00
Escherichia
albertii
coli
fergusonii
marmotae
sp. E4742
sp. SCLE84
Relative abundance, −log10
5
6
7
8



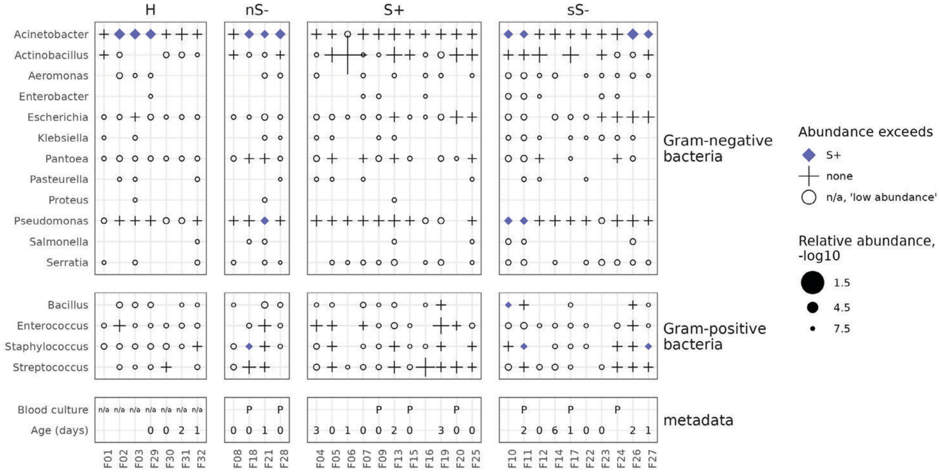
Supplementary Figure 7. Relative abundance of pathogenic bacterial species and genera.

a. The dot plot illustrates the relative abundance (x-axis) of species from four selected Gram-negative pathogenic bacterial genera. It also shows the total relative abundance of these genera (sum of species) across samples (represented by black circles). Each dot color represents a different species, while the dot size indicates their relative abundance. For species with no observations, relative abundance is indicated as infinity (Inf). **b.** The dot plot illustrates the relative abundance (x-axis) of species from four selected Gram-negative pathogenic bacterial genera. It also shows the total relative abundance of these genera (sum of species) across samples (represented by black circles). Each dot color represents a different species, while the dot size indicates their relative abundance. For species with no observations, relative abundance is indicated as infinity (Inf).



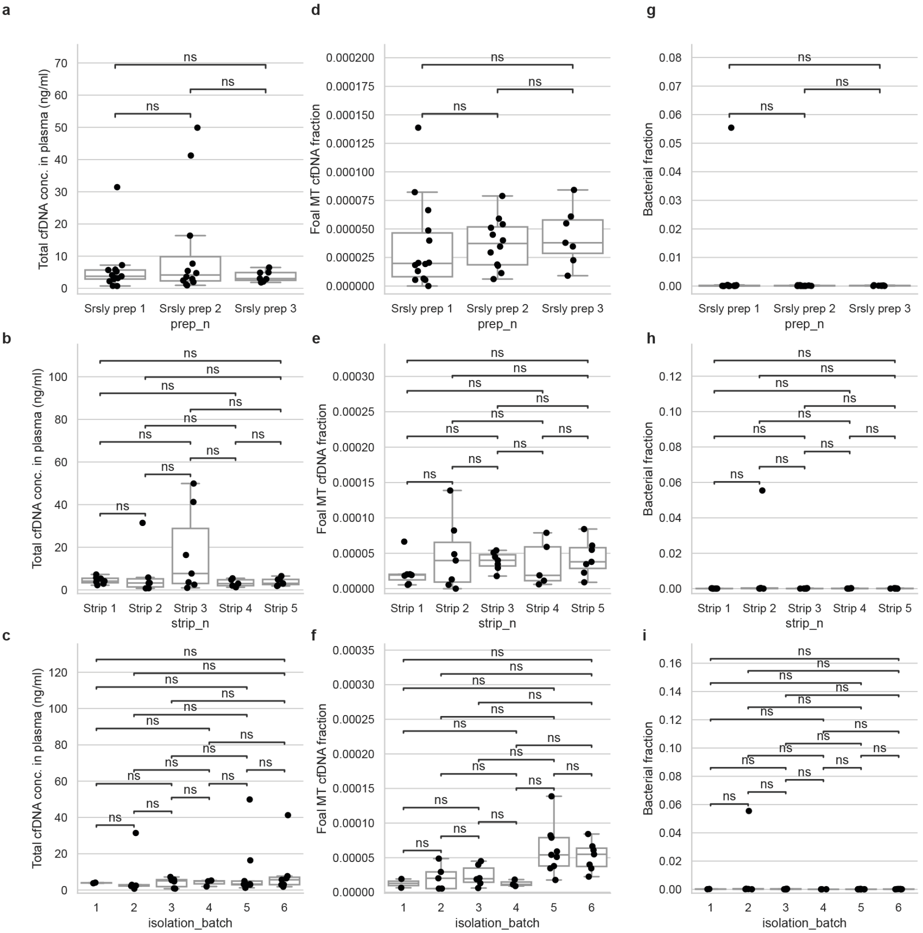
Supplementary Figure 8. Detection of pathogenic bacteria in sick foals

Dotplot displaying the detection of the 16 most frequently cultured pathogenic genera. Gram-negative species are shown at the top, and Gram-positive species are shown in the middle. Each dot on the plot represents a genus that was detected with at least 1 read. The different symbols and colors convey the relative abundance of the genera in comparison to healthy (H) foals and nS- foals: Blue circles indicate genera with a relative abundance higher than in H foals. Light-gray squares indicate genera with a relative abundance higher than in nS- foals. Red triangles denote genera with a relative abundance higher than in H and nS- foals. White circles represent genera detected at low abundance (fewer than 10 reads) and therefore not compared to either H or nS- foals. Plus signs ('+') represent genera that are not elevated compared to either healthy (H) or non-suppurative (nS-) foals. Metadata is displayed at the bottom, including blood culture results ('P' for positive, 'n/a' for not performed) and age at hospital presentation, if known.



Supplementary Figure 9. Relative abundance of pathogenic bacterial species and genera compared to S+.

Dotplot displaying the detection of the 16 most frequently cultured pathogenic genera. Gram-negative species are shown at the top, and Gram-positive species are shown in the middle. Each dot on the plot represents a genus that was detected with at least 1 read. The different symbols and colors convey the relative abundance of the genera in comparison to healthy (H) foals and nS- foals: Blue diamonds indicate genera with a relative abundance higher than in S+ foals. White circles represent genera detected at low abundance (fewer than 10 reads) and therefore not compared to either S+ foals. Plus signs (‘+’) represent genera that are not elevated compared to S+ foals. Metadata is displayed at the bottom, including blood culture results (‘P’ for positive, ‘n/a’ for not performed) and age at hospital presentation, if known.



Supplementary Figure 10. Influence of various batch effects on total cfDNA concentration in plasma, foal mitochondrial cfDNA fraction, and bacterial fractions.

Batch effects include different sample preparation methods, strip variations, and isolation batches. Each box plot displays raw data points and represents the 25th percentile (bottom), median (middle), and 75th percentile (top), with whiskers extending to the rest of the distribution within 1.5 times the interquartile range. A Mann-Whitney U Test with Bonferroni correction was performed, revealing no significant differences (ns) among all tests, indicating the robustness and reproducibility of the cfDNA measurement procedures. **a.** Total cfDNA concentration in plasma (ng/mL) for three different library preparation batches, SRSLY prep 1, SRSLY prep 2, and SRSLY prep 3. **b.** Total cfDNA concentration in plasma (ng/mL) across five different strip locations (Strip 1 to Strip 5). **c.** Total cfDNA concentration in plasma (ng/mL) across six isolation batches. **d.** Fraction of MT cfDNA for the three library preparation batches. **e.** Fraction of MT cfDNA across the five strip locations. **f.** Fraction of MT cfDNA across the six isolation batches. **g.** Bacterial fraction of cfDNA for the three library preparation batches. **h.** Bacterial fraction of cfDNA across the five strip locations. **i.** Bacterial fraction of cfDNA across the six isolation batches.

Supplementary table

Supplementary Tables 1-11 can be found here: https://www.biorxiv.org/content/10.1101/2024.10.31.620104v1

CHAPTER 5

Discussion

Emmy Wesdorp [1,2], Myrthe Jager [1,2], Jeroen de Ridder [1,2]

[1] Center for Molecular Medicine, University Medical Center Utrecht, Utrecht University, 3584 CX Utrecht, The Netherlands

[2] Oncode Institute, 3521 AL, Utrecht, The NetherlandS

Pathogen identification and microbial characterization through next- and third-generation sequencing has become fundamental practice over the past 15 years (see Box 2 of the *General Introduction*). Recently, metagenomic next-generation sequencing (NGS) of cell-free DNA (cfDNA) from plasma and other liquid biopsy samples has emerged as a promising new diagnostic tool. These diagnostic metagenomic cfDNA NGS approaches identify pathogens by analyzing their relative abundance in patient liquid biopsy samples compared to control samples. Notable commercial applications include the Karius™ test and DISQVER® from Noscendo. While minimally invasive and offering unbiased insights, the widespread implementation of microbial cfDNA NGS into routine clinical practice requires overcoming several technical, clinical and validation challenges.

In this thesis, we aimed to address several key technical challenges in the field of pathogen identification through NGS of cfDNA. In collaboration with clinical partners, our focus was on optimizing second-generation NGS workflows, spanning both wet lab preparation and computational analysis, to enhance pathogen detection, in the hope to ultimately advance the metagenomic cfDNA NGS techniques. Additionally, we explored the clinical potential of cfDNA NGS in two relatively underexplored contexts with urgent diagnostic needs, with the goal of expanding insights in its diagnostic utility.

In *Chapter 2*, we developed a DNA NGS workflow for diagnosing invasive mold diseases (IMD) in immunocompromised children, optimizing fungal DNA capture and identification in plasma and BAL samples, to ultimately evaluate the potential of cfDNA NGS in addressing the critical need for reliable and rapid pediatric IMD diagnostics. In *Chapter 3*, we present a detailed fragmentomics analysis of cfDNA from immunocompromised IMD patients and controls, aiming to establish a method for enriching cfDNA from *Aspergillus fumigatus*, a fungus frequently responsible for IMD[1,2]. In *Chapter 4* we shifted focus to bacterial infections, investigating the application of single-stranded cfDNA NGS for sepsis diagnosis in neonatal foals.

This discussion is divided into two parts. **Part I** focuses on the technical challenges, outlining our contributions to workflow development, emphasizing key considerations for future studies, and providing insights into what we think would be potential avenues for further improvements. **Part II** contextualizes the pathogen identification results from *Chapters 2* (IMD) and *4* (sepsis) by comparing them with established diagnostic standards. It explores the potential value of cfDNA NGS as a diagnostic tool in these settings and beyond, discussing its broader clinical utility and offering recommendations for future advancements.

Part I

Contributions of our work to addressing technical challenges

This thesis aimed to tackle several technical challenges in cfDNA-based diagnostics, with a particular emphasis on optimizing sample preparation and computational workflows. Below, we highlight the advancements made in these areas, contributing to the ongoing incremental progress within the field.

Improved sample workup for sequencing

To address the typically low quantities of microbial cfDNA, our studies utilized customized workflows that enhanced the abundance of targeted microbial taxa through improved wet lab methodologies. In *Chapter 2*, we evaluated various library preparation strategies (double-stranded from genomic DNA and cfDNA, as well as single-stranded library preparation from cfDNA) and found that single-stranded cfDNA (ss-cfDNA) preparation from BAL fluid and plasma samples improved the recovery of pathogenic fungi, particularly enhancing the relative abundance of *Aspergillus*, although the increase was not statistically significant. While similar enhancements have been noted for bacterial cfDNA[3], we have now demonstrated this for fungal detection for the first time, which may prompt a shift from conventional double-stranded cfDNA library preparation to ss-cfDNA methods upon validation. In *Chapter 4*, we further refined the ss-cfDNA methodology to enrich for short bacterial DNA fragments using the small-fragment retention version of the ss-cfDNA SRSLY library preparation protocol. Together with bead-based size selection, we were able to obtain higher bacterial cfDNA yields. This approach could also benefit future fungal detection, as *in silico* analyses presented in *Chapter 3* suggest that additional size selection after ss-cfDNA may enhance fungal abundance.

In *Chapter 3*, we also explored the potential of end-motif selection as a novel approach. Although CG end-motif selection shows promise for enriching fungal reads, this strategy is likely to require increased cfDNA input to offset the diagnostic test performance losses observed during *in silico* end-motif selection. *In vitro* validation is therefore essential for confirming its efficacy. If validated, end-motif-based enrichment could provide a foundation for broader pathogen detection strategies. Specifically, CG motifs, which are underrepresented in host genomic DNA, may be a more general target for enhancing pathogen abundance. Literature indicates that GC content is higher in bacteria, protozoa, and fungi than in most mammalian vertebrates[4]. More generally, motif frequency analysis of both host and pathogen genomes could guide the design of similar enrichment strategies, based on different motifs, for pathogen detection in other clinical contexts.

Optimizing computational workflows

A central aspect of cfDNA microbial NGS is minimizing false positives while maximizing true positive diagnostic results, which relies heavily on accurately determining the taxonomic origin of each sequencing read. Both the cfSPI pipeline (*Chapter 2*) and the cfFBI pipeline (*Chapter 4*) try to achieve this through a combination of host-genome read alignment and kraken2-based taxonomic classification.

To enhance origin identification through host mapping, we followed recent recommendations from the cancer microbiome research field[5], refining the standard reference genome mapping step by performing dual mapping to both GRCh38 and CHM13v2. While previous work focused on preventing inflated bacterial counts, our work in *Chapter 2* is the first to demonstrate that this approach is also crucial for preventing inflated fungal counts. Building on these findings, we explored whether the same strategy could be applied to non-human host diagnostics. In *Chapter 4*, we aligned cfDNA reads to a compendium of host (equine) genomes, which not only increased the identification of host reads but also helped prevent inflated bacterial counts in this context. This marked the first suggestion that this approach is effective across multiple mammalian diagnostic settings. To further enhance host-read identification after host-genome alignment, we incorporated the complete set of host reference genomes into our kraken2 taxonomic database, as described in *Chapters 2* and *4*. In *Chapter 2*, we demonstrated that this integration is helpful to counter inflated microbial counts. While these steps represent improvements in taxonomic origin identification accuracy, a recent study[6] has suggested that host-read identification might be improved by aligning against the human pangenome[7]. Together, these host identification strategies provide a framework for improving microbial identification in mammalian cfDNA sequencing.

In *Chapter 2*, we specifically focused on refining fungal read classification by kraken2, with an emphasis on *Aspergillus* species-level identification. We extended the kraken2 database to include additional fungal reference genomes from both pathogenic and closely related non-pathogenic species, expanding *Aspergillus* coverage from 7 to over 250 species, approaching the estimated total of 300+ species. This expansion revealed that no single database guarantees optimal results: smaller databases enable precise species-level classification but risk false negatives due to limited species representation, while larger databases enhance taxonomic coverage but may obscure species-level resolution, increasing false negatives. To overcome these challenges, we implemented a dual-database strategy within the cfSPI pipeline, using one database for genus-level identification and another for species-level classification. While a single-database solution would be ideal for practical applications due to its simplicity and ease of implementation, our dual-database approach takes advantage of simultaneous processing within the snakemake pipeline. This innovative strategy improves *Aspergillus* detection across multiple taxonomic levels, representing a novel solution in the metagenomic cfDNA NGS field by effectively balancing taxonomic coverage and classification precision.

Additionally, we adjusted kraken2 classification thresholds to enhance *Aspergillus* detection, drawing from earlier work[8] but taking it further through optimization via *in silico* evaluations of detection limits, expressed in molecules per million. Our approach proved effective by identifying an optimal threshold based on control background and classification rates from our kraken2 database, with the threshold found to be dependent on the specific database used.

Limitations and ongoing challenges in our works

Like other NGS-based methods, cfDNA-based diagnostic approaches are continuously evolving, driven by new insights, advances in sequencing technologies, and bioinformatic developments. In the following section, we reflect on the limitations of our own work. We then relate these limitations to challenges highlighted in previous studies, and speculate on directions for future advancements.

Dealing with contamination

Non-source DNA can contaminate samples at various stages, originating from sources such as extraction kits, laboratory consumables, researcher DNA, and cross-sample contamination[9–11]. In our work in *Chapter 4*, we observed human genomic sequencing contamination (0,33%) in the foal dataset, likely due to either contamination during sample collection, during wet-lab handling or computational spillover from sequencing foal samples alongside human cfDNA. Our decontamination efforts suggested that 6,42% of the bacterial reads were likely contaminants. While these findings, as discussed in detail in *Chapter 4*, underscore the widespread presence of contamination in our samples, they also reemphasize the importance of careful contamination control, especially in low microbial biomass liquid biopsy samples[10–12].

Another notable example of contamination control in our work involved *Komagataella phaffii*, which we traced to a specific ss-cfDNA library preparation kit batch. Contamination levels increased from a maximum of 1 read per million in earlier batches to consistent detection of *K. phaffii* across all samples, reaching up to 60 reads per million. Discussions with the supplier confirmed the issue, which was attributed to a new reagent supplier. This highlights the critical need for rigorous in-house quality control of new batches to detect and address batch-specific pathogen abundance changes.

By identifying these sources, we can better avoid misinterpreting contaminant signals as true biological findings, which would otherwise have far-reaching implications. These efforts are essential to maintaining integrity and accuracy of cfDNA-based metagenomic approaches.

Alternative read classification methods

As mentioned previously, the accuracy of microbial sequencing-based diagnostics hinges e.g. on accurately identifying the taxonomic origin of each sequenced cfDNA molecule, indicating the microbial source of the cfDNA. Several taxonomic classification tools have been developed for this purpose (see *General Introduction*). In this thesis, we used kraken2[13], a DNA-to-DNA classifier, selected for its speed, sensitivity, and ability to identify the lowest common ancestor (LCA) at varying taxonomic levels. The LCA can range from high-level classifications such as kingdom or phylum to more specific levels like species or strain. In our clinical application settings, however, we were mostly interested in identifying genus and species levels, as these provide the necessary resolution for diagnosing microbial infections and guiding treatment decisions. Given the lack of consensus on the optimal tool for this task, this choice was based on informed judgment. However, we now recognize that direct mapping to a compendium of relevant fungal genomes might have been a simpler and more computationally efficient alternative in *Chapter 2* and *3*. This insight emerged retrospectively, after reflecting on *Chapter 2*, where we observed that kraken2-classified *Aspergillus fumigatus* reads did not map well to non-fumigatus *Aspergillus* genomes (as shown in *Chapter 2*, Supplementary Fig. 15). Direct genome mapping to a selected set of consensus genomes could thus offer an appealing, straightforward alternative. Furthermore, tools like PathoScope[14] could regain relevance in such scenarios. While PathoScope can be computationally demanding when working with large reference genome sets and requires a pre-built reference database, it could operate efficiently for applications involving a limited number of genomes, such as in the case of *Aspergillus*. Its capability to reassign ambiguously mapped reads makes it particularly useful for focused analyses. If mapping-based approaches can be shown to minimize misclassification effectively, they could provide a computationally efficient pathway for future targeted diagnostic research

— particularly in fungal diagnostics — extending beyond the scope of this thesis. However, this reconsideration is specific to applications where shotgun sequencing is combined with a targeted focus, such as the single-genus analysis in *Chapter 2*, but would not be appropriate for broader diagnostic settings, like those described in *Chapter 4*, where a very diverse array of potential pathogens must be considered.

Addressing host cfDNA variability by normalization

To define elevated levels of pathogen species, we assessed microbial abundance relative to the total number of quality-filtered reads, a common approach in the field. However, we did not account for variability in host cfDNA concentrations across samples. Factors like exercise[15–17] and certain pathological conditions (e.g., hematological cancers[18]) can significantly influence (host) cfDNA levels. This is particularly relevant in *Chapter 2*, where patients undergoing hematopoietic stem cell transplantation or those with hematological malignancies or graft-versus-host disease were included — conditions (that tend to be) accompanied by an increase in cell death. A potentially more accurate approach would be to quantify microbial organisms in terms of reads per microliter of cfDNA source fluid, similar to the approach used by Karius in their work (as described recently again in the work of Huygens *et al.*[19]). This emphasizes the need to document source fluid volumes, which we did not do, thereby limiting the implementation of this strategy across our studies. Despite this missing normalization step, our approach still yielded high specificity, with only one false positive in *Chapter 2*, demonstrating that our normalization strategy, although imperfect, was effective in the current setting.

Looking ahead, it would be valuable to compare normalization by volume versus normalization based on the total number of quality-controlled reads. To assess this, testing various normalization strategies using samples where the abundance of at least one or a few microbes is known could provide insight. Specifically, spiking sheared genomic DNA into liquid biopsy samples devoid of background signals for these pathogens could help determine whether normalization by volume effectively addresses the challenges outlined above.

Statistical challenges for pathogen identification: not just a matter of increasing cohort size

In cfDNA analysis, distinguishing true pathogen elevations from background microbial levels is crucial. However, defining a “normal” background remains an unresolved issue in the field. In this context, the potential existence of a circulating cell-free “microbiome”[20] composed of microbial cfDNA sequences, warrants consideration. This concept is supported by the identification of microbial taxa in one cfDNA study, with findings corroborated by subsequent research[21]. Some studies have shown that the circulating cell-free microbiome is predominantly made up of *Proteobacteria phylum*, followed by *Actinobacteria*, *Firmicutes*, and *Bacteroidetes* species[20,22,23]. However, the existence of a stable circulating microbiome remains debated, particularly as other studies have highlighted differences in cfDNA microbial genus abundance between healthy individuals and patients with immune-related diseases[24]. These differences raise two possibilities: either the circulating cfDNA microbiome does not exist as a stable entity, or its diagnostic potential is restricted to a (yet undefined) set of taxa that remain consistent regardless of disease states. Given these uncertainties, the most practical approach at present is to use control groups composed of patients with similar demographics and clinical backgrounds but without signs of infection. This strategy provides the most reliable baseline for distinguishing pathogen-derived cfDNA from background microbial levels.

Obtaining a sufficient number of control samples with similar clinical status and demographic characteristics for our cfDNA work was challenging. In *Chapter 4*, we combined two distinct control groups to increase the number of background samples. Notwithstanding this pooling, in both *Chapters 2* and *4*, the limited sample size hindered robust statistical testing for pathogen identification, and small variations in the control set could significantly affect the results. To address this challenge, we adopted two different, both conservative approaches in this thesis to identify elevated levels of pathogenic species/genera. One, in *Chapter 4* we compared (aggregated) normalized read counts of pathogenic bacteria against the highest observations in control background samples. Two, in *Chapter 2*, we utilized the average from pairwise one-tailed Fisher's exact tests between controls and individual patient samples for similar reasons.

Expanding cohort sizes is essential for future studies. For context, a follow-up study on cfSPI's plasma-based *Aspergillus* testing would require 150 patients (based on a 15% prevalence, with 22 plasma samples from invasive pulmonary aspergillosis patients and 128 from immunocompromised controls), significantly larger than the cohort described in this thesis. A larger cohort would allow the use of other, more common, (i.e. 97.5th percentile) statistical thresholds based on healthy control samples[25] or SIQ scores (Sepsis Indicating Quantifier), which incorporate both abundance and statistical significance to identify pathogens [26].

However, it is important to reemphasize that simply increasing cohort size or statistical power will not resolve all challenges, especially the specificity issues as brought up in previous works[27] where issues persisted even with larger cohorts. In studies validating Karius cfDNA NGS in sepsis patients, for example, it was found that 22.8% of their 167 asymptomatic controls also tested positive for at least one pathogen[25]. Similarly, a recent multicenter study conducted in the Netherlands and Belgium found that among patients suspected of invasive microbial disease who tested positive for microbial cfDNA on the Karius test, over 50% were found to have non-relevant pathogens identified by the tests[19]. These studies, encompassing hundreds of plasma samples, raise a critical question: Are these false positives, or does the elevated pathogen abundance mean that the detected pathogens are present, but not causally linked to the disease?

Beyond detection: establishing causality in infectious diseases

Currently, cfDNA sequencing cannot differentiate between live, dead, or dying microorganisms, nor can it identify their anatomical location. The presence of microbial cfDNA can result from various scenarios, such as active secretion, release during colonization, microbial translocation, or ongoing infection. Therefore, detecting elevated microbial DNA does not necessarily indicate an active infection. For instance, a patient may show increased levels of a pathogen's cfDNA due to past exposure or colonization, without it being linked to current disease symptoms — or the absence thereof. This could help explain the pathogen reads observed in healthy individuals in the Karius test[25] and in some nSIRS-negative foals in *Chapter 4*. For future work a key challenge lies in distinguishing between harmless colonizers (which may not pose an immediate threat), transient microbes (which are not necessarily problematic), and pathogens actively involved in disease pathogenesis (which require treatment). Achieving this distinction would be a significant advancement, and if realized, could offer a major advantage for cfDNA over other diagnostic techniques, such as PCR, by providing a more nuanced understanding of microbial presence and its relevance to patient health.

Future technical perspectives

Despite the numerous challenges, the field of cfDNA NGS for pathogen identification is in a developing stage, and poised for significant technical advancements. In this part of the discussion, I highlight two key areas that, in my view, warrant particular attention in future research:

Reducing costs and improving turnaround times

The high cost associated with Illumina-based cfDNA sequencing, such as the Karius test priced at approximately $2,000 per sample[28], remains a barrier to clinical adoption. This is largely due to the need for high sequencing output — typically around 50 million reads per sample — because liquid biopsies contain a majority of host DNA and only minimal microbial DNA. In our own work (Chapter 2), the library preparation and sequencing costs (excluding additional expenses such as personnel, computational analysis, etc.) amounted to €24,500 for analyzing 12 liquid biopsy samples from patients with IMD and several controls. However, only a small fraction (€0.30) of these costs was attributed to sequencing fungal DNA, highlighting inefficiencies in extracting meaningful pathogen data. A similar pattern was observed in Chapter 4, where about €3,150 was spent on cfDNA sequencing, of which less than €10 was allocated to bacterial cfDNA readout.

In contrast, traditional diagnostic methods like blood cultures or PCR typically cost between €50 and €500 per test. It is expected that the cost of sequencing will keep decreasing. For instance, newly emerging sequencing platforms such as the Element Biosciences' AVITI system are good contenders to reduce the cost per sample to a range of €250-€400. This, along with other innovations, could substantially decrease per-sample costs and could make cfDNA-based diagnostics more accessible for routine clinical use.

To further reduce costs, enhancing cfDNA sequencing efficiency through pathogen enrichment and/or host DNA depletion strategies presents an interesting alternative. Techniques such as size selection and CG end-motif selection (*Chapter 3*) show promise in increasing pathogen relative abundance. Additionally, biotinylated RNA baits and magnetic beads can selectively capture pathogen DNA[29]. Host DNA depletion methods from metagenomic sequencing[29], like MBD2-based capture[30] and methylation-dependent restriction endonucleases, could also be adapted for cfDNA. CRISPR/Cas9-based methods present potential alternatives, either by enriching pathogen regions using inactive Cas9 or by using guide RNAs to cut host sequences[31]. While these strategies may require more input material, potentially posing challenges in certain clinical settings, they could reduce sequencing output per sample and improve turnaround times.

Currently, Illumina-based cfDNA sequencing, such as the Karius test and theoretically our work if applied in a diagnostic setting, typically takes around 3-4 days from sample collection to result[32]. However, real-time sequencing technologies such as Oxford Nanopore Technologies (ONT) offer the potential to reduce both costs and turnaround times, though with lower output. Recent studies have demonstrated a rapid 6-hour turnaround for cfDNA-based pathogen identification in bloodstream infections using ONT[33], highlighting its promise for clinical microbiological diagnostics.

Complementary readout to mere microbial cfDNA elevation

The concept that microbial cfDNA offers insights beyond mere pathogen presence — extending to its breakdown fragmentomic characteristics — is supported by several studies. For instance, viral cfDNA sequencing can reveal unique fragmentation patterns indicative of viral latency and immune responses[34]. Similarly, bacterial genome coverage could help identify fast-dividing species, with replication origins and termination points offering insights into pathogen growth[35]. Past research has also shown that contaminant and pathogenic bacteria exhibit distinct end-motifs[36], which could serve as markers for pathogen identification. Investigating fragmentomic characteristics, such as these end-motifs, could furthermore open new avenues for understanding pathogen behavior and causality in infectious diseases.

Microbial identification using cfDNA-based diagnostics, combined with host-response profiling, could offer deeper insights into a patient's condition, as previously demonstrated[37,38]. In this context, we explored mitochondrial host DNA levels and host end-motifs in *Chapter 4* as indicators of SIRS and its prognosis in nSIRS-positive foals. Additionally, in *Chapter 2*, BAL samples with positive *Aspergillus* cfDNA NGS results, clustered based on host end-motifs, suggesting that variations in these motifs may correlate with clinical status. Although exploratory, these findings suggest valuable information may be embedded in these markers, and combining them with metagenomic analysis could be a promising next step for these applications.

Looking ahead, integrating host infectious disease status markers, such as CRP (C-reactive protein) and white blood cell count, along with temporal data from traditional diagnostics like PCR or blood cultures, could help to improve our ability to differentiate between active, past, or latent infections, based on both the microbial and host-reads. Localization data from imaging or biopsies could further help pinpoint the anatomical site of infection. This integrated approach would offer a more comprehensive diagnostic picture, ultimately enhancing our ability to track disease progression and optimize patient care.

Part II

While the primary focus of this thesis has been on developing methodologies, we have worked closely with the clinic, driven by the urgent need for improved diagnostics in clinical practice. The work presented in this thesis is still in the proof-of-concept stage. Therefore, real-world validation will be essential in determining whether these methods can make a meaningful impact in clinical settings. It is important to acknowledge that the broader applicability of untargeted cfDNA readouts for microbial detection in clinical settings remains unestablished and even debated[39]. A study at the Children's Hospital of Chicago found that microbial NGS provided clinically relevant information in 56% of tested samples[40]. In contrast, a structured review of 82 tests across five U.S. centers found that, while 50 of 82 samples tested positive (with 25 containing multiple organisms), only 6 of 82 tests (7.3%) positively impacted patient management[41]. Moreover, 3 tests (3.7%) had a negative effect on patient outcomes. This later study concluded that the real-world value of microbial NGS was limited, possibly due to factors such as varying reasons for test orders and patients already undergoing conventional diagnostic tests.

Given the considerations above, we conclude that it is important to sketch the current ("gold standard") diagnostic workup for sepsis and IMD, alongside our cfDNA NGS work results and that of others. While doing so, we speculate on how and whether the cfDNA NGS approach could contribute to meaningful clinical advancements in the diagnosis and management of these patient groups.

Pathogen detection in SIRS or sepsis

For bloodstream infections such as sepsis, blood cultures remain the primary diagnostic tool in veterinary foal clinics, and culture-based diagnostics play a central role in human medicine[42–44]. Although these cultures are essential for pathogen identification (facilitated by technologies like MALDI-TOF) and antimicrobial susceptibility testing, they come with significant limitations. Blood cultures have a turnaround time of several days and a sensitivity of less than 50%[45–47]. Moreover, they are restricted by the range of pathogens that can be grown in culture and are heavily influenced by prior antimicrobial (pre)treatment. Several recent human studies have highlighted the advantages of circulating microbial cfDNA sequencing over conventional blood cultures for diagnosing bloodstream infections (BSIs) and sepsis in humans[48–52]. Compared to traditional blood cultures, microbial cfDNA sequencing demonstrated a higher positivity rate for pathogen identification (and the circulating cfDNA from the causative pathogen remained detectable for a significantly longer period[51,52]). A recent study in neonatal sepsis, executed by the commercial DISQVER® from Noscendo, suggested that cfDNA NGS may enhance sensitivity in sepsis diagnosis[51], further bolstering its promise in critical care settings. In neonatal foals, we report that cfDNA NGS for bacterial detection, as detailed in *Chapter 4*, demonstrates high sensitivity — superior to culture-based methods and comparable to RT-PCR (87% for RT-PCR[53,54] and 100% for cfDNA NGS). While these findings are promising, the approach is not without limitations. Similar to RT-PCR, cfDNA NGS is confined to detecting a predefined set of pathogens based on existing knowledge. A broader limitation of cfDNA NGS is its inability to reliably predict antimicrobial resistance (AMR)[25,55]. Additionally, cfDNA sequencing for bacterial detection via cfFBI's presents some challenges that need to be addressed. In some cases, results have shown discrepancies with culture-based diagnostics, with overlapping pathogenic bacteria detected in about half of the cases. Factors such as high background levels in control populations, contamination risks, and the need for further statistical validation have highlighted areas for improvement in clinical applicability. However, these challenges also offer valuable opportunities for refinement, underscoring the potential for cfDNA NGS tools to be better aligned with gold-standard diagnostics. With continued optimization, the insights provided by cfDNA sequencing could become an even more actionable and reliable addition to clinical practice.

IMD diagnostics

When evaluating the potential role of cfDNA NGS in the diagnostic landscape of IMD, it is clear that current diagnostic patient classification relies on a diverse criteria and complex microbiological toolkit. This includes histopathology, culture, and antigen detection (e.g., galactomannan, beta-D-glucan) to molecular assays like *Aspergillus* PCR, and resistance profiling RT-PCRs. According to the EORTC/MSG criteria[56], "proven" IMD requires definitive fungal confirmation via culture, histopathology, or microscopic examination. In contrast, a "probable" IMD classification combines clinical findings, imaging suggestive of fungal infection, and host factors with microbiological evidence, such as positive antigen tests or PCR results. However, these microbiological diagnostic tools like BAL fluid analysis and serum antigen testing often suffer from limited sensitivity and specificity, especially after prior antifungal treatment, with BAL galactomannan demonstrating a sensitivity of only 60%[57]. PCR-based assays show sensitivity and specificity values of 0.17 (95% CI, 0.05–0.45) and 0.87 (95% CI, 0.82–0.92)[58].

At the Princess Máxima Center, our clinical collaborator, only 5-10% of suspected IMD cases in cancer patients are consequently classified as "proven," while 50-70% are labeled "possible", meaning they present clinical findings and suspected lesions but lack microbiological confirmation. If cfDNA NGS could move even a small fraction of these "possible" cases to a "probable" category while providing species-level evidence, it would influence treatment and could greatly improve patient outcomes. Although not established in pediatric cases[50], cfDNA NGS has demonstrated effectiveness in adults of delivering molecular evidence in probable IMD cases[19], showing a potential future for reclassifying such instances. Likewise, the cfSPI workflow appears promising in pediatric cases, with a high positive predictive value (PPV) anticipated for *Aspergillus* species. This while the negative predictive value (NPV) may be less robust, as false negatives have been observed in both bronchoalveolar lavage (BAL) and plasma samples — a phenomenon also seen with standard diagnostics such as culture, galactomannan testing as well as RT-PCR. These findings indicate that cfSPI could serve as a valuable complement to existing diagnostic methods by confirming infections and identifying pathogenic species. By doing so, cfSPI has the potential to provide additional diagnostic certainty, improving decision-making in managing IMD cases.

Moving forward: collaboration, ethics, and sustainability

Establish a collaborative consortium

Establishing an international, interdisciplinary consortium or even multiple working groups would be highly beneficial for facilitating large-scale, multicenter clinical studies on cfDNA sequencing for specific clinical applications. The primary goals of such collaborative clinical study should be to validate cfDNA NGS for targeted clinical use and to develop clear, evidence-based guidelines for its deployment in specific clinical settings. These guidelines should address when to request cfDNA NGS, how to interpret its results, and how to adjust treatment strategies based on the detection of unexpected pathogens — areas where guidance is currently lacking. Furthermore, I believe this initiative would be even more impactful by prioritizing the development of open-source workflows to enhance interpretability, promote collaborative innovation tailored to specific diagnostic applications, and standardize results across institutions. This approach would ensure the reliable and consistent use of cfDNA NGS in clinical practice, minimizing reliance on proprietary, black-box testing by commercial entities.

Ethical and environmental considerations

The adoption of cfDNA sequencing raises both ethical and environmental challenges. Ethically, cfDNA analysis risks revealing sensitive genetic information, such as hereditary disease predispositions or residual tumor DNA (as noted in *Chapter 2*), underscoring the need for robust data protection, informed consent, and safeguards against third-party misuse. Lessons from prenatal cfDNA screening, or other shotgun NGS diagnostic settings, can hereby offer guidance. In terms of accessibility and equity, cfDNA sequencing presents significant barriers due to its high cost and reliance on advanced infrastructure. These challenges are likely to exacerbate health disparities, especially in resource-limited settings where access to such technology is limited. Efforts to drive down costs through technological innovation and expand global infrastructure are essential for ensuring equitable access.

Environmentally, cfDNA sequencing is resource-intensive, contributing to energy use and e-waste due to the computational demands and extensive data storage requirements. For instance, executing the cfSPI pipeline on the entire sample set generates a carbon footprint comparable to a flight from Paris to London, as calculated by the Green Algorithms 4[59]). The cfFBI pipeline is not far off in terms of environmental impact. Reflecting on the development phases of both pipelines, which included multiple reruns, it is evident that considerable environmental costs have already been incurred. Exploring more energy-efficient computational approaches, and improving data storage practices (an identified challenge) are crucial to minimizing environmental impacts. It might also be interesting to use tools such as the Carbon Aware Task Scheduler, which tries to minimize predicted carbon intensity of running the process by scheduling cluster jobs (works currently only in the UK)[60] in the research setting. Ultimately, achieving a balance between the benefits of cfDNA sequencing and the imperative to minimize its ethical and environmental costs will be crucial for its sustainable and responsible implementation.

Concluding remarks

The evolution of microbiological molecular diagnostic tools for infectious diseases continues to advance, with cfDNA NGS serving as a prime example of such an innovation. cfDNA NGS introduces novel diagnostic methods that are still under development. While challenges persist, particularly in pan-microbial pathogen identification, advancements in cfDNA-based strategies show promise.

Our application also encountered serious roadblocks, such as deviations from gold standards and the detection of multiple pathogens (as observed in *Chapter 4*), highlighting areas where applications may not yet be fruitful. For promising uses like fungal detection (*Chapter 2*), larger-scale validation and continuous optimization (as targeted in *Chapter 3*) will be crucial to ensuring these tools deliver meaningful clinical outcomes. Interdisciplinary collaboration will remain essential for advancing the technology and facilitating its adoption and integration of cfDNA NGS into clinical practice.

References

1. Pana, Z. D., Roilides, E., Warris, A., Groll, A. H. & Zaoutis, T. Epidemiology of invasive fungal disease in children. *J. Pediatric Infect. Dis. Soc*. 6, S3–S11 (2017).
2. Hill, J. A. et al. Liquid biopsy for invasive mold infections in hematopoietic cell transplant recipients with pneumonia through next-generation sequencing of microbial cell-free DNA in plasma. Clin. Infect. Dis. 73, e3876–e3883 (2021).
3. Burnham, P. *et al.* Single-stranded DNA library preparation uncovers the origin and diversity of ultrashort cell-free DNA in plasma. *Sci. Rep.* **6**, 27859 (2016).
4. A, M., Valle I, F. do, G, C. & D, R. Statistical modelling of CG interdistance across multiple organisms. *BMC Bioinformatics* **19**, 355 (2018).
5. Gihawi, A. *et al.* Major data analysis errors invalidate cancer microbiome findings. *MBio* **14**, e0160723 (2023).
6. Sepich-Poore, G. D. *et al.* Robustness of cancer microbiome signals over a broad range of methodological variation. *Oncogene* **43**, 1127–1148 (2024).
7. Liao, W.-W. *et al.* A draft human pangenome reference. *Nature* **617**, 312–324 (2023).
8. Wright, R. J., Comeau, A. M. & Langille, M. G. I. From defaults to databases: parameter and database choice dramatically impact the performance of metagenomic taxonomic classification tools. *Microb. Genom.* **9**, (2023).
9. Laurence, M., Hatzis, C. & Brash, D. E. Common contaminants in next-generation sequencing that hinder discovery of low-abundance microbes. *PLoS One* **9**, e97876 (2014).
10. Salter, S. J. *et al.* Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* **12**, 87 (2014).
11. Eisenhofer, R. *et al.* Contamination in low microbial biomass microbiome studies: Issues and recommendations. *Trends Microbiol.* **27**, 105–117 (2019).
12. Davis, N. M., Proctor, D. M., Holmes, S. P., Relman, D. A. & Callahan, B. J. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome* **6**, 226 (2018).
13. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).
14. Hong, C. *et al.* PathoScope 2.0: a complete computational framework for strain identification in environmental or clinical sequencing samples. *Microbiome* **2**, 33 (2014).
15. Stawski, R. *et al.* Repeated bouts of exhaustive exercise increase circulating cell free nuclear and mitochondrial DNA without development of tolerance in healthy men. *PLoS One* **12**, e0178216 (2017).
16. Breitbach, S., Tug, S. & Simon, P. Circulating cell-free DNA: an up-coming molecular marker in exercise physiology: An up-coming molecular marker in exercise physiology. *Sports Med.* **42**, 565–586 (2012).
17. Haller, N., Tug, S., Breitbach, S., Jörgensen, A. & Simon, P. Increases in circulating cell-free DNA during aerobic running depend on intensity and duration. *Int. J. Sports Physiol. Perform.* **12**, 455–462 (2017).
18. Lim, J. K., Kuss, B. & Talaulikar, D. Role of cell-free DNA in haematological malignancies. *Pathology* **53**, 416–426 (2021).
19. Huygens, S. *et al.* Diagnostic value of microbial cell-free DNA sequencing for suspected invasive fungal infections: A retrospective multicenter cohort study. *Open Forum Infect. Dis.* **11**, ofae252 (2024).
20. Whittle, E., Leonard, M. O., Harrison, R., Gant, T. W. & Tonge, D. P. Multi-method characterization of the human circulating microbiome. *Front. Microbiol.* **9**, 3266 (2018).
21. Xiao, Q. *et al.* Alterations of circulating bacterial DNA in colorectal cancer and adenoma: A proof-of-concept study. *Cancer Lett.* **499**, 201–208 (2021).
22. Païssé, S. *et al.* Comprehensive description of blood microbiome from healthy donors assessed by 16S targeted metagenomic sequencing. *Transfusion* **56**, 1138–1147 (2016).
23. Kowarsky, M. *et al.* Numerous uncharacterized and highly divergent microbes which colonize humans are revealed by circulating cell-free DNA. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 9623–9628 (2017).
24. Zhao, T. *et al.* Cell free bacterial DNAs in human plasma provide fingerprints for immune-related diseases. *Med. Microecol.* **5**, 100022 (2020).
25. Blauwkamp, T. A. *et al.* Analytical and clinical validation of a microbial cell-free DNA sequencing test for infectious disease. *Nat. Microbiol.* **4**, 663–674 (2019).
26. Grumaz, S. *et al.* Next-generation sequencing diagnostics of bacteremia in septic patients. *Genome Med.* **8**, 73 (2016).
27. Hu, Y. *et al.* Cell-free DNA: a promising biomarker in infectious diseases. *Trends Microbiol.* (2024) doi:10.1016/j.tim.2024.06.005.
28. O’Grady, J. A powerful, non-invasive test to rule out infection. *Nat. Microbiol.* **4**, 554–555 (2019).
29. Islam Sajib, M. S. *et al.* Advances in host depletion and pathogen enrichment methods for rapid sequencing-based diagnosis of bloodstream infection. *J. Mol. Diagn.* **26**, 741–753 (2024).
30. Feehery, G. R. *et al.* A method for selectively enriching microbial DNA from contaminating vertebrate host DNA. *PLoS One* **8**, e76096 (2013).
31. Gu, W. *et al.* Depletion of Abundant Sequences by Hybridization (DASH): using Cas9 to remove unwanted high-abundance species in sequencing libraries and molecular counting applications. *Genome Biol.* **17**, 41 (2016).

32. Francisco, D. M. A. *et al.* The effect of a plasma next-generation sequencing test on antimicrobial management in immunocompetent and immunocompromised patients-A single-center retrospective study. *Antimicrob. Steward. Healthc. Epidemiol.* **3**, e31 (2023).
33. Nielsen, M. E. *et al.* A new method using rapid Nanopore metagenomic cell-free DNA sequencing to diagnose bloodstream infections: a prospective observational study. *medRxiv* 2024.05.09.24307053 (2024) doi:10.1101/2024.05.09.24307053.
34. Linthorst, J., Welkers, M. R. A. & Sistermans, E. A. Distinct fragmentation patterns of circulating viral cell-free DNA in 83,552 non-invasive prenatal testing samples. *Extracellular Vesicles and Circulating Nucleic Acids* **2**, 228–237 (2021).
35. Burnham, P. *et al.* Urinary cell-free DNA is a versatile analyte for monitoring infections of the urinary tract. *Nat. Commun.* **9**, 2412 (2018).
36. Wang, G. *et al.* Fragment ends of circulating microbial DNA as signatures for pathogen detection in sepsis. *Clin. Chem.* **69**, 189–201 (2023).
37. Burnham, P., Khush, K. & De Vlaminck, I. Myriad applications of circulating cell-free DNA in precision organ transplant monitoring. *Ann. Am. Thorac. Soc.* **14**, S237–S241 (2017).
38. Cheng, A. P. *et al.* Cell-free DNA profiling informs all major complications of hematopoietic cell transplantation. *Proc. Natl. Acad. Sci. U. S. A.* **119**, e2113476118 (2022).
39. Babady, N. E. Clinical metagenomics for bloodstream infections: Is the juice worth the squeeze? *Clin. Infect. Dis.* **72**, 246–248 (2021).
40. Rossoff, J. *et al.* Noninvasive diagnosis of infection using plasma next-generation sequencing: A single-center experience. *Open Forum Infect. Dis.* **6**, (2019).
41. Hogan, C. A. *et al.* Clinical impact of metagenomic next-generation sequencing of plasma cell-free DNA for the diagnosis of infectious diseases: A multicenter retrospective cohort study. *Clin. Infect. Dis.* **72**, 239–245 (2021).
42. Churpek, M. M., Snyder, A., Sokol, S., Pettit, N. N. & Edelson, D. P. Investigating the impact of different suspicion of infection criteria on the accuracy of quick sepsis-related Organ Failure Assessment, systemic inflammatory response syndrome, and early warning scores. *Crit. Care Med.* **45**, 1805–1812 (2017).
43. Yealy, D. M. *et al.* Early care of adults with suspected sepsis in the emergency department and out-of-hospital environment: A consensus-based task force report. *Ann. Emerg. Med.* **78**, 1–19 (2021).
44. Book, M., Lehmann, L. E., Zhang, X. & Stüber, F. Monitoring infection: from blood culture to polymerase chain reaction (PCR). *Best Pract. Res. Clin. Anaesthesiol.* **27**, 279–288 (2013).
45. Giancola, S. & Hart, K. A. Equine blood cultures: Can we do better? *Equine Vet. J.* **55**, 584–592 (2023).
46. Russell, C. M., Axon, J. E., Blishen, A. & Begg, A. P. Blood culture isolates and antimicrobial sensitivities from 427 critically ill neonatal foals. *Aust. Vet. J.* **86**, 266–271 (2008).
47. Hytychová, T. ʼana & Bezděková, B. Retrospective evaluation of blood culture isolates and sepsis survival rate in foals in the Czech Republic: 50 cases (2011-2013): Blood culture and sepsis survival rate. *J. Vet. Emerg. Crit. Care (San Antonio)* **25**, 660–666 (2015).
48. Jing, Q., Leung, C. H. C. & Wu, A. R. Cell-free DNA as biomarker for sepsis by integration of microbial and host information. *Clin. Chem.* **68**, 1184–1195 (2022).
49. Goggin, K. P. *et al.* Evaluation of plasma microbial cell-free DNA sequencing to predict bloodstream infection in pediatric patients with relapsed or refractory cancer. *JAMA Oncol.* **6**, 552–556 (2020).
50. Armstrong, A. E. *et al.* Cell-free DNA next-generation sequencing successfully detects infectious pathogens in pediatric oncology and hematopoietic stem cell transplant patients at risk for invasive fungal disease. *Pediatr. Blood Cancer* **66**, e27734 (2019).
51. Grumaz, S. *et al.* Enhanced performance of next-generation sequencing diagnostics compared with standard of care microbiological diagnostics in patients suffering from septic shock. *Crit. Care Med.* **47**, e394–e402 (2019).
52. Eichenberger, E. M. *et al.* Microbial cell-free DNA identifies etiology of bloodstream infections, persists longer than conventional blood cultures, and its duration of detection is associated with metastatic infection in patients with Staphylococcus aureus and gram-negative bacteremia. *Clin. Infect. Dis.* **74**, 2020–2027 (2022).
53. Elmas, C. R. *et al.* Evaluation of a broad range real-time polymerase chain reaction (RT-PCR) assay for the diagnosis of septic synovitis in horses. *Can. J. Vet. Res.* **77**, 211–217 (2013).
54. Oeser, C. *et al.* PCR for the detection of pathogens in neonatal early onset sepsis. *PLoS One* **15**, e0226817 (2020).
55. Chien, J.-Y., Yu, C.-J. & Hsueh, P.-R. Utility of metagenomic next-generation sequencing for etiological diagnosis of patients with sepsis in intensive care units. *Microbiol. Spectr.* **10**, e0074622 (2022).
56. Donnelly, J. P. *et al.* Revision and update of the consensus definitions of invasive fungal disease from the European Organization for research and Treatment of cancer and the Mycoses Study Group education and research consortium. *Clin. Infect. Dis.* **71**, 1367–1376 (2020).
57. de Mol, M. *et al.* Diagnosis of invasive pulmonary aspergillosis in children with bronchoalveolar lavage galactomannan: BAL Galactomannan Aspergillosis Children. *Pediatr. Pulmonol.* **48**, 789–796 (2013).

58. Reinwald, M. *et al.* Therapy with antifungals decreases the diagnostic performance of PCR for diagnosing invasive aspergillosis in bronchoalveolar lavage samples of patients with haematological malignancies. *J. Antimicrob. Chemother.* **67**, 2260–2267 (2012).
59. Lannelongue, L., Grealey, J. & Inouye, M. Green Algorithms: Quantifying the carbon footprint of computation. *Adv. Sci. (Weinh.)* **8**, 2100707 (2021).
60. *CATS: The Climate-Aware Task Scheduler*. (Github).
61. Grumaz, C. *et al.* Rapid next-generation sequencing-based diagnostics of bacteremia in septic patients. *J. Mol. Diagn.* **22**, 405–418 (2020).
62. Agudelo-Pérez, S., Fernández-Sarmiento, J., Rivera León, D. & Peláez, R. G. Metagenomics by next-generation sequencing (mNGS) in the etiological characterization of neonatal and pediatric sepsis: A systematic review. *Front. Pediatr.* **11**, 1011723 (2023).
63. Chen, L. *et al.* Metagenomic next-generation sequencing for the diagnosis of neonatal infectious diseases. *Microbiol. Spectr.* **10**, e0119522 (2022).

Nederlandse samenvatting

Author contributions per chapter

List of publications

Dankwoord

Curriculum vitae

Nederlandse samenvatting

Pathogenen en hun rol in infectieziekten

Pathogenen zijn micro-organismen die infecties en ziekten kunnen veroorzaken bij mensen, dieren en planten. Ze omvatten bacteriën, virussen, schimmels en andere micro-organismen die in staat zijn het lichaam binnen te dringen, zich daar te vermenigvuldigen en een breed scala aan ziektes te veroorzaken. Virussen, zoals het influenzavirus, het COVID-19-virus en HIV, kunnen ernstige ziektebeelden veroorzaken en hebben vaak verstrekkende gevolgen voor de gezondheid van de bevolking. Bacteriën, zoals *Escherichia coli* en *Streptococcus pneumoniae*, kunnen infecties veroorzaken die variëren van milde maagklachten tot aandoeningen zoals longontsteking. In sommige gevallen kunnen bacteriën zelfs sepsis veroorzaken, een levensbedreigende reactie op een infectie waarbij de bacteriën in de bloedbaan terechtkomen en een ontstekingsreactie in het hele lichaam uitlokken, wat kan leiden tot orgaanfalen. Schimmels kunnen ernstige infecties veroorzaken, die in sommige gevallen zelfs dodelijk kunnen zijn, vooral bij mensen met een verzwakt immuunsysteem. Deze infecties, die vaak in de longen beginnen, kunnen zich via de bloedbaan naar andere organen verspreiden, inclusief de hersenen. Dit vergroot het risico op complicaties en kan leiden tot ernstige gezondheidsproblemen, met name bij kwetsbare patiënten die al lijden aan onderliggende ziekten.

Pathogeen detectie

Gezien de ernstige gevolgen van het niet adequaat behandelen van infecties, is een snelle herkenning van het ziektebeeld en nauwkeurige pathogeen identificatie van cruciaal belang. Dit is essentieel omdat de behandeling afhankelijk is van het type ziekteverwekker: antivirale middelen zijn gericht op virussen, antibiotica op bacteriën en antischimmelmedicatie op schimmels. Het is echter belangrijk te benadrukken dat deze middelen specifiek zijn: niet alle antibiotica werken tegen elke bacterie, niet alle antischimmelmiddelen zijn effectief tegen elke schimmel, en antivirale middelen werken alleen tegen bepaalde virussen.

Alleen door het identificeren van het pathogeen kan met zekerheid worden vastgesteld welke effectieve behandeling moet worden gestart. Diagnostische methoden zoals kweektests, PCR-tests en serologische onderzoeken spelen hierbij een cruciale rol. Kweektests en PCR kunnen specifieke micro-organismen identificeren, terwijl serologische tests pathogeen-specifieke eiwitten of antistoffen in het bloed aantonen. Beeldvormende technieken, zoals röntgenfoto's, bieden inzicht in de gevolgen van de infectie en ondersteunen het monitoren van het ziekteverloop.

De snelheid en nauwkeurigheid van de diagnostiek, gericht op het identificeren van de ziekteverwekker, zijn van cruciaal belang, vooral in levensbedreigende situaties en bij immuun gecompromitteerde patiënten die moeite hebben met het opruimen van pathogenen. Een tijdige identificatie van het pathogeen maakt gerichte behandeling mogelijk, vergroot de kans op een positieve uitkomst, versnelt het herstel en vermindert het risico op resistentie en complicaties.

Genetische sequencing

Genetische sequencing is een geavanceerde technologie die wordt gebruikt om de genetische code van mensen, dieren, planten en micro-organismen te analyseren. Elke cel in een organisme bevat genetisch materiaal, en hetzelfde geldt voor verschillende soorten pathogenen, zoals bacteriën, virussen en schimmels. Door sequencing kan de genetische code van deze organismen worden gelezen en geanalyseerd, wat van cruciaal belang is voor het begrijpen van hun genetische samenstelling en het onderzoeken van biologische processen. Sequencing, ook wel DNA-sequencing genoemd, is het proces waarbij de volgorde van nucleotiden (de bouwstenen van DNA) in een molecuul wordt bepaald. Dit stelt wetenschappers in staat om de genetische informatie van een organisme nauwkeurig in kaart te brengen. Er zijn verschillende sequencing technieken, waarvan Next-Generation Sequencing (NGS) een van de meest geavanceerde methoden is.

NGS is een techniek waarmee miljoenen DNA-moleculen tegelijk kunnen worden geanalyseerd. Dit maakt het mogelijk om snel en efficiënt de genetische samenstelling van verschillende pathogenen te onderzoeken. In tegenstelling tot traditionele sequencing methoden, die sequenties één voor één de DNA sequentie uitlezen, kan NGS parallel vele moleculen tegelijkertijd uitlezen. Dit versnelt het proces aanzienlijk en stelt onderzoekers in staat om grote hoeveelheden gegevens te verwerken, wat van cruciaal belang is voor toepassingen zoals de snelle identificatie van ziekteverwekkers, of het opsporen van kanker.

Bij genomic DNA sequencing wordt de genetische code van een organisme — zoals de mens, een dier, virus of bacterie — volledig in kaart gebracht, wat gedetailleerde en diepgaande informatie oplevert. In tegenstelling tot genomic DNA sequencing richt cfDNA (celvrij DNA) sequencing zich op het isoleren en analyseren van DNA uit de bloedbaan of andere lichaamsvloeistoffen. Dit cfDNA bestaat uit korte fragmenten die vaak vrijkomen wanneer cellen afsterven. Het cfDNA kan afkomstig zijn van alle cellen in het lichaam, inclusief tumorcellen (indien aanwezig), micro-organismen en geïnfecteerde cellen, wat het een waardevolle bron maakt voor het monitoren van ziekteprocessen.

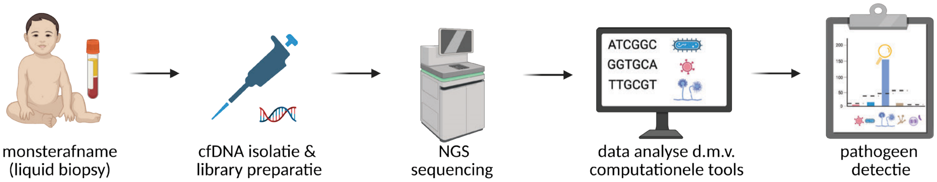
De toepassing en optimalisatie van cfDNA NGS voor pathogeen detectie

Het proces van cfDNA NGS voor pathogeen detectie is echter complex. Er zijn verschillende stappen tussen het afnemen van het monster bij de patiënt en het bepalen van de verhoogde aanwezigheid van mogelijke pathogenen (Figuur 1). Het begint met het verzamelen van een lichaamsvloeistof monster, zoals bloed, urine of longvloeistof. Daarna wordt het cfDNA uit het monster geïsoleerd. De verkregen cfDNA-moleculen worden vervolgens voorbereid voor NGS-analyse via een speciaal proces, 'library preparatie', waarbij de DNA-fragmenten geschikt worden gemaakt voor de sequencer, zodat ze efficiënt en nauwkeurig gelezen kunnen worden. Vervolgens worden miljoenen cfDNA-moleculen gelijktijdig uitgelezen met NGS.

De resulterende data moeten met geavanceerde computationele methoden worden verwerkt om de enorme hoeveelheden informatie effectief te analyseren. De belangrijkste stap daarbij is het achterhalen van de origine van het cfDNA-molecuul. Aangezien het merendeel van de cfDNA moleculen afkomstig is van de patiënt zelf en slechts een klein percentage (<1%) van microben, en een nog kleiner deel van de pathogenen die de infectie veroorzaken, is het proces vergelijkbaar met het zoeken naar een speld in een hooiberg. De uitdaging wordt verder vergroot doordat de cfDNA-sequenties kort zijn en vaak niet specifiek voor een bepaald micro-organisme.

Om dit proces te vergemakkelijken, wordt in de praktijk vaak eerst het lichaams-eigen DNA geïdentificeerd, waarna gedifferentieerd kan worden tussen cfDNA van pathogene microben en microben die wel aanwezig zijn, maar niet ziekteverwekkend. Om de oorsprong van de vaak korte cfDNA-fragmenten nauwkeurig vast te stellen, worden de cfDNA sequenties vergeleken met de genetische codes van duizenden pathogenen, waarvan de volledige genetische informatie bekend is. Deze stap vereist het gebruik van geavanceerde computationele methoden. Door dit proces voor elk cfDNA molecuul afzonderlijk uit te voeren, kan de oorsprong van het overgrote merendeel van deze miljoenen fragmenten nauwkeurig worden vastgesteld. Het resultaat is een gedetailleerd overzicht van de hoeveelheid patiënt specifiek, microbiële en pathogeen-specifieke cfDNA.

Gezien de complexiteit van deze stappen, maar ook van andere fasen in het proces, is het niet verwonderlijk dat de NGS-werkstromen gepaard gaan met specifieke uitdagingen die zorgvuldig moeten worden aangepakt. Hoewel er al veelbelovende onderzoeksresultaten zijn die de effectiviteit van cfDNA sequencing voor pathogeen detectie aantonen, richt dit proefschrift zich op verdere optimalisatie van deze methode.


ATCGGC
GGTGCA
TTGCGT
200
150
100
50
0
monsterafname (liquid biopsy)
cfDNA isolatie & library preparatie
NGS sequencing
data analyse d.m.v. computationele tools
pathogeen detectie

Figuur 1: Werkstroom pathogeen detectie door middel van cfDNA NGS

Monsterafname (liquid biopsy): Een (bloed)monster wordt afgenomen van een patiënt met symptomen die wijzen op een infectie, met als doel cfDNA (celvrij DNA) te isoleren en te analyseren om de veroorzakende pathogeen te identificeren. ***cfDNA isolatie***: Het cfDNA wordt geïsoleerd uit het bloedmonster. Een kwaliteitscontrole wordt uitgevoerd door zowel de hoeveelheid als de integriteit van het cfDNA te meten, om te garanderen dat het geschikt is voor verdere analyse. ***Library preparatie***: Het cfDNA wordt omgezet in een sequencing-library (door middel van adapter ligatie en PCR-amplificatie). Dit maakt het mogelijke om het cfDNA efficiënt te sequencen. ***NGS sequencing***: De cfDNA sequencing-libraries worden geanalyseerd met sequencing via Next-Generation Sequencing (NGS), waarbij miljoenen cfDNA-fragmenten gelijktijdig en nauwkeurig worden gelezen. ***Data analyse d.m.v. computationele tools:*** De NGS-sequencing data worden verwerkt met geavanceerde computationele methoden die essentieel zijn voor het effectief analyseren van de enorme hoeveelheden sequencing data. Een belangrijke stap hierin is het bepalen van de oorsprong van elk cfDNA-molecuul, waarvoor geavanceerde bioinformatica-methoden worden ingezet. Zodra de oorsprong van elk cfDNA-molecuul is geïdentificeerd, kan een gedetailleerd overzicht worden verkregen van patiëntspecifiek cfDNA, evenals microbiële en pathogeen-specifieke cfDNA. ***Pathogeen detectie***: Een verhoogde aanwezigheid van cfDNA van een pathogeen maakt nauwkeurige identificatie mogelijk als waarschijnlijke veroorzaker van de klachten. Deze informatie biedt waardevolle inzichten voor de diagnose en gerichte behandeling van infecties. Afbeelding gemaakt met BioRender.com (verkregen op 28-12-2024).

Hoofdstukken van het onderzoek

Dit proefschrift behandelt technische uitdagingen in pathogeen identificatie via cfDNA NGS, zowel op het gebied van laboratoriumtechnieken (library preparatie) als computationele data-analyse, met als doel de detectie van pathogenen te verbeteren. In samenwerking met klinische partners werd de nadruk gelegd op het optimaliseren van cfDNA NGS-werkstromen voor klinische toepassingen. Specifiek werd de klinische potentie van cfDNA NGS onderzocht in twee onontgonnen domeinen met dringende diagnostische behoeften: schimmelinfecties bij immuungecompromitteerde kinderen en bacteriële sepsis bij pasgeboren veulens.

Hoofdstuk 2: Dit hoofdstuk richt zich op de ontwikkeling van cfDNA-gebaseerde diagnostiek voor invasieve schimmelziekten bij immuungecompromitteerde kinderen, omdat het aantal betrouwbare en snelle diagnostische opties momenteel beperkt is. Om schimmel-cfDNA te detecteren, hebben we verschillende library voorbereiding strategieën getest en geavanceerde computationele tools geoptimaliseerd voor de betrouwbare identificatie van schimmel-cfDNA. Dit werd bereikt door enerzijds het menselijk cfDNA nauwkeurig te onderscheiden en anderzijds de referentie genetische informatie van relevante schimmels te gebruiken om de oorsprong van het niet-menselijke cfDNA vast te stellen.

In onze proof-of-principle studie, uitgevoerd met 7 patiënten die vermoedelijk een invasieve *Aspergillus* schimmelinfectie in de long hadden, en 18 immuungecompromitteerde controlepersonen, ontdekten we verhoogde hoeveelheden cfDNA van het pathogeen *Aspergillus fumigatus* in 6 van de 7 patiënten met schimmelinfectie. Specifiek detecteerden we verhoogde *A. fumigatus* hoeveelheden in 5 van de 7 longvloeistof monsters en 3 van de 5 bloedmonsters. Deze resultaten benadrukken de veelbelovende toepassing van cfDNA NGS-testen in longvloeistof voor de detectie van *Aspergillus*. Plasma-monsters bieden een waardevolle, minder invasieve optie, hoewel minder gevoelig dan longvloeistof. Ondanks de veelbelovende bevindingen, is verdere validatie noodzakelijk, en er is momenteel subsidie aangevraagd om de volgende cruciale stap in dit onderzoek te ondersteunen.

Hoofdstuk 3: In dit hoofdstuk heranalyseerden we de cfDNA NGS data van immuungecompromitteerde patiënten uit hoofdstuk 2, met als doel een methode te ontwikkelen voor het verrijken van cfDNA afkomstig van *Aspergillus fumigatus*. Uit onze bevindingen in hoofdstuk 2 bleek dat *Aspergillus*-cfDNA slechts in zeer kleine hoeveelheden aanwezig is in bloed- en longvloeistof monsters, wat het detecteren van dit pathogeen uitdagend maakt. We veronderstelden dat het cfDNA van *Aspergillus*, qua fragment lengte en eind-motief, verschilt van menselijk cfDNA, wat mogelijk een kans biedt om voor *Aspergillus*-cfDNA te verrijken.

Met behulp van computeranalyses onderzochten we of selectie op basis van deze fysieke kenmerken — zoals de lengte of eind-motief van de DNA-moleculen — de hoeveelheid gedetecteerd *Aspergillus*-cfDNA zou kunnen verhogen. Onze bevindingen ondersteunen het idee dat deze selectie strategieën de detectie van *Aspergillus* verbeteren door de relatieve hoeveelheid pathogen-specifiek cfDNA te verhogen in vergelijking met het menselijke cfDNA, dat doorgaans in veel grotere hoeveelheden aanwezig is.

De volgende stap is het testen van deze methoden in het laboratorium om te zien of de geselecteerde fysieke kenmerken daadwerkelijk kunnen bijdragen aan een efficiëntere verrijking van *Aspergillus*-cfDNA. We verwachten dat het benutten van deze fysieke eigenschappen in de toekomst niet alleen de detectie nauwkeuriger zal maken, maar ook het uitlezen van het cfDNA efficiënter, sneller en kosteneffectiever kan maken.

Hoofdstuk 4: In dit hoofdstuk onderzochten we de toepassing van cfDNA NGS voor de diagnose van sepsis bij pasgeboren veulens, voortbouwend op eerdere studies die het potentieel van cfDNA NGS bij mensen, inclusief neonaten, hebben aangetoond. We onderzochten of ditzelfde potentieel ook voor pasgeboren veulens, die vatbaar zijn voor bacteriële infecties en sepsis, geldt.

Om bacterieel cfDNA effectief te detecteren, optimaliseerden we de library preparatie om zo veel mogelijk bacterieel cfDNA uit de monsters te verkrijgen. Door bacterieel cfDNA te verrijken, wilden we de sequencing resultaten verbeteren en de kans vergroten om bacteriële pathogenen te identificeren.

We analyseerden 11 veulens met symptomen van neonatale systemische inflammatoire respons syndroom, een kenmerk van sepsis. Bij deze veulens identificeerden we verhoogde hoeveelheden cfDNA van specifieke bacteriële soorten, wat suggereert dat deze bacteriën betrokken kunnen zijn bij de ontstekingsreactie en sepsis. De resultaten benadrukken de potentie van cfDNA NGS voor het karakteriseren van de bacteriële samenstelling bij pasgeboren veulens met sepsis of het risico daarop.

Toekomstige perspectieven en uitdagingen

De bevindingen in dit proefschrift benadrukken het potentieel van cfDNA voor de diagnostiek van infecties. Verdere ontwikkeling en validatie van deze technologie kunnen resulteren in snellere en nauwkeurigere diagnostische methoden, wat aanzienlijke voordelen zou kunnen bieden in de behandeling van infectieziekten.

Toch blijven er belangrijke uitdagingen bestaan, zoals het verbeteren van de kostenefficiëntie en het beter vaststellen van causaliteit bij infectieziekten. Momenteel wordt aangenomen dat een verhoogde aanwezigheid van cfDNA van een pathogeen erop wijst dat dit pathogeen de veroorzaker is van de infectie. Echter, de vraag is of we deze aanname zomaar kunnen maken. Het onderscheid tussen een pathogeen dat daadwerkelijk de ziekte veroorzaakt en een pathogeen dat toevallig aanwezig is, vereist aanvullende validatie en een grondiger begrip van de complexe interacties tussen gastheer en micro-organismen. Mogelijk kan cfDNA afkomstig uit (geïnfecteerde) gastheercellen in de toekomst een waardevolle rol spelen bij het ontrafelen van deze causaliteits vraagstukken.

Author contributions per chapter

Chapter 1: General introduction

EW wrote the text and prepared the figures, with feedback provided by MJ, LC, and JdR.

Chapter 2: NGS-based *Aspergillus* detection in plasma and lung lavage of children with invasive pulmonary aspergillosis

EW, LR, LC and MJ conceived experiments and wrote the article. LR searched for and collected samples, managed patients, provided clinical information and sample data and interpreted with EW sequencing data to clinical data. EW, LC and NB performed experiments. EW and LC contributed to constructing the data analysis pipeline and conducting the data analysis. EW curated tables and created figures. MJ, CV, FH, TvdB, CL, TW, LB and JdR designed experiments, contributed to the writing of the article and/or provided (clinical) information.

Chapter 3: Exploring cell-free DNA fragmentomics to improve *Aspergillus* detection in invasive mold infections

EW authored the chapter and designed the analysis in collaboration with MJ. EW carried out the analysis, with feedback on the chapter provided by ARH, MJ, and JdR.

Chapter 4: Bacterial cell-free DNA profiling reveals co-elevation of multiple bacteria in newborn foals with suspected sepsis

LC, EW, MJ, ES, MT, CV and JdR conceptually designed the study. ES and MT collected samples, gathered clinical information, and gave clinical input. EW, ES, NB, CV and MT prepared samples. EW, NB and CV optimized protocols and generated the sequencing libraries. LC, EW and MJ contributed to data analysis. CV, AZ, EB and JW provided input on the experiments and analyses. LC, EW and MJ wrote the manuscript. JdR coordinated the study.

Chapter 5: Discussion

EW authored the discussion with input from MJ and feedback from MJ, and JdR.

List of publications

Wesdorp AE, Rotte L*, Chen LT*, Jager M*, Besselink N, Vermeulen C, Hagen F, van der Bruggen T, Lindemans C, Wolfs T, Bont L, de Ridder J. NGS-based Aspergillus detection in plasma and lung lavage of children with invasive pulmonary aspergillosis. NPJ Genom Med. 2025 Mar 17;10(1):24. doi: 10.1038/s41525-025-00482-8.

Chen LT*, **Wesdorp AE***, Jager M, Siegers EW, Theelen MJP, Besselink N, Vermeulen C, Vermeulen C, Zomer AL, Broens EM, Wagenaar JA, de Ridder J. Bacterial cell-free DNA profiling reveals co-elevation of multiple bacteria in newborn foals with suspected sepsis. *Provisionally accepted in iScience*.

Lopes R, Sprouffske K, Sheng C, Uijttewaal ECH, **Wesdorp AE**, Dahinden J, Wengert S, Diaz-Miyar J, Yildiz U, Bleu M, Apfel V, Mermet-Meillon F, Krese R, Eder M, Olsen AV, Hoppe P, Knehr J, Carbone W, Cuttat R, Waldt A, Altorfer M, Naumann U, Weischenfeldt J, deWeck A, Kauffmann A, Roma G, Schübeler D, Galli GG. Systematic dissection of transcriptional regulatory networks by genome-scale and single-cell CRISPR screens. *Sci Adv*. 2021 Jul 2;7(27):eabf5733. doi: 10.1126/sciadv.abf5733.

*Equal contribution

Dankwoord

Na meer dan vijf jaar is het zover: het proefschrift is een feit. Een traject waarin talloze mensen, direct of indirect, een waardevolle bijdrage hebben geleverd.

Jeroen, bedankt voor je vertrouwen! Mij aannemen zonder substantiële ervaring in de bio-informatica was ongetwijfeld een gewaagde stap. Toch is de thesis er - zij het via wat omwegen - gekomen. De vrijheid die je me gaf om zowel een nieuw onderzoeksveld (*microbial cfDNA*) als een nieuwe stad (Barcelona) te verkennen - *I've never taken that for granted!* Ik ben me er ook terdege van bewust dat ik het je niet altijd makkelijk heb gemaakt, mede omdat ik soms nogal mijn eigen plannen volg ;) Maar steeds vaker denk ik terug aan je goede adviezen, en daar ben ik je dankbaar voor. We spreken elkaar snel weer - want hoe leuk is het dat jij betrokken blijft als scientific advisor bij de schimmel-cfDNA follow-up! Voor jou persoonlijk hoop ik dat de reis die voor je ligt als full professor gevuld zal zijn met plezier, veel energie (Huel) en inspirerende ontdekkingen. Met daarnaast hopelijk ook de nodige vaaruren in je vrije tijd!

Myrthe, zonder jouw hulp, betrokkenheid en luisterend oor was deze thesis nooit op deze manier tot stand gekomen. Extra bijzonder - en stiekem een kleine eer - is het dan ook dat je nu als co-promotor officieel deel uitmaakt van dit geheel. Ik ben je enorm dankbaar voor je inzet, je steun aan zowel Li-Ting als mij, en je rustige begeleiding. Je liet ons vaak zelf proberen en uitzoeken, maar stuurde bij waar nodig - een vaardigheid die je vast hebt geleerd met je drie jongens ;) - en herinnerde ons regelmatig dat we ergens *echt geen tijd meer voor hadden*. Tot slot verdient jouw geduld met mijn schrijfproces een op zichzelf staand dankwoord; dit talent is nog volop in ontwikkeling (helaas!).

Peter de Keizer en **Madelon Maurice**, dank voor jullie scherpe blik tijdens de jaarlijkse commissievergaderingen. Jullie kritische vragen en betrokkenheid hebben wezenlijk bijgedragen aan de totstandkoming van deze thesis. **Edwin Cuppen**, dank dat je in het eerste deel van mijn PhD de rol van promotor op je wilde nemen.

Lude Franke, Michael Seidl, Wim Tissing en **José Borghans**, hartelijk dank voor het zorgvuldig doornemen van mijn werk en voor het mogelijk maken van mijn verdediging.

Li-Ting, thank you so much for sharing the pain and joy of the PhD journey. We may have had different working styles, but we really learned a lot together - something reflected in our shared chapters. You introduced me to Git and Snakemake, and I'm really grateful — using them so often now! Beyond science, I'm so glad we shared Almadrava, the Stelvio, barbecues in the park behind your house, and countless Amelisweerd pancakes. Here's to many more in the future!

Nicolle, dankjewel voor je grote bijdrage aan deze thesis. Jij en ik weten allebei maar al te goed, hoeveel energie en tijd er heeft gezeten in de Illumina multiplexing runs, welke de basis vormen van deze hele thesis! Ik ben jaloers op je pipetting skills en waardeer erg je inzit, energie en betrokkenheid bij mijn PhD. De eerste periode in Barcelona werd een stuk leuker dankzij jouw maandagochtend Slack-berichtjes en regelmatige belletjes. Daardoor voelde ik me toch écht onderdeel van het Ridder-lab. Nu we allebei naar de oranje overkant zijn overgestoken, zien we elkaar minder, maar gelukkig is de koffie in het Máxima Centrum goed, dus laten we er snel weer eentje doen!

Ook veel dank aan alle (voormalige) **Ridder-teamleden**, in het bijzonder: **Carlo**, het was een eer om van jou te leren hoe je (cfDNA) sequencing-projecten opzet, van mijn eerste plasma- en cfDNA-isolaties tot de single-strand DNA-preparaties. Je kritische blik hield ons steeds scherp. Mooi om te zien dat diezelfde scherpte je nu verder brengt in je carrière, als assistant professor! Blijf vooral de jonge generatie begeleiden en met hen samenwerken – juist dát maakt wetenschap zo leuk. **Joske**, jouw energie binnen de groep was (en is!) inspirerend. Je introduceerde me in treatment benefit prediction - een project dat we helaas niet hebben voortgezet, maar je bent een blijvende bron van inspiratie. Zeker ook door je wandelavonturen! **Marc**, gracias por todo lo compartido: los cafés en Barcelona Sants y Utrecht Centraal, la experiencia en Almadrava, el tour por Tarragona (con horchata incluida, still one of Axel's favorites) ¡i molt més! Espero que sigamos brindando con vermut de vez en cuando. **Roy**, jouw oog voor detail is indrukwekkend. Blijf dat vasthouden! **Joanna von Berg**, dank voor de leuke discussies over poolcoördinaten. **Jasmine**, the online focus hour you organized at the beginning of my PhD truly helped me navigate the challenges of the early corona period. **Lucia**, keep shining - your energy is infectious! You're a strong woman, fantastic addition to the group, qué alegría tenerte en el grupo de Ridder! **Adrien**, wat een sprankelende bron van positiviteit ben jij! Ik snap nog steeds niet hoe je zó vrolijk weet te blijven in dit koude, regenachtige land. Je warmte en enthousiasme werken aanstekelijk – ik hoop dat je die blijft doorgeven aan nieuwe generaties studenten, zoals je dat ooit bij mij deed. Jij maakt de bioinformatica persoonlijk! **Marta**, it was such a pleasure having you as part of our group. Officially you joined as a student, but from the start you felt more like a full-fledged colleague. Your dedication and work ethic are truly admirable. I have no doubt you'll complete your own PhD with great success! ¡Éxito en todo lo que hagas!

Ook veel dank aan de vele fijne collega's binnen het **CMM** en de afdeling **Genetica** die werkzaam zijn in het **Stratenum**. De lunchtafel, spontane ontmoetingen in de gang en de (soms welkome, soms afleidende) gesprekken tijdens mijn tijd in het wetlab hebben écht bijgedragen aan het werkplezier. Daarnaast mijn bijzondere dank voor het sequencen, zorgvuldig uitgevoerd door de collega's van **USEQ**!

Deze thesis had niet tot stand kunnen komen zonder de fijne multidisciplinaire samenwerking. Vanuit het **PMC/WKZ**: **Laura**, wat bijzonder dat we elkaar via online meetings leerden kennen. Twee verschillende werelden - jij vanuit de kliniek, ik vanuit de bio-informatica - maar juist daardoor was het leren van elkaar zo waardevol. Des te meer kijk ik uit naar onze verdere samenwerking en goede sparsessies in ons vervolgonderzoek. Ik bewonder enorm hoe je je PhD, klinische carrière én gezinsleven weet te combineren. Echt bewonderenswaardig! **Janine**, wat ben jij een fijne en gezellige meid. Ik denk met veel plezier terug aan de tapas in Barcelona. Nog even doorzetten, en straks mag ook jij met trots de doctorstitel aan je naam toevoegen. Er ligt een wereld aan mogelijkheden voor je open - binnen én/of buiten het ziekenhuis! **Louis**, **Tom** en **Caroline**, heel erg bedankt voor jullie vertrouwen, hulp en de mooie wetenschap die we samen hebben kunnen doen. Ik kijk ernaar uit om onze samenwerking voort te zetten, met de financiële ondersteuning vanuit **KiKa**! **Ferry Hagen** en **Tjomme van der Bruggen**, ook jullie hartelijk bedankt voor jullie waardevolle, inhoudelijke input en jullie bijdrage aan dit gezamenlijke werk. Vanuit de **Faculteit Diergeneeskunde**: **Mathijs** en **Esther**, jullie tomeloze inzet en passie voor het verbeteren van de zorg voor septische veulens zijn simpelweg inspirerend. Ik bewonder hoe jullie elk jaar weer de tropendagen op de veulen-IC met zoveel toewijding en doorzettingsvermogen doorstaan. Dankzij jullie zorgvuldige sample collectie (vaak op de meest onmogelijke momenten), scherpe inhoudelijke input, ook van **Aldert**, en aanstekelijk enthousiasme heeft ons gezamenlijk onderzoek mede mogelijk gemaakt. Hoewel cfDNA binnen deze complexe sepsis diagnostiek nog veel vragen oproept, hebben we samen veel geleerd. **Els**, hartelijk dank voor het tot stand brengen van het contact met het eerder genoemde team! Vanuit **Skyline: Martin**, bedankt voor de vele online brainstormsessies en de kans om nieuwe machine learning vaardigheden te ontwikkelen. Hoewel deze efforts niet hebben geleid tot ‘treatment benefit prediction’, ben ik ervan overtuigd dat de voortgezette samenwerking met de Ridder groep zal leiden tot mooie nieuwe ontdekkingen.

Lieve en **Tytgat-collega’s**, dank voor de warme welkom, de inspiratie en de kansen om te blijven groeien – én om met cfDNA te blijven werken (wat wil je nog meer?). Lieve en het hele **liquid biopsy-team**: ontzettend bedankt voor jullie enthousiasme en steun!

Martina, you run such a wonderful coworking - full of love, care, and great energy! You’ve made remote working so much more relaxed and enjoyable. It’s become more than just a place to work and focus - it’s a space to connect, socialize, and truly feel at home. Thank you for creating that for all of us. And how amazing is it that by the time this is printed, you’ll have *three* of them!

Josephina, dear ex-coworking-coworker! I’m so grateful for our conversations, the laughter, and even the occasional tear (not to mention all those PhD frustrations - haha). On top of that, this thesis truly wouldn’t have looked the same without you. Your creativity, endless patience (!), and keen eye for detail made designing the cover a joy – and the result is simply beautiful. I couldn’t have asked for anyone better to bring it to life!

I can’t thank every member of the **Loft 153 co-working** individually, but each of you has contributed in your own unique and special way. **Fran,** gracias por trabajar juntos, incluso los fines de semana! **(bike) Marc**, our shared love for coffee, food, and Austral is unmatched! There’s no better place to chat, drink (way too many) cappuccinos (“with leche de vaca”), enjoy some delicious cookies relleno, and, of course, to keep up-to-date about bikes and bikeride adventures. **Marc,** you’re such a genuinely sweet guy — it’s always a pleasure sharing lunch with you on the days you’re at the office! **Diego** (Viladomat)**,** I’m so glad I finally convinced you that Barceloneta is a gem! I hope you continue to thrive, travel, enjoy the mountains, spot wildlife, and eat (I) Can Pizza at regular times! **Jeromy,** wat heb je een geweldig slechte humor — loving it! “Er zitten drie mensen op een scooter...” **Loïc, Caio,** and **Sofia:** Thank you for the coffee talks! And to **everyone else** who taught me Spanish, shared lunches, and brought food to the table - muchas gracias!

Silvina! You are such an inspiring, cool, and sporty woman — beautiful, smart, and an incredible listener and advisor. You feel like a big BCN sister to me, and I truly hope we stay in touch for many years to come. You bring thoughtful conversations, a sharp perspective, and so much joy. Let's definitely plan another hike, via ferrata, ski day, or just a relaxed day with **nosotres** soon! **Nico**, how kind can one person be? To me, you're like a yoga teacher (just with a very different job) — so balanced, warm, and calm. I often feel like I talk too much around you, but I hope you don't mind!

Keish, muchas gracias para todos! ¡tanto por ser un gran jefe de Loft Viladomat, compañero de escalada y gran amigo! Tienes una personalidad maravillosa, te preocupas mucho por los demás y, además, eres muy fuerte: ¡esas cualidades son realmente admirables! Keep up the good vibes, be proud of yourself and of your country (arepas are simply the best)! Let's keep going on many more outdoor adventures together, whether it's skiing, climbing, barranquismo, or anything else that's a bit scary but still (kind of) fun!

Barcelona - what can I say? You've taught me to be (more) humble, to take life a little less seriously, to handle the heat (a bit) better, and above all, to simply be myself. I adore your charmingly imperfect, delightfully un-Dutch sense of organization, your mix of Catalan and Latin spirit, your stunning mountains and the Mediterranean right next door, your endless sports opportunities, and so much more. Honestly, you have surprised me in so many wonderful ways - just like the amazing people I've been lucky to meet and call together **my Barcelona friends**. **Fede**! Thank you for your great friendship, openness, and positive spirit. You made our time in Barcelona so much richer, also as we met so early on in this adventure and directly connected. I truly treasure our volleyball games (now fading into distant memories of when I was still young and fit - huge thanks to the whole social volley crew for those amazing times!), bouldering sessions (I'll never forget your iconic "walk" in pie de gatos), burger nights, and countless coffee moments - not to mention your ever-evolving hairstyles! You've been the best coach and sometimes a mirror, reflecting back what I was doing and why. Keep giving and caring, including to yourself! **Christina**, chica, you happy, crazy soul! I still miss you here in BCN: the good (girly) talks, the snorkeling adventures… but hey, at least the fish and pulpos are still around. Hope to see you again soon – whether in Barcelona, Utrecht or Vienna! **Silke**, chica, wat een lekkere (gekke) diva ben jij toch, we spreken elkaar snel weer! **Chiara & Federica**, thanks for the wonderful friendship, football games, good coffee, and even introducing me to my very first moka pot! **Edu & Ingrid**, so crazy we met in Coma de Vaca. Looking forward to many more mountain adventures with you. Your thirst for adventure is inspiring – keep it up, especially now you're back from your incredible trip around the world! **Karol** and **Karla**, you are both so sweet! Truly lovely to have you as friends. **Eelco & Valeria**, such a fun coincidence that you 'followed' us to BCN! I hope you'll soon get to enjoy plenty of adventurous Spanish outdoor fun with your little one.

Bouldergang: **Gine**, it has been such a pleasure to climb with you since we met that random Sunday at Bloc, I'm still amazed by your (professional! haha) climbing skills and your infectious enthusiasm - keep that vibe going strong! **Luca**, what a burst of Italian energy you bring! **Alvar**, as one of the newest (though by now not-so-new) members of the Loft & Bouldergang, thank you for the outdoor adventures, great (and often political) conversations, and the typically

Catalan spirit you bring (you said it yourself..), which is lovely! Who knows, maybe one day we'll climb a 6b outdoors, buy a climbing rope together, or I'll even join you in a protest ;) **Vincent** & **Line**, thank you for being such a source of energy and climbing inspiration, and for leading me up to Montserrat!

Josien, wat een geschiedenis hebben wij samen! In 2011 leerde ik je kennen als een blij ei, en sindsdien zijn we vriendinnen - en hebben we samen al heel wat meegemaakt en doorleefd. Eerst waren we onafscheidelijk, inmiddels zijn er flink wat kilometers tussen ons – en nu zelfs tijdelijk héél veel! Wat een bijzonder avontuur zijn jullie aangegaan! Veel plezier daar met Anne en jullie kleine tamāhine! Wat leiden we allebei mooie, volle levens. Natuurlijk waren het niet alleen maar hoogtepunten de afgelopen jaren, helaas, maar juist die momenten laten ons groeien. En zoals we vroeger als jonge (en nog zó naïeve) studenten altijd zeiden: als we nu af en toe struggelen, slaan we die midlifecrisis gewoon over! Proost op dat - en op nog vele jaren van een steeds veranderende, maar altijd waardevolle vriendschap!

Karen, sportieveling! Wat bijzonder dat we al zó vaak samen op vakantie en avontuur zijn geweest, meestal wandelend, maar ook rennend (poeh, wat was dat destijds een plan! Ik weet nog steeds niet wat zwaarder was: het trainen voor die ene vakantie of het doen van een PhD...?) Maar wat een onvergetelijk corona-project (in de regen) is het geworden! En wat is Nederland eigenlijk verrassend mooi (ook in de regen). Ik ben je ontzettend dankbaar dat je ook later in mijn PhD-tijd mee wilde naar Ordesa National Parc om daar samen een stukje van de GR11 te lopen, zo leuk en zo veel herinneringen. En Corsica staat natuurlijk nog steeds op onze lijst. Dus als Jos en de kleine het toelaten, gaan we er gewoon weer een keer tussenuit, ergens in de komende jaren, de bergen in! Jij houdt mij dan netjes op het pad (geen ochtendomwegen!), en ik zorg dat we iets meer meenemen dan alleen mueslirepen als snacks. Op nog veel meer goeie gesprekken en mooie kilometers samen!

Caroline, wat een bijzondere vriendschap delen we. Hoewel we elkaar niet vaak spreken, ben ik je ontzettend dankbaar voor de waardevolle gesprekken, je wijze raad en het open delen van gedachten, zorgen en twijfels tijdens onze wandelingen en koffiemomenten (althans ik met koffie, jij met thee). Zoals we helaas maar al te goed weten, kan het leven soms oneerlijk zijn. Juist in die momenten hoop ik dat onze vriendschap je een beetje steun mag bieden, naast alles wat je zo dierbaar is: je lieve familie, de honden en dat prachtige paard. Ik hoop van harte dat je, tegen de tijd dat dit boekje gedrukt wordt, in een betere plek bent dan nu - ik schrijf dit namelijk een week voor je operatie. Weet dat ik enorm veel vertrouwen in je heb; dat vertrouwen was er al vanaf de allereerste dag dat ik je ontmoette in het UMC Utrecht!

Luan, I'm not exactly sure when or where we exactly met, but I guess we got to know each other as colleagues when I started the PhD - though we rarely talked about work (even though Axel sometimes tried, haha - you did recommend me StatQuest though!). Despite our very different tastes in music, movies, video games and what more, we quickly connected over our shared love for dogs, but more importantly the best sport there is: climbing. Back then, I often had to bail because of busy Dutch (family) obligations - sorry again about that - but I always enjoyed our sessions at the climbing wall! As well as your great cooking! Now that we're separated by half the planet, we're lucky to still catch up at least once a year (thanks, Hartwig!), and when we do, it feels just like old times. Although I'm sad you're so far away, I'm mostly happy that you've found your place - and your girl!

Pipetten, wat zijn jullie toch geweldige mensen, en wat hebben we samen veel mooie momenten gedeeld! **Lianne** (& Benno), Galicië, Porto en het dolfijnen spotten (en het zeeziek zijn) zal ik nooit meer vergeten! **Eline B** (& Jornel), het was fantastisch om samen door Andorra te wandelen! **Lisanne**, wat een toffe meid ben je - altijd eerlijk en open, luv it! **Eline K**, wat fijn dat we elkaar de afgelopen jaren zo goed hebben leren kennen tijdens de gezellige etentjes op de Griftstraat (dank ook aan Joran!), je bent een schat! **Naneth** en **Inge**, ook jullie zijn echte toppers! **Hester**, we zullen je altijd blijven missen. Soms vraag ik me nog vaak af wat je had gevonden van ons buitenlandavontuur, of van deze thesis. Maar hoewel we dat nooit zullen weten, draag je gelukkig altijd een plekje in ons hart als sterretje!

Zeeuwse meiden, **Dorien** en **Tirza**, wat is het geweldig dat we na al die jaren nog steeds vriendinnen zijn! Ondanks onze verschillende levensstijlen ben ik ontzettend trots op jullie - en wat hebben jullie samen een prachtig gezinsleven in het Zeeuwse opgebouwd! Basel girls, **Alisa** and **Mirjam**, thanks for our friendship! I'll never forget swimming in the Rhine and your visit to Barcelona. Hope to see you again soon in Basel! **Melanie**, science girl, it was great to share some of the PhD struggles with you. Looking forward to catching up again soon, whether in Switzerland or somewhere else in Europe! **David**, oude kilometervreter! Het was super om je hier over de vloer te hebben (overigens: het is een wonder dat jij en je fiets überhaupt in ons huisje pasten). En, de volgende keer fiets ik gewoon mee hoor! **My-Anh** & **Hieu**, you've been such an inspiration of how to live life abroad. I'm truly grateful for the time and food we shared in San Francisco and Utrecht. You have a wonderful family; ik wens jullie samen het aller-allerbeste!

Bram en **Ingrid**, dank jullie wel voor alle goede zorgen (ook voor onze meubels in de 'opslag', oepsie). Jullie belangstelling, het meeleven met werk, en natuurlijk de heerlijke bolussen waarvan jullie ons regelmatig voorzien, worden enorm gewaardeerd. Samen met **Joke** (wat ben jij toch een creatief talent!), **Eva** (altijd gezellig!) en **Mio** (wat een heerlijke spring-in-'t-veld!), vormen jullie samen een huis vol. Hopelijk vinden jullie daarnaast dan ook af en toe tijd om, bijvoorbeeld met de camper, buiten het Kapelse op ontdekkingstocht te gaan! Leuk dat ook jullie langs zijn geweest!

Lieve Marlies en Mieke, mijn allerliefste **zussen**, ontzettend bedankt voor alles wat jullie mij hebben geleerd en voor hoe jullie mij als klein zusje hebben geholpen te groeien tot wie ik vandaag ben - met een beetje van jullie beiden in mijn persoonlijkheid (leuk, hè!). Tijdens mijn PhD, maar natuurlijk ook daarbuiten, stonden jullie zo vaak voor me klaar - juist in moeilijke momenten waarin een traantje mocht vallen, maar ook tijdens de ontelbare blije momenten: thuis, op vakantie, met de kinderen, tijdens logeerpartijtjes bij jullie in het 'grote mensenhuis' en natuurlijk ook bij ons in Barceloneta, in ons knusse kleine huisje (wat was het fijn dat jullie hier allebei hebben kunnen logeren!). Er zijn zoveel mooie momenten geweest dat ik ze hier niet allemaal kan beschrijven. Wat ik wél wil zeggen, is dat ik me niets mooiers kan wensen dan zulke toppers als zussen!

Marlies, jij als een ongelofelijke harde werker, scherpe denker, sterke vrouw én een liefdevolle moeder. En alsof dat nog niet genoeg is, ben je ook nog eens een echte schat: altijd zorgzaam voor iedereen om je heen, en telkens weer kom je met die attente cadeaus (hoe je dat toch telkens doet, snap ik nog steeds niet)! **Sjoerd**, zorg jij maar goed voor die slimme, krachtige dame van je en vergeet niet: ook jij bent een echte (hockey)topper! Samen met jullie jongens maken jullie er altijd een vrolijke, gezellige buitenbende van. Het is dan ook zó leuk om met jullie op (natuur)expedities te gaan, zoals (berg)wandelen in Vall d'Aran, bakfietsen door Barcelona (Parc Güell is makkelijk aan te fietsen, oepsie) en (voor het eerst) een waddeneiland te bezoeken. En jongens, die bonte avond in Salardú zal ik zeker niet snel vergeten!

Mieke, wat ben jij een lieverd. Altijd hard aan het werk, druk in de weer met of voor je kids en vriendinnen, en dat nooit zonder iets lekkers in huis. Samen met **Steven** vormen jullie zo'n warm team. **Steven**, ook jij: dank je wel voor de talloze keren dat ik tijdens mijn verblijf in Nederland bij jullie in Utrecht mocht logeren. De koffie, altijd aan mogen schuiven met eten - het voelt bijna als een hotel, zó goed zorgen jullie voor me! Hoe kan ik dat ooit terugbetalen? En jullie toetjes zijn misschien een tikkeltje ongezond, maar absoluut onovertroffen! En dan jullie lieve kindjes (sorry dat ik jullie slaapkamer soms als logeerkamer kaap!), wat brengen jullie een heerlijke, vrolijke chaos met je mee. Jullie maken die weken nóg gezelliger en specialer!

Annie, dank je wel dat je ons vanaf jongs af aan de liefde voor avontuur hebt meegegeven - iets met kapotte auto's en onverharde wegen in Tsjechië ;) Ook je enthousiasme voor de natuur, een bewust(er) leven (there is no planet B!) en het genieten van het zonnetje werkt aanstekelijk. Hoewel... ik kan nog steeds niet tippen aan het aantal uren dat jij in de zon kunt doorbrengen (zelfs niet vanuit Spanje!), maar ik kom steeds dichter in de buurt! Misschien is dat ook wel waarom jij zo'n fan bent van Spanje en al vier keer in Barcelona bent geweest? Daarmee ben jij een absolute topscore-bezoeker!

Axel, liefje, wat kan ik zeggen... Soms kan ik me bijna niet meer voorstellen hoe het was vóór we elkaar kenden. We hebben al zoveel samen meegemaakt — en wie weet wat er nog allemaal op ons pad komt! Laat het vooral maar veel moois zijn!! In ieder geval geen nieuwe PhD's meer; dat tijdperk ligt nu toch echt achter ons ;) Ook al zijn we soms nog zoekende naar de beste stap richting ons volgende avontuur, ik kan me een leven zonder jouw liefde, steun, ontbijt-koffie (ooit maak ik er één voor jóu!) en knuffels echt niet voorstellen. Met jou aan mijn zijde kan ik alles nét wat beter aan – en wordt het leven ook nog eens een tikkie mooier!

Curriculum vitae

Adriana Emma Wesdorp, known as Emmy Wesdorp, was born on March 25, 1994, in Goes, Zeeland, The Netherlands. After completing her secondary education at the Buys Ballot College (later known as Ostrea Lyceum) in 2011, she began a bachelor's degree in Medicine the same year. In 2013, she also started a bachelor's degree in Biomedical Sciences, completing both degrees in 2015 and 2016, respectively.

Following an initial year in the Medicine master's program, she decided to pivot fully to Biomedical Sciences and began a master's in Molecular and Cellular Life Sciences with an integrated bioinformatics profile at the Utrecht University, The Netherlands. During her master's studies, she completed internships with Scott Zamvil at the University of California, San Francisco; Marvin Tanenbaum at the Hubrecht Institute, The Netherlands; and the Giorgio Galli group at the Novartis Institutes for Biomedical Research, Switzerland. Additionally, she participated in two honors programs, U/Select and Graduate Honours Interdisciplinary Seminars, and wrote an ethics-oriented thesis under the supervision of Karen Jongsma at the University Medical Center Utrecht (UMCU).

Though primarily experienced in wet lab and medical fields, she joined the bioinformatics research group of Jeroen de Ridder at the UMCU in early 2020. There, her research focused on shotgun next-generation sequencing-based pathogen detection methods, involving both wet lab work and the development and bioinformatics analysis of (meta)genomics data pipelines, the results of which are presented in this thesis.

Currently, Emmy Wesdorp is part of the research team at the Princess Máxima Center for Pediatric Oncology, where she works under the guidance of Lieve Tytgat on developing liquid biopsy-based cancer diagnostic and detection methods.

In her free time, Emmy enjoys bouldering, hiking in the mountains, snorkeling, and has a deep passion for drinking coffee in sunny places — a habit that, as she notes, helped her power through the challenges of her PhD.

Pathogens — bacteria, viruses, fungi, and parasites — are major causes of disease in humans and animals, yet can go undetected due to diagnostic limitations. Despite their clinical significance, timely and accurate identification remains challenging in certain settings.

This thesis explores next-generation sequencing of cell-free DNA as a minimally invasive diagnostic approach in two such contexts: detecting Aspergillus in immunocompromised pediatric patients and bacterial pathogens in septic newborn foals.