universität wien

DISSERTATION | DOCTORAL THESIS

Titel | Title

Short cfDNA fragments: Bioinformatic analysis and diagnostic applications

verfasst von | submitted by

Jan Müller B.Sc. M.Sc.

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Doctor of Philosophy (PhD)

Wien | Vienna, 2025

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 794 620 490
Dissertationsgebiet lt. Studienblatt | Field of study as it appears on the student record sheet:	Molekulare Biologie
Betreut von | Supervisor:	Dr. Arndt von Haeseler

Acknowledgements

Completing this PhD thesis has been a significant journey, and I am truly grateful to those who have supported me throughout the process.

First, I would like to express my sincere gratitude to my supervisors, Kai Sohn and Arndt von Haeseler, for the opportunities they created, their guidance, feedback, and support. Their expertise and encouragement have been instrumental in shaping this research. I also wish to thank the third member of my thesis advisory committee, Karl Kuchler, for his valuable suggestions, which have improved the quality of this work.

My heartfelt thanks go to my colleagues and friends whom I met during my time at IGB and CIBIV for their contribution to a very pleasant working environment. Also, a big thanks to Iris and Gerlinde, for helping with all administrative obstacles.

Lastly, I would like to thank our cooperation partners, funders, clinics, and patients without which this research would not have been possible. Thank you all for your support.

Contents

1 Abstract

The development of minimally-invasive diagnostic tools is crucial for improving outcomes and managing complex diseases such as cancer or sepsis. Cell-free DNA (cfDNA), fragments of DNA circulating in body fluids, holds promise as a basis for such diagnostic tools. While previous research has focused on nucleosomal cfDNA fragments, emerging evidence suggests that shorter double-stranded cfDNA (short cfDNA) fragments, between 20 and 60 base pairs in length, may offer unique insights due to their origin from interactions with non-histone DNA binding proteins, including transcription factors, which are key regulators of gene transcription and human physiology. Consequently, short cfDNA may provide a comprehensive view of the body's response mechanisms to physiological alterations, from which diagnostic potential can be derived. However, this potential has remained underexplored. This thesis aims to develop a technology for identifying novel biomarkers based on short cfDNA fragments and to establish a tailored bioinformatic approach for their corresponding analysis.

Chapter 4 presents a method for enriching short cfDNA from blood plasma. This method, combined with a custom bioinformatic pipeline, enables high-throughput sequencing and quantification of such cfDNA fragments. The study demonstrates the accumulation of these fragments at regulatory genomic elements, confirms their origin from transcription factor binding events, and characterizes the relationship to chromatin status and gene transcription. This method facilitates an untargeted identification of transcription factors potentially implicated in clinical conditions with the discovery of putative biomarkers. Analyses of a pilot cohort demonstrated the ability to distinguish between patient samples from two different gastrointestinal carcinoma types and between septic and non-septic patients as well as healthy individuals. Chapter 5 further explores the clinical potential of short cfDNA fragments for sepsis diagnostics. Sepsis, a severe disease characterized by a dysregulated host response to an infection, presents significant diagnostic challenges. The analysis of short cfDNA fragments provided insights into the host response and functionality of organ systems, as well as superior discrimination of bacterial or viral sepsis and early prediction of mortality risk compared to established metrics.

In conclusion, this thesis presents an approach for identifying and analyzing short cfDNA as a new biomarker class for different clinical indications. The findings underscore the potential of these short cfDNA fragments to improve diagnostics of complex diseases, offering a new approach for more accurate and minimally invasive clinical diagnostic tools. This work lays the foundation for developing targeted analysis approaches of the identified biomarkers and, thereby, the clinical application in precision medicine.

2 Zusammenfassung

Die Entwicklung minimal-invasiver Diagnoseverfahren ist entscheidend für die Verbesserung der Überlebenschancen und Behandlung komplexer Krankheiten wie Krebs oder Sepsis. Zellfreie DNA (cfDNA), d.h. DNA-Fragmente, die in Körperflüssigkeiten zirkulieren, sind eine vielversprechende Grundlage für solche Diagnoseverfahren. Während sich die bisherige Forschung auf nukleosomale cfDNA-Fragmente konzentrierte, deuten neue Erkenntnisse darauf hin, dass kürzere doppelsträngige cfDNA-Fragmente (short cfDNA), die zwischen 20 und 60 Basenpaaren lang sind, einzigartige Einblicke bieten können, da sie aus Wechselwirkungen mit nicht-histonischen DNA-bindenden Proteinen stammen, darunter Transkriptionsfaktoren, die wichtige Regulatoren der Gentranskription und der menschlichen Physiologie sind. Folglich kann kurze cfDNA einen umfassenden Einblick in die Reaktionsmechanismen des Körpers auf physiologische Veränderungen geben, woraus sich wiederum ein diagnostisches Potenzial ableiten lässt. Dieses Potenzial ist bisher jedoch noch nicht ausreichend erforscht worden. Ziel dieser Arbeit ist es, eine Technologie zur Identifizierung neuartiger Biomarker auf der Grundlage kurzer cfDNA-Fragmente zu entwickeln und eine darauf abgestimmte bioinformatische Analyse zu etablieren.

In Kapitel 4 wird eine Methode zur Anreicherung kurzer cfDNA aus Blutplasma vorgestellt. Diese Methode, kombiniert mit einer speziellen bioinformatischen Pipeline, ermöglicht die Hochdurchsatz-Sequenzierung und Quantifizierung solcher cfDNA-Fragmente. Die Studie zeigt die Anhäufung dieser Fragmente an regulatorischen genomischen Elementen, bestätigt ihren Ursprung aus Transkriptionsfaktor-Bindungsereignissen und charakterisiert die Beziehung zum Chromatinstatus und zur Gentranskription. Diese Methode ermöglicht eine ungerichtete Identifizierung von Transkriptionsfaktoren, die bei Erkrankungen eine Rolle spielen könnten, und eröffnet die Möglichkeit, neue Biomarker zu entdecken. Die Analysen einer Pilotkohorte haben gezeigt, dass es möglich ist zwischen Patientenproben von zwei verschiedenen gastrointestinalen Karzinomtypen und septischen und nicht-septischen Patienten sowie gesunden Personen zu unterscheiden. In Kapitel 5 wird das klinische Potenzial von kurzen cfDNA-Fragmenten für die Sepsisdiagnostik weiter untersucht. Sepsis, eine schwere Erkrankung, die durch eine dysregulierte Reaktion des Wirts auf eine Infektion gekennzeichnet ist, birgt erhebliche diagnostische Herausforderungen. Die Analyse kurzer cfDNA-Fragmente ermöglichte Einblicke in die Wirtsreaktion und die Funktionalität von Organsystemen sowie eine genauere Unterscheidung von bakterieller oder viraler Sepsis und eine frühzeitige Vorhersage des Mortalitätsrisikos im Vergleich zu etablierten Metriken.

Zusammenfassend wird in dieser Arbeit ein Ansatz zur Identifizierung und Analyse kurzer cfDNA als neue Biomarkerklasse für verschiedene klinische Indikationen vorgestellt. Die Ergebnisse unterstreichen das Potenzial dieser kurzen cfDNA-Fragmente zur Verbesserung

der Diagnostik komplexer Krankheiten und bieten einen neuen Ansatz für genauere und minimalinvasive klinische Diagnoseverfahren. Diese Arbeit legt den Grundstein für die Entwicklung gezielter Analyseansätze für die identifizierten Biomarker und damit für die klinische Anwendung in der Präzisionsmedizin.

3 Introduction

3.1 Chromatin and gene regulation

Deoxyribonucleic acid (DNA) serves as the molecular blueprint for the development, functioning, growth, and reproduction of all known living organisms and many viruses. Its central role in the storage and transmission of genetic information makes it the fundamental building block of life, and regulating its correct use is of utmost importance (Graw, 2015, p. 22).

DNA consists of a double helix formed by two complementary strands of nucleotides. Each DNA nucleotide consists of a phosphate group, a deoxyribose sugar, and one of the four nitrogenous bases: adenine (A), thymine (T), cytosine (C) and, guanine (G) (Graw, 2015, p. 23). The double helix structure is stabilized by hydrogen bonds between complementary bases, with the bond of the base pair (bp) G-C being stronger than that of A-T (Graw, 2015, p. 29). The helical structure of DNA creates two distinct grooves along the exterior, the major and minor grooves, which are critical for the binding of proteins and other molecules that regulate the function of DNA (Graw, 2015, p. 23). Due to its size and the need to regulate physical access to genetic information, the DNA in the cell nucleus is highly compacted with the help of proteins, forming chromatin. The fundamental unit of chromatin is the nucleosome, which consists of a segment of DNA wrapped around an octamer of histone proteins (Histone H2A, Histone H2B, Histone H3, and Histone H4) (Graw, 2015, p. 216). Additional proteins such as histone H1 and DNA topoisomerase, as well as modifications of the DNA or the histone proteins, can increase compaction, resulting in more condensed chromatin known as heterochromatin. The less condensed state is referred to as euchromatin (Graw, 2015, 220–221, 231-232).

The genetic information stored in the DNA is transcribed into messenger RNA and then into proteins by translation, a process called gene expression (Graw, 2015, p. 64). Gene regulation involves the control of timing, location, and amount of gene expression, which is essential for cellular differentiation, development, and response to environmental stimuli (Graw, 2015, p. 279). Chromatin structure is a key aspect of gene regulation. The organization of chromatin and nucleosome occupancy regulates gene accessibility and expression. Heterochromatin and high nucleosome occupancy is typically transcriptionally inactive, while euchromatin and low nucleosome occupancy are associated with active transcription (Figure 1 a (1, 2) and b) (Graw, 2015, pp. 286–287). Chromatin structure is dynamic and can be altered through various mechanisms that either promote or inhibit gene transcription. Chemical modifications of histone proteins, such as acetylation and methylation, can either promote or inhibit gene transcription. Histone acetylation is typically associated with active transcription, while histone methylation can either activate or

repress transcription depending on the specific histone residue that is methylated. DNA methylation is generally associated with inactive regulatory elements and gene repression, while gene bodies of actively transcribed genes are methylated (Figure 1 – a (3) and b) (Isbel et al., 2022). Regulatory genome elements, i.e., *cis*-regulatory elements, such as promoters or enhancers, influence the transcription of neighboring genes by providing specific DNA sequences (motifs) for the binding of the transcription machinery, e.g., the RNA polymerase, and additional regulatory proteins, known as transcription factors (TF) (Figure 1 – a (4, 5) and b). Transcription factors play a pivotal role in orchestrating the gene regulatory network. They can function as activators, repressors, or, depending on the context, switch between these roles. TFs usually contain one or more DNA-binding domains (DBD) that enable the interaction with their respective DNA motif. These DNA-binding domains are also used to classify TFs into groups (Wingender et al., 2018). Common types of DNA-binding domains include C2H2-zinc finger, Homeodomain, basic helix-loop-helix, basic leucine zipper, and nuclear hormone receptor (Lambert et al., 2018). In addition to the DBD, transcription factors often possess effector domains that mediate their regulatory function to the target, such as the recruitment of the transcriptional machinery, such as RNA polymerase II and its co-regulators. Transcription factors that are part of the transcription machinery, like general transcription factor IIA (GTF2A) or GTF2B, are often referred to as 'general' TFs. 'Specific' transcription factors regulate a defined set of genes as a response to a signal, like steroid hormone receptors or nuclear factor kappa B (NF-κB) (Eun, 1996). Pioneer factors, like forkhead box A1 (FOXA1) or SRY-box transcription factor 2 (SOX2), are a special class of transcription factors that can bind to compacted chromatin and initiate its opening, thereby making it accessible (Barral and Zaret, 2024). TFs, like CCCTC-binding factor (CTCF) or SATB homeobox 1 (SATB1), are also involved in the organization of the genome into higher-order structures. They play a critical role in the formation of chromatin loops and insulator functions, influencing gene expression by shaping the three-dimensional architecture of the genome (Holwerda and de Laat, 2013). Transcription factors are crucial for various biological processes, including cell growth, differentiation, response to external stimuli, and underlie many different aspects of human physiology or disease (Lambert et al., 2018). Altered TF regulation can be used as a response to a disease or be the source of it, as dysregulation of transcription factor activity can lead to numerous diseases, including cancer, autoimmune disorders, and developmental abnormalities. For instance, the gene of transcription factor p53 is the single most frequently mutated gene in human cancer (Vaddavalli and Schumacher, 2022). These diverse and critical functions of TFs highlight the valuable insight about the physiological state of an individual that can be gained from them.

3.2 Genomics and Bioinformatics

All the factors of gene regulation and chromatin described, along with those not covered, such as the crucial non-coding RNAs, constitute a highly complex gene regulatory net-

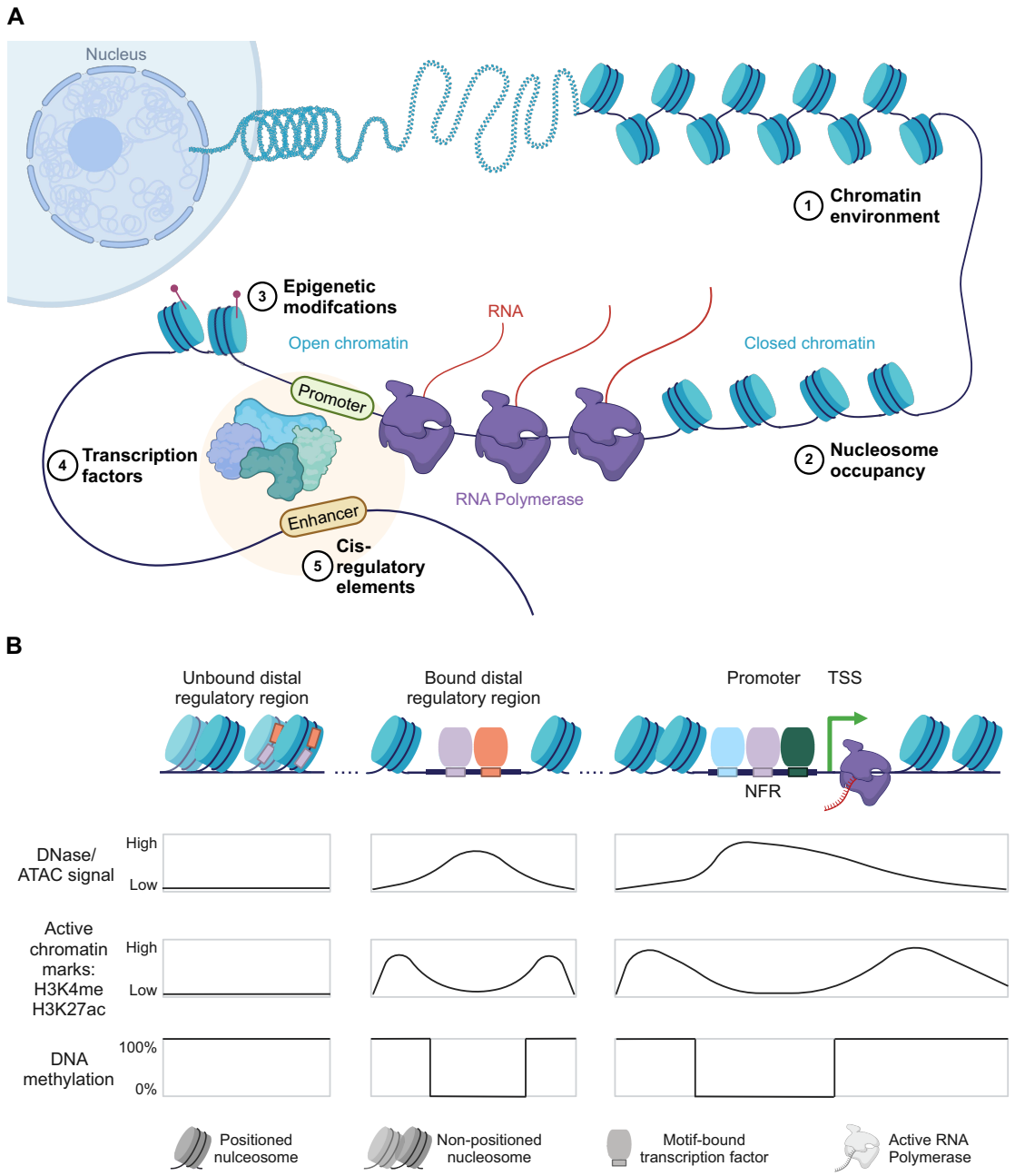


Figure 1: a) Schematic representation of the mechanisms that regulate gene expression at the chromatin level. The higher-order chromatin environment (hetero- or euchromatin state; 1) and the local degree of nucleosome occupancy (2) determine whether the chromatin is in a compacted (closed chromatin) or an accessible state (open chromatin). Epigenetic modifications (3), such as the addition or removal of chemical groups (e.g., acetylation, methylation) on histones or DNA, play a crucial role in modulating the chromatin structure. These modifications can either promote a relaxed chromatin state that favors transcription or maintain a tightly packed, transcriptionally inactive state. In regions of open chromatin, transcription factors (4) bind to specific DNA sequences and interact with *cis*-regulatory elements (5), such as promoters and enhancers, to recruit and position RNA polymerases at the start of a gene. RNA polymerases will then transcribe the DNA into RNA to produce proteins, the functional units, after translation and subsequent processing. b) Chromatin features at active regulatory elements. Gene activation is accompanied by TF binding and clear nucleosome organization, with strong positioning near the transcription start site (TSS) and nucleosome-free region (NFR) and around high-affinity TF binding sites. The assay for transposase-accessible chromatin (ATAC) and DNase detect open chromatin and are located at NFRs, i.e., at the promoter and at bound regulatory regions. Activation is accompanied by a loss of DNA methylation at upstream regulatory elements of the promoter and an increase in active chromatin marks such as monomethylated, dimethylated, and trimethylated histone H3-lysine 4 (H3K4) or acetylated histone H3-lysine 27 (H3K27ac) besides these elements. Adapted from Isbel et al. (2022) and created in BioRender by Müller, J. [2025, https://BioRender.com/t85e075, https://BioRender.com/w45a104].

work. In this network, numerous components influence each other and thus enable the finely tuned biological functions that are necessary for life but also create an exceptionally complex system. Studying this system has gained a lot of momentum since the development of high-throughput, next-generation sequencing (NGS) of DNA and RNA (Bentley et al., 2008). Today, these technologies enable multiplexed sequencing and, thereby, the quantification of billions of DNA fragments, consisting of up to 2-3 trillion nucleotides, in a single experiment with continuously reducing prices per sequenced base. This advancement has introduced the field of genomics, which involves analyzing the structure, function, evolution, mapping, and editing of genomes, offering profound insights into the genetic basis of diseases and the mechanisms of complex biological systems. An essential aspect of genomics is the ability to handle and analyze data from the NGS platforms using bioinformatic methods.

The simplest sequencing assay, the sequencing of fragmented pieces of genomic DNA, known as whole genome sequencing, also provides the basic bioinformatics methods used in most subsequent sequencing assays. The quality control of sequencing data and the removal of the platform- and procedure-specific artifacts, such as adapter sequences, using tools like FastQC and Trimmomatic, is the first step (Andrews, 2010; Bolger et al., 2014; Koboldt et al., 2010). Quality-filtered reads are then aligned to a reference genome using mapping tools like the Burrows-Wheeler Alignment tool or NextGenMap (Koboldt et al., 2010; Li and Durbin, 2009; Sedlazeck et al., 2013). Following whole genome sequencing, more specialized sequencing assays were developed to analyze more detailed aspects of genomics. Chromatin immunoprecipitation assays (ChIP-Seq) are used to identify binding sites of TFs across the genome or to determine the positioning of nucleosomes with specific modifications, like H3K27ac (Johnson et al., 2007). Open chromatin and, therefore, nucleosome-free regions can be identified with enzymes that only interact with freely accessible DNA like DNase I (DNase-Seq) or today's more common transposase (ATAC-Seq) (Boyle et al., 2008; Buenrostro et al., 2013). Since these assays generate non-uniformly distributed sequencing reads across the genome, local enrichments are identified with peak callers like MACS or more tailored ones like HINT (Li et al., 2019; Zhang et al., 2008). Such peaks or the mapped reads in general are used in a wide range of downstream analyses, with two important examples being the identification of enriched TF binding motifs using tools like MEME or differential read count enrichment analysis between experimental condition, mostly with tools developed for RNA-Seq, like DESeq (Bailey et al., 2015; Love et al., 2014; Nakato and Sakata, 2021; Yan et al., 2020). Whilst ChIP-Seq directly maps single TF's binding sites, DNase-Seq and ATAC-Seq can also capture traces of multiple TF's binding, so-called footprints. Here, a footprint refers to a pattern where a TF binds to DNA and prevents enzymatic cleavage within a detected region of open chromatin. This leaves a relative depletion within a peak ('peak-dip-peak') but also requires a comparably large sequencing depth to effectively retrieve footprints (Liu et al., 2019; Yan et al., 2020). These *in vitro* assays are proven reliable and powerful tools for studying bi-

ological processes and the importance of TFs but require intact genomic DNA, which can only be obtained from cell cultures or primary tissue samples. Hence, their application in diagnostic use cases is limited as these source materials are only accessible through invasive procedures. This makes it challenging to routinely obtain the necessary samples for diagnostic purposes, thereby restricting the broader clinical application of these assays. As a result, there is a need for alternative methods that can provide similar insights with less invasive sampling techniques.

3.3 Cell-free deoxyribonucleic acids

Nucleic acids were first detected in circulation in human blood plasma in 1948 by Mandel and Metais. Since then, and especially with the advent of modern high-throughput sequencing technologies for nucleic acids, much has been learned about what is now called cell-free deoxyribonucleic acid (cfDNA) or cell-free ribonucleic acid (cfRNA). CfDNA is abundant in blood plasma, originates from a wide range of tissues and cell types, contains a plethora of biological information, and provides valuable insights into human physiology *in vivo* with easily accessible sample material (Kustanovich et al., 2019).

The DNA released into the bloodstream is digested by nucleases and rapidly cleared from the bloodstream, with a half-life of less than 30 minutes (Lo et al., 1999). Assuming a steady state, the rapid degradation is balanced by the rapid supply of cfDNA from the sources. Healthy subjects' cfDNA originates predominantly from leukocytes: granulocytes (30%), megakaryocytes (30%), monocytes or macrophages (20%), erythrocyte progenitors (5%), and lymphocytes (3%). Solid tissues contributing to cfDNA include vascular endothelial cells (6%) and hepatocytes (3%) and others in lower percentages, as identified through DNA methylation-based deconvolution of tissues by Loyfer et al.. These proportions may vary depending on the patient's current physiology and disease status. CfDNA is mainly produced by cell death processes, including apoptosis, pyroptosis, ferroptosis, or necrosis, but also by excretory mechanisms such as NETosis, the activation, and release of Neutrophil Extracellular Traps, or cell-to-cell communication via extracellular vesicles (Mouliere, 2022). The majority of cfDNA originates from human cells, but pathogenic microbes also contribute to the circulating nucleic acids, especially during infections (Figure 2). Various forms of cell-free nucleic acids have been characterized, most of which are nucleosomal cfDNA, i.e., DNA that is wrapped around a nucleosome core particle and thus protected from nuclease digestion. Other DNA-bound proteins, vesicles, or secondary structures also slow down degradation. Unprotected nucleic acids are also in circulation, albeit in much smaller quantities (Mouliere (2022); Figure 2). In addition to cfDNA, cfRNA is also found in the bloodstream, but it has not yet been studied and characterized as intensively (Zhong et al., 2024).

Most cfDNA in blood plasma is present as the nucleosomal type (regular cfDNA), with fragment lengths ranging from 147 bp to 167 bp, owing to the binding of core histones and

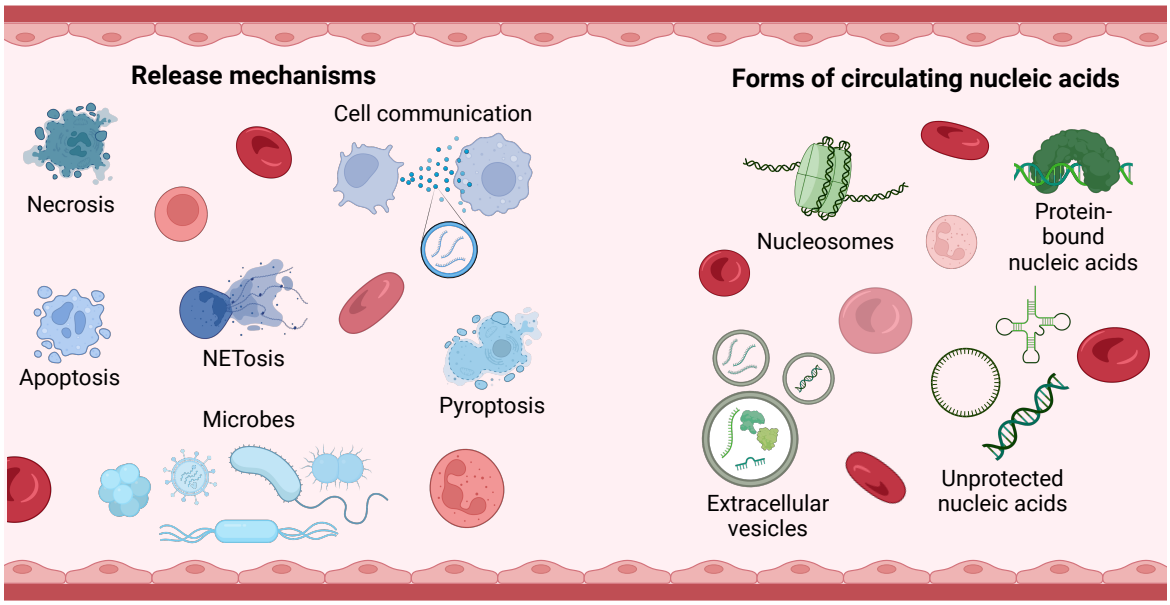


Figure 2: Release mechanisms and forms of circulating nucleic acids in the bloodstream. Various cellular processes such as necrosis, apoptosis, NETosis, pyroptosis, and microbial activity contribute to the release of nucleic acids into the bloodstream. In addition, cell communication pathways can release nucleic acids through extracellular vesicles. Once the nucleic acids enter the bloodstream, they can be present in different forms: protected by proteins, either histones or other proteins, encapsulated in extracellular vesicles, or as unprotected nucleic acids with or without secondary structures. Adapted from Loy et al. (2024) and created in BioRender by Müller, J. [2025, https://BioRender.com/m77e911].

linker histone 1. Incompletely digested genomic DNA can comprise multiples of the single nucleosome unit, resulting in larger fragments at intervals. (Fan et al., 2010). In addition to the nucleosomal cfDNA, the presence and potential origins of shorter cfDNA fragments were first reported in 2016 by Burnham et al. and Snyder et al.. They found that sequencing of regular cfDNA provides a genomewide map of *in vivo* nucleosome occupancy, while short cfDNA fragments, ranging from 35 to 80 bp, directly map transcription factor occupancy. However, due to the tiny amount of short cfDNA fragments, extremely deep sequencing depths (10^9) per single sample were required (Snyder et al., 2016). A crucial step in high-throughput DNA sequencing is converting DNA molecules into a library of fragments suitable for sequencing on an NGS platform. The objective of DNA sequencing library preparation is to attach adapters to the ends of these fragments, allowing for their amplification and sequencing. The two most commonly used methods involve either ligating double-stranded DNA adapters to both ends of the DNA fragment or ligating a single-stranded DNA adapter to one end. These two approaches fundamentally affect the analysis of short cfDNA (Burnham et al., 2016). As a first solution, single-stranded DNA sequencing library preparation methods yielded more shorter fragments but also a wider range of cfDNA types, including single-stranded cfDNA and severely degraded cfDNA (Burnham et al., 2016). Subsequent research showed that single-stranded cfDNA combined with *in vitro* size selection using gel electrophoresis of fragments yields this mixture of short cfDNA fragments that demonstrated a diagnostic potential for cancer diagnostics (Hudecova et al., 2022). Since these methods also extract fragments of other biologi-

cal origin than those protected by DNA-binding proteins, e.g., DNA secondary structures like G4 quadruplexes, the direct mapping of transcription factor occupancy described by Snyder et al. is blurred. As an alternative, Ulz et al. developed a bioinformatic method to extract TF occupancy using nucleosomal cfDNA. This approach enabled accurate colorectal tumor subtype prediction and tumor detection (Ulz et al., 2019). However, the quantification of TF binding via nucleosome positioning is an indirect measurement, as it quantifies the absence of nucleosomes rather than the direct binding of the TF itself. This indirect approach still necessitates deep genomewide read coverage. The substantial amount of data, resources, and financial investment required for achieving sufficient sequencing depth, ranging from approximately 50 to 500 million paired-end sequencing reads, makes it relatively costly for diagnostic procedures. Therefore, obtaining information about TF occupancy *in vivo* is most readily and efficiently achieved through direct analysis of fragments protected by TFs, which is intact short double-stranded cfDNA (from now on referred to as 'short cfDNA', unless explicitly stated otherwise, e.g., short single-stranded cfDNA).

3.4 Liquid biopsy and cfDNA biomarker

Much of the research on cfDNA aims to develop so-called liquid biopsy applications. Liquid biopsy represents a disruptive approach to diagnostics, enabling the detection and monitoring of disease through minimally invasive sampling of body fluids, such as blood. In this way, liquid biopsy approaches offer a minimally invasive and easily repeatable alternative to established tissue biopsies. Liquid biopsies can use a wide range of analytes known as biomarkers or biological markers, i.e., any measurable indicator of a biological state or condition. In addition to metabolites or proteins in the bloodstream, biomarkers based on cfDNA have been gaining importance over the last 25 years.

The detection of fetal cfDNA in maternal blood plasma by Lo et al. in 1997 laid the foundation for non-invasive prenatal testing (NIPT). With the advancements in NGS, NIPT quickly moved from concept to clinical application as an alternative to invasive tests and serum protein screens. Initially intended for high-risk pregnancies, NIPT has now become a primary screening tool, with tens of millions of applications annually (Samura, 2020). Expedited trials have consistently demonstrated its impressive accuracy in diagnosing aneuploidies such as trisomy 21 (Norton et al., 2015). Similar to the distinction between fetal and maternal cfDNA, Lo et al. showed in 1998 that cfDNA from organ transplants (donor-derived cfDNA, dd-cfDNA) circulates in the blood plasma of liver and kidney transplant recipients. With further improvements of the methodology by (Snyder et al., 2011) and following studies of the field, the utility of dd-cfDNA as a marker of rejection is by now firmly established in heart, kidney, liver, lung, and other transplants (Loy et al., 2024). Dd-cfDNA has become one of the best-studied biomarkers in solid organ transplantation (Knight et al., 2019). As cancer treatment shifts more toward precision

medicine with patient-tailored therapies, cfDNA has become an increasingly important biomarker. NGS cancer panels aim to detect driver mutations, copy number alterations, tumor mutational burden, and microsatellite instability. If such a panel has sufficient sensitivity for lower tumor fractions, cfDNA-based liquid biopsy methods are a logical enhancement over tissue-based analyses that have found their way into clinical application (Ignatiadis et al., 2021). Aside from the guidance of personalized treatments, cfDNA is used to assess the treatment efficacy and also to predict relapse (Moding et al., 2021). The field of oncology diagnostics based on cfDNA is expanding rapidly and shifting to focus on a variety of non-genetic cfDNA traits, such as epigenetic alterations or TF interactions (Moser et al., 2023; Ulz et al., 2019). A prominent example of the shift towards non-genetic cfDNA biomarkers is the development of multi-cancer early detection tests by Illumina and GRAIL, which primarily rely on cfDNA methylation patterns. These tests aim to revolutionize cancer screening by enabling the detection of multiple cancer types in a large, asymptomatic population using a single, minimally invasive test (Jamshidi et al., 2022).

The evolution of cfDNA diagnostics from prenatal testing to oncology and beyond demonstrates its profound impact on modern diagnostic medicine. As the field advances, the development of non-genetic cfDNA biomarkers is set to enhance diagnostic capabilities for a broader range of conditions, including multifactorial diseases, such as oncological or immune-mediated diseases. By enabling the detection of physiological alterations beyond the scope of genetic markers alone, non-genetic cfDNA biomarkers open new avenues for diagnosing complex diseases and physiology changes, highlighting their potential in the realm of diagnostics.

3.5 Sepsis and its diagnostics

Sepsis is a severe and potentially fatal condition resulting from a dysregulated immune response to an infection, leading to systemic inflammation and organ dysfunction (Singer et al., 2016). The underlying infections are typically pulmonary, intra-abdominal (gastrointestinal or urogenital), or bloodstream-related and caused mostly by bacteria, but also from viruses, fungi, or rarely parasites (Evans et al., 2021; Vincent et al., 2009). Sepsis and septic shock are serious public health problems that affect millions of people around the world every year, with one in three to six of those affected dying (Fleischmann-Struzek et al., 2020; Rhee et al., 2017). Risk factors independently associated with a higher risk of hospital mortality include comorbid cancer, heart failure, immunosuppression, cirrhosis, infections with *Pseudomonas*, *Enterococcus*, or *Acinetobacter* species, older age, and the need for mechanical ventilation or renal replacement therapy (Vincent et al., 2009). Early diagnosis and prompt correct treatment are critical for improving the prognosis of sepsis patients (Evans et al., 2021). To achieve this, the identification of causative pathogens and the assessment of the patient's current physiology and response to the in-

fection are essential to initiate early and adequate interventions for optimal treatment.

Traditional diagnostic methods to identify sepsis-causing pathogens are blood cultures and polymerase chain reaction (PCR) assays. Both come with significant limitations. Whilst PCR assays are fast in turnaround time and accurate, they are limited to a few molecular targets, i.e., pathogens, that can be detected (Li et al., 2022). Given the broad spectrum of putative sepsis pathogens, this limitation makes it challenging to comprehensively test the possible causative agents of sepsis in a timely manner. Blood culture testing, the current standard of care, suffers from even more limitations. Only 40% of patients diagnosed with sepsis or septic shock have positive blood cultures (Li et al., 2021). Additionally, the use of antibiotics further reduces the probability of identifying the causative pathogen (Scheer et al., 2019). While some bacterial species and various fungi do not reliably grow in blood cultures, viruses cannot be detected at all (Duncan et al., 2021). With a turnaround time of at least 48-72 hours to diagnosis, there is still an urgent need for improved diagnostic solutions (Vincent et al., 2015).

The presence of microbial DNA in the bloodstream, coupled with the fact that most tissues in the body are connected to blood circulation and the minimally invasive nature of blood sampling, makes blood plasma an attractive sample type and cfDNA an appealing analyte for sepsis diagnostics. Sequencing-based approaches utilizing microbial cfDNA allow for hypothesis-free screening of all DNA pathogens in a single test. With improved turnaround times of around 24 hours, significantly increased sensitivity (60 to 70%), and comparable specificity, these methods have the potential to become the new standard for sepsis pathogen diagnosis (Blauwkamp et al., 2019; Brenner et al., 2021; Grumaz et al., 2019; Park et al., 2023). This may lead to reductions in intensive care unit (ICU) or intermediate care unit length of stay, mortality rate, acute kidney injury incidence, duration of antimicrobial therapy, and costs from complications and outpatient aftercare, along with an improvement in patient quality of life (Brenner et al., 2021). The host response plays a crucial role in distinguishing between mere colonization and active infection, as well as in gauging the severity of sepsis. Recent advances suggest that the limitations of current diagnostic methods can be addressed by integrating metagenomic cfDNA sequencing with host response profiling.

Recent studies evaluated the potential of the integrative analysis of non-genetic cfDNA or cfRNA biomarkers and microbial cfDNA to trace the tissue origins of cfDNA, quantify the presence of various viral and bacterial pathogens, and assess the extent of host tissue damage due to infection (Cheng et al., 2019, 2021; Jing et al., 2022; Kalantar et al., 2022). Combining pathogen detection with host response evaluation may change sepsis diagnostics, enhancing the precision and relevance of clinical assessments. Comparable to these advancements, there is the opportunity to explore short cfDNA-based biomarkers and transcription factor occupancy, which could offer further insights into the host-pathogen interaction. These innovations have the potential not only to refine the diagnostic process

but also to improve the treatment of sepsis and patient outcomes.

3.6 Aims of the Thesis

Understanding the pivotal role of chromatin and gene regulation in essential cellular processes, alongside significant advancements in genomics and bioinformatics, sets the stage for exploring the potential of cfDNA as a minimally invasive biomarker for various diseases. Current diagnostic methodologies, particularly for conditions such as sepsis, face limitations necessitating innovative approaches to improve diagnostics and patient outcomes. Non-genetic cfDNA biomarkers, like cfDNA methylation, have emerged as promising candidates for novel biomarker developments. Short cfDNA fragments are believed to originate from non-histone DNA binding proteins, such as transcription factors, offering a unique insight into gene regulatory networks and cellular states. Moreover, their presence in easily accessible blood plasma samples enhances their potential for clinical diagnostic applications. Consequently, this thesis harnesses the properties of short cfDNA fragments identify non-genetic biomarkers using dedicated bioinformatics analyses. The main objective of the dissertation is to establish a technology for identifying novel biomarkers based on short cfDNA fragments for clinical applications. This thesis emphasizes the establishment of a dedicated bioinformatic analysis for this novel data. The first part of the work aims to develop a robust bioinformatics and molecular biology protocol for the enrichment of short cfDNA (approx. 40 base pairs long) from blood plasma for high-throughput DNA sequencing to ensure the specific quantification of this class of cfDNA. The developed bioinformatics pipeline will ensure that putative sequencing artifacts are removed and systematic biases are addressed to achieve the most accurate quantification of short cfDNA fragments. In addition, the properties of short cfDNA and its supposed origin from non-histone DNA binding proteins, like transcription factors, as well as the interaction with other genetic, epigenetic, and transcriptional information, will be characterized. The analysis of pilot phase cohorts of clinical samples will provide first insights into the diagnostic potential of short cfDNA and, thus, a first indication of the approach's utility for precision medicine. The pilot phase cohorts will include patients with oncologic diseases as well as infectious or immune-mediated diseases, specifically sepsis. The second part of the work will focus on a deeper assessment of the diagnostic potential of short cfDNA, especially in the example of sepsis. Specialized bioinformatic analysis will be used for biomarker identification. With these biomarkers, predictive and diagnostic performance will be compared to established clinical metrics to diagnose infection status (bacterial vs. viral) and predict early death with larger cohort sizes than in the first pilot phase. This will also yield putative short cfDNA biomarkers for deepened investigation in future research. Moreover, this part aims to evaluate what information short cfDNA can provide about a patient's current physiological state. In combination, this research will position short cfDNA in the field of precision diagnostic research and will also enable the development of new diagnostic tools.

4 A novel approach for *in vivo* DNA footprinting using short double-stranded cell-free DNA from plasma

4.1 Preamble

This publication introduces a novel method to enrich short double-stranded cfDNA fragments from blood plasma for high-throughput sequencing and the dedicated bioinformatic analysis of the generated data. The novel data is thoroughly characterized and incorporated within the field of cfDNA research.

The study confirms that short cfDNA sequencing data provides information on the genome-wide interaction of non-histone proteins with the genome from liquid biopsy samples. Associations to transcription factors, chromatin status, and gene transcription are described. Moreover, the study provides a proof-of-concept for the diagnostic potential of short cfDNA. First, putative biomarkers that distinguish two gastrointestinal carcinomas or septic patients from non-septic controls have been identified from a few pilot samples.

These results form the basis for the use of short cfDNA sequencing as a source of biomarkers for minimally-invasive diagnostics of complex diseases.

4.2 Contribution of the Thesis Author

The thesis author is the first author of this publication. The thesis author and Kai Sohn conceptualized and designed the study. The thesis author performed all bioinformatics analyses and prepared all figures. The thesis author, along with Kai Sohn, Yevhen Vainshtein, and Arndt von Haeseler, interpreted the data analysis results. The thesis author, Kai Sohn, and Mirko Sonntag wrote the manuscript.

4.3 Publication Status

Published in Genome Research (Cold Spring Harbor Laboratory Press) on September 13, 2024: **Müller, J.**, Hartwig, C., Sonntag, M., Bitzer, L., Adelmann, C., Vainshtein, Y., Glanz, K., Decker, S. O., Brenner, T., Weber, G. F., von Haeseler, A., & Sohn, K. (2024). A novel approach for in vivo DNA footprinting using short double-stranded cell-free DNA from plasma. Genome research. https://doi.org/10.1101/gr.279326.124

Method

A novel approach for in vivo DNA footprinting using short double-stranded cell-free DNA from plasma

Jan Müller,[1,2,3,4] Christina Hartwig,[1,5] Mirko Sonntag,[1,6] Lisa Bitzer,[1] Christopher Adelmann,[1] Yevhen Vainshtein,[1] Karolina Glanz,[1] Sebastian O. Decker,[7] Thorsten Brenner,[8] Georg F. Weber,[9,10] Arndt von Haeseler,[11,12] and Kai Sohn[1]

[1]*Innovation Field In-vitro Diagnostics, Fraunhofer Institute for Interfacial Engineering and Biotechnology IGB, 70569 Stuttgart, Germany;* [2]*Max Perutz Labs, Vienna Biocenter Campus, 1030 Vienna, Austria;* [3]*University of Vienna, Max Perutz Labs, Department of Structural and Computational Biology, Center of Integrative Bioinformatics Vienna, 1030 Vienna, Austria;* [4]*Vienna BioCenter PhD Program, Doctoral School of the University of Vienna and Medical University of Vienna, 1030 Vienna, Austria;* [5]*Institute for Interfacial Engineering and Plasma Technology, University of Stuttgart, 70569 Stuttgart, Germany;* [6]*Interfaculty Graduate School of Infection Biology and Microbiology, Eberhard Karls University Tübingen, 72074 Tübingen, Germany;* [7]*Heidelberg University, Medical Faculty Heidelberg, Department of Anesthesiology, 69120 Heidelberg, Germany;* [8]*Department of Anesthesiology and Intensive Care Medicine, University Hospital Essen, University Duisburg-Essen, 45147 Essen, Germany;* [9]*Department of Surgery, Friedrich-Alexander University Erlangen-Nürnberg and Universitätsklinikum Erlangen, 91054 Erlangen, Germany;* [10]*Comprehensive Cancer Center Erlangen-EMN, Friedrich-Alexander University Erlangen-Nürnberg and Universitätsklinikum Erlangen, 91054 Erlangen, Germany;* [11]*Center of Integrative Bioinformatics Vienna, Max Perutz Labs, University of Vienna and Medical University of Vienna, Vienna BioCenter, 1030 Vienna, Austria;* [12]*University of Vienna, Faculty of Computer Science Bioinformatics and Computational Biology, 1090 Vienna, Austria*

Here, we present a method for enrichment of double-stranded cfDNA with an average length of ~40 bp from cfDNA for high-throughput DNA sequencing. This class of cfDNA is enriched at gene promoters and binding sites of transcription factors or structural DNA-binding proteins, so that a genome-wide DNA footprint is directly captured from liquid biopsies. In short double-stranded cfDNA from healthy individuals, we find significant enrichment of 203 transcription factor motifs. Additionally, short double-stranded cfDNA signals at specific genomic regions correlate negatively with DNA methylation, positively with H3K4me3 histone modifications and gene transcription. The diagnostic potential of short double-stranded cell-free DNA (cfDNA) in blood plasma has not yet been recognized. When comparing short double-stranded cfDNA from patient samples of pancreatic ductal adenocarcinoma with colorectal carcinoma or septic with post-operative controls, we identify 136 and 241 differentially enriched loci, respectively. Using these differentially enriched loci, the disease types can be clearly distinguished by principal component analysis, demonstrating the diagnostic potential of short double-stranded cfDNA signals as a new class of biomarkers for liquid biopsies.

[Supplemental material is available for this article.]

Liquid biopsies are based on various types of analytes, including circulating extracellular nucleic acids like cell-free DNA (cfDNA), extracellular vesicles, or circulating tumor cells, for example (Bronkhorst et al. 2019; Kustanovich et al. 2019). Physiologically, cfDNA is to a large extent released from the hematopoietic system by apoptosis, necrosis, or active secretion from almost all cell types and tissues into the bloodstream (Grabuschnig et al. 2020). In addition to release from normal physiological turnover, cancer cells or microbial pathogens are also known to release their DNA into bloodstream circulation (Heitzer et al. 2020). Released genomic DNA is then degraded by DNA-digesting enzymes (nucleases), producing fragments mainly 147 to 167 bp in size, corresponding to a single nucleosome (Han et al. 2020; Lo et al. 2021). By high-throughput sequencing of cfDNA fragments, nucleosome positioning can be inferred at base pair resolution (Fan et al. 2008; Snyder et al. 2016). The exact positions of nucleosomes and chromatin structure play a key role in regulating gene expression by providing access to DNA for the transcription machinery. Open chromatin structures depleted of nucleosomes make DNA more accessible for key regulators of transcription, including transcription factors, enhancers, or repressors. However, the routine and efficient measurement of genome-wide protection through regulatory DNA-binding proteins (DBPs) is not yet established. Recently, a minor fraction of double-stranded cfDNA that is significantly shorter than normal cfDNA was found, ranging from 35 to 80 nucleotides (nt) using ultradeep sequencing of total cfDNA (Burnham et al. 2016; Snyder et al. 2016). It has been proposed that this short cfDNA might be protected by DNA-binding factors and therefore could represent direct transcription factor binding.

Here, we established an enrichment approach for short double-stranded cfDNA fragments (20–60 bp) from blood plasma (further referred to as "short cfDNA sequencing"). Corresponding

Corresponding author: kai.sohn@igb.fraunhofer.de
Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.279326.124. Freely available online through the *Genome Research* Open Access option.

short double-stranded cfDNA fragments (further referred to as "short cfDNA") accumulate at open chromatin as well as gene regulatory elements. Differential enrichment of short cfDNA at genomic loci facilitated discrimination of colorectal and pancreatic cancer patient samples as well as between septic patients and clinical controls, showing potential for diagnostic applications.

Results

Enrichment of cfDNA footprints at regulatory genomic regions

Enrichment of short double-stranded cfDNA comprises extraction of total cfDNA from blood plasma by magnetic beads followed by double-stranded DNA-specific library preparation (Supplemental Fig. S1). To select short double-stranded cfDNA fragments of up to 60 bp, two size-selection steps were performed using a preparative gel electrophoresis device. For this, fragments in the range of 150 bp to 200 bp are enriched, which corresponds to double-stranded cfDNA fragments of up to 60 bp ligated to a sequencing adapter of 140 bp (Supplemental Fig. S2). After high-throughput sequencing of size-selected libraries, reads were quality-checked to ensure a size range between 20 bp and 60 bp (Supplemental Fig. S3). The protocol revealed sequencing reads with a mean read length of 37.9 bp across the patient conditions and comparable read length distributions per condition despite varying sample storage durations (SD = 6.6 bp, $n = 2 \times 10^7$ reads uniformly sampled from the total reads of 20 individuals: four healthy individuals, four patients with pancreatic ductal adenocarcinoma [PDAC], four patients with colorectal carcinoma [CRC], four patients with sepsis, and four nonseptic postoperative clinical control patients [Post-OP]) (Supplemental Fig. S4; Supplemental Table S1). Short cfDNA reads revealed an elevated average GC content of 57.8% (SD = 1.9%) (Supplemental Fig. S5) in contrast to 40.9% for average human genomic DNA (Piovesan et al. 2019). We mapped short cfDNA to the human genome and compared it with the mapping of regular cfDNA of four additional healthy individuals. Coverage profiles of short cfDNA and regular cfDNA differed considerably as short cfDNA tends to accumulate either in single narrow peaks or in clusters of narrow peaks (clustered peaks), which are frequently located at transcriptional start sites (TSSs) or ChIP-seq-validated transcription factor binding sites (TFBSs) (Fig. 1A). Nucleosome-free regions (NFRs) were analyzed in comparison to clustered peaks, because the absence of regular cfDNA can be an indicator of the presence of other DBPs. Assignment of clustered peaks from short cfDNA and NFRs to annotated functional elements of the genome shows that an approximately four- and a six-fold higher proportion of clustered peaks are found in promoters (>1 kb upstream of the TSS) and 5′ UTRs of genes, respectively. Moreover, the proportion of clustered peaks assigned to exons is

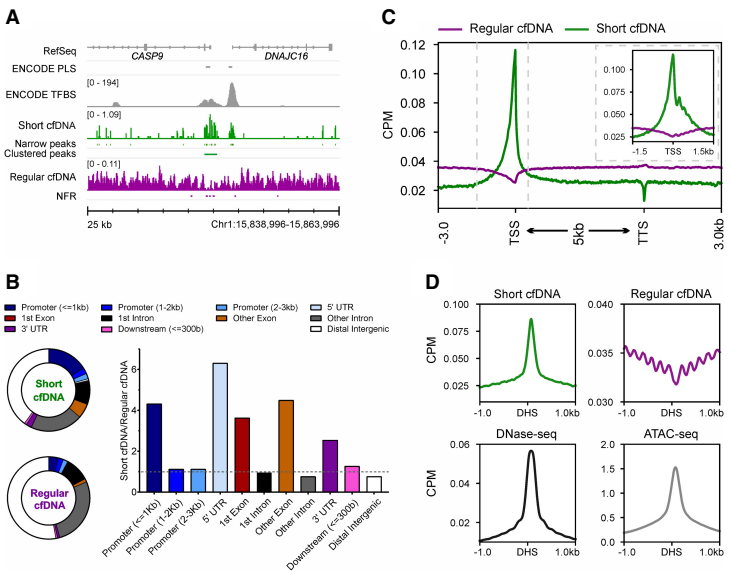


Figure 1. Short cfDNA is enriched in regulatory regions of genes and open chromatin. (*A*) Coverage profile of short cfDNA (green; S03) and regular cfDNA (violet; average of S05–S08). The short cfDNA profile shows narrow and clustered peaks, whereas regular cfDNA shows nucleosome-free regions (NFRs; depletion of reads). RefSeq genes, ENCODE promoter-like structures (PLSs), and ENCODE transcription factor binding sites (TFBSs) based on ChIP-seq experiments are included as references. (*B*) Pie charts display the proportions of annotated genomic features for clustered peaks from short cfDNA and NFRs from regular cfDNA. The bar plot shows the ratio between clustered peaks and NFRs for each genomic feature. (*C*) Average coverage profiles for short cfDNA (S03) and regular cfDNA (S06) along all annotated protein-coding genes. The gene bodies of all genes are scaled to 5 kb. Dashed vertical lines indicate the interval shown in the subplot. The subplot shows average profiles at the transcription start site without rescaling of the gene body. (*D*) Average coverage profiles for short cfDNA, regular cfDNA, and publicly available ATAC-seq data at DNase I hypersensitive sites (DHSs) derived from publicly available DNase-seq data.

about four times higher than the proportion of NFRs (Fig. 1B). A more detailed examination of the genomic coverage for protein-coding genes reveals an opposite profile between short cfDNA and regular cfDNA. Although short cfDNA is enriched at TSSs, regular cfDNA is depleted at these sites. In addition, short cfDNA possesses a reciprocal pattern oscillation compared with regular cfDNA, with short cfDNA exhibiting an inverted pattern 1 kb downstream from the TSS in regular cfDNA (Fig. 1C, inset). In addition to enrichment at the TSS, short cfDNA also exhibits a prominent signal at DNase I hypersensitive sites (DHSs) from a reference annotation of the B cell line GM12878, whereas the regular cfDNA signal oscillates at neighboring genomic locations and is depleted at the DHS (Fig. 1D). Taken together, short cfDNA accumulates at open chromatin or TSS of genes, whereas regular cfDNA, namely, nucleosomes, was clearly depleted in these regions.

Short cfDNA sequencing detects binding of transcription factors

Because short cfDNA could be protected from nuclease digestion by binding to regulatory DBPs, peak regions were examined for encompassing transcription factor binding motifs. Accordingly, a consensus peak set was defined from peaks of short cfDNA sequencing data of four healthy individuals. Transcription factor motif enrichment analysis at the genomic locations of these consensus peaks revealed a significant enrichment of 203 transcription factor binding motifs out of 401 listed in the reference

motif database HOCOMOCO (Kulakovskiy et al. 2018). The 203 transcription factor motifs belong to 46 transcription factor families from nine transcription factor superclasses. ChIP-seq-validated TFBSs from ENCODE3 reveal clear enrichment signals for nuclear factor, erythroid 2 (NFE2), RE1 silencing transcription factor (REST), and Spi-1 proto-oncogene (SPI1) in short cfDNA, for example. On the other hand, regular cfDNA is depleted at these binding sites. Overall, the average profile of all ChIP-seq-based TFBSs shows a clear enrichment of short cfDNA in contrast to regular cfDNA independent of the transcription factor (Fig. 2A). ChIP-seq-validated

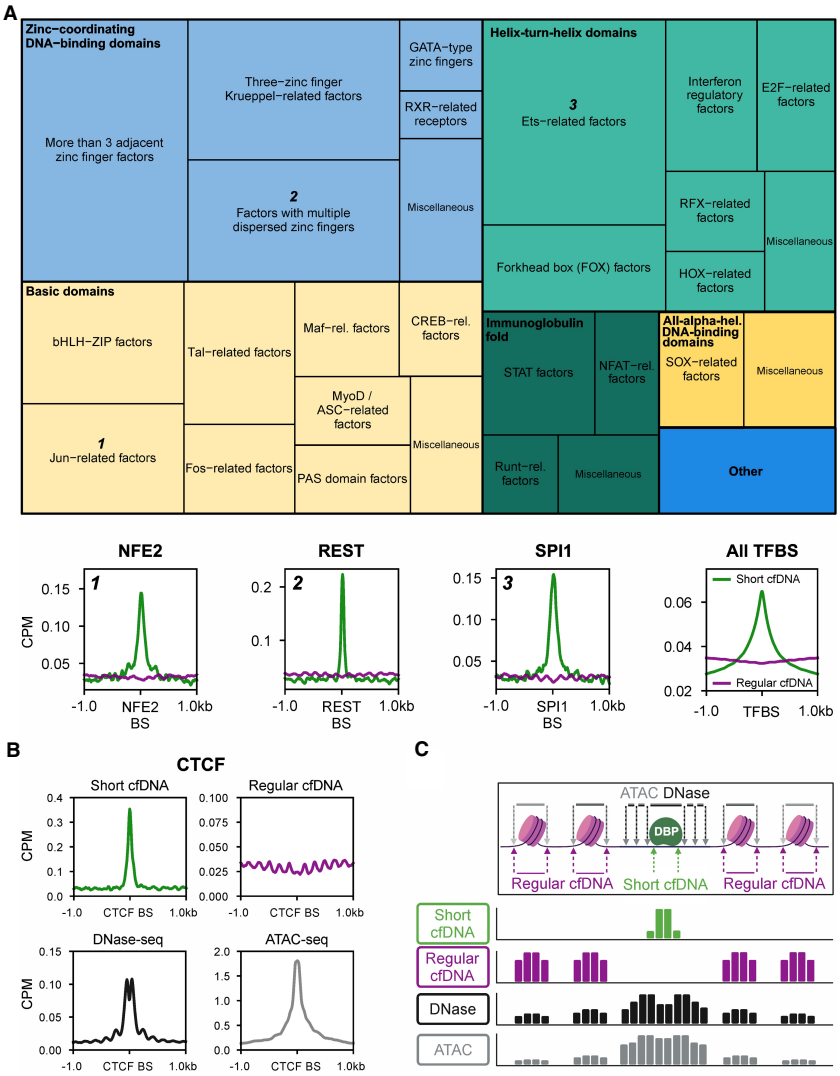


Figure 2. Short cfDNA sequencing directly captures the protection of DNA across the genome by various transcription factors. (*A*) Treemap showing transcription factors whose DNA motif was significantly enriched in short cfDNA consensus peaks from four healthy individuals. More than 200 transcription factor motifs of all nine transcription factor superclasses in the database were identified as enriched in short cfDNA peaks. The size of squares in the treemap encodes the relative number of transcription factor motifs per class (thin outlined boxes) and superclass (thick outlined boxes). As examples, average coverage profiles of short cfDNA (S03) and regular cfDNA (S06) are shown for three transcription factors (NFE2, REST, and SPI1) from three different transcription factor superclasses. Average profiles are based on the 1000 most robust binding sites annotated in the Gene Transcription Regulation Database (GTRD). In addition, the average profile of all ChIP-seq TFBSs annotated in ENCODE3 is shown for short cfDNA in comparison to regular cfDNA. (*B*) Average coverage profile of short cfDNA, regular cfDNA, DNase-seq, and ATAC-seq at CTCF binding sites (CTCF BSs). (C) Inferred molecular model for the formation of short cfDNA at the binding site of a DNA-binding protein (DBP) surrounded by two nucleosomes on either side. Arrows indicate the endpoints of DNA fragments for the respective sequencing technique. The resulting theoretical coverage tracks are depicted for each sequencing technique.

binding sites of CCCTC-binding factor (CTCF) show an even more pronounced signal for short cfDNA, whereas regular cfDNA exhibits a high-frequency oscillation occupancy adjacent to these binding sites (Fig. 2B). In good agreement, DNase-seq data show a combination of a footprint peak with adjacent oscillation patterns, whereas ATAC-seq detects a peak representative for open chromatin at the CTCF binding sites (Fig. 2B). Our data suggest that short cfDNA most specifically reveals protection of DNA through DBPs at regulatory sites with high resolution as exemplified by CTCF, NFE2, REST, or SPI1 (Fig. 2C). Previously, it has been shown that a significant fraction of short cfDNA exists as short single-stranded DNA (ssDNA). To analyze how short ssDNA signals compare to short cfDNA signals, we examined transcription start sites and TFBSs with data from Snyder et al. (2016) for short cfDNA sequencing (single-stranded library preparation). We found a superior signal-to-noise ratio for DBP footprints, allowing the detection of signals, for example, tumor protein p53 binding protein 1 (TP53BP1), that were not detectable in the short ssDNA of Snyder et al. (Snyder et al. 2016; Supplemental Fig. S6). These data suggest that short cfDNA represents a biological entity that captures DBP footprints.

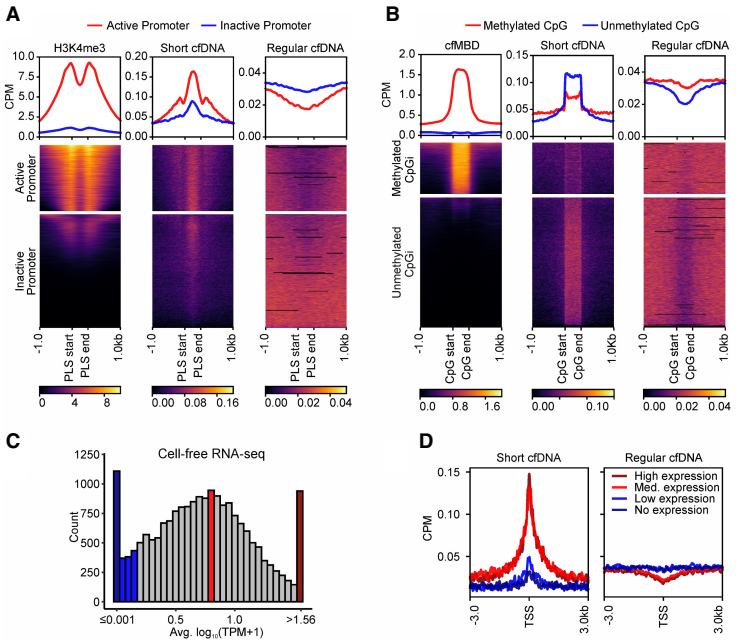


Figure 3. Short cfDNA is enriched at loci with markers of epigenetic activation and transcriptionally active regions. (*A*) Annotated PLSs from ENCODE are clustered based on publicly available cell-free H3K4me3 ChIP-seq signal strength into two clusters, representing active (red) or inactive (blue) promoters. Average coverage profiles for short cfDNA and regular cfDNA at active or inactive promoters demonstrate the influence of the promoter activation status. (*B*) Annotated CpG islands are clustered based on cell-free methyl-CpG-binding domain (cfMBD) sequencing signal strength into two clusters. Average coverage profiles for short cfDNA and regular cfDNA at methylated or unmethylated CpG islands reveal the influence of methylation levels at CpG islands. (*C*) Histogram showing average expression levels of protein-coding genes in publicly available cell-free RNA sequencing data. For each category, 5% of all analyzed genes were selected (938 genes each for no expression [dark blue], low expression [blue], medium expression [red], and high expression [dark red]). (*D*) Average coverage profiles for short cfDNA and regular cfDNA at the transcription start sites of the defined gene groups of *C* reveal the influence of transcriptional activity. All data in *A* and *D* were generated from samples of healthy individuals. S03 was used as the short cfDNA data set, and S06 was used as the regular cfDNA data set.

Correlation to epigenetic activation and gene transcription

Given that short cfDNA fragments are likely derived from the protection of regulatory DBPs and are overrepresented at open chromatin regions, we investigated a potential relationship between short cfDNA signal strength (i.e., local read enrichment) and promoter activation state. Annotated promoter-like structures were classified as active or inactive promoters based on cell-free histone 3 lysine 4 triple-methylation (H3K4me3) signals from publicly available ChIP-seq data of a healthy individual (Sadeh et al. 2021). Active promoters with a strong H3K4me3 signal showed a higher coverage of short cfDNA than did promoters with weak or no H3K4me3 signal, whereas regular cfDNA shows exactly the opposite behavior (Fig. 3A). Moreover, the strongest signals for H3K4me3, representing the nucleosome positions, occur at local minima of short cfDNA. DNA methylation is known to be an essential regulator of gene activity and is associated with transcription factor binding and thus, potentially, DNA footprint signals. Consequently, we classified CpG islands (CpGis) into methylated and unmethylated CpGis based on the signal strength of cell-free methyl-CpG-binding domain sequencing (cfMBD-seq) data from a healthy individual. Strongly methylated CpGis show a weaker accumulation of short cfDNA compared with unmethylated CpGis, whereas regular cfDNA again behaves the opposite (Fig. 3B), demonstrating the relationship between DNA methylation and short cfDNA. To further analyze a connection between localized short cfDNA signals and downstream gene transcription, we used publicly available cell-free RNA-seq data from five healthy individuals (Zhu et al. 2021). Based on the average transcript abundance level of protein-coding genes, four subsets of genes were defined: "no expression," "low expression," "medium expression," and "high expression" genes (Fig. 3C). Genes with medium or high expression show considerable enrichment of short cfDNA reads at their respective TSSs, whereas genes with low or no expression show no substantial enrichment. Regular cfDNA again behaves contrary to short cfDNA and shows a much less pronounced difference between active expression (high and medium) and low, or no, expression (Fig. 3D). Consistent with the definition of these gene subgroups, H3K4me3 histone modifications increase and DNA methylation levels decrease as expression levels increase at the respective TSS of the genes (Supplemental Fig. S7B,C). For a more detailed analysis of the joint effect of DNA footprint signals and DNA methylation in regulatory elements of genes on its transcription, we used matched sequencing data from four septic patients.

Different combinations of short cfDNA and DNA methylation signals also seem to affect the gene expression levels (Supplemental Fig. S8A). Genes with clearly decreased DNA methylation in their proximal CpGis and significantly elevated short cfDNA abundance at their core promoter seem to be associated with higher transcription levels (clusters 2 and 4), whereas genes with increased DNA methylation and reduced short cfDNA abundance seem to be associated with lower transcription levels (clusters 5, 8, and 9). Between these clear cases, the associations of regulatory signals with the resulting transcription levels of the genes are more complex and becoming less clear (Supplemental Fig. S8B). In summary, the signal strength of short cfDNA is higher in active promoters than in inactive promoters, higher in unmethylated CpGis than in methylated CpGis, and higher in TSSs of actively transcribed genes than in TSSs of untranscribed genes. Thus, short cfDNA is enriched at loci with markers for epigenetic activation and transcriptionally active genomic regions.

Disease-specific short cfDNA signatures in liquid biopsies

To identify disease-specific signatures, short cfDNA data were generated for four biological replicates of four different clinical indications: two types of gastrointestinal carcinomas (PDAC and CRC), as well as sepsis and Post-OP controls (Supplemental Table S1). Post-OP patients were selected as controls because this patient group is prone to develop a sepsis after surgery, and therefore, a tight control and separation are of clinical relevance (Chen et al. 2019b; Lukaszewski et al. 2022). Comparison of consensus peak sets for PDAC versus CRC revealed 136 differentially enriched loci (Fig. 4A; Supplemental Fig. S10) and 241 loci with a differential enrichment comparing sepsis with Post-OP (Fig. 4B; Supplemental Fig. S10). Principal component analysis (PCA) based on all differentially enriched regions (DERs) demonstrated a clear separation of all four clinical indications as well as the healthy samples by the first two principal components (Fig. 4C). Two exemplary

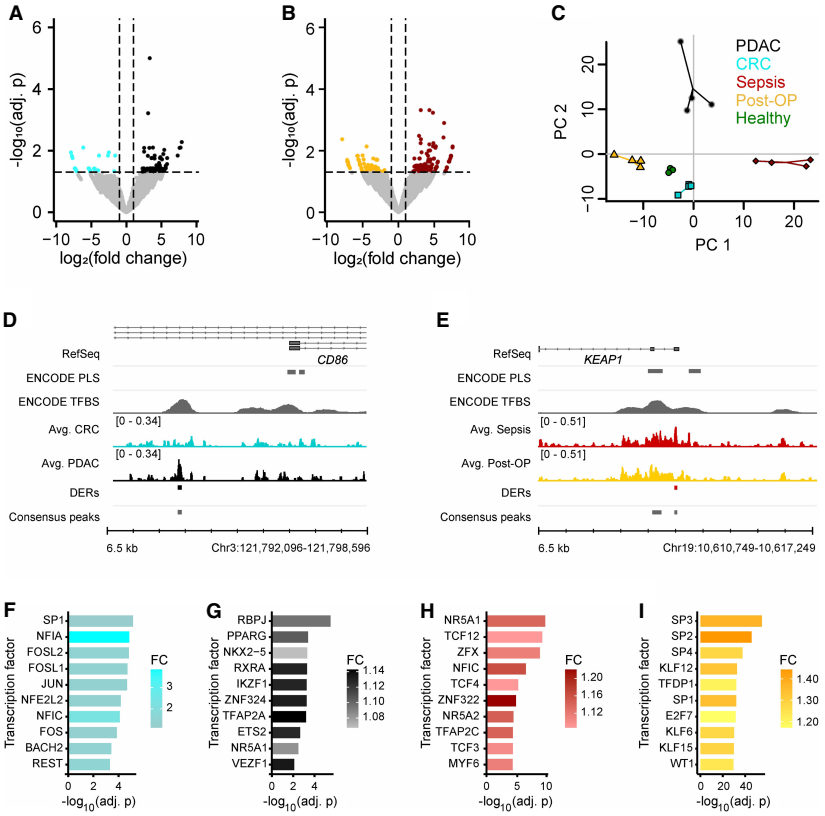


Figure 4. Comparison of short cfDNA data reveals condition-specific protection of TFBSs. (*A*) Differential enrichment analysis for pancreatic ductal adenocarcinoma (PDAC) samples with colorectal cancer (CRC) samples identifies differentially enriched TFBSs (Adj. *P*-value ≤ 0.05 and $|\log_2(FC)| \geq 1$). (*B*) Differential enrichment analysis for sepsis samples with postoperative (Post-OP) samples identifies differentially enriched TFBSs (Adj. *P*-value ≤ 0.05 and $|\log_2(FC)| \geq 1$). (*C*) Principal component analysis based on identified differential TFBSs separates all conditions and healthy individuals. Four biological replicates were used per condition. In addition to the samples, a centroid for each group is depicted as a larger data point. Variance explained: PC1 = 29.6%, PC2 = 20.9%. (*D*) Example for a differentially enriched TFBS in PDAC with novel TF binding near the alternative transcription start site of *CD86*. (*E*) Example for a differentially enriched TFBS in sepsis with novel TF binding in the promoter of *KEAP1*. (*F–I*) Top 10 differentially enriched TF motifs. (*F*) CRC consensus peaks in comparison to PDAC consensus peaks. CTCF and CTCFL were identified as well but were not included in the figure. (*G*) PDAC consensus peaks in comparison to CRC consensus peaks. (*H*) Sepsis consensus peaks in comparison to Post-OP consensus peaks. (*I*) Post-OP consensus peaks in comparison to sepsis consensus peaks. CTCF and CTCFL were identified as well but were not included in the figure.

DERs demonstrate a distinct differential DNA footprint near a TSS in the context of a larger locus with TFBSs (Fig. 4D,E). A differentially enriched TFBS was detected near the alternative transcription start site in the promoter of the *CD86* gene in PDAC patients compared with CRC patients (Fig. 4D), whereas another protected TFBS was detected near the promoter of the kelch-like ECH associated protein 1 (*KEAP1*) in sepsis patients in contrast to Post-OP patients (Fig. 4E). In addition to differential analysis of signal strength at defined consensus peaks, differential enrichment of transcription factor binding motifs was also analyzed in all consensus peaks. For this purpose, the relative abundance of transcription factor motifs in the consensus peaks of one condition was compared with the abundance in the consensus peaks of the respective reference condition. Enrichment of 14 different transcription factor motifs was detected for PDAC in comparison to CRC, 19 for CRC in comparison to PDAC, 14 for sepsis in comparison to Post-OP, and 126 for Post-OP in comparison to sepsis (E-value ≤ 10) (Fig. 4F–I). The binding motif with the strongest enrichment in PDAC patients compared with CRC patients belongs to the recombination signal binding protein for immunoglobulin kappa J region (RBPJ) (Fig. 4F), for example. In this context, one of the physiological roles of the transcription factor RBPJ is the regulation of early pancreatic cell development. Overall, short cfDNA sequencing enables the detection of condition-specific enrichment signatures in liquid biopsies, which can be used to discriminate different diseases for diagnostic purposes. In addition, short cfDNA sequencing might also help identify transcription factors that may have physiological relevance to the condition.

Discussion

In this study, we present a novel approach for comprehensive DNA footprinting in liquid biopsies by analysis of short double-stranded plasma cfDNA. Our short cfDNA sequencing approach comprises preparative gel electrophoresis to specifically enrich cfDNA fragments with a mean length of ~40 bp in combination with high-throughput sequencing. We observed an enrichment of short cfDNA at the TSS of genes, in which regular cfDNA is depleted. A closer inspection of the average signals at the TSS also showed an oscillation pattern reciprocal to that of regular cfDNA at the 5′ UTR. These findings indicate a regular shift of nucleosomes from the promoter to the 5′ UTR of genes and protection by regulatory DBPs between displaced nucleosomes as revealed by short cfDNA enrichment. In line with the enrichment of short cfDNA at TSS, we also found that short cfDNA fragments showed a higher GC content than the human genome on average (short cfDNA = 57.8%, human genome = 40.9%). Gene regulatory elements in humans possess an increased GC content. Consequently, short cfDNA fragments and TFBSs that are enriched in such regions of the genome have elevated GC contents. Short cfDNA also accumulates at DHSs. At these open chromatin locations, regular cfDNA exhibits a combined signal originating from high-frequency and low-frequency nucleosomal oscillation signals, as described by Ulz et al. (2019). Comparable to this finding, we observed two different types of short cfDNA signals created by DBPs with a narrow binding signal, such as CTCF, and multiple adjacent protein binding events that result in a broader signal. The observed narrow signal of regulatory proteins, like CTCF, SPI1, or REST, can be explained by their ability to initiate nucleosome displacement in closed chromatin (pioneer factors) (Fu et al. 2008; Heinz et al. 2010; Barozzi et al. 2014; Vanzan et al. 2021). Pioneer factors can bind directly to closed chromatin without the need or presence of auxiliary proteins. In contrast, the binding sites of regulatory proteins such as MYC proto-oncogene, bHLH transcription factor (MYC), or TP53BP1 exhibit a much broader signal. These regulatory proteins bind in open chromatin, accompanied by the presence of additional regulatory proteins. Consequently, we also found a striking correlation between DNA footprint signals and markers representative for transcriptional regulation. In this context, we determined that promoters with H3K4me3 histone modifications, a characteristic for active gene transcription, revealed stronger short cfDNA signals than those without. Furthermore, we observed an opposing correlation between methylation levels of CpGis and corresponding enrichment of short cfDNA. Because unmethylated CpGis are considered markers of epigenetic activation, depletion of regular cfDNA and the presence of DNA footprinting signals indicate a protection from regulatory activating factors at such loci. Regarding gene expression, short cfDNA signals at the TSS are increased at actively transcribed genes compared with genes with low expression. However, short cfDNA does not strictly capture continuous dynamics of gene expression levels, rather more it exhibits a binary switch between high and low expression. Comparable findings were obtained through the analysis of the joint effect of DNA footprint signals and DNA methylation in regulatory elements of genes on their transcription. Here, the change between low and high gene expression levels can be detected to a certain extent, but not so precisely to exactly quantify gene expression levels. These aberrations are likely owing to the multitude of other epigenetic regulatory mechanisms that fine-tune gene expression levels. Although this holds also true for other regulatory signals of gene expression, such as DNA methylation, this might reflect a potential limitation of our approach to quantitatively predict an individual's physiology exactly. Further targeted studies are needed to assess the extent of this limitation. Nevertheless, short cfDNA is clearly correlated with DNA-methylation signatures, H3K4me3 histone modifications, and transcriptional activity of downstream gene loci. We found consistent enrichment of short cfDNA where regular cfDNA is depleted. Hence, we propose that the molecular origin of short cfDNA most likely is double-stranded cfDNA, which has been protected from enzymatic digestion by regulatory proteins, including transcription factors. In agreement with this model, we found that short cfDNA consensus peaks comprise a plethora of transcription factor motifs and that known TFBSs revealed clear enrichment for short cfDNA. Taken together, our data provide evidence that short cfDNA represents a characteristic subset of cfDNA that is distinct from nucleosomal cfDNA and therefore most likely does not originate from the degradation of regular cfDNA.

In contrast to short cfDNA, which is characterized by short double-stranded DNA (short dsDNA), publications by Snyder et al. (2016), Hisano et al. (2021), and Hudecova et al. (2022) reported on the presence of short ssDNA. Contrary to short dsDNA, short ssDNA fragments may represent a significant fraction of total cfDNA, comprising as much as 20% of a sequencing library. However, short ssDNA demonstrated that the relative signal strength at transcription start sites or TFBSs of proteins, such as MYC, is very low compared with short dsDNA. Therefore, we assume that cfDNA originally protected by regulatory DBPs is short dsDNA, whereas short ssDNA might be of other biological origin with higher signal-to-noise ratios. Hisano et al. (2021) and Hudecova et al. (2022) suggested that much of the short ssDNA could originate from noncanonical DNA structures, such as G4 quadruplexes, which may largely superimpose signals from regulatory protein DNA-binding events.

Transcription factor motifs from short cfDNA revealed sets of differentially enriched transcription factors between different conditions including samples from septic, CRC, and pancreatic cancer (PDAC) patients, suggesting a diagnostic potential for liquid biopsy applications. Sepsis is characterized by a complex interplay of proinflammatory and anti-inflammatory processes orchestrated by regulators of the immune system. Seven out of the top 10 differentially enriched transcription factor motifs identified in septic samples are linked to regulatory pathways of the immune system. For example, transcription factor 4 (TCF4; also known as E2-2 or ITF2) regulates genes for the differentiation of dendritic cells into interferon-producing plasmacytoid dendritic cells (Reizis 2010). TCF4 also regulates the immune response as a downstream target of toll-like receptor 2 (TLR2) signaling to induce the expression of immunoregulatory genes such as interleukin 10 (IL10) (Manoharan et al. 2014). The identified transcription factor zinc finger protein X-linked (ZFX) is known to be involved in the maintenance of peripheral T cells as well as expansion and maintenance during B cell development and peripheral homeostasis (Arenzana et al. 2009; Smith-Raska et al. 2018). For CRC patients, Sp1 transcription factor (SP1) was identified as the top hit, and several FOS/JUN family transcription factors were identified as differentially enriched. SP1 is a ubiquitous transcription factor and mediator of critical physiological pathways, including cell cycle, proliferation, and metastasis. SP1 plays a key role in regulating genes involved in CRC growth and metastasis (Bajpai and Nagaraju 2017). Two members of the FOS/JUN family—FOS-like 1, AP-1 transcription factor subunit (FOSL1), and FOS-like 2, AP-1 transcription factor subunit (FOSL2)—are known to promote tumorigenesis and metastasis in colon cancer (Li et al. 2018; Liu et al. 2021). For PDAC patients, RBPJ was identified as the corresponding top hit. RBPJ is known to form a heterocomplex with pancreas associated transcription factor 1a (PTF1A), and their interaction is required in the early stage of pancreatic growth, morphogenesis, and lineage fate decision (Masui et al. 2007). In addition, peroxisome proliferator-activated receptor gamma (PPARG) and retinoid X receptor alpha (RXRA) were identified, which are known to form a heterocomplex (Lehrke and Lazar 2005). PPARG is a key regulator of adipocyte differentiation, regulates insulin and adipokine production and secretion, and is associated with PDAC. In addition, RXRA promotes proliferation and inhibits apoptosis of pancreatic cancer cells (Chen et al. 2019a). Moreover, we found disease-specific differentially enriched loci that may enable clear separation of patients with PCA. For example, in PDAC patients, we found differential protection from DBPs near the alternative TSS of the *CD86* gene compared with CRC patients. In PDAC, the tumor microenvironment is highly immunosuppressive and characterized by a dense stroma and an abundance of immunosuppressive cells such as myeloid-derived suppressor cells and regulatory T cells, as well as increased levels of monocytes in peripheral blood (Gautam et al. 2023; Hansen et al. 2023). CD86 is an important costimulatory molecule involved in T cell activation. CD86, together with other immune markers such as integrin subunit alpha X (ITGAX; also known as CD11c) and CD274 molecule (CD274; also known as PD-L1), can serve as a prognostic indicator in PDAC, in which increased expression of CD86 on peripheral blood monocytes of PDAC patients significantly correlates with disease severity (Hansen et al. 2023). In septic patients, we found an additional signal of short cfDNA near the TSS of the *KEAP1* gene compared with postoperative controls. KEAP1 regulates oxidative stress and inflammation in sepsis via the NFE2-like bZIP transcription factor 2 (NFE2L2; also known as NRF2) signaling pathway. Inhibition of KEAP1 allows NRF2 to activate antioxidant responses, thereby reducing the damage caused by sepsis. Enhanced NRF2 activity is associated with better outcomes, whereas impaired signaling worsens sepsis in mouse studies. The use of protein–protein interaction (PPI) inhibitors targeting the NRF2-KEAP1 signaling pathway are considered and evaluated in various studies for the treatment of sepsis by increasing NRF2 activity (Gunne et al. 2020; Wang et al. 2023).

Taken together, analysis of short double-stranded cfDNA might provide the most accurate picture of a genome-wide transcription factor footprint in liquid biopsies. With the ability to identify disease-specific TFBS enrichment for patient classification, we see considerable potential in the application of short cfDNA sequencing for liquid biopsy applications.

Many studies explore the properties and information to be inferred from various cfDNA fragments and fragmentation patterns. Often the clinical relevance of the obtained information is not yet well evaluated. Although we started to explore the clinical relevance and potential of short cfDNA sequencing with clinical samples, the results presented in this study are based on a limited number of samples. Further studies with more clinical samples are still required to validate the potential of this approach. In addition to validity and utility, cost-effectiveness is an essential factor for clinical diagnostic methods. Although our workflow already represents a resource-efficient alternative to ultradeep high-throughput sequencing of total cfDNA, a targeted and even-more-economical approach might be advantageous for implementation. With the new short cfDNA sequencing approach, which directly maps the protection across the genome via regulatory proteins, we want to add a new tool to liquid biopsy diagnostics that could improve the detection of cancer or enable clinically relevant differential diagnosis of cancer types.

Methods

Blood samples

This study included blood samples from nine healthy individuals, four individuals with PDAC, four individuals with CRC, four individuals with sepsis, and four individuals that underwent major abdominal surgery (Supplemental Table S1). Blood from the healthy individuals was acquired commercially from Biomex. Septic patients participated in a previously published, prospective observational clinical study that was conducted in the surgical intensive care unit of Heidelberg University Hospital, Germany between November 2013 and January 2015 (S13–S16 and S25–S37; German Clinical Trials Register: DRKS00005463) (Decker et al. 2017). Treatment of these sepsis patients included early-goal directed therapy, elimination of the septic focus, and broad-spectrum antibiotic therapy. Identified pathogens of all sepsis patients and blood cell counts are included in Supplemental Table S2. Patients without cancer (nonseptic controls, i.e., Post-OP) that underwent major abdominal surgery and all individuals with cancer were recruited at the University Hospital of Erlangen with the approval of the local ethics committee and the clinical trial number 180_19 B (S09–S12 and S17–S24). All experiments were performed in accordance with the study protocol approved by the ethics committee.

cfDNA isolation

Plasma was prepared by centrifugation of whole blood for 10 min at 1600*g* and 4°C. Afterward, blood plasma was centrifuged again for 10 min at 16,000*g* and 4°C. Afterward, 1.1 mL of the supernatant was transferred into a fresh 1.5 mL DNA LoBind tube and

stored at −80°C. If necessary, the sample volume was filled up with freshly prepared 1 × phosphate-buffered saline solution. cfDNA was isolated with the QIAsymphony SP DNA preparation system and the QIAsymphony DSP circulating DNA kit (Qiagen) according to the manufacturer's advice. Eluted cfDNA was quantified with the Qubit dsDNA HS assay kit (Thermo Fisher Scientific), and the cfDNA quality was assessed by the fragment analyzer high-sensitivity DNA kit (Agilent).

High-throughput sequencing

Sequencing libraries for regular cfDNA were prepared with the NEBNext Ultra II DNA library prep kit (New England Biolabs) according to the manufacturer's protocol; 0.5 ng isolated cfDNA was used as input. The NEBNext adaptor was diluted 1:25 for all reactions. PCR was performed with 10 PCR cycles and 4 µL primers. Sequencing was performed on a NextSeq 2000 (Illumina) with 100 bp single-end reagent kits.

Sequencing libraries for short double-stranded cfDNA (short cfDNA) were prepared from 3 ng to 15 ng of cfDNA, depending on the clinical condition, using the NEXTFLEX cell free DNA-seq kit (V2; PerkinElmer) according to the manufacturer's advice with one exception: The final library after PCR amplification was eluted in 20 µL nuclease-free water. Library generation was performed with a Biomek FXP workstation (Beckman Coulter). Library quality was assessed by the fragment analyzer high-sensitivity DNA kit (Agilent), and the concentration was measured by the Qubit dsDNA HS assay kit (Thermo Fisher Scientific). Size selection of the cfDNA libraries was performed using a BluePippin instrument (Sage Science). To select the short cfDNA portion in the range of 150–200 bp, samples were applied to a 3% agarose BluePippin cassette according to the manufacturer's protocol. Briefly described, samples were filled up with water to 30 µL and were mixed with 10 µL of supplied internal marker (100 bp to 250 bp). Twenty-three microliters of the eluted, size-selected sample was reamplified with 25 µL of NEXTFLEX PCR master mix 2.0 and 2 µL of 1:10 diluted NEXTFLEX primer mix 2.0 as described in the NEXTFLEX cell free DNA-seq kit (V2; PerkinElmer), step C PCR amplification. Afterward, samples were purified with 1.2 × the volume of AMPureXP beads (Beckman Coulter) according to the manufacturer's advice. Size-selection performance was evaluated by the fragment analyzer high-sensitivity DNA kit (Agilent). If the sample still contained fragments outside the target range of 150 bp to 200 bp, size selection was repeated as previously described, with the exception that the input sample was adjusted to 30 µL with water and the reamplification step was performed with five to eight cycles. After size selection, sequencing was performed on a NextSeq 2000 (Illumina) with 100 bp single-end reagent kits.

Methylation enrichment of cfDNA was performed using the EpiMark methylated DNA enrichment kit (New England Biolabs) according to the manufacturer's instructions. Two reactions with 5–15 ng of cfDNA input each were performed in parallel per sample. Methylated cfDNA was eluted sequentially from both reactions using the same 50 µL of nuclease-free water to increase yield. Sequencing libraries were prepared with the NEBNext Ultra II DNA library prep kit (New England Biolabs) according to the manufacturer's protocol. Fifty microliters of methylated cfDNA from the EpiMark procedure was used as input. The NEBNext adaptor was diluted 1:25 for all reactions. PCR was performed with 17 PCR cycles and 3 µL primers. Sequencing was performed on a NextSeq 2000 (Illumina) with 100 bp single-end reagent kits.

For whole-blood RNA-seq, blood of patients was collected in PAXgene blood RNA tubes (BD Biosciences), incubated at room temperature for 2 h to achieve complete lysis of blood cells, and frozen at −80°C until further processing. Before nucleic acid isolation, tubes were thawed at room temperature for 2 h. Nucleic acid isolation was performed using the QIAcube (Qiagen) and the PAXgene blood miRNA kit according to the manufacturer's protocol to extract gene-encoding mRNAs. Nucleic acids were eluted in 2 × 40 µL buffer BR5. The quantity and quality of the isolated RNA were determined with a Qubit fluorometer (Thermo Fisher Scientific) and a fragment analyzer (Agilent), respectively. Library preparation and sequencing were performed using 250 ng RNA with the ScriptSeq kit v2 (Illumina). Sequencing of the libraries was performed with HiSeq 2500 (Illumina), with 100 bp single-end reagent kits.

Sequencing data processing

After initial quality control of short cfDNA sequencing of raw sequencing reads with FastQC (v0.11.8), the following steps were performed sequentially to remove sequencing artifacts: (1) removal of sequencing adapters, terminal poly(G) sequences (min 10 sequential G's), and quality trimming (BBTools bbduk.sh v38.67); (2) removal of terminal single A nucleotide added during library preparation (BBTools bbduk.sh v38.67); (3) size selection of sequenced fragments allowing read lengths >20 bp and <60 bp (BBTools bbduk.sh v38.67); (4) removal of sequencing reads with low complexity, that is, dust scores smaller than seven (prinseq-lite v0.20.4) (Andrews 2010; Schmieder and Edwards 2011; Bushnell et al. 2017); (5) processed reads mapped to the human reference genome assembly GRCh37 using NextGenMap (v0.5.5) with default settings (Sedlazeck et al. 2013); and (6) mapped reads deduplicated with SAMtools rmdup (v1.9), reads in blacklisted regions removed with BEDTools intersect (v2.30.0), and reads with a MAPQ value lower than one removed with SAMtools view (v1.9) (Li et al. 2009; Quinlan and Hall 2010; Amemiya et al. 2019). These reads were converted to bigWig format for visualization and other downstream analyses with deepTools (bin size = 10, normalization = counts per million [CPM], bamCoverage v3.5.1) (Ramírez et al. 2016). This workflow is graphically summarized in Supplemental Figure S3. Realigning the reads to GRCh38 or newer assemblies would not significantly affect the results and conclusions, as only Chromosomes 1–22, X, and Y were analyzed, which are highly conserved across different genome versions. Using the ENCODE blacklist with GRCh37 helps to avoid potential sequencing artifacts in poorly mappable regions of these chromosomes, such as centromeres and telomeres, by excluding such problematic regions.

After initial quality control of regular cfDNA raw sequencing reads with FastQC (v0.11.8), the following steps were performed (Andrews 2010): (1) removal of sequencing adapters, removal of terminal poly(G) sequences (min 10 sequential G's), removal of reads <50 bp, and quality trimming (BBTools bbduk.sh v38.67) (Bushnell et al. 2017); (2) processed reads mapped to the human reference genome assembly GRCh37 using NextGenMap (v0.5.5) with default settings (Sedlazeck et al. 2013); and (3) mapped reads deduplicated with SAMtools rmdup (v1.9), and reads in blacklisted regions removed with BEDTools intersect (v2.30.0) (Li et al. 2009; Quinlan and Hall 2010). Mapped reads were converted to bigWig format for visualization and other downstream analyses with deepTools (bin size = 10, normalization = CPM, bamCoverage v3.5.1) (Ramírez et al. 2016).

cfMBD-seq data were processed as regular cfDNA sequencing data.

Public sequencing data processing

FASTQ files of five publicly available cell-free RNA sequencing data sets of different healthy individuals from Zhu et al. (2021) were obtained from the NCBI Sequence Read Archive (SRA; https://www.ncbi.nlm.nih.gov/sra). Accession numbers and unique identifiers

are listed in Supplemental Table S3. All samples were sequenced in paired-end mode with a read length of 150 bp and about 10 million reads per sample on average. After initial quality control of raw sequencing reads with FastQC (v0.11.8), the following steps were performed to obtain read counts for individual genes (Andrews 2010): (1) removal of sequencing adapters, quality trimming, and removal of reads <100 bp (BBTools bbduk.sh v38.67) (Bushnell et al. 2017); (2) mapping of processed reads to the human reference genome assembly GRCh37 using NextGenMap (v0.5.5) with default settings, keeping only reads with a MAPQ value greater than two (Sedlazeck et al. 2013); (3) gene quantification in raw read counts using featureCounts (v2.0.1) with the UCSC Genes annotation (available at https://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/genes/hg19.knownGene.gtf.gz) (Liao et al. 2014), and only protein-coding genes on the autosomal chromosomes included in the final readcount matrix to reduce confounding by the gender of sample donors; and (4) if required, raw read counts converted to transcripts per million (TPM) using R (R Core Team 2024).

FASTQ files of cfDNA sequencing data sets of a healthy individual from Snyder et al. (2016) were obtained from the SRA. Accession numbers and unique identifiers are listed in Supplemental Table S3. After initial quality control of raw sequencing reads with FastQC (v0.11.8), the following steps were performed: (1) removal of sequencing adapters and quality trimming (cutadapt v4.0) (Martin 2011); (2) mapping of trimmed reads to the human reference genome assembly GRCh37 using NextGenMap (v0.5.5) with minimum map quality of 10 (Sedlazeck et al. 2013); (3) in silico size selection of the mapped reads with 35–80 nt for short cfDNA and 120–180 nt for regular cfDNA; (4) removal of low complexity reads and deduplication (prinseq-lite v0.20.4) (Schmieder and Edwards 2011); (5) removal of reads in blacklisted regions with BEDTools intersect (v2.30.0) (Quinlan and Hall 2010); and (6) for downstream analysis and visualization, bigWig files generated with deepTools (bin size = 10, normalization = CPM, bamCoverage v3.5.1) (Ramírez et al. 2016).

ATAC-seq and DNase-seq data from ENCODE (see Supplemental Table S3) were retrieved as processed BAM files and only converted to other data types, like bigWig for downstream analysis. Cell-free H3K4me3 ChIP-seq data from Sadeh et al. (2021) (see Supplemental Table S3) were retrieved as processed BED files and only converted to other data types, like BAM and bigWig for downstream analysis.

Peak calling and NFR calling

For short cfDNA sequencing data, narrow peaks and clustered peaks were called with MACS2 callpeak (narrow: --nomodel --extsize 32 --call-summits --min-length 30 -q 0.05, clustered: --broad --nomodel --extsize 32 --max-gap 100 --min-length 500 --broad-cutoff 0.05) (Zhang et al. 2008). Consensus peaks of a condition were identified using R and defined as genomic sites where a narrow peak was identified in at least three of four samples (R Core Team 2024). In addition, consensus peaks <31 nt apart were combined. For regular cfDNA, NFRs were called on the merged data set of four biological replicates to increase genome-wide coverage and, thus, reliability. NFRs were identified with the R packages NucDyn, utilizing nucleR, with default settings (Buitrago et al. 2019). Peaks and NFRs were annotated to genomic features in R with the ChIPseeker package and default settings (Yu et al. 2015).

Average coverage profiles and heatmaps

Average coverage profiles and heatmaps were created from bigWig files with deepTools (computeMatrix, plotHeatmap, and plotProfile; v3.5.1) (Ramírez et al. 2016). Different genomic reference locations were used for average coverage profiles. For Figure 2 and Supplemental Figure S2, validated binding sites of transcription factors were retrieved from the Gene Transcription Regulation Database (GTRD) (Kolmykov et al. 2021). Here, TFBSs were ranked according to their robustness across different cell lines and tissues, namely, the number of tissues and cell lines in which the respective binding site was found. Only the top 1000 binding sites were used for average coverage profiles. Promoter-like structures from ENCODE were converted from GRCh38 to GRCh37 with the liftOver tool from UCSC and used as reference regions in Figure 3A (available at https://www.encodeproject.org/files/ENCFF379UDA/) (Hinrichs et al. 2006). CpGis were used as reference regions in Figure 3B (available at UCSC TableBrowser, track name = cpgIslandExt) (Karolchik et al. 2004). The heatmaps in Figure 3, A and B, were clustered in two groups using the implemented *k*-means clustering based on the line-wise average. Missing data in heatmaps were plotted in black.

Composite effect of DNA methylation and DNA footprint signals on gene transcription

From four sepsis patients, matched short cfDNA, cfMBD, and whole-blood RNA sequencing data were acquired (S26–S37). For a combined analysis, different genomic annotations were combined. The same gene annotation as used with the cfRNA sequencing data from Zhu et al. (2021) was utilized. A core promoter annotation for GRCh37 was retrieved from the eukaryotic promoter database (EPD; version 6; available at https://epd.expasy.org/ftp/epdnew/H_sapiens/006/Hs_EPDnew_006_hg19.bb) (Dreos et al. 2015; Meylan et al. 2020). For each gene, only the best-matching core promoters were kept ("_1" flag). As for CpGis, the aforementioned CpGis from UCSC were used again. CpGis were assigned to genes based on their proximity to the EPD promoter of the gene in the genome, with a maximum absolute distance of 5 kb. With these assignments, from the roughly 20,000 protein-coding genes in the annotation, about 12,000 remained with an associated core promoter and CpGis. Counts of sequencing reads of each sequencing data type and each patient were retrieved for the curated gene bodies, core promoters, and CpGis with featureCounts (v2.0.1) and converted to TPM (Liao et al. 2014). From the patient replicates a mean TPM value was calculated per region and sequencing type. Additionally, a pseudocount of 0.01 was added. For better comparability of the signals, the cfMBD-seq data in CpGis and short cfDNA sequencing data in core promoters were additionally adjusted for GC content. The GC content of the respective regions was extracted from GRCh37 using BEDTools nuc (v2.30.0) (Quinlan and Hall 2010). The adjusted signal was calculated as follows: A locally weighted regression function was fitted with the mean readcount value and GC content for each annotation type (R lowess function). Each mean readcount value (observed) is divided by its fitted value (expected) to obtain an observed over expected (OoE) ratio. For better interpretability and visualization, the values are also $\log_2$-scaled. Genes with comparable composite signals in the core promoter and in the CpGis were identified by *k*-means clustering ($k = 10$) and visualized in a heatmap using R and ComplexHeatmap (Fig. 3E; Gu et al. 2016). The average gene expression values were added as an annotation but had no influence on clustering. For better visualization and comparison of the distributions of gene expression values, the RNA-seq data were also visualized in a ridge plot in R with ggplot2, according to the clusters from the heatmap, and sorted by descending median value (Fig. 3F). Clusters with fewer than 200 genes were excluded from this visualization as the sample size was considered too small for a representative distribution.

Transcription factor motif enrichment analysis

Enriched transcription factor motifs were identified with the AME tool from MEME suite (--scoring avg --method fisher; v5.4.1) (McLeay and Bailey 2010). Input DNA sequences were retrieved from consensus peak sets and analyzed for the enrichment of motifs listed in the HOCOMOCOv11_core_HUMAN_mono database (Kulakovskiy et al. 2018). DNA sequences of consensus peaks from short cfDNA sequencing data of healthy individuals were compared with control sequences, generated by shuffling the letters in the input while preserving the frequencies of *k*-mers (--shuffle). The proportions of identified transcription factor motif classes and their respective superclasses were summarized in a treemap plot using R and the treemap package (v.2.4.3); https://cran.r-project.org/web/packages/treemap/index.html). Short cfDNA sequencing data from other than healthy states were compared with each other. The DNA sequences underlying each consensus peak set were used as control sequences, for example, consensus peak DNA sequences from CRC as input compared with consensus peak DNA sequences from PDAC as control.

Differential enrichment analysis

Identification of DERs was performed with the R package DEBrowser (v1.2.0) using the implemented edgeR method with raw read counts in the combined consensus peak sets of two compared conditions, TMM normalization, a glmLRT, and dispersion = 0 (Robinson et al. 2010; Kucukural et al. 2019). The use of this analysis is based on the observation that the read counts in the consensus peak sets are best modeled with a negative binomial distribution in comparison to a Poisson of geometric distribution (Supplemental Fig. S9). Genes were considered as differentially expressed between two conditions (four biological replicates per condition) with an adjusted *P*-value smaller than 0.05 and a $\log_2$(fold change) ≤ -1 or ≥ 1. Volcano plots for the differential enrichment analysis were generated with the R package EnhancedVolcano (v1.8.0). The identified DERs of both comparisons were used for a PCA, and the first three principal components were visualized in R with pca3d (v0.10.2). Samples of each condition were linked to the centroid of the respective condition with a line.

Data access

The raw high-throughput sequencing data generated in this study have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession number PRJNA1033613. Processed high-throughput sequencing data of samples S01–S33 are available as genome coverage tracks in a UCSC Genome Browser session (https://genome.ucsc.edu/s/jnmllr/Short_cfDNA_seq_manuscript) or as files at Figshare (https://doi.org/10.6084/m9.figshare.25211525.v2). Custom data analysis code created for this work is available on GitHub (https://github.com/janmueller17/short_ds_cfDNA) and as Supplemental Code.

Competing interest statement

K.S. is a cofounder of Noscendo, a diagnostic company dedicated for detection of pathogens based on high-throughput sequencing. K.S., C.H., M.S., and J.M. are inventors of a patent application in bioinformatics algorithms for analyzing cfDNA fragmentomics.

Acknowledgments

We thank Drs. Jens and Susanne Freitag for their valuable support. This work was supported by internal funding from Fraunhofer-Gesellschaft.

Author contributions: J.M. and K.S. conceptualized and designed the study. C.H., M.S., L.B., and K.G. processed blood samples, isolated cfDNA, and established the size-selection method for short double-stranded cfDNA. J.M. performed bioinformatics analysis. C.A. analyzed the effect of library preparation methods on short and regular cfDNA in public data. J.M., K.S., Y.V., and A.v.H. analyzed and interpreted the sequencing data. G.F.W., S.O.D., and T.B. recruited patients and collected samples. J.M., K.S., and M.S. wrote the manuscript. All authors read and approved the final manuscript.

References

Amemiya H, Kundaje A, Boyle A. 2019. The ENCODE blacklist: identification of problematic regions of the genome. *Sci Rep* **9:** 9354. doi:10.1038/s41598-019-45839-z

Andrews S. 2010. *FASTQC. A quality control tool for high throughput sequence data.* Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom.

Arenzana T, Smith-Raska M, Reizis B. 2009. Transcription factor Zfx controls BCR-induced proliferation and survival of B lymphocytes. *Blood* **113:** 5857–5867. doi:10.1182/blood-2008-11-188888

Bajpai R, Nagaraju G. 2017. Specificity protein 1: its role in colorectal cancer progression and metastasis. *Crit Rev Oncol Hematol* **113:** 1–7. doi:10.1016/j.critrevonc.2017.02.024

Barozzi I, Simonatto M, Bonifacio S, Yang L, Rohs R, Ghisletti S, Natoli G. 2014. Coregulation of transcription factor binding and nucleosome occupancy through DNA features of mammalian enhancers. *Mol Cell* **54:** 844–857. doi:10.1016/j.molcel.2014.04.006

Bronkhorst A, Ungerer V, Holdenrieder S. 2019. The emerging role of cell-free DNA as a molecular marker for cancer management. *Biomol Detect Quantif* **17:** 100087. doi:10.1016/j.bdq.2019.100087

Buitrago D, Codó L, Illa R, de Jorge P, Battistini F, Flores O, Bayarri G, Royo R, Del Pino M, Heath S, et al. 2019. Nucleosome dynamics: a new tool for the dynamic analysis of nucleosome positioning. *Nucleic Acids Res* **47:** 9511–9523. doi:10.1093/nar/gkz759

Burnham P, Kim M, Agbor-Enoh S, Luikart H, Valantine H, Khush K, 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. doi:10.1038/srep27859

Bushnell B, Rood J, Singer E. 2017. BBMerge: accurate paired shotgun read merging via overlap. *PLoS One* **12:** e0185056. doi:10.1371/journal.pone.0185056

Chen G, Hu M, Wang X-C, Du S-H, Cheng Z-X, Xu J, Qu Y-K. 2019a. Effects of RXRα on proliferation and apoptosis of pancreatic cancer cells through TGF-β/Smad signaling pathway. *Eur Rev Med Pharmacol Sci* **23:** 4723–4729. doi:10.26355/eurrev_201906_18053

Chen P-Y, Luo C-W, Chen M-H, Yang M-L, Kuan Y-H. 2019b. Epidemiological characteristics of postoperative sepsis. *Open Med (War)* **14:** 928–938. doi:10.1515/med-2019-0110

Decker S, Sigl A, Grumaz C, Stevens P, Vainshtein Y, Zimmermann S, Weigand M, Hofer S, Sohn K, Brenner T. 2017. Immune-response patterns and next generation sequencing diagnostics for the detection of mycoses in patients with septic shock—results of a combined clinical and experimental investigation. *Int J Mol Sci* **18:** 1796. doi:10.3390/ijms18081796

Dreos R, Ambrosini G, Périer R, Bucher P. 2015. The Eukaryotic Promoter Database: expansion of EPDnew and new promoter analysis tools. *Nucleic Acids Res* **43:** D92–D96. doi:10.1093/nar/gku1111

Fan H, Blumenfeld Y, Chitkara U, Hudgins L, Quake S. 2008. Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. *Proc Natl Acad Sci USA* **105:** 16266–16271. doi:10.1073/pnas.0808319105

Fu Y, Sinha M, Peterson C, Weng Z. 2008. The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. *PLoS Genet* **4:** e1000138. doi:10.1371/journal.pgen.1000138

Gautam S, Batra S, Jain M. 2023. Molecular and metabolic regulation of immunosuppression in metastatic pancreatic ductal adenocarcinoma. *Mol Cancer* **22:** 118. doi:10.1186/s12943-023-01813-y

Grabuschnig S, Bronkhorst A, Holdenrieder S, Rosales Rodriguez I, Schliep K, Schwendenwein D, Ungerer V, Sensen C. 2020. Putative origins of cell-free DNA in humans: a review of active and passive nucleic acid release mechanisms. *Int J Mol Sci* **21:** 8062. doi:10.3390/ijms21218062

Gu Z, Eils R, Schlesner M. 2016. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32:** 2847–2849. doi:10.1093/bioinformatics/btw313

Gunne S, Heinicke U, Parnham M, Laux V, Zacharowski K, von Knethen A. 2020. Nrf2-A molecular target for sepsis patients in critical care. *Biomolecules* **10:** 1688. doi:10.3390/biom10121688

Han D, Ni M, Chan R, Chan V, Lui K, Chiu R, Lo Y. 2020. The biology of cell-free DNA fragmentation and the roles of DNASE1, DNASE1L3, and DFFB. *Am J Hum Genet* **106:** 202–214. doi:10.1016/j.ajhg.2020.01.008

Hansen F, David P, Akram M, Knoedler S, Mittelstädt A, Merkel S, Podolska M, Swierzy I, Roßdeutsch L, Klösch B, et al. 2023. Circulating monocytes serve as novel prognostic biomarker in pancreatic ductal adenocarcinoma patients. *Cancers (Basel)* **15:** 363. doi:10.3390/cancers15020363

Heinz S, Benner C, Spann N, Bertolino E, Lin Y, Laslo P, Cheng J, Murre C, Singh H, Glass C. 2010. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38:** 576–589. doi:10.1016/j.molcel.2010.05.004

Heitzer E, Auinger L, Speicher M. 2020. Cell-Free DNA and apoptosis: how dead cells inform about the living. *Trends Mol Med* **26:** 519–528. doi:10.1016/j.molmed.2020.01.012

Hinrichs A, Karolchik D, Baertsch R, Barber G, Bejerano G, Clawson H, Diekhans M, Furey T, Harte R, Hsu F, et al. 2006. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res* **34:** D590–D598. doi:10.1093/nar/gkj144

Hisano O, Ito T, Miura F. 2021. Short single-stranded DNAs with putative non-canonical structures comprise a new class of plasma cell-free DNA. *BMC Biol* **19:** 225. doi:10.1186/s12915-021-01160-8

Hudecova I, Smith C, Hänsel-Hertsch R, Chilamakuri C, Morris J, Vijayaraghavan A, Heider K, Chandrananda D, Cooper W, Gale D, et al. 2022. Characteristics, origin, and potential for cancer diagnostics of ultrashort plasma cell-free DNA. *Genome Res* **32:** 215–227. doi:10.1101/gr.275691.121

Karolchik D, Hinrichs A, Furey T, Roskin K, Sugnet C, Haussler D, Kent W. 2004. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* **32:** D493–D496. doi:10.1093/nar/gkh103

Kolmykov S, Yevshin I, Kulyashov M, Sharipov R, Kondrakhin Y, Makeev V, Kulakovskiy I, Kel A, Kolpakov F. 2021. GTRD: an integrated view of transcription regulation. *Nucleic Acids Res* **49:** D104–D111. doi:10.1093/nar/gkaa1057

Kucukural A, Yukselen O, Ozata D, Moore M, Garber M. 2019. DEBrowser: interactive differential expression analysis and visualization tool for count data. *BMC Genomics* **20:** 6. doi:10.1186/s12864-018-5362-x

Kulakovskiy I, Vorontsov I, Yevshin I, Sharipov R, Fedorova A, Rumynskiy E, Medvedeva Y, Magana-Mora A, Bajic V, Papatsenko D, et al. 2018. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res* **46:** D252–D259. doi:10.1093/nar/gkx1106

Kustanovich A, Schwartz R, Peretz T, Grinshpun A. 2019. Life and death of circulating cell-free DNA. *Cancer Biol Ther* **20:** 1057–1067. doi:10.1080/15384047.2019.1598759

Lehrke M, Lazar M. 2005. The many faces of PPARgamma. *Cell* **123:** 993–999. doi:10.1016/j.cell.2005.11.026

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25:** 2078–2079. doi:10.1093/bioinformatics/btp352

Li S, Fang X-D, Wang X-Y, Fei B-Y. 2018. Fos-like antigen 2 (FOSL2) promotes metastasis in colon cancer. *Exp Cell Res* **373:** 57–61. doi:10.1016/j.yexcr.2018.08.016

Liao Y, Smyth G, Shi W. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30:** 923–930. doi:10.1093/bioinformatics/btt656

Liu Y, Yue M, Li Z. 2021. FOSL1 promotes tumorigenesis in colorectal carcinoma by mediating the FBXL2/Wnt/β-catenin axis via Smurf1. *Pharmacol Res* **165:** 105405. doi:10.1016/j.phrs.2020.105405

Lo Y, Han D, Jiang P, Chiu R. 2021. Epigenetics, fragmentomics, and topology of cell-free DNA in liquid biopsies. *Science* **372:** eaaw3616. doi:10.1126/science.aaw3616

Lukaszewski R, Jones H, Gersuk V, Russell P, Simpson A, Brealey D, Walker J, Thomas M, Whitehouse T, Ostermann M, et al. 2022. Presymptomatic diagnosis of postoperative infection and sepsis using gene expression signatures. *Intensive Care Med* **48:** 1133–1143. doi:10.1007/s00134-022-06769-z

Manoharan I, Hong Y, Suryawanshi A, Angus-Hill M, Sun Z, Mellor A, Munn D, Manicassamy S. 2014. TLR2-dependent activation of β-catenin pathway in dendritic cells induces regulatory responses and attenuates autoimmune inflammation. *J Immunol* **193:** 4203–4213. doi:10.4049/jimmunol.1400614

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17:** 10–12. doi:10.14806/ej.17.1.200

Masui T, Long Q, Beres T, Magnuson M, MacDonald R. 2007. Early pancreatic development requires the vertebrate suppressor of hairless (RBPJ) in the PTF1 bHLH complex. *Genes Dev* **21:** 2629–2643. doi:10.1101/gad.1575207

McLeay R, Bailey T. 2010. Motif enrichment analysis: a unified framework and an evaluation on ChIP data. *BMC Bioinformatics* **11:** 165. doi:10.1186/1471-2105-11-165

Meylan P, Dreos R, Ambrosini G, Groux R, Bucher P. 2020. EPD in 2020: enhanced data visualization and extension to ncRNA promoters. *Nucleic Acids Res* **48:** D65–D69. doi:10.1093/nar/gkz1014

Piovesan A, Pelleri M, Antonaros F, Strippoli P, Caracausi M, Vitale L. 2019. On the length, weight and GC content of the human genome. *BMC Res Notes* **12:** 106. doi:10.1186/s13104-019-4137-z

Quinlan A, Hall I. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26:** 841–842. doi:10.1093/bioinformatics/btq033

Ramírez F, Ryan D, Grüning B, Bhardwaj V, Kilpert F, Richter A, Heyne S, Dündar F, Manke T. 2016. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* **44:** W160–W165. doi:10.1093/nar/gkw257

R Core Team. 2024. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/.

Reizis B. 2010. Regulation of plasmacytoid dendritic cell development. *Curr Opin Immunol* **22:** 206–211. doi:10.1016/j.coi.2010.01.005

Robinson M, McCarthy D, Smyth G. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26:** 139–140. doi:10.1093/bioinformatics/btp616

Sadeh R, Sharkia I, Fialkoff G, Rahat A, Gutin J, Chappleboim A, Nitzan M, Fox-Fisher I, Neiman D, Meler G, et al. 2021. ChIP-seq of plasma cell-free nucleosomes identifies gene expression programs of the cells of origin. *Nat Biotechnol* **39:** 586–598. doi:10.1038/s41587-020-00775-6

Schmieder R, Edwards R. 2011. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27:** 863–864. doi:10.1093/bioinformatics/btr026

Sedlazeck F, Rescheneder P, von Haeseler A. 2013. NextGenMap: fast and accurate read mapping in highly polymorphic genomes. *Bioinformatics* **29:** 2790–2791. doi:10.1093/bioinformatics/btt468

Smith-Raska M, Arenzana T, D'Cruz L, Khodadadi-Jamayran A, Tsirigos A, Goldrath A, Reizis B. 2018. The transcription factor Zfx regulates peripheral T cell self-renewal and proliferation. *Front Immunol* **9:** 1482. doi:10.3389/fimmu.2018.01482

Snyder M, Kircher M, Hill A, Daza R, Shendure J. 2016. Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues-of-origin. *Cell* **164:** 57–68. doi:10.1016/j.cell.2015.11.050

Ulz P, Perakis S, Zhou Q, Moser T, Belic J, Lazzeri I, Wölfler A, Zebisch A, Gerger A, Pristauz G, et al. 2019. Inference of transcription factor binding from cell-free DNA enables tumor subtype prediction and early detection. *Nat Commun* **10:** 4666. doi:10.1038/s41467-019-12714-4

Vanzan L, Soldati H, Ythier V, Anand S, Braun S, Francis N, Murr R. 2021. High throughput screening identifies SOX2 as a super pioneer factor that inhibits DNA methylation maintenance at its binding sites. *Nat Commun* **12:** 3337. doi:10.1038/s41467-021-23630-x

Wang Y, Tang B, Li H, Zheng J, Zhang C, Yang Z, Tan X, Luo P, Ma L, Wang Y, et al. 2023. A small-molecule inhibitor of Keap1–Nrf2 interaction attenuates sepsis by selectively augmenting the antibacterial defence of macrophages at infection sites. *EBioMedicine* **90:** 104480. doi:10.1016/j.ebiom.2023.104480

Yu G, Wang L-G, He Q-Y. 2015. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* **31:** 2382–2383. doi:10.1093/bioinformatics/btv145

Zhang Y, Liu T, Meyer C, Eeckhoute J, Johnson D, Bernstein B, Nusbaum C, Myers R, Brown M, Li W, et al. 2008. Model-based Analysis of ChIP-Seq (MACS). *Genome Biol* **9:** R137. doi:10.1186/gb-2008-9-9-r137

Zhu Y, Wang S, Xi X, Zhang M, Liu X, Tang W, Cai P, Xing S, Bao P, Jin Y, et al. 2021. Integrative analysis of long extracellular RNAs reveals a detection panel of noncoding RNAs for liver cancer. *Theranostics* **11:** 181–193. doi:10.7150/thno.48206

Received March 15, 2024; accepted in revised form August 12, 2024.

SUPPLEMENTARY MATERIALS

TITLE

A novel approach for *in vivo* DNA footprinting using short double-stranded cell-free DNA from plasma

AUTHORS

Jan Müller[1,2,3,4], Christina Hartwig[1,5], Mirko Sonntag[1,6], Lisa Bitzer[1], Christopher Adelmann[1], Yevhen Vainshtein[1], Karolina Glanz[1], Sebastian O. Decker[7], Thorsten Brenner[8], Georg F. Weber[9,10], Arndt von Haeseler[11,12] and Kai Sohn[1,*]

[1] Innovation Field In-vitro Diagnostics, Fraunhofer Institute for Interfacial Engineering and Biotechnology IGB, Stuttgart, Germany

[2] Max Perutz Labs, Vienna Biocenter Campus (VBC), Vienna, Austria

[3] University of Vienna, Max Perutz Labs, Department of Structural and Computational Biology, CIBIV, Vienna, Austria

[4] Vienna BioCenter PhD Program, Doctoral School of the University of Vienna and Medical University of Vienna, Vienna, Austria

[5] Institute for Interfacial Engineering and Plasma Technology (IGVP), University of Stuttgart, Stuttgart, Germany

[6] Interfaculty Graduate School of Infection Biology and Microbiology (IGIM), Eberhard Karls University Tübingen, Tübingen, Germany

[7] Heidelberg University, Medical Faculty Heidelberg, Department of Anesthesiology, Heidelberg, Germany

[8] Department of Anesthesiology and Intensive Care Medicine, University Hospital Essen, University Duisburg-Essen, Essen, Germany

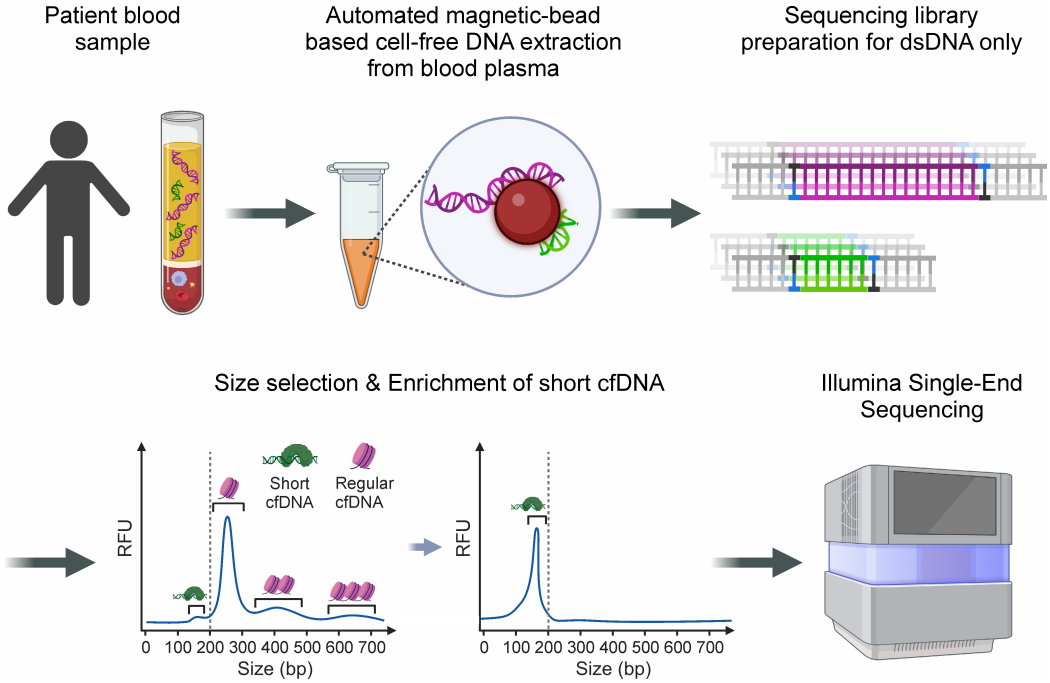
[9] Department of Surgery, Friedrich-Alexander University (FAU) Erlangen-Nürnberg and Universitätsklinikum Erlangen, Erlangen, Germany

[10] Comprehensive Cancer Center (CCC) Erlangen-EMN, Friedrich-Alexander University (FAU) Erlangen-Nürnberg and Universitätsklinikum Erlangen, Erlangen, Germany

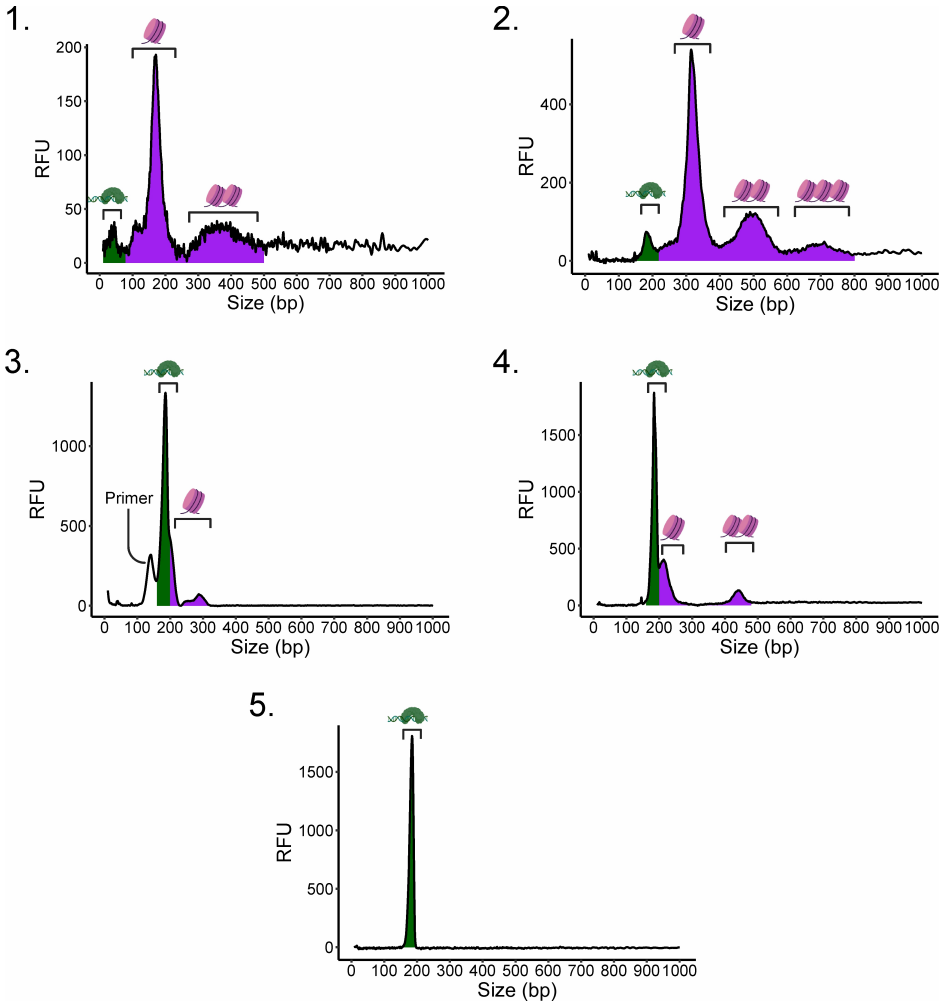
[11] Center of Integrative Bioinformatics Vienna (CIBIV), Max Perutz Labs, University of Vienna and Medical University of Vienna, Vienna BioCenter (VBC), Vienna, Austria

[12] University of Vienna, Faculty of Computer Science Bioinformatics and Computational Biology, Vienna, Austria

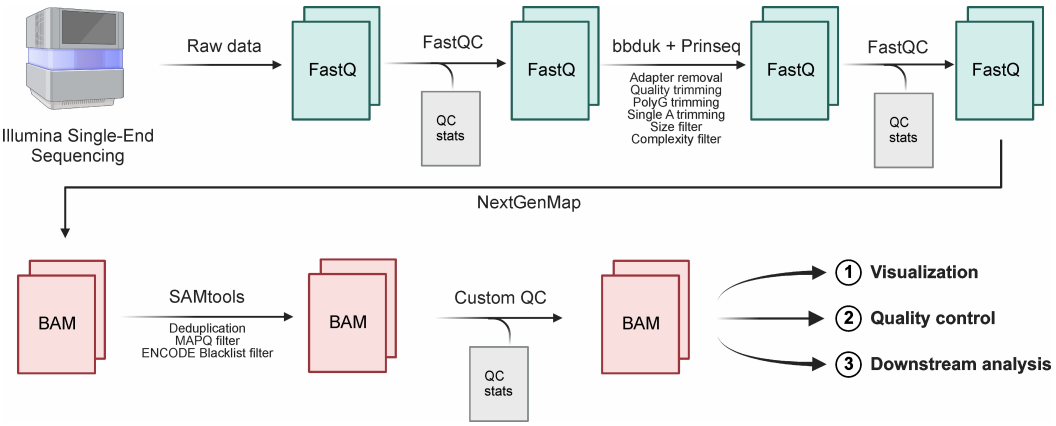
* To whom correspondence should be addressed. Tel: +49 711 970-4055; Email: kai.sohn@igb.fraunhofer.de



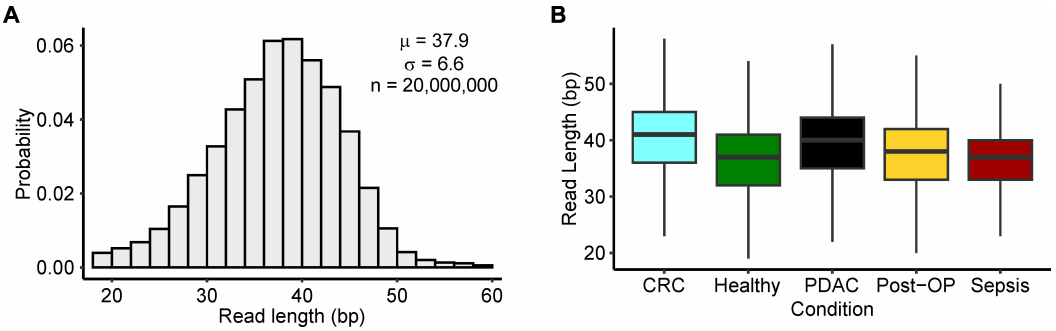
Supplementary Figure S1. Visual summary of short cfDNA extraction and sequencing from blood plasma. Cell-free DNA was extracted from the blood plasma of patients using an automated magnetic bead-based kit. Sequencing library preparation was performed by fragment end-repair and adapter ligation. Only intact double-stranded DNA fragments are enriched in the final sequencing library because of the PCR amplification. Short cfDNA is enriched by size selection from sequencing libraries. Short cfDNA sequencing libraries are sequenced on an Illumina platform in single-end mode. This figure was created with biorender.



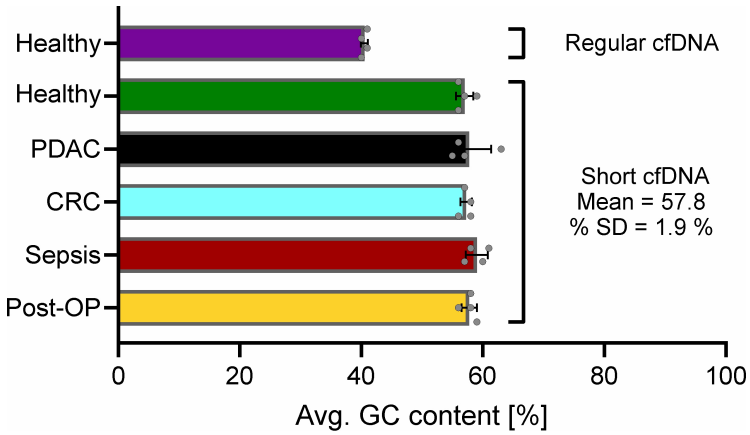
Supplementary Figure S2. Depiction of the full short cfDNA enrichment process from cell-free DNA for high-throughput sequencing. The steps shown correspond to: 1. Isolated cell-free DNA. 2. Double-stranded DNA sequencing library. 3. Size-selected sequencing library. 4. PCR-amplified sequencing library. 5. Sequencing library after second size selection. Size selection was performed using an automated preparative gel electrophoresis instrument. For each step, the Fragment Analyzer profiles of S19 are shown. The library preparation adds around 100 bps to the DNA fragments resulting in the fragment size shift seen from 2. onwards. The purple color indicates DNA fragments that can be assigned to nucleosomes or regular cfDNA, while the green color indicates DNA fragments that can be assigned to short cfDNA. The icons originally created with biorender for the supplementary figure S1 were reused in this figure.



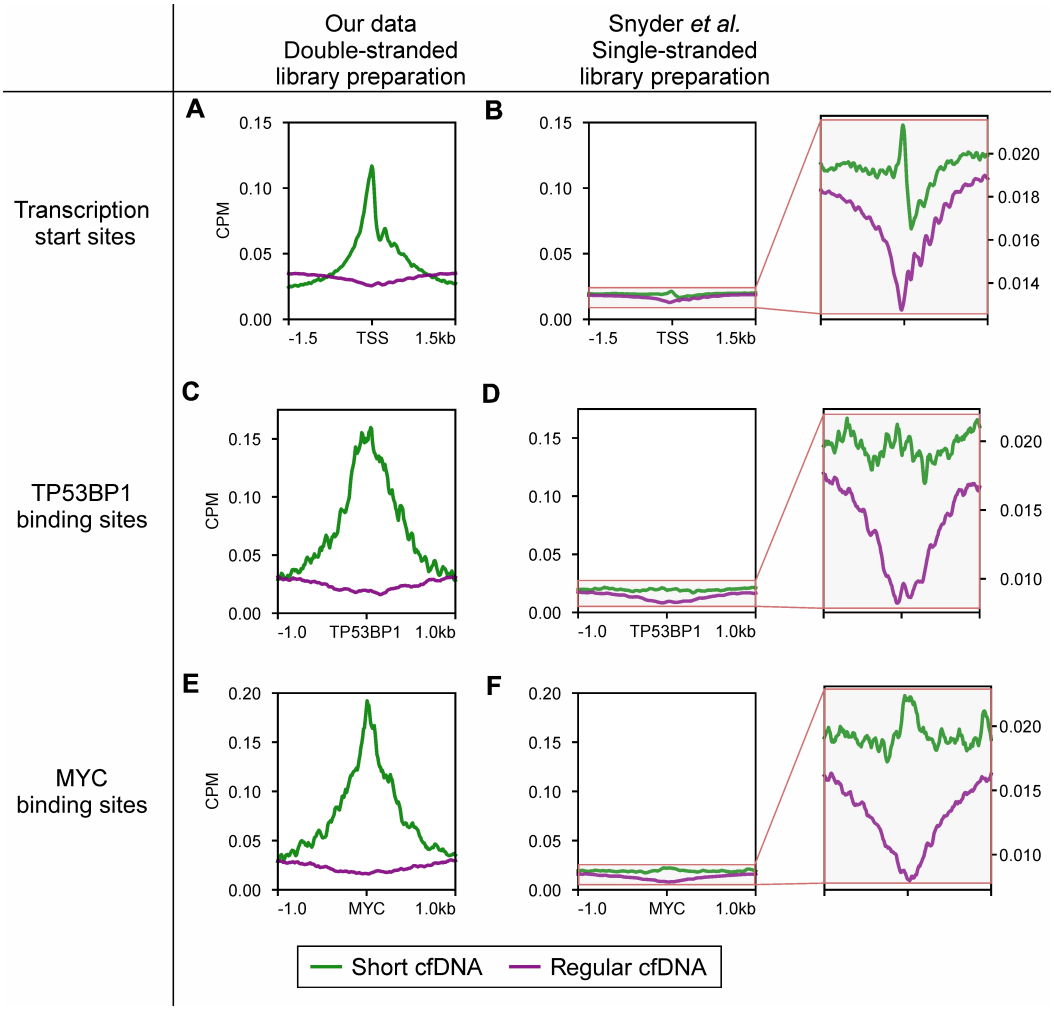
Supplementary Figure S3. Visual summary of the short cfDNA data processing pipeline. In short, the raw short cfDNA sequencing data is cleaned and filtered, mapped to a human reference genome, and filtered again before further downstream analysis. The individual steps are described in detail in the methods section. This figure was created with biorender.



Supplementary Figure S4. Analysis of read length distributions. (a) Histogram depicting the observed length of fully processed short cfDNA sequencing reads. One million random reads were taken from all twenty sequenced short cfDNA samples and their read lengths were plotted in a histogram (total n = 20 million). The observed distribution has a mean read length of = 37.9 bp (μ) and a standard deviation = 6.6 bps (σ). (b) Boxplot with data from (a) split by patient conditions.

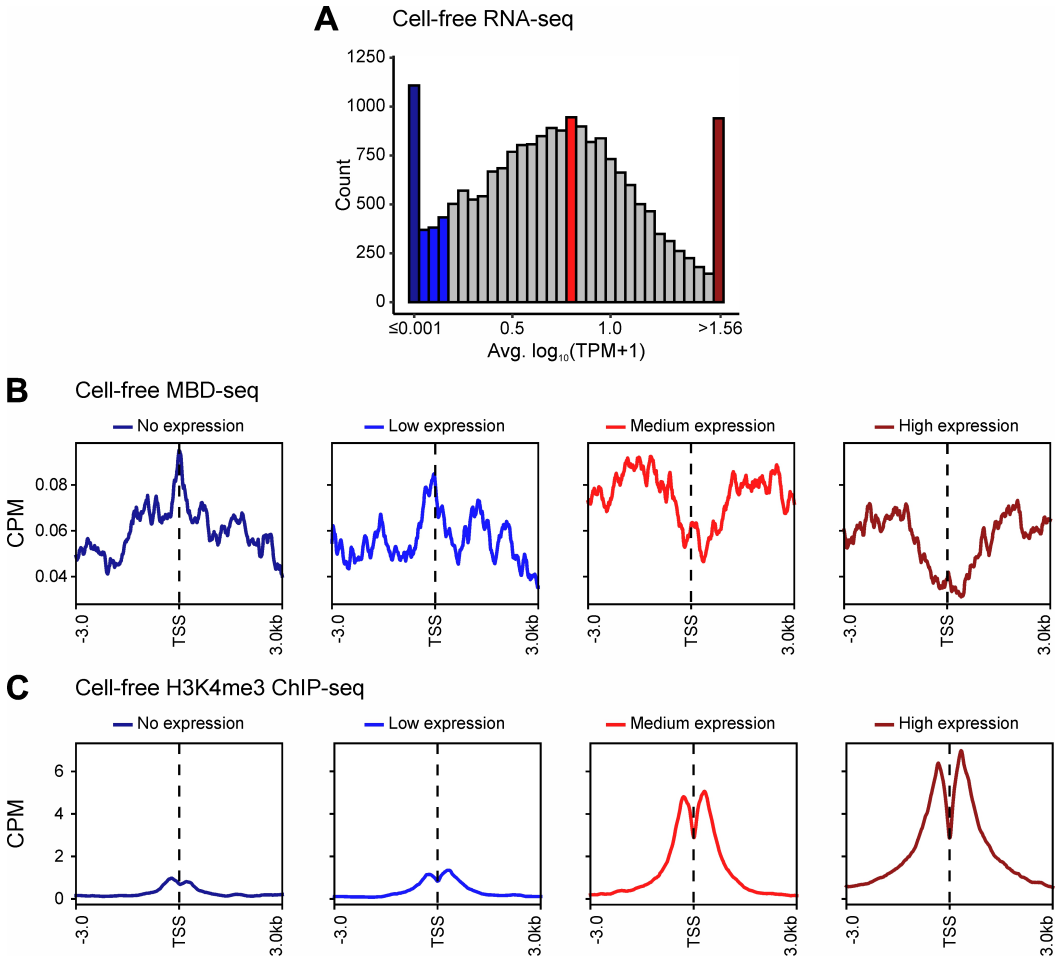


Supplementary Figure S5. Average GC content of processed sequencing reads for regular cfDNA and short cfDNA sequencing. Data from the samples S01 - S24 were used for this analysis (see Supplementary Table S1). The bar lengths represent the mean value, while error bars indicate standard deviations.

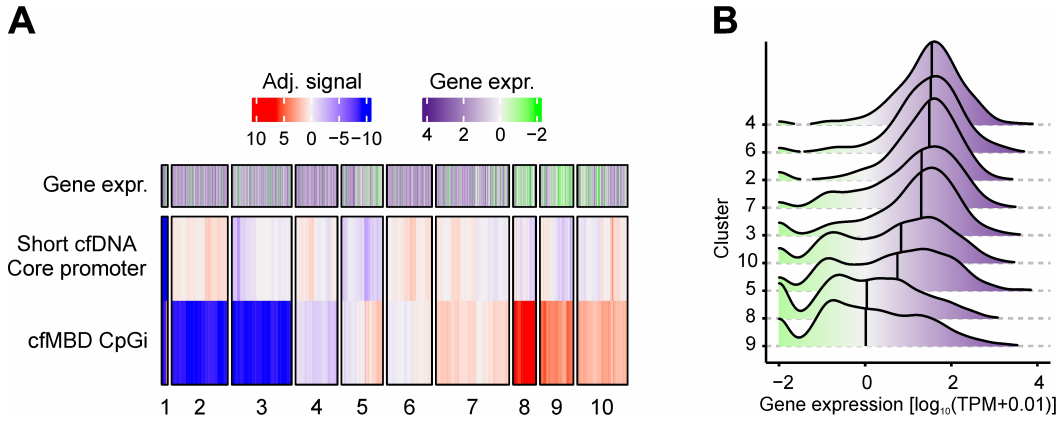


Supplementary Figure S6. Comparison of our short cfDNA sequencing data to single-stranded cfDNA sequencing data from Snyder *et al.* (Snyder *et al.* 2016). (a, c, e) Average coverage profiles based on our short cfDNA sequencing data (Sequencing depths: Short cfDNA (S03) = 8.29×10^6, Regular cfDNA (S06) = 2.60×10^7). (b, d, f) Average coverage profiles based on sequencing data from Snyder *et al.* generated with a single-stranded library preparation method (Sequencing depths: Short cfDNA = 1.33×10^8, Regular cfDNA = 3.84×10^8). These average coverage profiles are also plotted on a smaller scale to better visualize the dynamics of the data. (a and b) Average coverage profiles of annotated transcription start sites for short cfDNA and regular cfDNA. (c and d) Average coverage profiles of one thousand ChIP-seq validated tumor

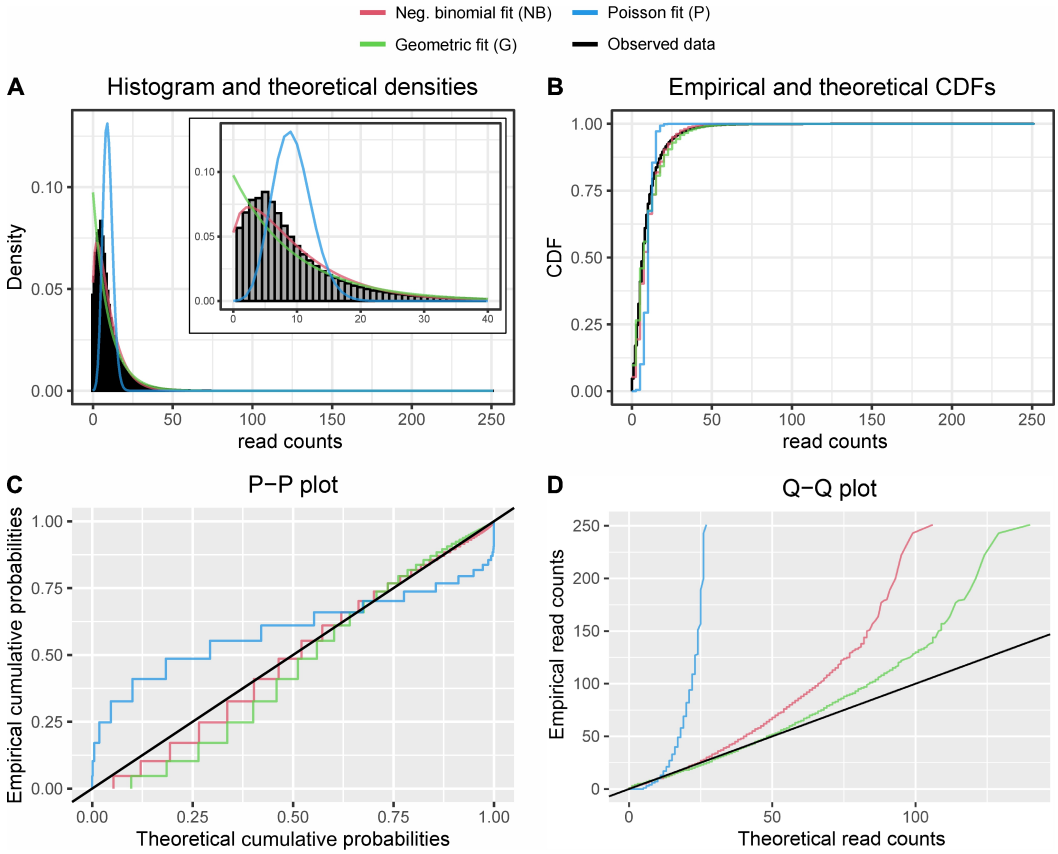
protein p53 binding protein 1 (TP53BP1) binding sites for short cfDNA and regular cfDNA. (e - f) Average coverage profiles of one thousand ChIP-seq validated MYC proto-oncogene, bHLH transcription factor (MYC) binding sites for short cfDNA and regular cfDNA. The purple color indicates data from regular cfDNA, while the green color indicates data from short cfDNA. Ultra-deep sequencing data by Snyder *et al*. was produced from the blood plasma of a healthy individual with a single-stranded library preparation method (Snyder *et al.* 2016). Raw sequencing data were retrieved from SRA (file ID = SRR2130051) and split into short cfDNA (35-80 nt) and regular cfDNA (120-180 nt) *in silico*.



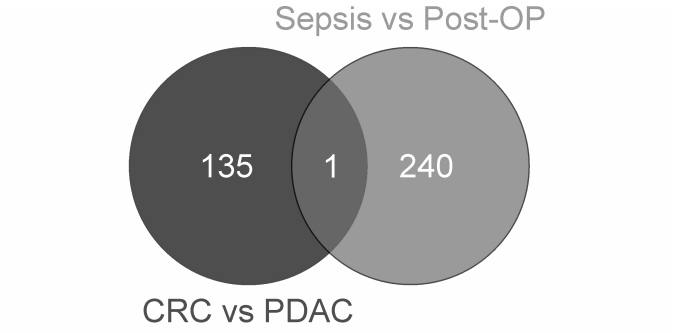
Supplementary Figure S7. Relationship of cell-free MBD-seq and cell-free H3K4me3 ChIP-seq with expression levels of genes. (a) Histogram showing average expression levels of protein-coding genes in publicly available cell-free RNA sequencing data. For each category 938 genes were selected (5 % of all analyzed genes): no expression (dark blue), low expression (blue), medium expression (red), and high expression (dark red). (b) Average coverage profiles for cell-free MBD-seq reads at selected transcription start sites. (c) Average coverage profiles for H3K4me3 ChIP-seq reads at selected transcription start sites.



Supplementary Figure S8. Analysis of joint influence of short cfDNA abundance and DNA methylation on gene expression levels. (a) Clustered heatmap with average short cfDNA signals at core promoters and average cfMBD signals at CpG islands of protein coding genes. Average gene expression levels from whole blood are annotated. (b) Sorted ridgeplot of the gene expression annotation per cluster from (a) shows the composite influence of short cfDNA abundance and DNA methylation on gene expression level distributions. Data were created from samples S26-S37.



Supplementary Figure S9. Distribution fitting and evaluation for observed read count data from short cfDNA. (A) Histogram of observed read count data (black), with three different fitted distributions as line plots (negative binomial (NB) = red, geometric (G) = green, poisson (P) = blue). An inset plot shows the read count range from 0 to 40. (B) Cumulative distribution function (CDF) plots for observed and fitted data. (C) Probability-probability plot (P-P plot) for observed and fitted data. (D) Quantile-quantile plot (Q-Q plot) for observed and fitted data.



Comparison	PDAC vs CRC		Sepsis vs Post-OP	
High in	PDAC	CRC	Sepsis	Post-OP
# DERs	105	30	155	85

Supplementary Figure S10. Comparison of loci with differential enrichment of short cfDNA. The table shows the number of differentially enriched regions (DERs) derived from short cfDNA sequencing data for the comparisons of pancreatic ductal adenocarcinoma (PDAC) versus colorectal carcinoma (CRC) and post-operative controls (Post-OP) versus sepsis. The Venn diagram shows the overlap of the two sets of DERs.

Supplementary Table S1. Metadata for the generated patient sequencing datasets included in this study. All sequenced datasets in this table are available at the SRA (PRJNA1033613).

Sequencing type	Sample	Condition	Age	Gender	Staging	Diagnosis
Short	S01	Healthy	50	male	-	-
Short	S02	Healthy	50	male	-	-
Short	S03	Healthy	45	female	-	-
Short	S04	Healthy	45	male	-	-
Regular cfDNA	S05	Healthy	47	female	-	-
Regular cfDNA	S06	Healthy	61	female	-	-
Regular cfDNA	S07	Healthy	50	male	-	-
Regular cfDNA	S08	Healthy	52	male	-	-
Short cfDNA	S09	Post-OP	20	female	-	Sigmoid colon perforation
Short	S10	Post-OP	38	female	-	Colitis ulcerosa
Short	S11	Post-OP	29	female	-	Gastroparesis
Short cfDNA	S12	Post-OP	56	female	-	Cholelithiasis with chronic cholecystitis
Short	S13	Sepsis	50	male	-	Sepsis
Short	S14	Sepsis	65	female	-	Sepsis
Short	S15	Sepsis	74	male	-	Sepsis
Short	S16	Sepsis	71	male	-	Sepsis
Short cfDNA	S17	PDAC	68	male	III	Pancreatic head ductal adenocarcinoma
Short	S18	PDAC	75	male	II	Pancreatic tail
Short cfDNA	S19	PDAC	60	male	I	Pancreatic ductal adenocarcinoma
Short cfDNA	S20	PDAC	63	female	III	Pancreatic corpus carcinoma

Short cfDNA	S21	CRC	84	male	II	Adenocarcinoma of colon ascendens
Short	S22	CRC	65	male	II	Coecal carcinoma
Short cfDNA	S23	CRC	72	female	III	Deep-seated rectal carcinoma
Short cfDNA	S24	CRC	86	male	0 (regression)	Deep-seated rectal carcinoma
cfMBD-seq	S25	Healthy	45	female	-	-
Short	S26	Sepsis	75	male	-	Sepsis
Short	S27	Sepsis	57	male	-	Sepsis
Short	S28	Sepsis	45	male	-	Sepsis
Short	S29	Sepsis	59	male	-	Sepsis
cfMBD-seq	S30	Sepsis	75	male	-	Sepsis
cfMBD-seq	S31	Sepsis	57	male	-	Sepsis
cfMBD-seq	S32	Sepsis	45	male	-	Sepsis
cfMBD-seq	S33	Sepsis	59	male	-	Sepsis
RNA-seq	S34	Sepsis	75	male	-	Sepsis
RNA-seq	S35	Sepsis	57	male	-	Sepsis
RNA-seq	S36	Sepsis	45	male	-	Sepsis
RNA-seq	S37	Sepsis	59	male	-	Sepsis

Supplementary Table S2. Detailed metadata of sepsis patients. Sepsis samples S30-S37 correspond to the same four individuals as S26-S29. Abbreviations: Erythrocytes = Erythro, Thrombocytes = Thrombo, Leukocytes = Leuko. Units: Erythrocytes = 10^{12} cells/L, Thrombocytes = 10^{9} cells/L, Leukocytes = 10^{9} cells/L.

Sample	**Age**	**Gender**	**Pathogens**	**Erythro**	**Thrombo**	**Leuko**
S13	50	male	*Bacteroides fragilis; Escherichia coli*	2.4	1127	13.67
S14	65	female	*Bordetella petrii; Human herpesvirus 5; Klebsiella pneumoniae; Pseudomonas mendocina; Salmonella enterica*	NA	NA	NA

S15	74	male	*Enterococcus faecium*	2.7	273	10.06
S16	71	male	*Enterococcus faecium*	3.3	338	24.13
S26	75	male	*Bifidobacterium animalis; Enterococcus faecium; Escherichia coli; Klebsiella pneumoniae; Lactobacillus fermentum; Staphylococcus warneri*	4	156	2.12
S27	57	male	*Haemophilus influenzae; Haemophilus parainfluenzae*	3.5	221	18.15
S28	45	male	NA	4	457	17.33
S29	59	male	*Propionibacterium spp.*	2.9	83	4.39

Supplementary Table S3. Metadata for the utilized public datasets in this study. Cell-free H3K4me3 ChIP-seq data are available from Zenodo: https://zenodo.org/records/4277001. ATAC-seq data from ENCODE were downloaded with the GRCh38 assembly and converted locally to the GRCh37 assembly with liftOver.

Data type	Sample type	Source	Project ID	File ID
ATAC-seq	Cell line (GM12878)	ENCODE	ENCSR095QNB	ENCFF646NWY
Cell-free RNA-seq	Plasma of healthy individuals	SRA	PRJNA598835	SRR10822588, SRR10822583, SRR10822579, SRR10822594, SRR10822591
Cell-free H3K4me3 ChIP-seq	Plasma of a healthy individual	Zenodo	-	H013.1
DNase-seq	Cell line (GM12878)	ENCODE	ENCSR000EMT	ENCFF783ZLL
DNase hypersensitive sites	Cell line (GM12878)	ENCODE	ENCSR000EMT	ENCFF273MVV
Double-stranded cell-free DNA	Plasma of a healthy individual	SRA	PRJNA291063	SRR2130050
Single-stranded cell-free DNA	Plasma of a healthy individual	SRA	PRJNA291063	SRR2130051

REFERENCES

Snyder M, Kircher M, Hill A, Daza R, Shendure J. 2016. Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. *Cell* **164:** 57–68. doi:10.1016/j.cell.2015.11.050.

5 Human *in vivo* footprints from blood plasma samples for improved diagnostics of septic patients

5.1 Preamble

Building on the initial proof-of-concept methodology introduced in the first publication, that distinguished septic from non-septic patients with a limited sample set, this publication extends the application of the newly established short double-stranded cfDNA sequencing to a cohort of 121 clinical patient samples to further evaluate its diagnostic potential.

This study reveals that sequencing of short cfDNA offers comprehensive insights into the host response in sepsis patients, providing information about key organ systems and physiological functions such as the liver, kidney, immune response, and oxygen transport. The results demonstrate the diagnostic potential of the novel biomarkers, enabling superior discrimination of infection types and prediction of early mortality risk in septic patients compared to established clinical metrics.

The publication substantiates the diagnostic and prognostic utility of short cfDNA biomarkers. The identified biomarkers hold promise for the development of novel, targeted, and more efficient analytical methods, opening the route for their translation into real-world diagnostic applications.

5.2 Contribution of the Thesis Author

The thesis author shares the first authorship of this publication. The thesis author, Mirko Sonntag, and Kai Sohn conceptualized the presented study. The thesis author performed the computational and statistical analysis of all data with guidance from Arndt von Haeseler. The thesis author contributed to the further interpretation of the results alongside Mirko Sonntag and Kai Sohn. Additionally, the thesis author prepared the figures. The thesis author, along with Mirko Sonntag and Kai Sohn, wrote and revised the manuscript.

5.3 Publication Status

Preprint published at medRxiv (Cold Spring Harbor Laboratory Press) on January 29, 2025: Sonntag, M., **Mueller, J.**, Brenner, T., Decker, S. O., Haeseler, A. von, Sohn, K., German Anaesthesia and Intensive Care Trials Group (2025). Human in vivo foot-

prints from blood plasma samples for improved diagnostics in septic patients. MedRxiv. https://doi.org/10.1101/2025.01.29.25320179

Human *in vivo* footprints from blood plasma samples for improved diagnostics in septic patients

Mirko Sonntag [1,2,†], Jan Müller [1,3,4,†], Thorsten Brenner [5], Sebastian O. Decker [6], Arndt von Haeseler [7] and Kai Sohn [1*] for the German Anaesthesia and Intensive Care Trials Group (GAIN-CARE)[¥]

1 - Innovation Field In-Vitro Diagnostics, Fraunhofer Institute for Interfacial Engineering and Biotechnology IGB, 70569 Stuttgart, Germany

2 - Interfaculty Graduate School of Infection Biology and Microbiology (IGIM), Eberhard Karls University Tuebingen, 72076 Tuebingen, Germany

3 - Center for Integrative Bioinformatics Vienna, Max Perutz Labs, University of Vienna, Medical University of Vienna, 1030 Vienna, Austria

4 - Vienna Biocenter PhD Program, a Doctoral School of the University of Vienna and the Medical University of Vienna, 1030 Vienna, Austria

5 - Department of Anesthesiology and Intensive Care Medicine, University Hospital Essen, University Duisburg-Essen, 45147 Essen, Germany

6 - Department of Anesthesiology, Medical Faculty Heidelberg, Heidelberg University, 69120 Heidelberg, Germany

7 - Ludwig Boltzmann Institute for Network Medicine, University of Vienna, Augasse 2-6, 1090 Vienna, Austria

* Correspondence: kai.sohn@igb.fraunhofer.de

† These authors contributed equally to this work

[¥]GAIN-CARE Collaborators:

Markus A. Weigand[1], Marc M. Berger[2], Yevhen Vainshtein[3], Silke Grumaz[3], Mehdi Manoochehri[3], Karolina Glanz[3], Manuel Feißt[4], Andrea Seidel-Glätzer[5], Mathias W. Pletz[6], Hendrik Bracht[7], Eberhart Bart[7], Kristina Fuest[8], Manfred Blobner[7,8], Friedhelm Bach[9], Onnen Moerer[10], Martin S. Winkler[10], Timo Brandenburger[11], Thomas Dimski[11], Klaudiusz Suchodolski[12], Hans-Jörg Gillmann[12], Ulrike Jäkel[13], Jana Zischkau[13], Helene Häberle[14], Peter Rosenberger[14], Tobias Schürholz[15,16], Simone Lindau[17], Stefan J. Schaller[8,18], Christian Putensen[19], Fabian Dusse[20], Sirak Petros[21], Max Gaasch[22], Christian Nusshag[23]

[1] Heidelberg University, Medical Faculty Heidelberg, Department of Anesthesiology, Im Neuenheimer Feld 420, 69120 Heidelberg, Germany

[2] Department of Anesthesiology and Intensive Care Medicine, University Hospital Essen. University Duisburg-Essen, Hufelandstr. 55, 45147 Essen, Germany

[3] Fraunhofer IGB, Nobelstr. 12, 70569 Stuttgart, Germany

[4] Institute of Medical Biometry and Informatics, University of Heidelberg, Im Neuenheimer Feld 130.3, 69120 Heidelberg, Germany

[5] Coordination Centre for Clinical Trials (KKS), Ruprecht-Karls-University, Berliner Straße 10, 69120 Heidelberg, Germany

[6] Center for Infectious Diseases and Infection Control, Jena University Hospital, Erlanger Allee 101, 07740 Jena, Germany

[7] University Ulm, Faculty of Medicine, Department of Anaesthesiology & Intensive Care Medicine, Ulm, Germany

[8] Technical University of Munich, School of Medicine and Health, Department of Anaesthesiology and Intensive Care Medicine, Munich, Germany

[9] Abteilung Klinische Infektiologie, Evangelisches Klinikum Bethel, Universitätsklinikum OWL der Universität Bielefeld, Campus Bielefeld Bethel, Bielefeld, Germany

[10] Department of Anesthesiology, University Medical Center Göttingen (UMG), Medical University of Göttingen, Göttingen, Germany

[11] Department of Anesthesiology, Medical Faculty and University Hospital Duesseldorf, Heinrich-Heine-University Duesseldorf, 40225 Duesseldorf, Germany

[12] Klinik für Anästhesiologie und Intensivmedizin, Medizinische Hochschule Hannover, Hannover, Deutschland

[13] Anästhesie und Intensivmedizin, Evangelisches Krankenhaus Luckau, Luckau, Deutschland

[14] Department of Anesthesiology and Intensive Care Medicine, University Hospital Tuebingen, Eberhard-Karls-University, Hoppe-Seyler-Strasse 3, 72076 Tuebingen, Germany

[15] Department of Anesthesia and Intensive Care, Rostock University Medical Center, Rostock, Germany

[16] Department of Intensive Care and Intermediate Care, University Hospital RWTH Aachen, Aachen, Germany

[17] Goethe University Frankfurt, Department of Anaesthesiology, Intensive Care Medicine and Pain Therapy, University Hospital Frankfurt, 60590 Frankfurt, Germany

[18] Charité - Universitätsmedizin Berlin, corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Department of Anesthesiology and Intensive Care Medicine (CCM/CVK), Berlin, Germany

[19] Department of Anesthesiology and Surgical Intensive Care Medicine, Division of Intensive Care Medicine, Universitätsklinikum Bonn, Bonn, Germany

[20] Department of Anesthesiology and Intensive Care Medicine, Faculty of Medicine and University Hospital of Cologne, University of Cologne, Kerpener Straße 62, 50937 Cologne, Germany

[21] Medical ICU, University Hospital Leipzig, Leipzig, Germany

[22] Department of Anaesthesiology and Intensive Care Medicine, General Hospital of Heidenheim, Heidenheim, Germany

[23] Department of Nephrology, Heidelberg University Hospital, Heidelberg, Im Neuenheimer Feld 162, Germany

Summary

Background

Liquid biopsy based on cell-free DNA (cfDNA) is an established approach in clinical diagnostics. In recent years, a fraction of cfDNA comprising short fragments has been discovered, that is enriched at gene promoters and binding sites of DNA-binding proteins. However, the diagnostic potential of such short double-stranded cell-free DNA (footprint DNA) remains to be fully explored. Therefore, we characterized the clinical utility of footprint DNA in septic patients.

Methods

We enriched for footprint DNA based on size selection and subsequent high-throughput DNA sequencing to receive an unbiased, genome-wide picture of the host response to the infection. Footprint DNA occupancies were analyzed for correlation with clinical metrics including urea, hemoglobin, or alanine aminotransferase (ALT). Additionally, footprint DNA markers were benchmarked by read and receiver operating curve (ROC) analysis against procalcitonin (PCT) as an established marker for infection status as well as against clinical parameters for early death prediction.

Findings

We found that levels of occupancy of footprint DNA at defined genomic loci semi-quantitatively correlated with physiological markers like ALT or urea from major organ systems including liver or kidney. In a small proof-of-concept cohort, differential signatures of DNA footprints distinguished between patient groups with bacterial and viral infections with an area under the ROC (AUROC) of 1.0, which is considerably better than PCT with an AUROC of 0.75. Likewise, footprint DNA could also predict early death in septic patients with an AUROC of 0.983, compared to the SOFA (Sepsis-related organ failure assessment) score with an AUROC of 0.76.

Interpretation

Our findings show that footprint DNA delivers quantitative information on physiology at the DNA level, demonstrating its diagnostic and prognostic potential. Identified footprint biomarker regions could be helpful in the clinical assessment of septic patients and other complex diseases outperforming current state-of-the-art clinical diagnostics.

Funding

This study was financed with internal funds from Fraunhofer society.

Keywords

Next-Generation diagnostics; Short cell-free DNA; Biomarker identification; Sepsis; Liquid biopsy

Introduction

The potential of liquid biopsy in clinical diagnostics has revolutionized the field of clinical diagnostics for multiple indications [1]. Especially cell-free DNA is widely used as a biomarker class for prognosis and diagnosis in non-invasive prenatal testing, cancer, sepsis and organ rejection, among others [2–4]. Exemplarily, several studies underline the potential of cfDNA based biomarkers for clinical sepsis diagnostics [5–7]. For pathogen detection in sepsis, it is well known that the clinical gold standard of culture-based analysis (e.g., blood culture) is associated with considerable limitations [8–10]. Using metagenomic high-throughput sequencing of plasma cfDNA, a six times higher sensitivity for pathogen detection than blood culture can be achieved [11]. More recently, in contrast to regular cfDNA, with an average fragment size of 167 bp, a very small subset of double-stranded cfDNA with a significantly shorter fragment length of 35 - 80 bp has been found to be protected by non-histone DNA-binding proteins, like transcription factors [12,13]. Using ultra-deep high-throughput DNA-

sequencing, it has been described that short cfDNA directly maps to genomic regions that are occupied by DNA binding proteins like transcription factors, indicating first potential for clinical application for example in breast cancer diagnostics [12,14]. Alternative approaches to infer the transcription factor occupancy make use of nucleosomal positioning therefore detecting open chromatin, where DNA regulatory binding protein may bind [15]. Quite recently, we have developed a workflow to specifically enrich short double-stranded cfDNA from plasma (footprint DNA) and could show that these fragments accumulate at open chromatin and gene regulatory elements [16]. The signals extracted from the footprint DNA were found to have the clinical potential to distinguish between two different types of gastrointestinal cancer as well as to discriminate sepsis from non-infected postoperative controls and healthy individuals [16].

Sepsis, defined as a “life-threatening organ dysfunction caused by a dysregulated host response to infection” [17], is a major health threat with high mortality rates and a global incidence of estimated 48.9 mio sepsis cases with 11 mio deaths per year [18]. However, precise and reliable biomarkers for sepsis diagnosis and prognosis are still in their infancies. One standard assessment for sepsis diagnosis represents the Sequential (Sepsis-related) Organ Failure Assessment (SOFA) score, which evaluates multiorgan failure in ICU patients [19] and has been described to have predictive value for mortality risk [20,21]. Another important clinical parameter frequently used in sepsis is procalcitonin (PCT) which is an indicator for bacterial sepsis and primarily suggested to guide antibiotic treatment [22–24]. Still, PCT and SOFA, although routinely used in clinical practice, do not represent reliable predictors of infection and outcome, thus indicating that further research is needed to overcome limitations and to establish more precise biomarkers [25,26].

In this study, we evaluated the diagnostic potential of footprint DNA in the context of sepsis. Consequently, we identified footprint DNA target regions that have the potential to semi-quantitatively retrace patient physiology. Based on the host response, footprint DNA allowed to more accurately distinguish between bacterial and viral sepsis and to predict early deaths versus survivors better than established biomarkers, including SOFA and PCT.

Methods

Ethics Approval and patient consent

Septic patient samples from the monocentric clinical study S-097/2013, conducted in the surgical intensive care unit of Heidelberg University Hospital (November 2013 - January 2015, German Clinical Trials Register: DRKS00005463) [11] and from the multicentric study S-084/2017 (March 2019 - August 2020, German Clinical Trials Register: DRKS00011911, ClinicalTrials.gov: NCT03356249) [27] were used, with study and control patients or their legal representatives signing written informed consent. Both studies were conducted in accordance with the Declaration of Helsinki and the professional code of conduct for physicians of the competent state medical association in the current versions, approved by the Institutional Ethics Committee of the Medical Faculty of Heidelberg University. Samples from S-097/2013 were used for correlating clinical metrics with footprint DNA signals [11,28]. For infection type and outcome prediction, septic samples from the study S-084/2017 [27] and postoperative, non-infection controls from S-097/2013 [11,28] were used.

Blood plasma preparation and cell-free DNA isolation

Plasma and cfDNA was isolated as described previously [11]. CfDNA was isolated with the QIAsymphony SP and the QIAsymphony DSP Circulating DNA Kit as well as the QiaCube connect and the QIAamp MinElute ccfDNA kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions. Eluted cfDNA was quantified with the Qubit dsDNA HS Assay Kit (Thermo Fisher, Waltham, USA) and cfDNA quality was assessed by the Fragment Analyzer High Sensitivity DNA Kit (Agilent, Santa Clara, California, USA).

Library generation and Next-Generation sequencing

Footprint DNA sequencing libraries from isolated cfDNA were prepared as already described in [16], with 15 ng of cfDNA as input and final elution in 20 µL of nuclease-free water. Library generation with the NEXTFLEX Cell free DNA-Seq Kit (V2) (Perkin Elmer, Pittsburgh, USA) was automated using the Biomek FXP (Beckman Coulter, Brea, USA). Library quality was

assessed with the Fragment Analyzer High Sensitivity DNA Kit (Agilent, Santa Clara, USA) and concentration was measured by the Qubit dsDNA HS Assay Kit (Thermo Fisher, Waltham, USA). Size selection of cfDNA libraries was performed as described in [16] using BluePippin and 3% agarose cassettes (Sage Science, Beverly, USA) with one minor modification: following the second size selection, samples were purified with 1.8 x the volume of AMPureXP beads (Beckman Coulter, Brea, USA). Size selection performance was evaluated by the Fragment Analyzer High Sensitivity DNA Kit (Agilent, Santa Clara, USA) and by Qubit dsDNA HS Assay Kit (Thermo Fisher, Waltham, USA). After size selection and molarity normalization, sequencing was performed with a NextSeq 2000 (Illumina, San Diego, USA) using 100 bp single read kits with a targeted sequencing depth of 50 million reads per sample.

Raw data processing

After quality control of raw sequencing reads with FastQC (v0.12.1), the following steps were performed to remove sequencing artifacts: 1) Removal of sequencing adapters, terminal polyG sequences (min 10 G's), and quality trimming (BBTools - bbduk.sh v39.01). 2) Removal of terminal single A introduced by library preparation (BBTools - bbduk.sh v39.01). 3) Size selection of sequenced reads allowing lengths greater than 20 bp and smaller than 60 bp (BBTools - bbduk.sh v39.01). 4) Removal of sequencing reads with dust scores smaller than 7 (prinseq-lite v0.20.4) [29–31]. 5) Processed reads were mapped to the human reference genome GRCh37 using NextGenMap (v0.5.5) with default settings [32]. 6) Mapped reads were deduplicated with samtools rmdup (v1.6), reads in blacklisted regions were removed, and reads with a MAPQ value < 1 were removed with samtools view (v1.6) [33,34]. Resulting reads were converted to BigWig for visualization and other downstream analyses with deeptools (bin size = 10, normalization = counts per million (CPM), bamCoverage v3.5.2) [35].

Peak calling and annotation

For footprint DNA data, peaks were called with MACS2 callpeak (narrow: --nomodel --extsize 32 --call-summits --min-length 30 -q 0.05; v2.2.9.1) [36]. Consensus peaks of a condition were identified with R if narrow peaks were identified in at least fifty percent of samples of a given

condition. In addition, consensus peaks less than 31 nucleotides apart were merged. Consensus peaks were annotated to genomic functional elements with bedtools (nuc; v2.30.0) [37] including gene bodies, CpG islands, cis-regulatory elements (CREs) and transcription factor binding sites (TFBS). References for functional elements are provided on github and were originally retrieved from UCSC (CpG islands: UCSC TableBrowser, track name=cpgIslandExt; TFBS: UCSC TableBrowser, track name=encRegTfbsClustered) and ENCODE (CREs: https://doi.org/doi:10.17989%2FENCSR461KLY, https://doi.org/doi:10.17989%2FENCSR471KRT, https://doi.org/doi:10.17989%2FENCSR676GWZ, https://doi.org/doi:10.17989%2FENCSR405FRQ).

Transcription factor motif enrichment analysis

Enriched transcription factor motifs were identified with MEME (AME; --scoring avg --method fisher; v5.4.1) [38]. Input DNA sequences were retrieved from consensus peaks with bedtools (nuc; v2.30.0) and analyzed for enrichment of motifs in the HOCOMOCOv11_core_HUMAN_mono database [37,39]. DNA sequences from each condition's consensus peak set served as test or control, e.g., consensus peak DNA sequences from bacterial sepsis as test for consensus peak DNA sequences from viral sepsis as control. Motifs with adjusted p-values < 0.05 were considered significantly enriched.

Identification of differentially occupied genomic regions

Differentially enriched genomic regions were identified from the respective consensus peak sets of two conditions of sepsis outcomes or sepsis infection types. Genomic GC content of the consensus peak set used in a comparison was extracted with bedtools (nuc; v2.30.0), for later normalization of read counts [37]. Raw read counts were obtained with deeptools (multiBamSummary; v3.5.2) [35]. The read count matrix, GC content information and experiment design were used as input for differential enrichment analysis in R with EDASeq and edgeR packages (EDASeq v2.32.0; withinLaneNormalization (y='GC', which='full'), betweenLaneNormalization (which='full'), glmFit, glmLRT) [40,41]. Genomic regions with an adjusted p-value < 0.05 and a |log2(fold change)| > 0.5 were considered as differentially enriched regions (DERs). In addition, regions were ranked by p-values per fold change

direction. From the ranked regions, the top 5 per fold change direction were selected. The top 5 DERs for the early death condition of the outcome comparison also includes regions with an adjusted p-value > 0.05. Normalized read counts from top regions were visualized as heatmaps with gplots (heatmap.2; v3.1.3; log2(normalized read count + 1), z-scaling per region, clustering by regions and samples using Euclidean distance and ward.D2 clustering). Normalized read counts of the top regions per condition were used as input for a principal component analysis (PCA) in R (prcomp; center=TRUE, .scale=TRUE). The first two principal components were evaluated for their discriminative power of the experimental conditions with receiver operating characteristic (ROC) analysis in R (pROC; v1.18.0).

For the infection type comparison, bacterial and viral sepsis samples were separately compared to non-infected control samples. All three comparisons, bacterial vs. viral, bacterial vs. no infection, and viral vs. no infection, yielded each 10 top DERs, in total 30. The GC content, sequencing depth, and region size normalized read counts from these regions were used for PCA.

Correlation analyses between clinical metrics and short cfDNA

For the correlations between footprint DNA and clinical metrics, two patient groups were used. The 'Identification' cohort (n = 28) included time course samples from six sepsis patients over 21 days, with eight samples missing due to unavailability or exclusion as described earlier [11]. Afterwards, identified correlates were validated with independent samples of varying time points from additional sepsis patients (‘Test’ cohort, n = 34). Correlations were derived from read counts of genomic bins to ensure an analysis that encompasses the entire genome, rather than focusing solely on regions with detectable peaks in the samples. This approach allows the detection of variations across the full dynamic range of read counts, including low or zero values, which might be missed by peak-calling methods. The genome was split in 30,962,143 bins of length 100nt and the short cfDNA reads per bin were counted. Read counts of bins were adjusted for GC content in R with EDASeq (EDASeq v2.32.0; withinLaneNormalization (y=’GC’, which=’full’)) [40]. After GC adjustment, values were transformed to transcripts per million (TPM), to account for the sequencing depth of each sample [42]. Pearson correlations

were calculated between normalized read counts of each bin and the following clinical metrics: activated partial thromboplastin time, alanine aminotransferase, albumin, alkaline phosphatase, aspartate aminotransferase, bilirubin, C-reactive protein, creatinine, erythrocytes, hemoglobin, leukocytes, procalcitonin, Quick`s value, thrombocytes, and urea. Correlation values were computed for each patient individually and all samples combined. For the latter , p-values were adjusted for multiple testing using the false discovery rate (FDR) [43]. For metrics with over 50 bins having an FDR < 0.05, the top 50 bins were selected based on the sum of the combined and patient-wise correlation values. The correlates from the identification cohort were evaluated for robustness with the test cohort. This evaluation was based on the correlation value in the test cohort and the mean absolute error (MAE) to a linear model fitted on the identification cohort. Outliers, defined as values outside the identification cohort's clinical metric data range, were excluded from thew test cohort. Correlates with the strongest correlation and smallest MAE were manually manual inspected to ensure that the correlation was not overestimated by a subset of data points. All correlation analyses were executed with python.

Statistics

Summary statistics for sepsis infection type and sepsis outcome cohorts were calculated in R (tidytlg; v0.1.4). For numerical metrics, mean with standard deviation, median, range, and interquartile range are reported. Categorical metrics are reported as percentages. Numerical metrics were tested for difference of medians using a two-sided Wilcoxon rank-sum test. Categorical metrics were tested for differences in proportions using a two-sided Fisher's exact test. The discriminatory power of numerical metrics for the patient groups in the two cohorts was evaluated using the area under the receiver operating characteristic (AUROC) curve calculated in R (pROC; v1.18.0). 95 % confidence interval values of AUROC values are reported and were calculated using DeLong's method. Differences between ROC curves were tested for significance with the bootstrap method of pROC. Optimal classification thresholds are provided that have been calculated using the Youden criterion.

Optimal classification thresholds derived from ROC analyses were used to categorize patients in the outcome cohort into high-risk or low-risk groups for mortality. Subsequently, time-to-event analyses were conducted to evaluate potential differences between these groups. Cox proportional hazards models were fitted using R, and significance of differences was assessed with pairwise_survdiff function from survminer (v0.4.9) and survival (v3.5.7). Hazard ratios were calculated and tested for statistical significance using the Wald test.

Role of the funding source

This study was financed with internal funds from the Fraunhofer society, which had no direct role in the study design, in the collection, analysis, and interpretation of data, in the writing of the manuscript, or in the decision to submit the paper for publication.

Results

Characterization of footprint DNA

Released DNA in the bloodstream gets rapidly degraded by DNA digesting enzymes, like DNase I, generating circulating cell-free DNA (cfDNA). Histones and other DNA binding proteins (DBP), like transcription factors, protect cfDNA from degradation (**Figure 1a**). Footprint DNA fragments originating from DBPs are generally shorter compared to cfDNA from histones, which facilitates specific enrichment and high-throughput sequencing of these fragments. Sequencing reads of footprint DNA are enriched at regulatory genomic sites like promoters, enhancers, transcription factor binding sites (TFBS), at CpG islands or in proximity to transcription start sites (TSS) [12,44] (**Figure 1a, b**, **Supplement Figure 1** and **Supplement Figure 2**). An exemplary genomic region with enrichment of footprint DNA reads is shown for the NOCT gene (**Figure 1b**). Enrichment and sequencing of footprint DNA provides genome-wide information on regulatory interactions between DBPs and the genome, which correlates with the host response of sepsis patients and thus can be used to identify valuable biomarkers. Accordingly, genomic regions with differentially enriched footprint DNA

reads, potentially revealing bona fide biomarkers, were inferred from peak calls, while correlations between footprint DNA signals and clinical metrics were identified from genomic partitions (bins) of 100 bp in size (**Figure 1b** zoom).

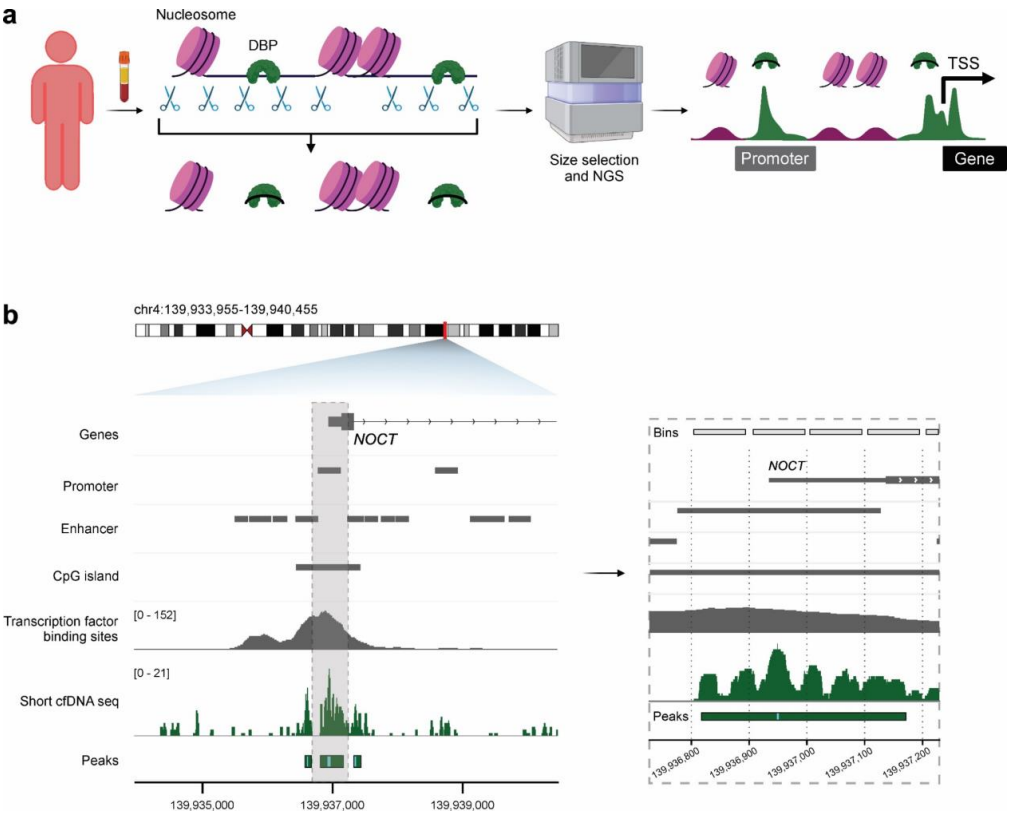


Figure 1: Origin of footprint DNA and genome-wide signal extraction approaches. a) CfDNA is protected from enzymatic digestion in the bloodstream either by histones (purple) or other DNA-binding proteins (DBP, footprint DNA, green). Isolating cfDNA from the plasma of patients, enriching DBP-bound footprint DNA according to size and sequencing it yields amplified signals at regulatory elements such as promoters or transcription start sites (TSS) of genes. Created in BioRender [Sonntag M., 2025, https://BioRender.com/i06x610]. Exemplary footprint DNA read enrichment at the *NOCT* gene promoter, enhancer, CpG island, and various TFBS. Either by peak calling (left) or by dividing the human genome into bins, footprint DNA signals can be extracted and analyzed (right, zoom in).

Footprint DNA contains information about sepsis patient physiology

Footprint DNA provides genome-wide information about the interaction of regulatory proteins with the genome. Next, we investigate whether this information can be linked to known physiological functions measured by established clinical parameters. Blood and plasma samples were collected from sepsis patients over a period of 21 days for routine clinical measurements and plasma extraction (**Figure 2a**). Patient characteristics and physiological parameter are depicted in **Supplement Table 1** and **Supplement Table 2**. Plasma was used for footprint DNA enrichment, sequencing, and extracted signals from the bin approach were evaluated for correlations to routine laboratory parameters. Parameters were evaluated for liver function (alanine aminotransferase (ALT), albumin, alkaline phosphatase (ALP), aspartate aminotransferase (AST), and bilirubin), kidney function (creatinine and urea), immune response (C-reactive protein (CRP), leukocytes, and procalcitonin (PCT)), blood coagulation system (activated partial thromboplastin time (aPTT), Quick`s value and thrombocytes), as well as oxygen transport (erythrocytes, hemoglobin). The correlations between footprint DNA signals and described metrics were tested both patient-centered and between patients ('identification' cohort) as well as validated in completely independent samples ('test' cohort).

For four physiological parameters, significant footprint DNA correlations were identified, including kidney function (urea; **Figure 2b**), immune response (PCT; **Figure 2c**), oxygen transport (hemoglobin; **Figure 2d**), and liver function (ALT; **Figure 2e**). Additionally, weaker correlates for additional liver functions (ALP, AST, and bilirubin; **Supplement Figure 3a-c**) and oxygen transport (erythrocytes; **Supplement Figure 3d**) were also identified. Patient physiology can change dynamically within a few days, leading to an exponential decrease of PCT in patient S33, for example (**Figure 2c**). Within the identification cohort for all six patients combined, footprint DNA signals correlate significantly and strongly with clinical parameters including urea (Pearson r: 0.766) or PCT (Pearson r: 0.806) (**Figure 2b-e**). At the individual patient level, correlations also showed similar trends (**Supplement Figure 4, Supplement Figure 7**). These correlations were then further validated using

samples from an independent test cohort (**Supplement Table 3**). Although extent of correlations decreased (Pearson r: 0.586 for urea and 0.440 for hemoglobin) and the MAEs of the linear fits increased, a moderate correlation with the clinical parameters was still observed (**Figure 2b-e** right plot). These results suggest that sequencing footprint DNA can provide semi-quantitative information to describe dynamic physiological changes of clinical metrices, which are used to assess the current status of the patient, e.g., organ function.

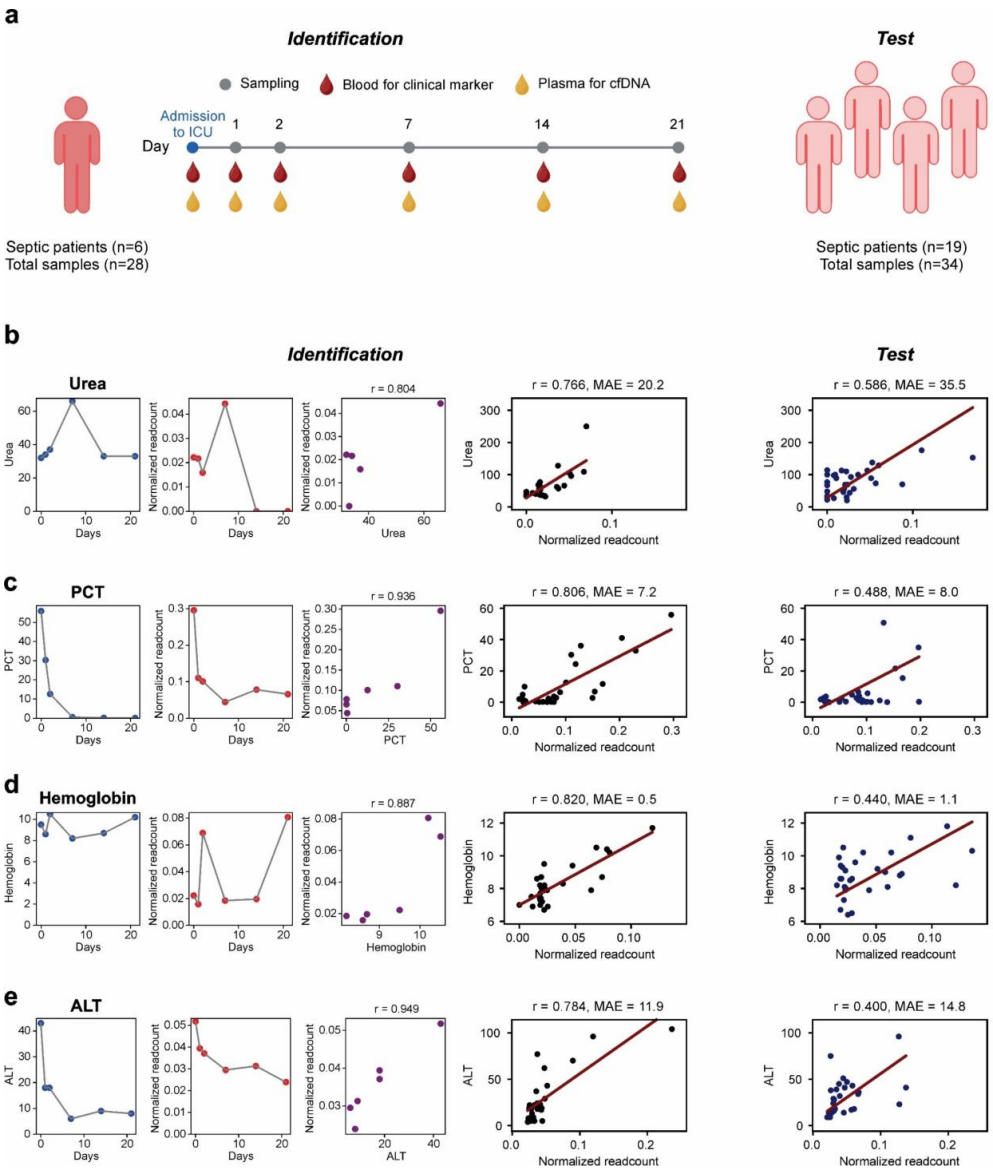


Figure 2: Correlation of footprint DNA signals to physiological markers. a) Experimental design for the identification and verification of footprint DNA correlates. Routine clinical metrics as well as short cfDNA sequencing data were acquired from six septic patients over 21 days (Identification, n=28 samples, 6 individuals). Independent samples from other septic patients were used for validation (Test, n=34 samples, 19 individuals). Created in BioRender [Sonntag M., 2025, https://BioRender.com/w86w257]. b-e) In the identification cohort, footprint DNA markers from the six patients were correlated patient wise to routine clinical metrics (left three plots) and across all patients (second right plot). In the test cohort, independent samples from sepsis patients were used for validation of identified correlation results. Reported correlations are pearson correlation values (r). Adjusted p values of correlations across patients: urea: r=0.766, p=0.0011; PCT: r=0.806, p=0.0163; hemoglobin: r=0.820, p=0.0424; ALT: r=0.784, p=0.0165. The red lines show the fitted linear function based on the identification cohort from which the MAE is calculated. PCT, procalcitonin; ALT, alanine transaminase: MAE, mean absolute error.

Footprint DNA accurately predicts type of infection in sepsis

In clinical sepsis diagnostic, PCT is routinely measured to monitor bacterial sepsis, with potential use in detecting bacterial sepsis. However, PCT not always provides reliable results to differentiate between bacterial and non-bacterial sepsis. Therefore, our cohort was used to determine whether footprint DNA markers are able to differentiate between sepsis infection types and uninfected controls already at the day of admission to the ICU (**Figure 3a** and **Supplement Table 4**). Bacterial and viral sepsis groups of the cohort do not differ significantly in any of the measured standard clinical parameters (**Supplement Table 5**). In clinical routine, procalcitonin (PCT) is often used as an indicator for bacterial sepsis, however, we found PCT values only significantly different between non-infected postoperative controls (POP) and infected patients, but not between bacterial and viral septic patients (**Supplement Figure 9**). Footprint DNA markers were identified by differential read coverage analysis from the consensus peaks of each condition (**Supplement Figure 10**). The top five differentially enriched regions (DER) per condition and per comparison (POP vs. viral, POP vs. bacterial, bacterial vs. viral; in total 30 DERs) separate the three infection types by the first and second principal components applying a principal component analysis (PCA) (**Figure 3b** and **Supplement Figure 11**). Separation of these groups based on the top 30 DERs from the three comparisons is also accomplished by hierarchical clustering (**Supplement Figure 12** and **Supplement Figure 13a-c**). Differential enrichment analysis for bacterial vs. viral sepsis revealed 220 DER in bacterial and 68 DER in viral sepsis, respectively (**Figure 3c**). Among significantly enriched bacterial DERs, a distinct differential DNA footprint in the promoter region of the *AP1B1* gene could be detected (**Supplement Figure 12**). Among significant viral DERs, differential footprint DNA signals could be exemplarily detected in the promoter of the RNA Polymerase III subunit B (*POLR3B*) or the solute carrier family 30 member 10 (*SLC30A10*), respectively (**Supplement Figure 12**). Receiver operating characteristics (ROC) analysis was used to compare the discriminatory and predictive power of the top DER with PCT. Results show that footprint DNA is a significantly better predictor of sepsis infection type than PCT (area under the ROC curve (AUROC): footprint DNA=1.000 vs. PCT=0.750;

p=0.0354) (**Figure 3d**). Additionally, differential enrichment of transcription factor binding motifs analysis was performed to identify regulatory proteins that may be involved in the discrimination of the conditions. In total, 37 significantly enriched transcription factor motifs for bacterial and 186 for viral sepsis were detected, respectively (**Figure 3e,f**). For viral sepsis the most significantly enriched binding motif belongs to SMAD family member 3 (SMAD3), while two Kruppel-like factors (KLF12 and KLF9) were detected among the top 10 most significant motifs in bacterial sepsis (**Figure 3e,f**).

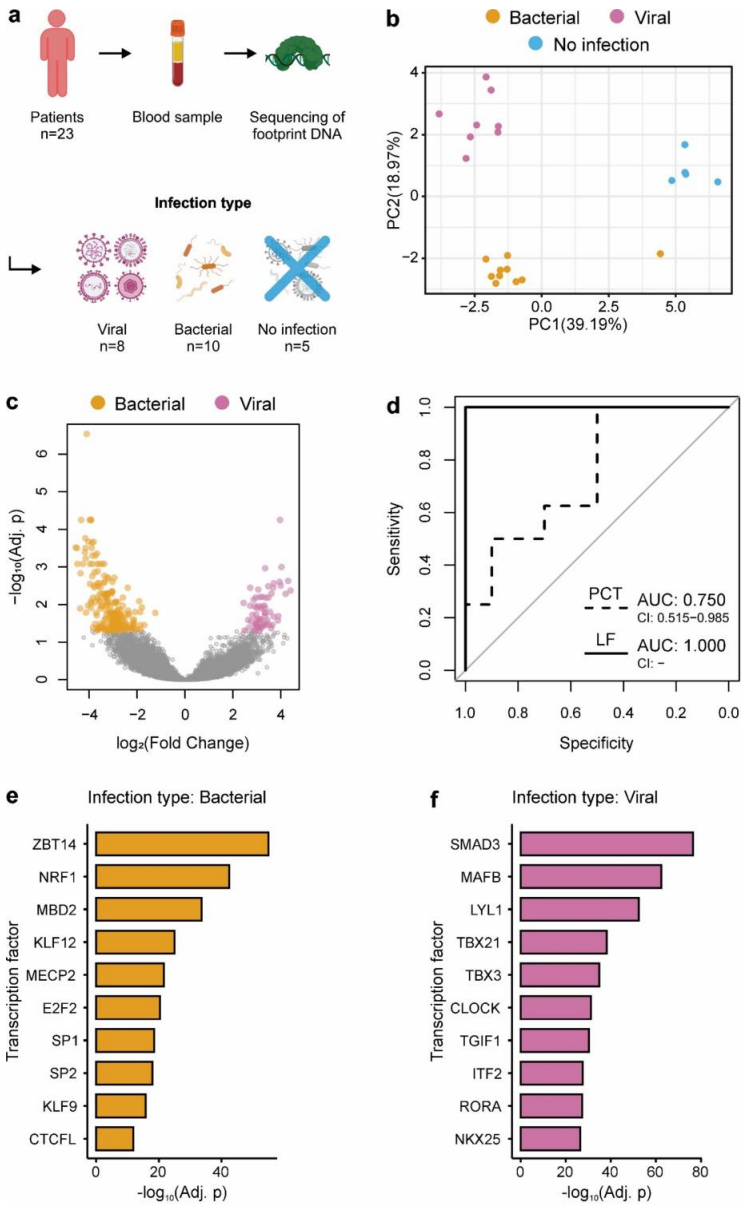


Figure 3: Discrimination of sepsis infection types by footprint DNA. a) Experimental design: For 23 samples (8 viral sepsis patients, 10 bacterial sepsis, 5 non-infected postoperative controls) plasma was prepared, footprint DNA enriched and sequenced. Created in BioRender [Sonntag M., 2025, https://BioRender.com/d50d259]. b) Principal component analysis based on the top 30 identified differentially enriched regions separates viral from bacterial sepsis and from non-infected postoperative controls. c) Differential enrichment analysis volcano plot for the comparison of viral and bacterial sepsis. Differentially enriched regions are colored according to their respective condition (adjusted p-value ≤ 0.05 and |log2 (fold change)| ≥ 0.5). d) Receiver operating characteristics analysis based on the top 10 identified Liquid Footprinting (LF) biomarkers from c) and for the routine clinical parameter PCT (95% confidence intervals are included for AUROC values; p=0.0354). e) and f) Top 10 most significant differentially enriched transcription factor motifs in bacterial sepsis (e) and in viral sepsis (f).

Footprint DNA accurately predicts early death in sepsis

Prognostic biomarkers for the prediction of outcome in sepsis can be crucial to administer the most suitable treatment options and to reduce adverse outcomes. We therefore defined a cohort of severe sepsis cases ('outcome' cohort) including 22 sepsis survivors (survival > 28 days) and 21 early death sepsis cases (death < 3 days). Based on the outcome cohort, we aim to identify footprint DNA biomarkers for identification of early death in sepsis (**Figure 4a** and **Supplement Table 7**). The outcome cohort groups differ significantly in eight of the 27 measured clinical metrics, with the SOFA score showing the lowest p value and the highest AUROC of 0.76 (**Supplement Table 8**). Differential enrichment analysis identified one DER for early death and seven for sepsis recovery, respectively (**Figure 4b**). Based on the top 5 footprint DNA signals per condition, early death and recovery from sepsis could be separated by the second principal component from PCA (**Figure 4c**), as well as by hierarchical clustering (**Supplement Figure 13d**). ROC analysis of the top 10 footprint DNA signals and the most established clinical metric SOFA yielded a significantly better predictive power for the footprint DNA biomarkers (AUROC: footprint DNA=0.983 vs. SOFA=0.760; p=0.002; **Figure 4d**). Additionally, time-to-event analysis revealed that the low-risk group identified by footprint DNA biomarkers had a significantly lower hazard ratio (HR) for death compared to the group identified by the SOFA score (HR: footprint DNA=0.042 (95 % confidence interval: 0.009-0.186), SOFA=0.364 (95 % confidence interval: 0.148-0.896); p=0.015; **Supplement Figure 14**). Differential enrichment analysis of transcription factor binding motifs identified 123 significantly enriched transcription factor motifs for early death and two for sepsis recovery, respectively (**Figure 4e,f**). For recovery in sepsis, the top binding motif reflects the transcription factor BTB domain and CNC homolog 1 (BACH1) (**Figure 4e,f**).

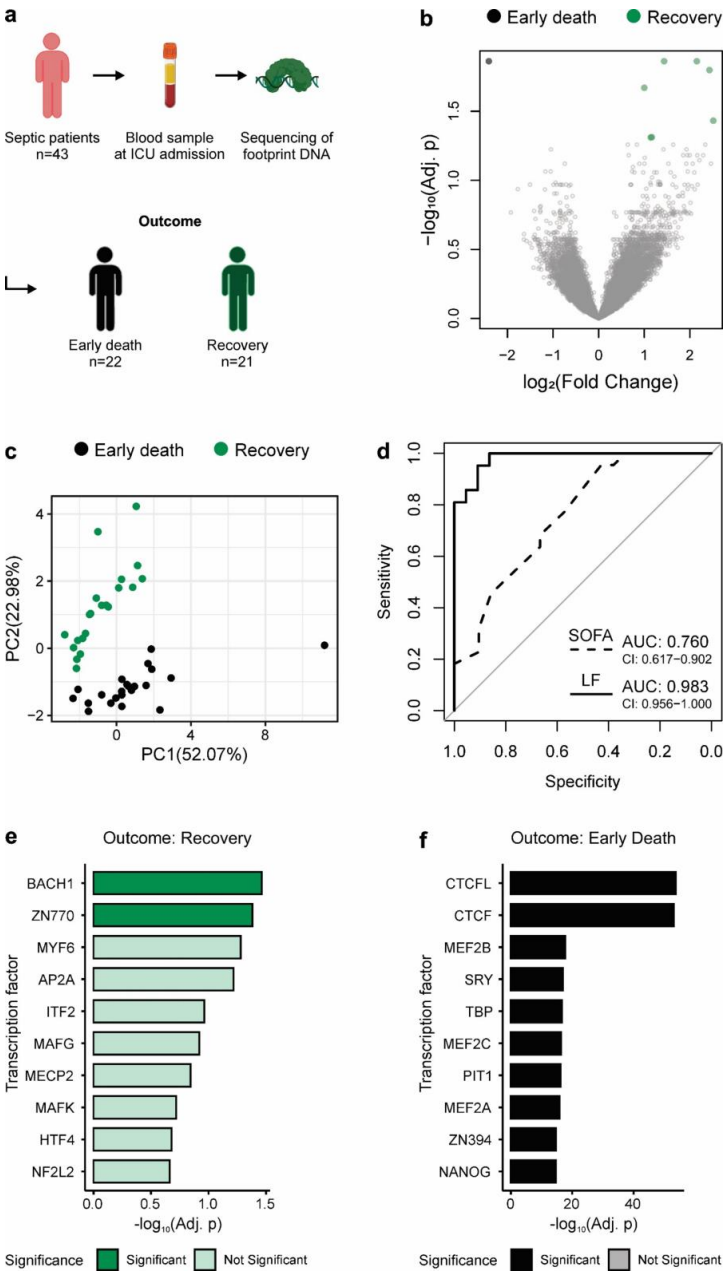


Figure 4: Discrimination of sepsis outcome by footprint DNA. a) Experimental design: From 43 septic patients samples (22 early death sepsis patients and 21 recovery patients) plasma was obtained, footprint DNA enriched and sequenced. Created in BioRender [Sonntag M., 2025, https://BioRender.com/b36q835]. b) Differential enrichment analysis volcano plot for the comparison of early death and recovery in sepsis. Differentially enriched regions are colored according to their respective condition (adjusted p-value ≤ 0.05 and |log2 (fold change)| ≥ 0.5). c) Principal component analysis based on the top 10 identified differentially enriched regions separates early death from recovery in sepsis. d) Receiver operating characteristics analysis based on the top 10 identified Liquid Footprinting (LF) biomarkers from b) and for the routine clinical metric SOFA (95% confidence intervals are included for AUROC values; p=0.002). e) and f) Top 10 most significant differentially enriched transcription factor motifs in recovery (e) and early death in sepsis (f).

Discussion

Sepsis remains a global health threat, in part due to the lack of specific diagnostic tools to assess the condition of septic patients. Monitoring the host immune response in sepsis offers valuable insights but is currently underutilized in clinical practice [45]. Several approaches aim to utilize host immune response information, for example gene expression data for pathogen detection or miRNA expression data for host response characterization [46,47]. However, dedicated sepsis biomarkers could facilitate early diagnosis, guide anti-infective therapy and track the severity of sepsis over time [48]. Accordingly, our approach aims to identify dedicated sepsis biomarkers based on cell-free DNA that capture the host response on the basis of a non-invasive liquid biopsy.

Footprint DNA sequencing provides genome-wide information on the interaction between regulatory proteins and the genome, hence signals mainly occur at regulatory genomic elements such as promoters or TFBS (**Figure 1** and **Supplement Figure 1**, **Supplement Figure 2**). Individual footprint signals correlate to clinical, physiological parameters, such as liver (for ALT, r = 0.784) and kidney (for urea, r = 0.766 (**Figure 2**). These footprint DNA markers capture rapid changes of the clinical metrics on single patient level as well as across all patients (**Supplement Figure 4**, **Supplement Figure 7**, **Supplement Figure 8**), underlining that the dynamic in patient's physiology can be assessed during disease progression, i.e., during a stay on the ICU. Despite the moderate correlation strength observed in the test cohort, the identified correlates suggest that DNA footprinting provides semi-quantitative data about dynamic changes in the physiological state of a patient. Footprint DNA allows body- and genome-wide the monitoring of different organ systems beyond the hematologic system by capturing the dynamics of clinical metrices. These results also imply that a DNA-based liquid biopsy marker can deliver physiological insights about the patient's status, similar to gene expression data from sources such as whole blood [49]. While this gene expression analysis only depicts signal mainly from the hematopoietic system, footprint DNA is capable deliver additionally information system-wide, e.g., from different organ systems or tumors. Advanced models that integrate multiple footprint DNA signals might further increase

correlation and therefore the prediction of established clinical metrics. However, directly replicating well-established, reliable, and inexpensive clinical metrics offers limited added value. Furthermore, the development of such models is limited by the comparably small cohort size.

Inadequate treatment is associated with increased mortality in septic patients, while adequate and early administration of anti-infective treatment leads to improved survival chances [50,51]. Based on the top 30 DERs, we could distinguish non-infected controls from bacterial and viral sepsis (**Figure 3b**). and with the top 10 DERs discriminate viral sepsis cases from bacterial cases with an AUROC of 1.000, outperforming all clinical routine biomarkers for infection including PCT (AUROC = 0.750) (**Figure 3c,d**; **Supplement Table 6**). Among significant DERs enriched in bacterial sepsis, a footprint signal at the promoter of AP1B1 was detected. AP1B1 is located at the Golgi complex to mediate recruitment and sorting of intracellular processes. It is a known regulator of Stimulator of interferon genes (STING) and thereby the cyclic GMP–AMP synthase (cGAS)-STING pathway. STING functions as a receptor for bacterial cyclic dinucleotides and small molecules that activates immunity during bacterial infection [52]. An enhancer in the AKT serine/threonine kinase 3 (AKT3) gene body was also identified as DER. AKT3 is known to be manipulated by bacterial infection and takes part in immune cell signaling including macrophages [53,54]. For viral sepsis, a significant DER in the promoter of POLR3B was detected. POLR3B is a part of the RNA polymerase III complex and senses common DNA viruses, such as cytomegalovirus, vaccinia, herpes simplex virus-1 and varicella zoster virus. This polymerase detects and transcribes viral genomic regions to generate AU-rich transcripts that bring to the induction of type I interferon [55,56]. Based on the transcription factor motif analysis, transcription factors enriched in either bacterial or viral sepsis were identified (**Figure 3e,f**). While the transcription factor SMAD3 is known to be activated during viral infection[57], the family of KLF transcription factors are induced during bacterial infection [58,59]. Also, footprint DNA sequencing outperforms common clinical routine metrics that were summarized in a recent review by Ahuja et al. [60]. Exemplarily for PCT, studies are described to reach an AUROC of 0.85 or lower discriminating bacterial sepsis,

which is in concordance with our infection cohort for PCT [61]. Other approaches using either miRNA signatures for host response profiling reach AUROC of 0.90 and 0.83 for bacterial or viral vs. non-infection, respectively [62], while integrated host response analysis for microbe detection reaches 0.96 [63]. Current clinical metagenomics detects microbial-derived cfDNA and is therefore limited in their analysis to bacteria and DNA viruses [4,11]. With footprint DNA biomarkers, also RNA-based microorganisms, e.g., RNA viruses, should be detected based on the host response analysis. Overall, our results show that footprint DNA biomarkers have the potential to discriminate sepsis infection types at an early time point solely on the host response, thereby delivering important information for adequate treatment decisions.

In addition, we used ten footprint DNA markers to prognostically differentiate between early death in sepsis and recovery cases on the day of ICU admission (**Figure 4a-c**). With an AUROC of 0.983, footprint DNA marker's discriminatory power is significantly better than that of clinical disease severity scoring tools such as the SOFA score (AUROC = 0.760) (**Figure 4d**, **Supplement Table 9**). The SOFA score for prediction of early death in critically ill patients had a sensitivity of 80% in a study by Ferreira et al. [64]. Also, other clinical parameters like lactate (AUROC = 0.66), lactate and SOFA (AUROC = 0.679 for initial sampling), presepsin (91.5 % sensitivity 28-day mortality) or a combination of several biomarkers including interleukin 6 and PCT (AUROC = 0.823) did not perform as promising as footprint DNA markers [65–68]. Recent approaches utilizing machine learning models with readily available clinical data were able to reach AUROCs between 0.83 to 0.89 depending on study and used data [69–71]. The top TF motifs enriched in early death cases, include different cell fate determining transcription factors including myocyte enhancer factor 2 (MEF2) and Nanog homeobox (NANOG). However, a direct causality to sepsis severity or survival remains elusive, yet.

Footprint DNA sequencing offers promising results for the quantitative assessment of the dynamic physiological changes and functionality of organ systems and for the prediction of the infection type and outcome of sepsis. At the current stage, the main limitation of this proof-of-concept study is its small sample size of the analyses. Follow up studies are essential to verify

the results. The complete procedure for footprint DNA sequencing and quantification has currently a turnaround time of two days, which is valuable time for critically ill patients. Therefore, a future improvement of the approach will involve the quantification of the most promising footprint DNA markers in a targeted approach to shorten the turnaround time for clinical applicability. Nonetheless, footprint DNA sequencing represents a promising new platform to identify liquid biopsy biomarkers, here presented for application in sepsis diagnostics. By combining the evaluation of the host response using footprint DNA with the evaluation of pathogens using clinical metagenomics, a more holistic assessment of sepsis could be achieved in the future. Footprint DNA biomarkers might be valuable not only for sepsis but also in other complex diseases including autoimmune diseases, cancer or other related fields [16].

Contributors

MS and KS conceptualized the presented ideas with input from JM. MS performed all laboratory experiments. JM performed computational and statistical analysis of all data with help of AvH. Further interpretation and analysis of the results was carried out by JM, MS and KS. JM prepared the figures. TB, SOD as well as the GAIN-CARE collaborators recruited the patients' samples used for this study. MS, JM, and KS wrote and revised the manuscript. KS supervised the project. All authors reviewed and approved the manuscript.

Data sharing statement

The raw high-throughput sequencing data used in this study have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject) under the accession number PRJNA1134400. Custom data analysis code created for this work is available on GitHub (https://github.com/janmueller17/short_cfDNA_for_sepsis).

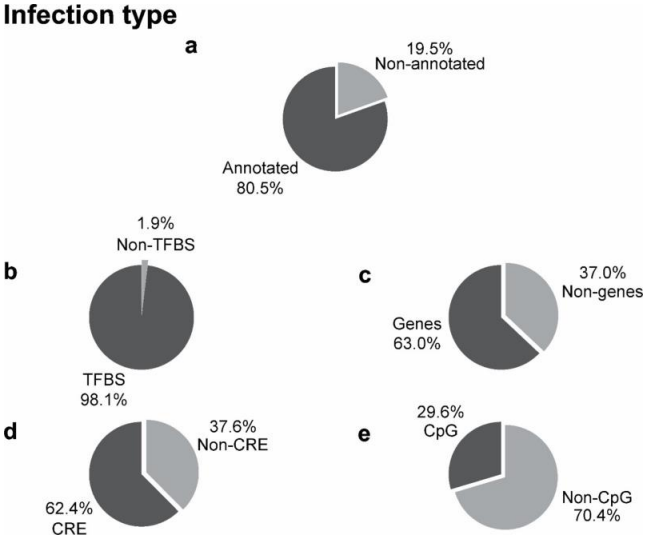
Declaration of interests

Kai Sohn is a co-founder of Noscendo GmbH, a diagnostic company that detects pathogens based on Next-generation sequencing. Moreover, Kai Sohn, Mirko Sonntag and Jan Müller filed for a patent regarding the described methods and bioinformatics algorithms for analyzing short double-stranded cfDNA (EP23203060.1).

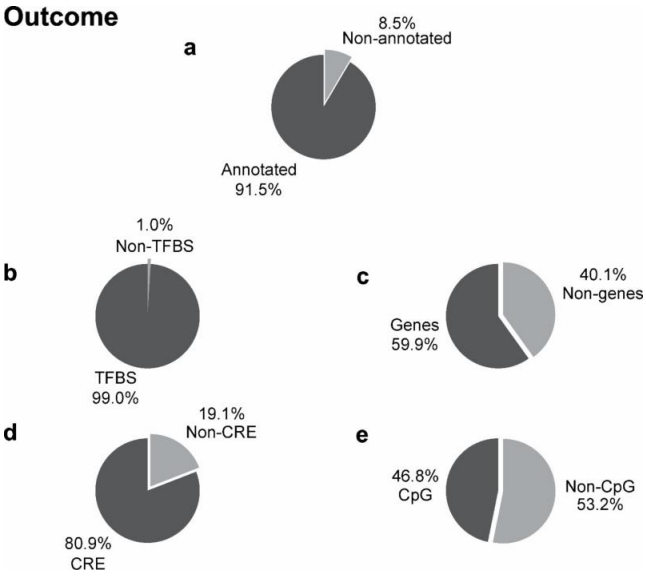
Acknowledgements

We want to thank all clinicians, nurses and patients of the participating intensive care units of both clinical studies.

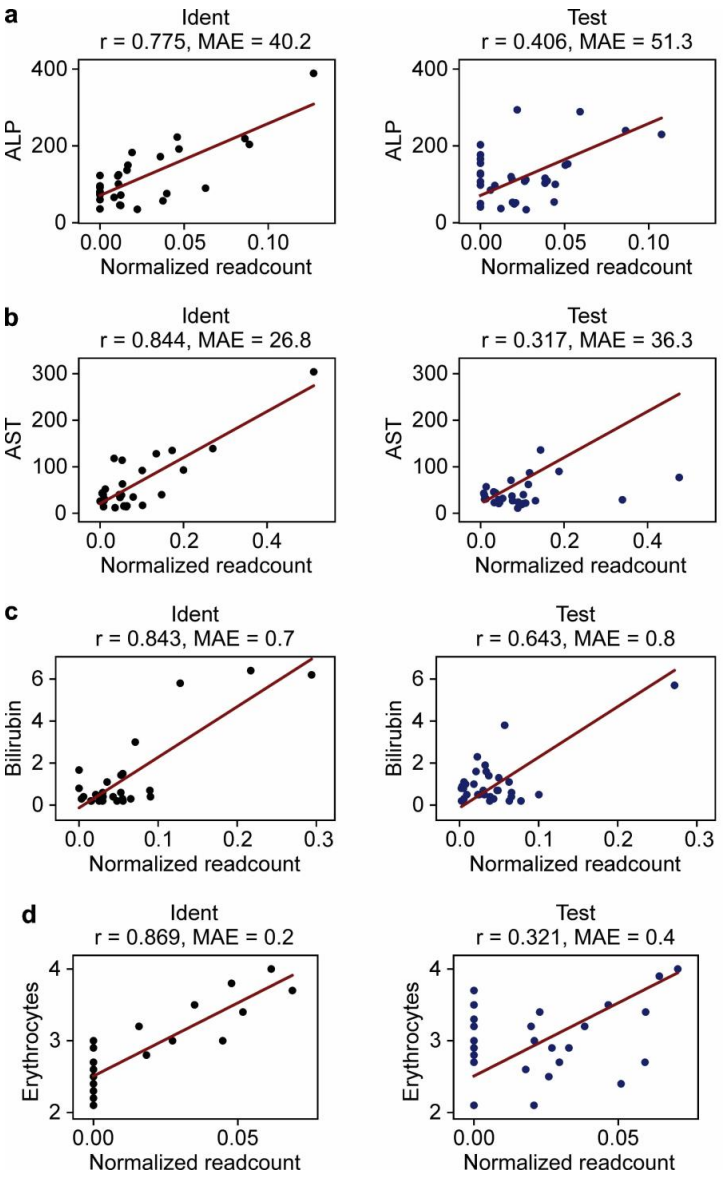
Supplementary Data



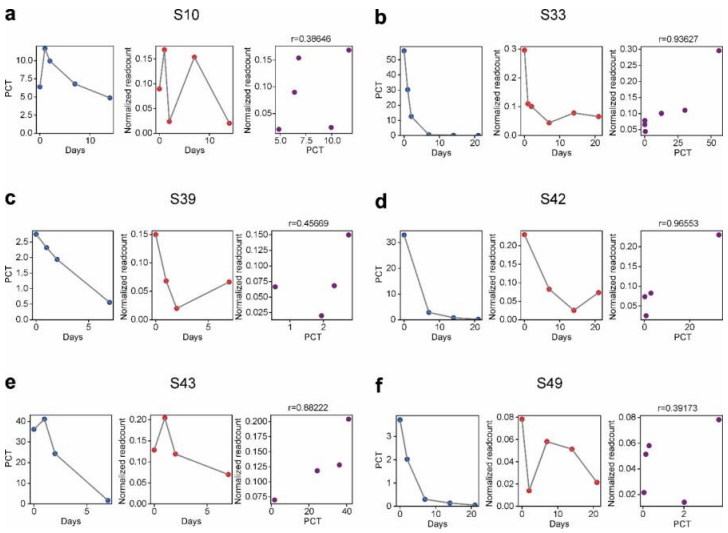
Supplement Figure 1: Characterizing of footprint DNA signals in regulatory elements for the infection type cohort. Obtained peaks from footprint DNA signals were mapped against the human genome (a). All annotated peaks were further mapped against annotated regulatory elements including transcription factors bindings sites (TFBS, b), genes (c), cis-regulatory elements (CRE, d) and CpG islands (e).



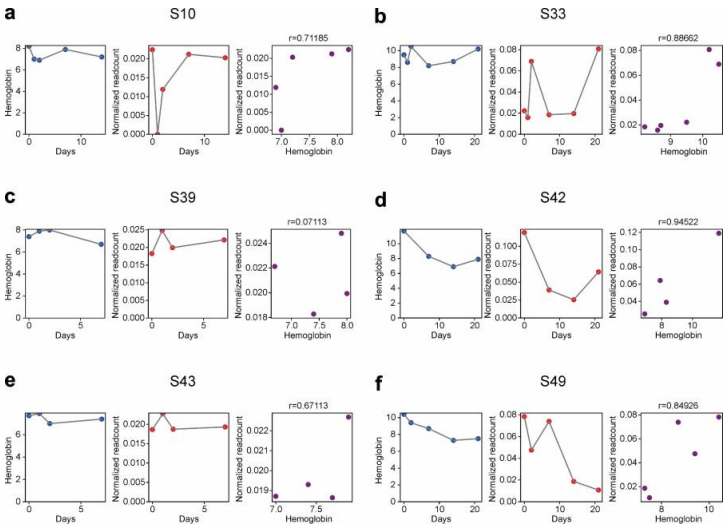
Supplement Figure 2: Characterizing of footprint DNA signals in regulatory elements for the outcome cohort. Obtained peaks from footprint DNA signals were mapped against the human genome (a). All annotated peaks were further mapped against annotated regulatory elements including transcription factors bindings sites (TFBS, b), genes (c), cis-regulatory elements (CRE, d) and CpG islands (e).



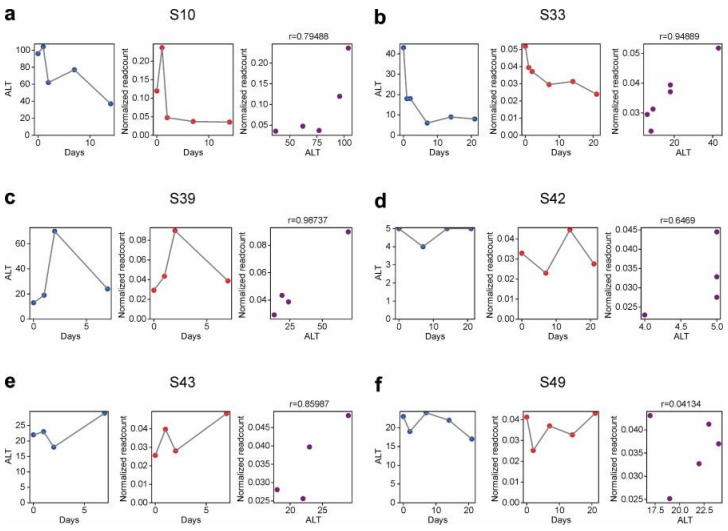
Supplement Figure 3: Correlation of footprint DNA signals to additional physiology markers. Routine clinical metrics as well as short cfDNA sequencing data were acquired from six septic patients over a 21 days timespan (Identification, n=28 samples). Independent samples from other septic patients were used for validation (Test, n=34 samples). In the identification cohort, footprint DNA markers from the six patients were correlated to routine clinical metrics across patients (left plot). In the test cohort, independent samples from sepsis patients were used for validation of identified correlation results (middle plot). Correlation of footprint DNA signals to a) Alkanine phosphatase (ALP), b) Aspartate aminotransferase (AST), c) Bilirubin, d) Erythrocytes. Reported correlations are pearson correlation values (r). Adjusted p values of correlations across patients: ALP: r=0.775, p=0.0011; AST: r=0.844, p=0.0163; bilirubin: r=0.843, p=0.0424; Erythrocytes: r=0.869, p=0.0165. The red lines show the fitted linear function based on the identification cohort from which the MAE is calculated. MAE, mean absolute error.



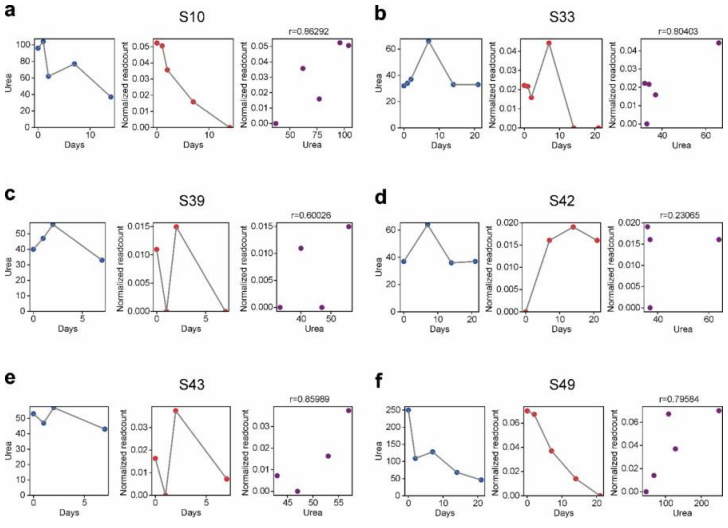
Supplement Figure 4: Correlation of footprint DNA markers to the physiological parameter PCT for single patients. a-f) Footprint DNA markers were analyzed for correlation to clinical metric PCT for all patients in the physiological cohort identification set. In the left and middle panel, PCT and normalized readcounts for the footprint, respectively, is depicted over time. The correlation of footprint DNA normalized counts and the clinical metric PCT is shown in the right panel. Results for patients are depicted the following: a) Patient S10, b) Patient S33, c) Patient S39, d) Patient S42, e) Patient S43 and f) Patient S49. PCT, procalcitonin.



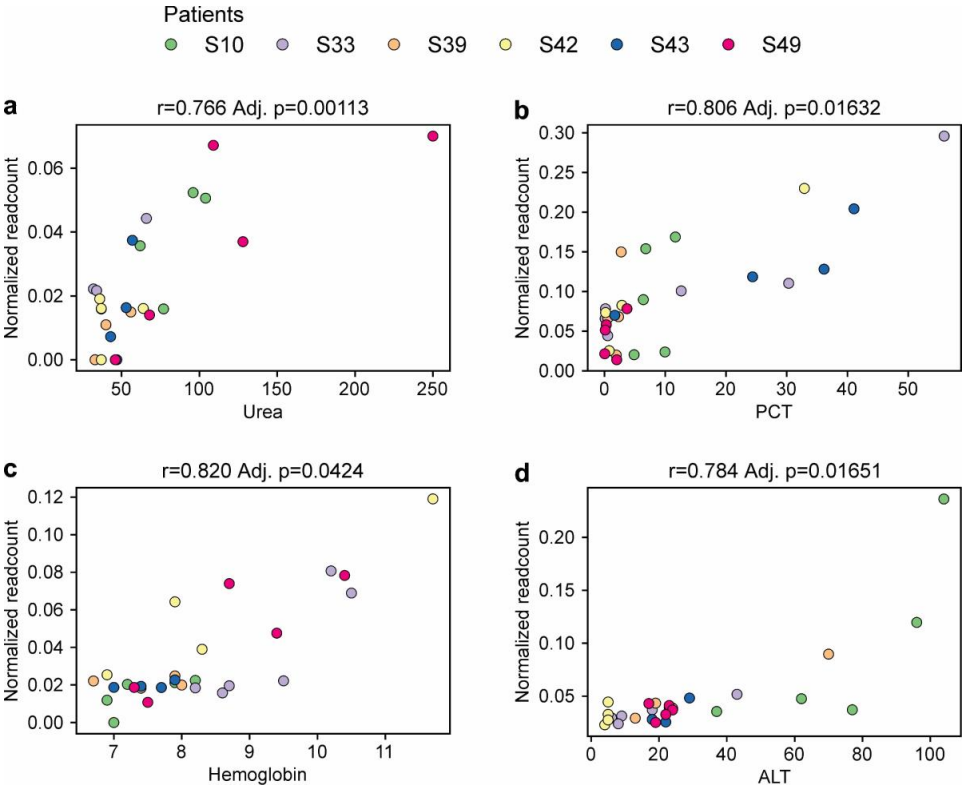
Supplement Figure 5: Correlation of footprint DNA marker to the physiological parameter hemoglobin for single patients. a-f) Footprint DNA markers were analyzed for correlation to the clinical metric hemoglobin for all patients in the physiological cohort identification set. In the left and middle panel, hemoglobin and normalized readcounts for the footprint, respectively, is depicted over time. The correlation of footprint DNA normalized counts and the clinical metric hemoglobin is shown in the right panel. Results for patients are depicted the following: a) Patient S10, b) Patient S33, c) Patient S39, d) Patient S42, e) Patient S43 and f) Patient S49.



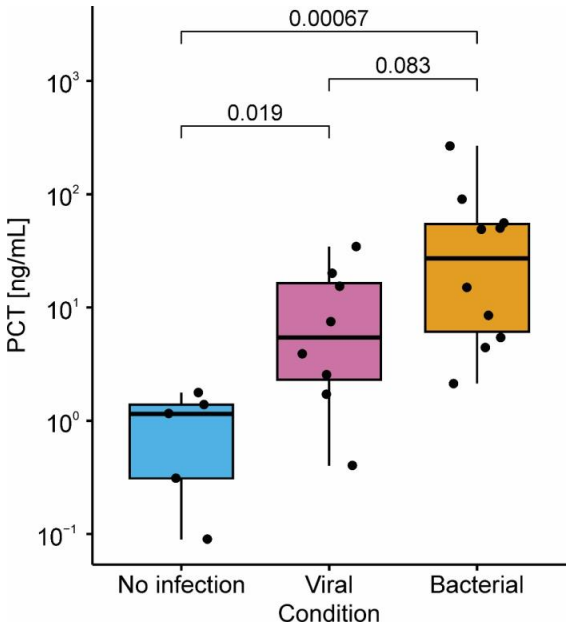
Supplement Figure 6: Correlation of footprint DNA marker to the physiological parameter ALT for single patients. a-f) Footprint DNA markers were analyzed for correlation to clinical metric ALT for all patients in the physiological cohort identification set. In the left and middle panel, ALT and normalized readcounts for the footprint, respectively, is depicted over time. The correlation of footprint DNA normalized counts and the clinical metric ALT is shown in the right panel. Results for patients are depicted the following: a) Patient S10, b) Patient S33, c) Patient S39, d) Patient S42, e) Patient S43 and f) Patient S49. ALT, alanine aminotransferase.



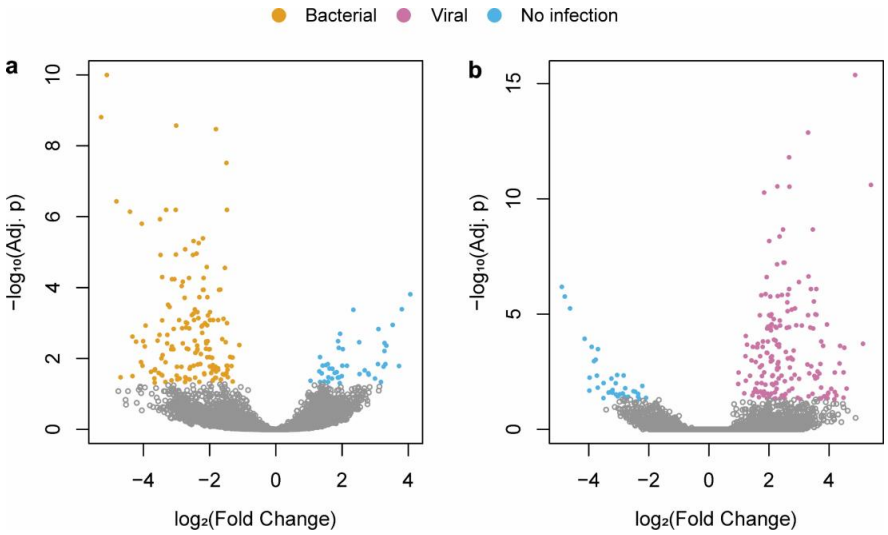
Supplement Figure 7: Correlation of footprint DNA marker to the physiological parameter urea for single patients. a-f) Normalized readcounts of footprint DNA markers were analyzed for correlation to the clinical metric urea for all patients in the physiological cohort identification set. In the left and middle panel, urea and normalized readcounts for the footprint, respectively, is depicted over time. The correlation of footprint DNA normalized counts and clinical metric urea is shown in the right panel. Results for patients are depicted the following: a) Patient S10, b) Patient S33, c) Patient S39, d) Patient S42, e) Patient S43 and f) Patient S49.



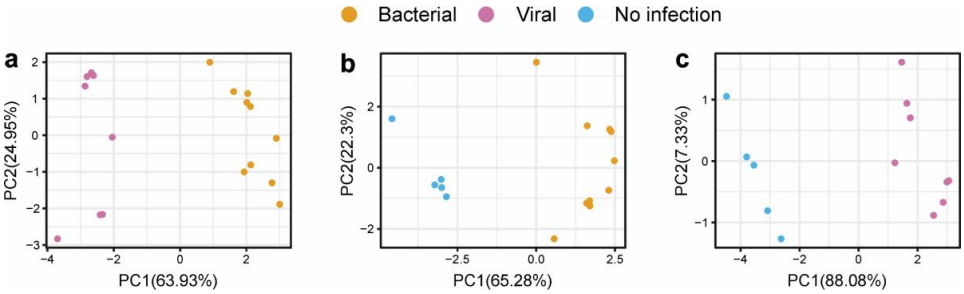
Supplement Figure 8: Patient-wide correlation of footprint DNA markers to physiological parameters. a-d) Normalized readcounts of footprint DNA markers were analyzed for correlation to each clinical metric combined for all patients in the physiological cohort identification set. Results for clinical metrics are depicted the following: a) urea, b) PCT, c) hemoglobin and d) ALT. PCT, procalcitonin; ALT, alanine aminotransferase.



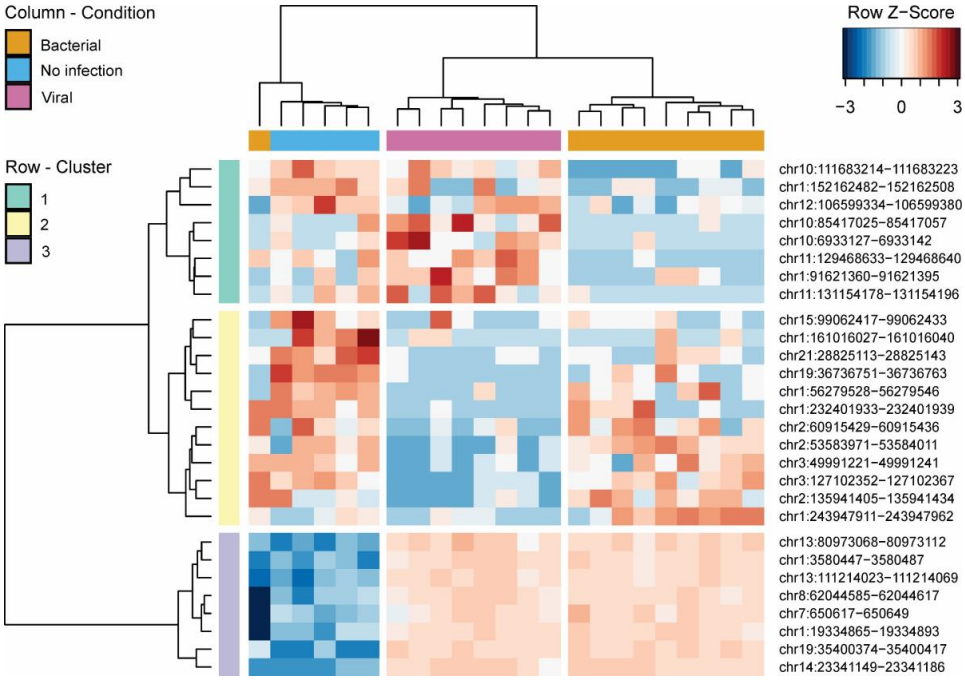
Supplement Figure 9: PCT concentration in the infection type cohort. The concentration of Procalcitonin (PCT) was measured for each sample and categorized to non-infection, postoperative controls (blue), viral sepsis (pink) and bacterial sepsis (yellow). Statistical testing was performed using a two-sided Wilcoxon rank-sum test. Resulting p-values are indicated above the brackets for each comparison.



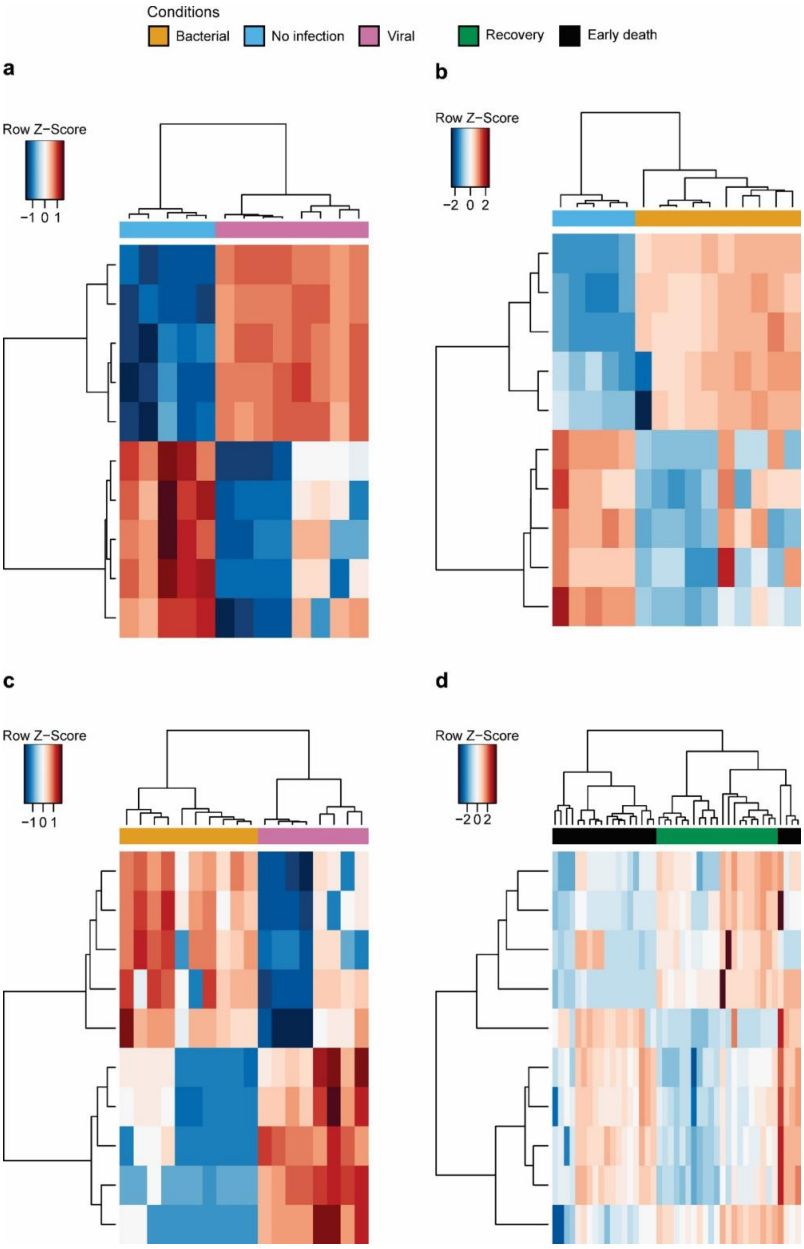
Supplement Figure 10: Differential enrichment analysis for non-infection controls and sepsis samples with viral or bacterial sepsis, respectively. Differentially enriched regions (DERs) were analyzed between two conditions for enriched transcription factor binding sites (log10 Adj. p-value ≤ 0.05 and log2 Fold change ≥ 2). Significant results are colored and depicted in a Volcano plot. a) non-infection controls (blue) were analyzed against bacterial sepsis cases (yellow) b) non-infection controls (blue) were analyzed against bacterial sepsis cases (pink).



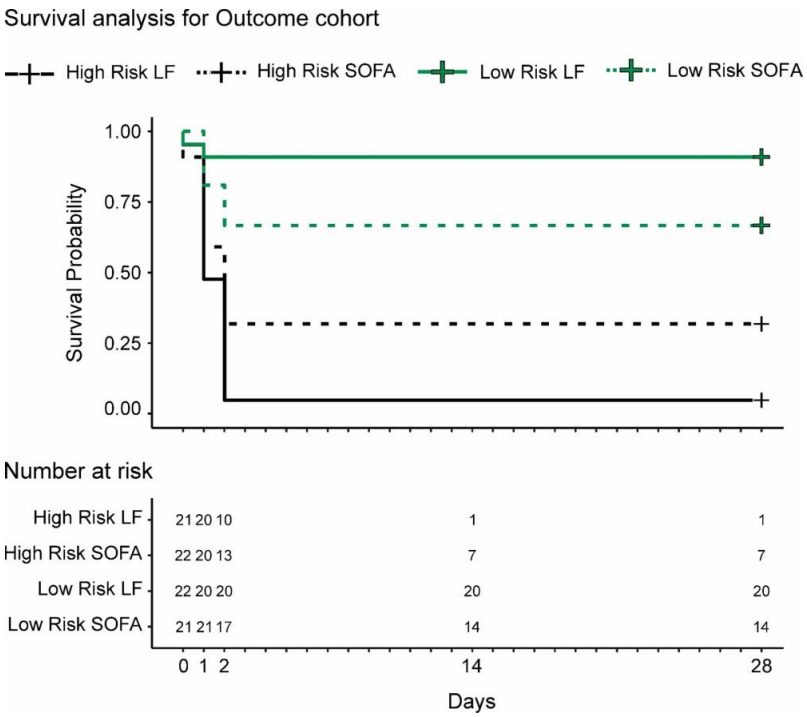
Supplement Figure 11: Discrimination of non-infection, postoperative controls, viral and bacterial sepsis via PCA. Based on the top 5 differential enriched regions (DERs) per condition, principal component analysis (PCA) was performed. The following conditions of the infection type cohort were investigated a) bacterial sepsis against viral sepsis b) non-infection control against bacterial sepsis c) non-infection control against viral sepsis.



Supplement Figure 12: Heatmap and hierarchical clustering of differential enrichment analysis of the infection type comparison of non-infected controls (blue), viral (pink) and bacterial sepsis (yellow). Based on the top 5 DER per condition, hierarchical clustering was performed. DER signals are normalized per row according to Z-score and color-coded.



Supplement Figure 13: Heatmap and hierarchical clustering of differential enrichment analysis of the infection type and outcome cohort. Based on the top 5 DER per condition, hierarchical clustering was performed. DER signals are normalized per row according to Z-score and color-coded. a) Heatmap for DERs for bacterial (yellow) in comparison to viral sepsis (pink). b) Heatmap for DERs for bacterial sepsis (yellow) in comparison to non-infected, postoperative controls (blue). c) Heatmap for DERs for viral sepsis (pink) in comparison to non-infected, postoperative controls (blue). d) Heatmap for DERs for early death (black) in compaison to recovery in sepsis (green).



Supplement Figure 14: Time-to-event analysis comparing footprint DNA and SOFA risk groups. Based on disease severity scoring with SOFA and footprint DNA signals, 43 patients from the outcome cohort were assigned to the low risk SOFA (SOFA ≤ 6; green dashed line) or high risk SOFA (Sofa > 6; black dashed line) and to the low risk footprint DNA (low risk LF, green line) or the high risk footprint DNA group (high risk LF, black line). Over the course of 28 days, the survival probability is depicted for each group.

Supplement Table 1: Patient metadata summary for sepsis cases over time (physiological status cohort).

Patient	Age	BMI	Gender	Height [cm]	RRT	Weight [kg]
S10	71-80	27.36	male	171	yes	80
S33	61-70	19.05	female	162	no	50
S39	71-80	24.34	male	172	no	72
S42	71-80	20.44	male	167	yes	57
S43	61-70	25.38	male	183	no	85
S49	61-70	37.11	female	160	yes	95

Key: BMI=Body mass index, RRT=Renal replacement therapy.

Supplement Table 2: Patient metadata summary for sepsis cases over time for each physiological parameter and time point (physiological status cohort identification).

Patient S10

Time point	To	T1	T2	T3	T4	T5
Albumin [g/L]	20.2	12.2	13.6	20.1	21.3	NA
ALT [U/L]	96	104	62	77	37	NA
AP [U/L]	73	60	78	96	389	NA
aPTT [s]	29.3	31.6	31.6	35.7	38.8	NA
AST [U/L]	40	139	304	92	135	NA
Bilirubin [µM]	10.2624	28.56368	13.6832	24.45872	51.312	NA
Creatinine [µM]	162.42	205.89	138.78	186.06	77.78	NA
CRP [mg/dL]	40.89	28.28	31.93	24.26	16.31	NA
Erythrocytes [1/pL]	2.9	2.4	2.3	2.7	2.4	NA
Hemoglobine [g/dL]	8.2	7	6.9	7.9	7.2	NA
Hematocrite [%]	25	21	20	24	21	NA
Leukocytes [gpt/I]	10.37	16.28	11.9	10.06	8.58	NA
PCT [ng/mL]	6.40	11.67	9.96	6.80	4.88	NA
Potassium [mM]	4.7	5.87	4.39	4.96	4.26	NA
Quick [%]	87.9	64.9	92.1	57.2	109	NA
Sodium [mM]	145	142	143	139	133	NA
SOFA	6	13	13	13	14	NA
Thrombocytes [gpt/L]	501	319	243	273	125	NA
Urea [mM]	15.984	17.316	10.323	12.8205	6.1605	NA

Patient S33

Time point	To	T1	T2	T3	T4	T5
Albumin [g/L]	5.8	23.2	23.3	20	22.9	24.1
ALT [U/L]	43	18	18	6	9	8
AP [U/L]	35	46	122	125	192	172
aPTT [s]	56.5	60.1	42.8	25.8	25.2	24.8
AST [U/L]	93	39	34	14	17	12
Bilirubin [µM]	3.4208	18.8144	8.552	3.4208	3.4208	3.4208
Creatinine [µM]	43.46	59.48	48.80	44.23	36.60	34.31
CRP [mg/dL]	2.83	15.87	29.83	15.49	15.63	11.95
Erythrocytes [1/pL]	3.4	3.2	3.7	3	3	3.8
Hemoglobine [g/dL]	9.5	8.6	10.5	8.2	8.7	10.2
Hematocrite [%]	29	26	32	25	26	31
Leukocytes [gpt/I]	2.07	7.46	19.68	23.16	21.61	20.66
PCT [ng/mL]	55.91	30.32	12.63	0.53	0.16	0.11
Potassium [mM]	3.98	3.94	NA	NA	NA	4.35
Quick [%]	30.7	40.5	66	63.4	61.8	63.4

Sodium [mM]	149	146	NA	NA	NA	141
SOFA	11	11	11	11	8	8
Thrombocytes [gpt/L]	345	160	105	207	612	634
Urea [mM]	5.328	5.661	6.1605	10.989	5.4945	5.4945

Patient S39

Time point	To	T1	T2	T3	T4	T5
Albumin [g/L]	19.9	24.9	22.3	27.4	NA	NA
ALT [U/L]	13	19	70	24	NA	NA
AP [U/L]	101	81	92	150	NA	NA
aPTT [s]	41.2	34.8	30.2	29	NA	NA
AST [U/L]	30	52	128	44	NA	NA
Bilirubin [µM]	99.2032	109.4656	106.0448	25.656	NA	NA
Creatinine [µM]	70.92	73.20	76.25	69.39	NA	NA
CRP [mg/dL]	21.66	20.76	13.45	11.65	NA	NA
Erythrocytes [1/pL]	2.5	2.5	2.5	2.2	NA	NA
Hemoglobine [g/dL]	7.4	7.9	8	6.7	NA	NA
Hematocrite [%]	23	23	23	21	NA	NA
Leukocytes [gpt/I]	16.74	9.33	4.77	3.46	NA	NA
PCT [ng/mL]	2.75	2.32	1.94	0.56	NA	NA
Potassium [mM]	4.73	4.81	4.37	4.15	NA	NA
Quick [%]	59.1	69.9	83.5	80.8	NA	NA
Sodium [mM]	138	143	147	139	NA	NA
SOFA	15	15	15	2	NA	NA
Thrombocytes [gpt/L]	92	69	78	197	NA	NA
Urea [mM]	6.66	7.8255	9.324	5.4945	NA	NA

Patient S42

Time point	To	T2	T3	T4	T5	T5
Albumin [g/L]	31	NA	NA	12.1	26.3	30.6
ALT [U/L]	5	NA	NA	4	5	5
AP [U/L]	45	NA	NA	223	183	137
aPTT [s]	25.3	NA	NA	47.9	43.6	25.3
AST [U/L]	14	NA	NA	15	16	16
Bilirubin [µM]	6.8416	NA	NA	5.1312	3.4208	3.4208
Creatinine [µM]	93.79	NA	NA	136.49	92.27	77.78
CRP [mg/dL]	15.16	NA	NA	19.33	15.37	10.37
Erythrocytes [1/pL]	4	NA	NA	2.8	2.3	2.7
Hemoglobine [g/dL]	11.7	NA	NA	8.3	6.9	7.9
Hematocrite [%]	34	NA	NA	25	21	24
Leukocytes [gpt/I]	2.12	NA	NA	13.04	9.04	10.95
PCT [ng/mL]	32.90	NA	NA	2.87	0.78	0.19
Potassium [mM]	3.71	NA	NA	4.46	4.14	4.01
Quick [%]	91	NA	NA	91	110	101.3
Sodium [mM]	136	NA	NA	141	134	132
SOFA	14	NA	NA	13	8	6
Thrombocytes [gpt/L]	156	NA	NA	147	468	350
Urea [mM]	6.1605	NA	NA	10.656	5.994	6.1605

Patient S43

Time point	To	T1	T2	T3	T4	T5
Albumin [g/L]	30.7	14.1	15.5	27.1	NA	NA
ALT [U/L]	22	23	18	29	NA	NA

AP [U/L]	36	76	57	90	NA	NA
aPTT [s]	50.2	37.8	38.7	27.2	NA	NA
AST [U/L]	35	40	27	40	NA	NA
Bilirubin [µM]	6.8416	6.8416	5.1312	6.8416	NA	NA
Creatinine [µM]	106.76	79.30	73.97	63.29	NA	NA
CRP [mg/dL]	43.03	49.99	43.01	18.11	NA	NA
Erythrocytes [1/pL]	2.1	2.6	2.2	2.4	NA	NA
Hemoglobine [g/dL]	7.7	7.9	7	7.4	NA	NA
Hematocrite [%]	22	25	22	22	NA	NA
Leukocytes [gpt/I]	2.15	12.41	11.31	13.55	NA	NA
PCT [ng/mL]	36.14	41.07	24.38	1.70	NA	NA
Potassium [mM]	4.17	4.23	4.09	3.95	NA	NA
Quick [%]	82.4	85.3	111	91	NA	NA
Sodium [mM]	145	150	152	141	NA	NA
SOFA	12	12	12	4	NA	NA
Thrombocytes [gpt/L]	187	128	152	165	NA	NA
Urea [mM]	8.8245	7.8255	9.4905	7.1595	NA	NA

Patient S49

Time point	**To**	**T1**	**T2**	**T3**	**T4**	**T5**
Albumin [g/L]	24.1	NA	17.6	NA	NA	NA
ALT [U/L]	23	NA	19	24	22	17
AP [U/L]	72	NA	66	123	204	219
aPTT [s]	32.3	NA	40.1	25.9	29.1	24
AST [U/L]	63	NA	118	114	43	26
Bilirubin [µM]	3.4208	NA	11.9728	10.2624	5.1312	6.8416
Creatinine [µM]	275.28	NA	128.87	50.33	39.65	30.50
CRP [mg/dL]	NA	NA	29.45	12.07	18.67	5.11
Erythrocytes [1/pL]	3.5	NA	3	2.9	2.4	2.4
Hemoglobine [g/dL]	10.4	NA	9.4	8.7	7.3	7.5
Hematocrite [%]	29	NA	25	25	21	21
Leukocytes [gpt/I]	4.13	NA	3.62	1.07	1.23	5.15
PCT [ng/mL]	3.71	NA	2.02	0.30	0.14	0.05
Potassium [mM]	4.6	NA	4.77	4.18	4.2	3.8
Quick [%]	50.9	NA	92.8	69.9	96.9	89.2
Sodium [mM]	140	NA	141	161	150	140
SOFA	17	NA	14	8	4	1
Thrombocytes [gpt/L]	70	NA	62	32	53	341
Urea [mM]	41.625	NA	18.1485	21.312	11.322	7.659

Supplement Table 3 – Patient metadata summary for sepsis cases over time for each physiological parameter and time point from the validation set (physiological status test validation cohort). NA=not available

	Patient S03			Patient S05			Patient S11
Time point	**T2**	**T3**	**T4**	**T1**	**T3**	**T4**	**T3**
Albumin [g/L]	13.9	23	28.3	21.4	20.5	NA	20.1
ALT [U/L]	29	19	43	39	41	NA	36
AP [U/L]	108	240	294	129	166	NA	203
aPTT [s]	40.2	25.2	26.8	30.6	22.4	NA	30
AST [U/L]	27	23	71	11	39	NA	27
Bilirubin [µM]	27.3664	17.104	6.8416	13.6832	8.552	NA	39.3392
Creatinine [µM]	80.83	39.65	32.79	91.50	38.13	NA	105.99
CRP [mg/dL]	21.4	14.32	9.85	26.85	1.98	NA	14.1
Erythrocytes [1/pL]	2.9	3.3	2.7	3.9	3	NA	2.9
Hemoglobine [g/dL]	9.1	10.2	8.1	11.8	9.4	NA	8.6
Hematocrite [%]	26	29	25	36	27	NA	29
Leukocytes [gpt/L]	16.87	24.13	14.36	26.32	17.84	NA	15.66
PCT [ng/mL]	1.78	0.20	0.08	2.62	0.13	0.04	1.29
Potassium [mM]	4.43	4.49	4.53	4.39	4.65	NA	5.07
Quick [%]	61.8	97.8	108.7	80.6	95.8	NA	73.7
Sodium [mM]	141	130	136	143	137	NA	153
SOFA	10	1	1	9	1	1	11
Thrombocytes [gpt/L]	148	338	845	193	572	NA	392
Urea [mM]	7.326	5.661	3.663	11.1555	4.995	NA	16.65

	Patient S13	Patient S14	Patient S16	Patient S17	Patient S21		Patient S29
Time point	**T3**	**T3**	**T2**	**T4**	**T2**	**T3**	**To**
Albumin [g/L]	NA	22.9	18.9	28.6	24.8	21.3	21.3
ALT [U/L]	45	41	14	75	14	9	18
AP [U/L]	116	177	54	120	103	85	50
aPTT [s]	33.7	24.5	31.3	26.9	28.5	30.3	41
AST [U/L]	20	57	21	62	22	19	43
Bilirubin [µM]	5.1312	8.552	8.552	3.4208	5.1312	3.4208	17.104
Creatinine [µM]	93.79	47.28	83.12	57.95	223.42	73.97	92.27
CRP [mg/dL]	11.6	8.65	17.14	16.97	20.29	13.65	4.75
Erythrocytes [1/pL]	2.9	3	2.7	2.1	3.2	2.6	2.8
Hemoglobine [g/dL]	8.8	8.5	7.9	6.7	9.9	8.2	8.2
Hematocrite [%]	27	26	24	20	32	25	25
Leukocytes [gpt/L]	19.92	19.66	4.55	8.14	17.23	14.69	2.58
PCT [ng/mL]	0.31	0.42	1.32	0.18	50.81	2.56	15.4
Potassium [mM]	4.7	4.37	4.26	4.28	4.83	4.67	3.82
Quick [%]	100.2	111.4	85.4	100.2	100.8	92.3	65.1
Sodium [mM]	145	148	148	135	136	140	148
SOFA	7	10	12	1	12	9	13
Thrombocytes [gpt/L]	207	347	173	347	71	149	143
Urea [mM]	18.981	18.8145	16.317	4.4955	18.8145	11.4885	16.65

	Patient S27			Patient S30		Patient S32	
Time point	**To**	**To**	**To**	**T2**	**T4**	**T2**	**T3**
Albumin [g/L]	17.1	17.1	23.3	20.6	24.1	22.7	22.2

ALT [U/L]	16	16	17	188	32	47	34
AP [U/L]	34	34	230	153	289	113	126
aPTT [s]	35.9	35.9	27	23.9	23.6	24.4	24.9
AST [U/L]	29	29	29	322	31	90	30
Bilirubin [µM]	6.8416	6.8416	18.8144	13.6832	10.2624	8.552	11.9728
Creatinine [µM]	137.26	137.26	66.34	124.29	47.28	74.73	68.63
CRP [mg/dL]	18.58	18.58	30.89	26.01	15.4	24.16	14.97
Erythrocytes [1/pL]	3.2	3.2	2.4	2.5	2.7	3	2.7
Hemoglobine [g/dL]	9	9	6.5	7.3	8	9.6	8.6
Hematocrite [%]	29	29	20	22	24	30	26
Leukocytes [gpt/L]	17.81	17.81	13.67	34.96	9.52	19.34	15.98
PCT [ng/mL]	178.72	178.72	1.14	6.55	0.13	2.72	0.26
Potassium [mM]	5.18	5.18	3.81	4.43	4.05	5.17	4.47
Quick [%]	49.1	49.1	83.7	106	78.6	106	87.6
Sodium [mM]	143	143	136	143	143	148	144
SOFA	9	9	10	12	4	12	4
Thrombocytes [gpt/L]	248	248	1127	505	751	258	327
Urea [mM]	9.324	9.324	4.995	7.659	3.33	11.655	12.8205

	Patient S35		**Patient S41**		**Patient S44**	
Time point	T3	T4	T3	**To**	**T1**	**T6**
Albumin [g/L]	24.3	19.1	24.3	17.3	15.2	19.9
ALT [U/L]	9	23	9	282	287	51
AP [U/L]	112	98	112	37	41	109
aPTT [s]	29.2	40.9	29.2	38.2	53.3	39.9
AST [U/L]	21	37	21	458	415	136
Bilirubin [µM]	11.9728	11.9728	11.9728	32.4976	22.2352	8.552
Creatinine [µM]	57.19	42.70	57.19	269.94	232.57	237.15
CRP [mg/dL]	13.72	8.7	13.72	1.92	28.18	6.32
Erythrocytes [1/pL]	3.2	2.7	3.2	3.5	2.7	2.1
Hemoglobine [g/dL]	9.3	7.9	9.3	10.2	8.1	6.4
Hematocrite [%]	28	25	28	30	24	20
Leukocytes [gpt/L]	10.26	11.2	10.26	18.15	16.17	20.27
PCT [ng/mL]	3.98	0.47	3.98	75.36	90.63	3.67
Potassium [mM]	4.2	4.58	4.2	5.47	5.89	4.41
Quick [%]	82.4	87.1	82.4	67.2	84	51.4
Sodium [mM]	143	159	143	144	143	155
SOFA	4	12	4	15	13	11
Thrombocytes [gpt/L]	263	53	263	221	203	161
Urea [mM]	9.324	22.977	9.324	18.315	14.8185	43.956

	Patient S45		**Patient S46**		**Patient S47**		**Patient S51**
Time point	**To**	**T3**	**To**	**T3**	**To**	**T3**	To
Albumin [g/L]	27.3	29.7	25.7	28	NA	NA	NA
ALT [U/L]	18	14	96	205	24	NA	NA
AP [U/L]	155	150	53	107	52	NA	49
aPTT [s]	28.7	30.5	31.5	53.3	27.1	22.8	39.4
AST [U/L]	32	23	87	77	44	NA	24
Bilirubin [µM]	3.4208	3.4208	64.9952	97.4928	18.8144	NA	15.3936
Creatinine [µM]	240.96	247.06	158.61	222.66	465.91	143.36	130.39
CRP [mg/dL]	23.93	17.61	16.71	30.03	33.64	NA	36.72

Erythrocytes [1/pL]	4	3.5	2.9	3.4	4.6	3.7	4.5
Hemoglobine [g/dL]	10.5	8.9	8.6	10.3	14.5	11.1	14.2
Hematocrite [%]	34	28	24	30	40	35	41
Leukocytes [gpt/L]	17.33	26.83	4.39	17.6	8.23	15.01	2.57
PCT [ng/mL]	0.40	0.36	35.00	5.71	64.40	1.35	21.61
Potassium [mM]	4.73	4.98	4.23	4.82	3.6	5.16	4.35
Quick [%]	101.3	100.9	82.4	73.5	65.2	76.3	27.4
Sodium [mM]	141	145	146	154	137	155	138
SOFA	11	12	17	15	14	4	12
Thrombocytes [gpt/L]	457	656	83	277	281	348	172
Urea [mM]	11.4885	15.4845	7.992	25.4745	29.304	16.65	12.1545

Key: ALT=Alanine aminotransferase, AP=Alkaline phosphatase, AST=Aspartate aminotransferase, aPTT= activated Partial Thromboplastin Time, CRP=C-reactive protein, PCT=Procalcitonin, SOFA=Sequential organ failure assessment, RRT=Renal replacement therapy; Note: NA reflects missing values (not available)

Supplement Table 4 – Patient descriptive data, pathogen burden, outcome and infection type summary for sepsis cases in infection type cohort. Early death reflects patients that died within two days after ICU admission. Survivors reached the end of the study at 28 days after ICU admission.

Patient	Age	Sex	Pathogen burden	Outcome	Infection type
01-061	61-70	Female	*Escherichia coli, Eubacterium rectale*	Recovery	Bacterial
01-078	71-80	Male	*Bacteroides spp., E. coli*	Recovery	Bacterial
01-001	71-80	Male	*Morganella morganii, Bacteroides fragilis*	Recovery	Bacterial
06-016	61-70	Male	*Propionibacterium acnes*	Early death	Bacterial
03-004	71-80	Female	*Proteus mirabilis, E. coli*	Recovery	Bacterial
01-113	71-80	Male	*Enterococcus faecium, Ruminococcus sp. SR1/5, Bacteroides vulgatus*	Early death	Bacterial
06-017	51-60	Female	*Haemophilus parainfluenzae, Veillonella parvula*	Recovery	Bacterial
06-037	71-80	Male	*Escherichia coli, Enterococcus faecium*	Recovery	Bacterial
01-075	61-70	Male	*Enterococcus faecium, Klebsiella pneumoniae, Enterobacter cloacae*	Recovery	Bacterial
01-118	71-80	Male	*Enterobacter aerogenes, Escherichia coli*	Recovery	Bacterial
01-104	71-80	Male	Human herpesvirus 5	Recovery	Viral
03-030	71-80	Male	Human herpesvirus 6B	Recovery	Viral
03-046	81-90	Male	Human herpesvirus 1	Recovery	Viral
05-015	41-50	Female	Human herpesvirus 6A	Recovery	Viral
08-015	51-60	Male	Human mastadenovirus C	Recovery	Viral
09-047	81-90	Male	Human herpesvirus 6B	Recovery	Viral
11-014	61-70	Male	Human herpesvirus 5, Human herpesvirus 1	Recovery	Viral
01-002	51-60	Male	Human herpesvirus 6B	Recovery	Viral
POP6_T1	N.A.	N.A.	n.d.	Recovery	Non-infected
POP6_T2	N.A.	N.A.	n.d.	Recovery	Non-infected
POP1_T1	N.A.	N.A.	n.d.	Recovery	Non-infected
POP11_T2	N.A.	N.A.	n.d.	Recovery	Non-infected
POP8_T1	N.A.	N.A.	n.d.	Recovery	Non-infected

Supplement Table 5 – Patient metadata summary statistics for bacterial and viral sepsis cases (infection type cohort). SD = Standard deviation, IQ = Inter quartile

Infection type	**Bacterial**	**Viral**
Age		
N	10	8
Mean (SD)	71.1 (6.79)	66.4 (15.69)
Median	71.5	69.5
Range	(57; 79)	(41; 86)
IQ range	(68.0; 76.0)	(54.0; 78.5)
Albumin [g/L]		
N	9	6
Mean (SD)	27.73 (2.310)	24.78 (7.776)
Median	26.60	26.20
Range	(25.7; 32.2)	(11.0; 32.5)
IQ range	(26.00; 29.00)	(21.75; 31.00)
ALT [U/L]		
N	10	7
Mean (SD)	148.1 (158.39)	465.7 (734.08)
Median	67.0	268.5
Range	(22; 473)	(24; 2086)
IQ range	(43.0; 291.0)	(48.0; 494.0)
AP [U/L]		
N	10	6
Mean (SD)	120.1 (66.60)	106.7 (62.46)
Median	104.5	89.0
Range	(63; 270)	(48; 192)
IQ range	(76.0; 122.5)	(48.6; 174.0)
aPTT [s]		
N	10	8
Mean (SD)	39.0 (10.08)	34.9 (7.94)
Median	38.5	33.8
Range	(25; 58)	(24; 48)
IQ range	(33.0; 47.0)	(29.5; 40.5)
AST [U/L]		
N	10	6
Mean (SD)	240.6 (316.32)	751.2 (648.67)
Median	117.0	685.0
Range	(34; 1080)	(72; 1747)
IQ range	(68.5; 294.0)	(128.0; 1190.0)
Bilirubin [µM]		
N	10	8
Mean (SD)	31.984 (24.2498)	19.087 (16.5312)
Median	28.222	10.981
Range	(4.28; 80.39)	(8.55; 47.89)
IQ range	(12.828; 53.878)	(8.980; 28.164)
BMI		
N	10	8
Mean (SD)	26.12 (3.748)	26.16 (4.787)
Median	27.20	26.99
Range	(17.9; 30.1)	(18.3; 31.8)
IQ range	(24.62; 29.07)	(22.33; 30.27)
Creatinine [µM]		

Infection type		
	Bacterial	**Viral**
N	10	8
Mean (SD)	137.600 (91.0104)	220.083 (139.6112)
Median	88.836	179.197
Range	(56.05; 269.94)	(61.00; 494.50)
IQ range	(73.204; 264.220)	(129.043; 294.340)
CRP [mg/dL]		
N	10	7
Mean (SD)	19.04 (12.300)	19.47 (10.791)
Median	17.25	15.53
Range	(3.1; 39.4)	(5.8; 37.1)
IQ range	(12.17; 27.02)	(11.90; 29.03)
Sex		
N	10	8
Male	7 (70.0%)	7 (87.5%)
Female	3 (30.0%)	1 (12.5%)
GFR [mL/min]		
N	9	8
Mean (SD)	53.28 (23.440)	34.10 (22.929)
Median	57.50	26.90
Range	(16.4; 87.0)	(10.8; 81.0)
IQ range	(52.95; 59.70)	(18.28; 45.33)
Hematocrit [%]		
N	10	8
Mean (SD)	28.9 (4.13)	30.9 (5.24)
Median	27.9	30.1
Range	(24; 38)	(25; 41)
IQ range	(26.0; 29.9)	(26.8; 33.5)
INR		
N	10	8
Mean (SD)	1.540 (0.3237)	1.326 (0.4704)
Median	1.450	1.253
Range	(1.26; 2.37)	(0.90; 2.40)
IQ range	(1.305; 1.600)	(1.025; 1.375)
Lactate [mM]		
N	10	8
Mean (SD)	6.473 (5.3764)	2.866 (1.9934)
Median	4.249	2.077
Range	(1.00; 18.97)	(0.65; 5.98)
IQ range	(3.258; 9.100)	(1.325; 4.750)
Leukocytes [gpt/I]		
N	10	8
Mean (SD)	16.018 (8.9481)	16.486 (6.9768)
Median	16.223	14.860
Range	(2.17; 34.80)	(9.55; 30.73)
IQ range	(8.900; 20.400)	(11.085; 19.860)
PCT [ng/mL]		
N	10	8
Mean (SD)	55.01 (80.582)	10.73 (11.803)
Median	31.93	5.74
Range	(2.1; 268.4)	(0.4; 34.3)
IQ range	(5.45; 55.96)	(2.14; 17.68)

Infection type		
	Bacterial	**Viral**
Potassium [mM]		
N	10	8
Mean (SD)	4.75 (0.439)	4.48 (0.542)
Median	4.77	4.57
Range	(4.0; 5.4)	(3.6; 5.2)
IQ range	(4.55; 4.95)	(4.08; 4.88)
Protein [g/L]		
N	7	5
Mean (SD)	45.37 (6.583)	48.20 (9.549)
Median	48.00	50.70
Range	(32.5; 51.0)	(37.7; 59.6)
IQ range	(40.45; 49.10)	(39.00; 54.00)
Quick [%]		
N	10	8
Mean (SD)	50.0 (10.87)	70.3 (27.90)
Median	52.5	64.3
Range	(28; 62)	(30; 111)
IQ range	(44.5; 58.0)	(53.0; 93.8)
RRT		
N	10	8
No	6 (60.0%)	7 (87.5%)
Yes	4 (40.0%)	1 (12.5%)
Septic shock		
N	10	8
No	2 (20.0%)	3 (37.5%)
Yes	8 (80.0%)	5 (62.5%)
Sodium [mM]		
N	10	8
Mean (SD)	144.65 (5.088)	141.94 (8.287)
Median	145.00	140.00
Range	(136.5; 151.5)	(132.0; 155.0)
IQ range	(140.00; 148.00)	(135.50; 148.75)
SOFA		
N	10	7
Mean (SD)	9.7 (3.53)	8.9 (3.13)
Median	10.0	8.0
Range	(4; 15)	(5; 13)
IQ range	(7.0; 12.0)	(6.0; 12.0)
Thrombocytes [gpt/I]		
N	10	8
Mean (SD)	194.7 (129.37)	195.1 (141.16)
Median	143.8	181.5
Range	(42; 392)	(52; 494)
IQ range	(105.0; 330.0)	(85.5; 240.5)
Urea [mM]		
N	10	8
Mean (SD)	17.210 (12.5304)	20.324 (6.1455)
Median	12.446	20.754
Range	(5.89; 43.00)	(9.74; 28.50)
IQ range	(8.242; 27.056)	(16.209; 25.213)
Vasopressors or inotropes		

Infection type		
	Bacterial	**Viral**
N	10	8
No	1 (10.0%)	1 (12.5%)
Yes	9 (90.0%)	7 (87.5%)

Key: ALT=Alanine aminotransferase, AP=Alkaline phosphatase, AST=Aspartate aminotransferase, BMI=Body-Mass-Index, CRP=C-reactive protein, GFR=Glomerular filtration rate, INR=International normalized ratio, PCT=Procalcitonin, SOFA=Sequential organ failure assessment, RRT=Renal replacement therapy; Note: N reflects non-missing values

Supplement Table 6 - Patient metadata differences between bacterial and viral sepsis cases (infection type cohort). Categorical features are marked with an asterisk ('*'). P values were obtained from two-sided Wilcoxon rank-sum test for numerical features, and Fisher's exact test for categorical features. Area under ROC curves ('AUROC') is reported with 95% confidence interval boundaries ('AUROC_ci_l' = lower boundary, 'AUROC_ci_h' = higher boundary). From the ROC analysis per metric, the decision threshold ('Thresh') that maximizes the sum of specificity ('Spec') and sensitivity ('Sens') is indicated with the corresponding specificity and sensitivity. ALT=Alanine aminotransferase, AP=Alkaline phosphatase, AST=Aspartate aminotransferase, BMI=Body-Mass-Index, CRP=C-reactive protein, GFR=Glomerular filtration rate, INR=International normalized ratio, PCT=Procalcitonin, SOFA=Sequential organ failure assessment, RRT=Renal replacement therapy.

Infection type – Bacterial vs. Viral							
Metric	**P value**	**AUROC**	**AUROC_ci_l**	**AUROC_ci_h**	**Thresh**	**Spec**	**Sens**
Age	0.689	0.563	0.236	0.889	64.500	0.900	0.500
Albumin	0.596	0.593	0.209	0.977	25.575	1.000	0.500
ALT	0.407	0.629	0.335	0.922	189.000	0.700	0.571
AP	0.447	0.625	0.299	0.951	103.000	0.600	0.667
aPTT	0.374	0.631	0.359	0.903	32.000	0.800	0.500
AST	0.093	0.767	0.497	1.000	446.750	0.900	0.667
Bilirubin	0.142	0.713	0.451	0.974	13.683	0.700	0.750
BMI	0.689	0.438	0.133	0.742	26.454	0.600	0.500
Creatinine	0.198	0.688	0.415	0.960	132.284	0.700	0.750
CRP	0.961	0.486	0.188	0.783	14.485	0.700	0.429
Sex*	0.588	NA	NA	NA	NA	NA	NA
GFR	0.112	0.736	0.460	1.000	51.800	0.778	0.875
Hematocrit	0.477	0.606	0.324	0.888	30.775	0.800	0.500
INR	0.061	0.769	0.511	1.000	1.375	0.700	0.750
Lactate	0.142	0.713	0.456	0.969	2.779	0.800	0.625
Leukocytes	1.000	0.500	0.211	0.789	15.890	0.600	0.625
PCT	0.083	0.750	0.515	0.985	8.068	0.700	0.625
Potassium	0.424	0.619	0.338	0.899	4.425	0.800	0.500
Protein	0.516	0.629	0.215	1.000	49.900	0.857	0.600
Quick	0.068	0.763	0.503	1.000	63.000	1.000	0.625
RRT*	0.314	NA	NA	NA	NA	NA	NA
Septic_shock*	0.608	NA	NA	NA	NA	NA	NA
Sodium	0.477	0.606	0.312	0.901	142.500	0.700	0.625
SOFA	0.659	0.571	0.280	0.863	8.500	0.600	0.571
Thrombocytes	0.894	0.525	0.232	0.818	151.500	0.600	0.625
Urea	0.230	0.675	0.390	0.960	15.751	0.700	0.875
Vasopressors_inotropes*	1.000	NA	NA	NA	NA	NA	NA

Supplement Table 7 – Patient descriptive data, pathogen burden, infection type and outcome for sepsis cases in outcome cohort. Early death reflects septic patients that died within two days after ICU admission. Recovery reached the end of the study at 28 days after ICU admission.

Patient	Age	Sex	Pathogen burden	Infection type	Outcome
01-001	71-80	Male	*Morganella morganii, Bacteroides fragilis*	Bacterial	Early death
01-002	51-60	Male	Human herpesvirus 6B	Viral	Early death
01-005	81-90	Male	*Enterobacter cloacae, Escherichia coli, Cronobacter sakazakii*	Bacterial	Early death
01-025	71-80	Female	*Escherichia coli, Klebsiella pneumoniae*	Bacterial	Early death
01-036	71-80	Male	*Enterobacter cloacae*	Bacterial	Early death
01-049	81-90	Male	*Escherichia coli*	Bacterial	Early death
01-061	61-70	Female	*Escherichia coli, Eubacterium rectale*	Bacterial	Early death
01-075	61-70	Male	*Enterococcus faecium, Klebsiella pneumoniae, Enterobacter cloacae*	Bacterial	Early death
01-078	71-80	Male	*Bacteroides spp., Escherichia coli*	Bacterial	Early death
02-037	51-60	Female	*Escherichia coli, Bacteroides fragilis, Bacteroides vulgatus, Alistipes shahii*	Bacterial	Early death
03-004	71-80	Female	*Proteus mirabilis, Escherichia coli*	Bacterial	Early death
03-009	51-50	Male	*Escherichia coli, Bacteroides fragilis*	Bacterial	Early death
06-017	51-60	Female	*Haemophilus parainfluenzae, Veillonella parvula*	Bacterial	Early death
06-037	71-80	Male	*Escherichia coli, Enterococcus faecium*	Bacterial	Early death
08-005	71-80	Male	*Enterobacter cloacae*	Bacterial	Early death
08-015	51-60	Male	Human mastadenovirus C	Viral	Early death
08-026	71-80	Male	*Escherichia coli*	Bacterial	Early death
08-056	71-80	Male	*Enterococcus faecalis*	Bacterial	Early death
08-059	81-90	Male	*Escherichia coli*	Bacterial	Early death
09-031	71-80	Male	*Escherichia coli*	Bacterial	Early death
16-004	51-60	Female	*Escherichia coli*	Bacterial	Early death
01-006	71-80	Male	*Staphylococcus aureus,* Torque teno virus	Bacterial, Viral	Early death
01-012	61-70	Female	*Pseudomonas aeruginosa, Lactobacillus fermentum*	Bacterial	Recovery
01-064	51-60	Female	*Escherichia coli*	Bacterial	Recovery
01-070	71-80	Female	*Streptococcus thermophilus, Propionibacterium freudenreichii*	Bacterial	Recovery
01-071	81-90	Male	*Enterococcus faecium, Candida glabrata, Escherichia coli*	Bacterial, Fungal	Recovery
01-074	61-70	Female	*Candida glabrata, Klebsiella oxytoca*	Bacterial, Fungal	Recovery
01-077	81-90	Male	*Klebsiella oxytoca, Streptococcus pneumoniae, Lactobacillus salivarius,* Human herpesvirus 6B	Bacterial, Viral	Recovery
01-086	81-90	Female	*Staphylococcus aureus*	Bacterial	Recovery
01-090	61-70	Female	*Escherichia coli*	Bacterial	Recovery
01-092	61-70	Male	*Escherichia coli*	Bacterial	Recovery
01-113	71-80	Male	*Enterococcus faecium, Ruminococcus sp. SR1/5, Bacteroides vulgatus*	Bacterial	Recovery
02-018	81-90	Male	*Proteus mirabilis*	Bacterial	Recovery

02-043	81-90	Male	*Escherichia coli*	Bacterial	Recovery
03-029	81-90	Male	*Escherichia coli*	Bacterial	Recovery
03-047	51-60	Female	*Pseudomonas aeruginosa, Escherichia coli*	Bacterial	Recovery
04-004	41-50	Male	*Clostridium perfringens*	Bacterial	Recovery
05-005	51-60	Male	*Roseburia intestinalis, Escherichia coli, Bacteroides vulgatus*	Bacterial	Recovery
06-003	61-70	Male	*Staphylococcus aureus*	Bacterial	Recovery
06-016	61-70	Male	*Propionibacterium acnes*	Bacterial	Recovery
08-034	61-70	Male	*Streptococcus spp.*	Bacterial	Recovery
08-044	81-90	Female	*Staphylococcus aureus, Lactobacillus fermentum*	Bacterial	Recovery
22-001	71-80	Male	*Escherichia coli, Bacteroides spp.*	Bacterial	Recovery

Supplement Table 8 – Patient metadata summary statistics for early death and recovery sepsis cases (outcome cohort). SD = Standard deviation, IQ = Inter quartile

Outcome	Early death	Recovery
Age		
N	21	22
Mean (SD)	69.9 (10.76)	70.5 (11.73)
Median	74.0	70.5
Range	(50; 84)	(45; 87)
IQ range	(57.0; 77.0)	(65.0; 82.0)
Albumin [g/L]		
N	15	19
Mean (SD)	27.61 (3.414)	27.43 (5.033)
Median	26.60	26.20
Range	(21.8; 33.0)	(19.3; 39.0)
IQ range	(25.65; 31.10)	(23.00; 30.70)
ALT [U/L]		
N	16	19
Mean (SD)	87.3 (106.37)	98.8 (145.46)
Median	44.5	38.0
Range	(7; 333)	(5; 509)
IQ range	(26.0; 71.0)	(21.0; 112.0)
AP [U/L]		
N	15	18
Mean (SD)	100.9 (52.90)	132.2 (94.21)
Median	82.0	114.3
Range	(35; 206)	(23; 394)
IQ range	(60.0; 143.0)	(63.0; 194.0)
aPTT [s]		
N	21	22
Mean (SD)	33.1 (8.98)	42.9 (13.90)
Median	32.0	41.3
Range	(24; 58)	(24; 79)
IQ range	(26.5; 36.5)	(32.0; 49.5)
AST [U/L]		
N	15	17
Mean (SD)	156.1 (295.78)	216.4 (305.26)
Median	63.0	72.0
Range	(8; 1190)	(20; 1080)
IQ range	(33.5; 149.0)	(37.5; 177.0)
Bilirubin [µM]		
N	18	21
Mean (SD)	18.280 (12.9571)	19.542 (13.9339)
Median	14.966	17.000
Range	(4.28; 54.73)	(3.42; 53.88)
IQ range	(10.262; 25.656)	(10.262; 22.235)
BMI		
N	21	21
Mean (SD)	27.62 (7.028)	27.86 (10.385)
Median	27.43	25.39
Range	(17.9; 51.2)	(15.9; 69.2)
IQ range	(23.88; 29.38)	(23.88; 27.78)
Creatinine [µM]		

Outcome		
	Early death	**Recovery**
N	21	22
Mean (SD)	155.931 (130.4415)	199.389 (110.8686)
Median	90.361	191.969
Range	(32.79; 494.50)	(30.50; 513.00)
IQ range	(73.204; 181.485)	(136.495; 269.558)
CRP [mg/dL]		
N	20	20
Mean (SD)	18.48 (12.349)	19.17 (10.833)
Median	15.36	15.98
Range	(3.1; 54.2)	(4.7; 44.9)
IQ range	(12.31; 25.95)	(12.24; 23.42)
Sex		
N	21	22
Male	15 (71.4%)	14 (63.6%)
Female	6 (28.6%)	8 (36.4%)
GFR [mL/min]		
N	20	19
Mean (SD)	52.44 (30.011)	28.22 (16.158)
Median	55.50	22.50
Range	(10.5; 101.9)	(9.1; 60.0)
IQ range	(21.58; 76.75)	(16.35; 39.65)
Hematocrit [%]		
N	21	22
Mean (SD)	29.9 (6.25)	30.7 (4.75)
Median	28.5	30.0
Range	(19; 41)	(23; 38)
IQ range	(26.0; 36.0)	(25.9; 35.0)
INR		
N	21	22
Mean (SD)	1.394 (0.3392)	1.681 (0.5847)
Median	1.305	1.523
Range	(0.90; 2.26)	(1.03; 3.70)
IQ range	(1.150; 1.500)	(1.330; 1.900)
Lactate [mM]		
N	21	22
Mean (SD)	4.795 (4.5987)	7.386 (5.2218)
Median	2.800	6.368
Range	(0.75; 18.97)	(1.00; 18.70)
IQ range	(1.277; 6.000)	(2.581; 12.300)
Leukocytes [gpt/I]		
N	21	22
Mean (SD)	17.791 (10.1078)	14.615 (12.6624)
Median	15.820	10.550
Range	(4.44; 43.68)	(0.66; 43.62)
IQ range	(10.320; 22.270)	(5.575; 23.500)
PCT [ng/mL]		
N	21	22
Mean (SD)	52.22 (86.922)	65.10 (89.933)
Median	5.87	14.39
Range	(0.1; 268.4)	(0.6; 279.2)
IQ range	(2.14; 50.36)	(7.12; 100.00)

Outcome	Early death	Recovery
Potassium [mM]		
N	21	22
Mean (SD)	4.36 (0.703)	4.67 (0.480)
Median	4.35	4.71
Range	(3.3; 5.4)	(3.5; 5.4)
IQ range	(3.95; 4.93)	(4.45; 5.00)
Protein [g/L]		
N	14	13
Mean (SD)	48.58 (8.706)	44.80 (9.309)
Median	49.20	43.70
Range	(32.5; 60.1)	(29.7; 58.0)
IQ range	(40.45; 55.70)	(36.90; 55.00)
Quick [%]		
N	21	22
Mean (SD)	61.2 (20.57)	51.9 (17.14)
Median	58.0	54.3
Range	(31; 111)	(15; 96)
IQ range	(49.5; 77.0)	(37.0; 60.0)
RRT		
N	21	22
No	18 (85.7%)	16 (72.7%)
Yes	3 (14.3%)	6 (27.3%)
Septic shock		
N	21	22
No	7 (33.3%)	2 (9.1%)
Yes	14 (66.7%)	20 (90.9%)
Sodium [mM]		
N	21	22
Mean (SD)	143.40 (5.183)	140.41 (6.607)
Median	144.00	138.75
Range	(132.0; 151.5)	(128.0; 160.5)
IQ range	(140.00; 147.00)	(136.00; 143.50)
SOFA		
N	21	22
Mean (SD)	7.9 (3.68)	11.4 (3.11)
Median	8.0	11.0
Range	(1; 14)	(6; 18)
IQ range	(5.0; 11.0)	(9.0; 13.0)
Thrombocytes [gpt/I]		
N	21	22
Mean (SD)	250.6 (143.36)	161.0 (136.67)
Median	240.0	110.0
Range	(86; 697)	(40; 566)
IQ range	(140.0; 330.0)	(82.0; 193.5)
Urea [mM]		
N	21	22
Mean (SD)	14.119 (9.8640)	22.721 (15.3671)
Median	9.990	21.006
Range	(5.00; 43.00)	(3.41; 55.85)
IQ range	(7.243; 18.200)	(11.405; 30.137)
Vasopressors or inotropes		

Outcome		
	Early death	**Recovery**
N	21	22
No	6 (28.6%)	1 (4.5%)
Yes	15 (71.4%)	21 (95.5%)

Key: ALT=Alanine aminotransferase, AP=Alkaline phosphatase, AST=Aspartate aminotransferase, BMI=Body-Mass-Index, CRP=C-reactive protein, GFR=Glomerular filtration rate, INR=International normalized ratio, PCT=Procalcitonin, SOFA=Sequential organ failure assessment, RRT=Renal replacement therapy; Note: N reflects non-missing values

Supplement Table 9 - Patient metadata differences between early death and survival sepsis cases (outcome cohort). Categorical features are marked with an asterisk ('*'). P values were obtained from two-sided Wilcoxon rank-sum test for numerical features, and Fisher's exact test for categorical features. Metrics with a p value <= 0.05 are highlighted in bold font. Area under ROC curves ('AUROC') is reported with 95% confidence interval boundaries ('AUROC_ci_l' = lower boundary, 'AUROC_ci_h' = higher boundary). From the ROC analysis per metric, the decision threshold ('Thresh') that maximizes the sum of specificity ('Spec') and sensitivity ('Sens') is indicated with the corresponding specificity and sensitivity. ALT=Alanine aminotransferase, AP=Alkaline phosphatase, AST=Aspartate aminotransferase, BMI=Body-Mass-Index, CRP=C-reactive protein, GFR=Glomerular filtration rate, INR=International normalized ratio, PCT=Procalcitonin, SOFA=Sequential organ failure assessment, RRT=Renal replacement therapy.

Outcome – Early death vs. Recovery							
Metric	**P value**	**AUROC**	**AUROC_ci_l**	**AUROC_ci_h**	**Thresh**	**Spec**	**Sens**
Age	0.990	0.498	0.318	0.678	71.500	0.571	0.591
Albumin	0.835	0.523	0.323	0.723	25.525	0.800	0.368
ALT	0.716	0.538	0.340	0.736	42.000	0.563	0.579
AP	0.459	0.578	0.376	0.779	106.500	0.733	0.556
aPTT	**0.011**	**0.727**	**0.574**	**0.881**	**37.250**	**0.810**	**0.636**
AST	0.571	0.561	0.355	0.767	70.250	0.600	0.529
Bilirubin	0.877	0.516	0.328	0.704	16.650	0.611	0.524
BMI	0.753	0.529	0.349	0.710	26.398	0.571	0.619
Creatinine	0.091	0.652	0.474	0.829	125.747	0.667	0.773
CRP	0.797	0.525	0.341	0.709	15.605	0.550	0.550
Sex*	0.747	NA	NA	NA	NA	NA	NA
GFR	**0.018**	**0.724**	**0.556**	**0.892**	**47.275**	**0.650**	**0.842**
Hematocrit	0.601	0.548	0.368	0.727	29.175	0.571	0.636
INR	**0.039**	**0.685**	**0.523**	**0.847**	**1.323**	**0.571**	**0.773**
Lactate	0.050	0.675	0.511	0.839	6.050	0.762	0.545
Leukocytes	0.178	0.621	0.447	0.795	13.485	0.667	0.591
PCT	0.220	0.610	0.434	0.786	9.780	0.619	0.682
Potassium	0.138	0.633	0.459	0.808	4.398	0.524	0.773
Protein	0.244	0.635	0.416	0.853	44.900	0.714	0.615
Quick	0.198	0.616	0.445	0.787	58.500	0.476	0.727
RRT*	0.457	NA	NA	NA	NA	NA	NA
Septic_shock*	0.069	NA	NA	NA	NA	NA	NA
Sodium	**0.030**	**0.694**	**0.531**	**0.857**	**139.250**	**0.810**	**0.545**
SOFA	**0.003**	**0.760**	**0.617**	**0.902**	**9.500**	**0.667**	**0.682**
Thrombocytes	**0.008**	**0.738**	**0.585**	**0.891**	**118.250**	**0.857**	**0.591**
Urea	**0.048**	**0.677**	**0.510**	**0.845**	**11.280**	**0.571**	**0.773**
Vasopressors_inotropes*	**0.046**	**NA**	**NA**	**NA**	**NA**	**NA**	**NA**

References

[1] Alix-Panabières C, Marchetti D, Lang JE. Liquid biopsy: from concept to clinical application. Scientific Reports 2023 13:1 2023;13:1–4. https://doi.org/10.1038/s41598-023-48501-x.

[2] Cisneros-Villanueva M, Hidalgo-Pérez L, Rios-Romero M, Cedro-Tanda A, Ruiz-Villavicencio CA, Page K, et al. Cell-free DNA analysis in current cancer clinical trials: a review. British Journal of Cancer 2022 126:3 2022;126:391–400. https://doi.org/10.1038/s41416-021-01696-0.

[3] Aucamp J, Bronkhorst AJ, Badenhorst CPS, Pretorius PJ. The diverse origins of circulating cell-free DNA in the human body: a critical re-evaluation of the literature. Biological Reviews 2018;93:1649–83. https://doi.org/10.1111/brv.12413.

[4] Grumaz S, Stevens P, Grumaz C, Decker SO, Weigand MA, Hofer S, et al. Next-generation sequencing diagnostics of bacteremia in septic patients. Genome Med 2016;8:73. https://doi.org/10.1186/s13073-016-0326-8.

[5] Grumaz C, Hoffmann A, Vainshtein Y, Kopp M, Grumaz S, Stevens P, et al. Rapid Next-Generation Sequencing–Based Diagnostics of Bacteremia in Septic Patients. Journal of Molecular Diagnostics 2020;22:405–18. https://doi.org/10.1016/j.jmoldx.2019.12.006.

[6] Long Y, Zhang Y, Gong Y, Sun R, Su L, Lin X, et al. Diagnosis of Sepsis with Cell-free DNA by Next-Generation Sequencing Technology in ICU Patients. Arch Med Res 2016;47:365–71. https://doi.org/10.1016/J.ARCMED.2016.08.004.

[7] Sonntag M, Elgeti VK, Vainshtein Y, Jenner L, Mueller J, Brenner T, et al. Suppression PCR-Based Selective Enrichment Sequencing for Pathogen and Antimicrobial Resistance Detection on Cell-Free DNA in Sepsis—A Targeted, Blood Culture-Independent Approach for Rapid Pathogen and Resistance Diagnostics in Septic Patients. International Journal of Molecular Sciences 2024, Vol 25, Page 5463 2024;25:5463. https://doi.org/10.3390/IJMS25105463.

[8] Rhodes A, Evans LE, Alhazzani W, Levy MM, Antonelli M, Ferrer R, et al. Surviving Sepsis Campaign: International Guidelines for Management of Sepsis and Septic Shock: 2016. Intensive Care Med 2017;43:304–77. https://doi.org/10.1007/s00134-017-4683-6.

[9] Scheer CS, Fuchs C, Gründling M, Vollmer M, Bast J, Bohnert JA, et al. Impact of antibiotic administration on blood culture positivity at the beginning of sepsis: a prospective clinical cohort study. Clinical Microbiology and Infection 2019;25:326–31. https://doi.org/10.1016/j.cmi.2018.05.016.

[10] Vincent JL, Brealey D, Libert N, Abidi NE, O'Dwyer M, Zacharowski K, et al. Rapid Diagnosis of Infection in the Critically Ill, a Multicenter Study of Molecular Detection in Bloodstream Infections, Pneumonia, and Sterile Site Infections. Crit Care Med 2015;43:2283–91. https://doi.org/10.1097/CCM.0000000000001249.

[11] Grumaz S, Grumaz C, Vainshtein Y, Stevens P, Glanz K, Decker SO, 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 2019;47:e394–402. https://doi.org/10.1097/CCM.0000000000003658.

[12] Snyder MW, Kircher M, Hill AJ, Daza RM, Shendure J. Cell-free DNA Comprises an in Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. Cell 2016;164:57–68. https://doi.org/10.1016/j.cell.2015.11.050.

[13] Burnham P, Kim MS, Agbor-Enoh S, Luikart H, Valantine HA, Khush KK, et al. Single-stranded DNA library preparation uncovers the origin and diversity of ultrashort cell-free DNA in plasma. Scientific Reports 2016 6:1 2016;6:1–9. https://doi.org/10.1038/srep27859.

[14] Hudecova I, Smith CG, Hänsel-Hertsch R, Chilamakuri CS, Morris JA, Vijayaraghavan A, et al. Characteristics, origin, and potential for cancer diagnostics of ultrashort plasma cell-free DNA. Genome Res 2022;32:215–27. https://doi.org/10.1101/GR.275691.121/-/DC1.

[15] Ulz P, Perakis S, Zhou Q, Moser T, Belic J, Lazzeri I, et al. Inference of transcription factor binding from cell-free DNA enables tumor subtype prediction and early detection. Nat Commun 2019;10. https://doi.org/10.1038/s41467-019-12714-4.

[16] Müller J, Hartwig C, Sonntag M, Bitzer L, Adelmann C, Vainshtein Y, et al. A novel approach for in vivo DNA footprinting using short double-stranded cell-free DNA from plasma. Genome Res 2024. https://doi.org/10.1101/GR.279326.124.

[17] Singer M, Deutschman CS, Seymour C, Shankar-Hari M, Annane D, Bauer M, et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA 2016;315:801. https://doi.org/10.1001/JAMA.2016.0287.

[18] Rudd KE, Johnson SC, Agesa KM, Shackelford KA, Tsoi D, Kievlan DR, et al. Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the Global Burden of Disease Study. The Lancet 2020;395:200–11. https://doi.org/10.1016/S0140-6736(19)32989-7.

[19] Vincent JL, Moreno R, Takala J, Willatts S, De Mendonça A, Bruining H, et al. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. Intensive Care Med 1996;22:707–10. https://doi.org/10.1007/BF01709751.

[20] Lopes Ferreira F, Peres Bota D, Bross A, Mélot C, Vincent JL. Serial evaluation of the SOFA score to predict outcome in critically ill patients. J Am Med Assoc 2001;286:1754–8. https://doi.org/10.1001/jama.286.14.1754.

[21] Liu C, Suo S, Luo L, Chen X, Ling C, Cao S. SOFA Score in relation to Sepsis: Clinical Implications in Diagnosis, Treatment, and Prognostic Assessment. Comput Math Methods Med 2022;2022. https://doi.org/10.1155/2022/7870434.

[22] Evans L, Rhodes A, Alhazzani W, Antonelli M, Coopersmith CM, French C, et al. Surviving sepsis campaign: international guidelines for management of sepsis and septic shock 2021. Intensive Care Medicine 2021 47:11 2021;47:1181–247. https://doi.org/10.1007/S00134-021-06506-Y.

[23] Jekarl DW, Lee S, Kim M, Kim Y, Woo SH, Lee WJ. Procalcitonin as a prognostic marker for sepsis based on SEPSIS-3. J Clin Lab Anal 2019;33:22996. https://doi.org/10.1002/JCLA.22996.

[24] Samsudin I, Vasikaran SD. Clinical Utility and Measurement of Procalcitonin. Clin Biochem Rev 2017;38:59.

[25] Vijayan AL, Ravindran S, Saikant R, Lakshmi S, Kartik R, Manoj G. Procalcitonin: a promising diagnostic marker for sepsis and antibiotic therapy. J Intensive Care 2017;5. https://doi.org/10.1186/S40560-017-0246-8.

[26] Gregoriano C, Heilmann E, Molitor A, Schuetz P. Role of procalcitonin use in the management of sepsis. J Thorac Dis 2020;12:S5. https://doi.org/10.21037/JTD.2019.11.63.

[27] Brenner T, Decker SO, Grumaz S, Stevens P, Bruckner T, Schmoch T, et al. Next-generation sequencing diagnostics of bacteremia in sepsis (Next GeneSiS-Trial): Study protocol of a prospective, observational, noninterventional, multicenter, clinical trial. Medicine 2018;97. https://doi.org/10.1097/MD.0000000000009868.

[28] Decker SO, Sigl A, Grumaz C, Stevens P, Vainshtein Y, Zimmermann S, et al. Immune-Response Patterns and Next Generation Sequencing Diagnostics for the Detection of Mycoses in Patients with Septic Shock—Results of a Combined Clinical and Experimental Investigation. Int J Mol Sci 2017;18. https://doi.org/10.3390/IJMS18081796.

[29] Bushnell B, Rood J, Singer E. BBMerge - Accurate paired shotgun read merging via overlap. PLoS One 2017;12. https://doi.org/10.1371/JOURNAL.PONE.0185056.

[30] Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. Bioinformatics 2011;27:863–4. https://doi.org/10.1093/BIOINFORMATICS/BTR026.

[31] Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data n.d. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed July 23, 2024).

[32] Sedlazeck FJ, Rescheneder P, Von Haeseler A. NextGenMap: fast and accurate read mapping in highly polymorphic genomes. Bioinformatics 2013;29:2790–1. https://doi.org/10.1093/BIOINFORMATICS/BTT468.

[33] Amemiya HM, Kundaje A, Boyle AP. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. Scientific Reports 2019 9:1 2019;9:1–5. https://doi.org/10.1038/s41598-019-45839-z.

[34] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009;25:2078–9. https://doi.org/10.1093/BIOINFORMATICS/BTP352.

[35] Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, et al. deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res 2016;44:W160–5. https://doi.org/10.1093/NAR/GKW257.

[36] Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol 2008;9. https://doi.org/10.1186/GB-2008-9-9-R137.

[37] Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010;26:841–2. https://doi.org/10.1093/BIOINFORMATICS/BTQ033.

[38] McLeay RC, Bailey TL. Motif Enrichment Analysis: A unified framework and an evaluation on ChIP data. BMC Bioinformatics 2010;11:1–11. https://doi.org/10.1186/1471-2105-11-165/FIGURES/5.

[39] Kulakovskiy I V., Vorontsov IE, Yevshin IS, Sharipov RN, Fedorova AD, Rumynskiy EI, et al. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. Nucleic Acids Res 2018;46:D252–9. https://doi.org/10.1093/NAR/GKX1106.

[40] Risso D, Schwartz K, Sherlock G, Dudoit S. GC-Content Normalization for RNA-Seq Data. BMC Bioinformatics 2011;12:1–17. https://doi.org/10.1186/1471-2105-12-480/FIGURES/7.

[41] Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 2010;26:139–40. https://doi.org/10.1093/BIOINFORMATICS/BTP616.

[42] Wagner GP, Kin K, Lynch VJ. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory in Biosciences 2012;131:281–5. https://doi.org/10.1007/S12064-012-0162-3/FIGURES/2.

[43] Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Series B Stat Methodol 1995;57:289–300. https://doi.org/10.1111/J.2517-6161.1995.TB02031.X.

[44] Müller J, Hartwig C, Sonntag M, Bitzer L, Adelmann C, Vainshtein Y, et al. Liquid footprinting: A novel approach for DNA footprinting using short double-stranded cell-free DNA from plasma. BioRxiv 2024:2024.02.09.579588. https://doi.org/10.1101/2024.02.09.579588.

[45] Rubio I, Osuchowski MF, Shankar-Hari M, Skirecki T, Winkler MS, Lachmann G, et al. Current gaps in sepsis immunology: new opportunities for translational research. Lancet Infect Dis 2019;19:e422–36. https://doi.org/10.1016/S1473-3099(19)30567-5.

[46] Kalantar KL, Neyton L, Abdelghany M, Mick E, Jauregui A, Caldera S, et al. Integrated host-microbe plasma metagenomics for sepsis diagnosis in a prospective cohort of critically ill adults. Nature Microbiology 2022 7:11 2022;7:1805–16. https://doi.org/10.1038/s41564-022-01237-2.

[47] Poore GD, Ko ER, Valente A, Henao R, Sumner K, Hong C, et al. A miRNA host response signature accurately discriminates acute respiratory infection etiologies. Front Microbiol 2018;9:414039. https://doi.org/10.3389/FMICB.2018.02957/BIBTEX.

[48] Llitjos J-F, Carrol ED, Osuchowski MF, Bonneville M, Scicluna BP, Payen D, et al. Enhancing sepsis biomarker development: key considerations from public and private perspectives. Critical Care 2024 28:1 2024;28:1–15. https://doi.org/10.1186/S13054-024-05032-9.

[49] Cano-Gamez E, Burnham KL, Goh C, Allcock A, Malick ZH, Overend L, et al. An immune dysfunction score for stratification of patients with acute infection based on whole-blood gene expression. Sci Transl Med 2022;14. https://doi.org/10.1126/SCITRANSLMED.ABQ4433/SUPPL_FILE/SCITRANSLMED.ABQ4433_MDAR_REPRODUCIBILITY_CHECKLIST.PDF.

[50] Retamar P, Portillo MM, López-Prieto MD, Rodríguez-López F, De Cueto M, García M V., et al. Impact of inadequate empirical therapy on the mortality of patients with bloodstream infections: A propensity score-based analysis. Antimicrob Agents Chemother 2012;56:472–8. https://doi.org/10.1128/AAC.00462-11/ASSET/F651B608-BF9D-4308-B380-1C6D136A5524/ASSETS/GRAPHIC/ZAC9991004550001.JPEG.

[51] Bloos F, Rüddel H, Thomas-Rüddel D, Schwarzkopf D, Pausch C, Harbarth S, et al. Effect of a multifaceted educational intervention for anti-infectious measures on sepsis mortality: a cluster randomized trial. Intensive Care Med 2017;43:1602–12. https://doi.org/10.1007/S00134-017-4782-4.

[52] Liu Y, Xu P, Rivara S, Liu C, Ricci J, Ren X, et al. Clathrin-associated AP-1 controls termination of STING signalling. Nature 2022 610:7933 2022;610:761–7. https://doi.org/10.1038/s41586-022-05354-0.

[53] Faherty CS, Maurelli AT. Staying alive: bacterial inhibition of apoptosis during infection. Trends Microbiol 2008;16:173. https://doi.org/10.1016/J.TIM.2008.02.001.

[54] Guerau-de-Arellano M, Piedra-Quintero ZL, Tsichlis PN. Akt isoforms in the immune system. Front Immunol 2022;13. https://doi.org/10.3389/FIMMU.2022.990874.

[55] Carter-Timofte ME, Paludan SR, Mogensen TH. RNA Polymerase III as a Gatekeeper to Prevent Severe VZV Infections. Trends Mol Med 2018;24:904–15. https://doi.org/10.1016/J.MOLMED.2018.07.009.

[56] Jarrous N, Rouvinski A. RNA polymerase III and antiviral innate immune response. Transcription 2021;12:1. https://doi.org/10.1080/21541264.2021.1890915.

[57] Pokharel SM, Shil NK, Bose S. Autophagy, TGF-β, and SMAD-2/3 signaling regulates interferon-β response in respiratory syncytial virus infected macrophages. Front Cell Infect Microbiol 2016;6:234300. https://doi.org/10.3389/FCIMB.2016.00174/BIBTEX.

[58] Herta T, Bhattacharyya A, Rosolowski M, Conrad C, Gurtner C, Gruber AD, et al. Krueppel-Like Factor 4 Expression in Phagocytes Regulates Early Inflammatory Response and Disease Severity in Pneumococcal Pneumonia. Front Immunol 2021;12:726135. https://doi.org/10.3389/FIMMU.2021.726135/BIBTEX.

[59] McConnell BB, Yang VW. Mammalian Krüppel-Like Factors in Health and Diseases. Physiol Rev 2010;90:1337. https://doi.org/10.1152/PHYSREV.00058.2009.

[60] Ahuja N, Mishra A, Gupta R, Ray S. Biomarkers in sepsis-looking for the Holy Grail or chasing a mirage! World J Crit Care Med 2023;12:188. https://doi.org/10.5492/WJCCM.V12.I4.188.

[61] Gregoriano C, Heilmann E, Molitor A, Schuetz P. Role of procalcitonin use in the management of sepsis. J Thorac Dis 2020;12:S5. https://doi.org/10.21037/JTD.2019.11.63.

[62] Bauer W, Kappert K, Galtung N, Lehmann D, Wacker J, Cheng HK, et al. A Novel 29-Messenger RNA Host-Response Assay From Whole Blood Accurately Identifies Bacterial and Viral Infections in Patients Presenting to the Emergency Department With Suspected Infections: A Prospective Observational Study. Crit Care Med 2021;49:1664–73. https://doi.org/10.1097/CCM.0000000000005119.

[63] Langelier C, Kalantar KL, Moazed F, Wilson MR, Crawford ED, Deiss T, et al. Integrating host response and unbiased microbe detection for lower respiratory tract infection diagnosis in critically ill adults. Proc Natl Acad Sci U S A 2018;115:E12353–62. https://doi.org/10.1073/PNAS.1809700115/SUPPL_FILE/PNAS.1809700115.SD09.XLSX.

[64] Lopes Ferreira F, Peres Bota D, Bross A, Mélot C, Vincent JL. Serial Evaluation of the SOFA Score to Predict Outcome in Critically Ill Patients. JAMA 2001;286:1754–8. https://doi.org/10.1001/JAMA.286.14.1754.

[65] Noparatkailas N, Inchai J, Deesomchok A. Blood Lactate Level and the Predictor of Death in Non-shock Septic Patients. Indian J Crit Care Med 2023;27:93. https://doi.org/10.5005/JP-JOURNALS-10071-24404.

[66] Park H, Lee J, Oh DK, Park MH, Lim CM, Lee SM, et al. Serial evaluation of the serum lactate level with the SOFA score to predict mortality in patients with sepsis. Scientific Reports 2023 13:1 2023;13:1–11. https://doi.org/10.1038/s41598-023-33227-7.

[67] Kyriazopoulou E, Leventogiannis K, Tavoulareas G, Mainas E, Toutouzas K, Mathas C, et al. Presepsin as a diagnostic and prognostic biomarker of severe bacterial infections and COVID-19. Scientific Reports 2023 13:1 2023;13:1–8. https://doi.org/10.1038/s41598-023-30807-5.

[68] Xie Y, Li B, Lin Y, Shi F, Chen W, Wu W, et al. Combining Blood-Based Biomarkers to Predict Mortality of Sepsis at Arrival at the Emergency Department. Med Sci Monit 2021;27:e929527-1. https://doi.org/10.12659/MSM.929527.

[69] Hu C, Li L, Huang W, Wu T, Xu Q, Liu J, et al. Interpretable Machine Learning for Early Prediction of Prognosis in Sepsis: A Discovery and Validation Study. Infect Dis Ther 2022;11:1117–32. https://doi.org/10.1007/S40121-022-00628-6.

[70] Karlsson A, Stassen W, Loutfi A, Wallgren U, Larsson E, Kurland L. Predicting mortality among septic patients presenting to the emergency department–a cross sectional analysis using machine learning. BMC Emerg Med 2021;21:1–8. https://doi.org/10.1186/S12873-021-00475-7/FIGURES/3.

[71] Park JY, Hsu TC, Hu JR, Chen CY, Hsu WT, Lee M, et al. Predicting Sepsis Mortality in a Population-Based National Database: Machine Learning Approach. J Med Internet Res 2022;24(4):E29982 Https://WwwJmirOrg/2022/4/E29982 2022;24:e29982. https://doi.org/10.2196/29982.

6 Discussion and Outlook

Non-genetic cfDNA biomarkers opened up promising opportunities for novel diagnostic approaches (Loy et al., 2024). The primary objective of this thesis was to explore the potential of short double-stranded cfDNA fragments, and thus *in vivo* TF footprints, as novel biomarkers for minimally invasive diagnostics in complex diseases, particularly sepsis.

Chapter 4 of the thesis concentrated on developing a robust method for enriching and quantifying short cfDNA fragments from blood plasma, complemented by a custom bioinformatics pipeline (Müller et al., 2024). This pipeline includes specialized read trimming and filtering prior to mapping, such as sequence complexity filtering, removal of terminal base additions from library preparation, and in silico size selection, in addition to standard base quality trimming and adapter removal. Post-mapping, duplicate reads, multimapped reads, and reads in blacklisted regions are excluded. Peak calling parameters and thresholds were optimized separately for narrow peaks and clusters of these peaks. The differential enrichment analysis method incorporates GC content, region size, and sample sequencing depth adjustments during normalization, ensuring the most accurate estimation of short cfDNA fragment abundances and thereby enhancing the reliability of biomarker identifications. Our approach facilitated high-throughput sequencing and provided valuable insights into the origin of these fragments, the most appropriate data analysis strategy, and the relations of short cfDNA signals to other epigenetic regulatory factors as well as gene transcription. The results confirmed that short cfDNA fragments accumulate at regulatory genomic sites, predominantly originating from transcription factor binding regions. We outline the advantages of short double-stranded cfDNA over short single-stranded cfDNA and nucleosomal cfDNA. Double-stranded fragments show a higher relative enrichment of TF-derived cfDNA compared to single-stranded fragments, which is due to the additional extraction of degraded fragments. This leads to a higher signal-to-noise ratio for TF footprints when analyzed with double-stranded fragments. Nucleosomal cfDNA can be used to analyze TF footprints, but it is an indirect measurement that actually assesses the depletion of nucleosomes. Therefore, the dynamic range of the signals is highly dependent on the sequencing depth, resulting in a higher required sequencing depth compared to short cfDNA. Thus, our methodology, which enables the efficient analysis of short double-stranded cfDNA, overcomes these limitations of the established methods in the field. While chapter 4 demonstrates the method's capability to distinguish between exemplary small patient cohorts, chapter 5 demonstrates the use of this methodology with a larger cohort of sepsis patients (Müller et al., 2024; Sonntag et al., 2025). The analysis revealed that short cfDNA fragments offer comprehensive insights into the host response during sepsis, including information about key organ functions and physiology. The results show superior discrimination of infection types and early pre-

diction of mortality risk compared to established clinical markers, underlining the clinical relevance and potential of short cfDNA biomarkers as well as non-genetic cfDNA biomarkers in sepsis diagnostics.

The publications, Müller et al. (2024) and Sonntag et al. (2025), directly address the development of a methodology for identifying novel non-genetic biomarkers based on short cfDNA fragments and establish a tailored bioinformatic approach for their analysis. Additionally, the publications provide a sophisticated initial evaluation of the diagnostic utility of short cfDNA fragments using sepsis as a case study. While these findings are promising, the current limitations of this research must be acknowledged. One of the primary limitations is the size of the clinical cohorts. Although sufficient for initial validation, these cohorts need expansion to strengthen the claims and robustness of both the method and the identified biomarkers. This expansion would enable a more robust assessment of diagnostic performance statistics and facilitate the use of advanced machine learning (ML) prediction tools to combine information from several short cfDNA biomarkers into a single classifier. Utilizing multiple short cfDNA biomarkers in a fine-tuned ML model is also likely to improve diagnostic performance (Moser et al., 2023). In chapter 5 (Sonntag et al., 2025), the correlation strengths of short cfDNA to physiological parameters in the test cohort were only moderate. Larger clinical cohorts would enhance the statistical power and reliability of the findings. Furthermore, they would enable the use of nested cross-validation cohort splits, a common practice in ML, which could identify and validate short cfDNA signals with more robust correlation strengths on unseen data. The current model aims to identify straightforward, simple relationships with physiological parameters for proof of concept. Thus, a simple linear model was used. However, non-linear and more complex dependencies could also be captured by ML models, potentially offering enhanced accuracy. Additionally, comparing these short cfDNA biomarkers with other cfDNA markers or any other relevant biomarker will be essential to highlight their advantages, disadvantages, or potential synergies. Embracing a multi-omics approach by integrating cfDNA biomarkers with genetic cfDNA biomarkers, other non-genetic cfDNA biomarkers, cfRNA biomarkers, proteomics data, or established clinical metrics could provide a richer, more comprehensive capturing of disease mechanisms, thereby enhancing diagnostic and prognostic capabilities (Xiao et al., 2022).

Developing clinically viable assays is a crucial next step to transition the current findings from the laboratory to real-world applications in clinics. Current methodologies, while effective in a research context, need to be adapted for clinical settings. This includes making the assays simpler, cheaper, faster, and ensuring they can be easily integrated into existing diagnostic workflows (Drucker and Krapfenbauer, 2013). Size-selective PCR assays for single biomarkers are a promising avenue, offering the potential for rapid and accurate quantification of specific short cfDNA fragments. An example of such an assay has been proposed by Alcaide et al. (2020) and could be adopted for short cfDNA biomarkers in future studies. With an assay that meets clinical application requirements and proves

the clinical utility of these biomarkers, additional clinical trials will be needed to define their potential benefits. These trials must be meticulously designed to test the efficacy and reliability of the biomarkers across diverse patient populations and clinical settings. A yet underexplored aspect of the short cfDNA is the influence of factors like age, gender, pre-existing conditions, or physical exercise. For regular cfDNA, such factors have been evaluated, and for example, physical exercise was found to have a significant impact on cfDNA physiology (Yuwono et al., 2021). Hence, a comparable evaluation study of potential confounding factors should be performed focusing on short cfDNA. Only such rigorous validation can ensure that these biomarkers are truly ready for clinical application. While this thesis has focused on sepsis, the potential applications of short cfDNA biomarkers extend beyond this single condition. Many oncological diseases, immune-mediated disorders, and other complex systemic conditions may benefit from the insights provided by short cfDNA biomarkers due to the extensive involvement of TFs in health and disease (Lambert et al., 2018). By broadening the scope, the full potential in diagnostics can be evaluated. Despite the remaining challenges for the novel short cfDNA biomarkers, the research in this thesis provides the necessary foundation to overcome these challenges and opens avenues for this new class of promising biomarkers.

7 Abbreviations

Table 1: List of abbreviations and their meanings

Abbreviation	Meaning
ATAC-Seq	Assay for transposase-accessible chromatin sequencing
bp	Base pairs
cfDNA	Cell-free deoxyribonucleic acids
cfRNA	Cell-free ribonucleic acids
ChIP-Seq	Chromatin immunoprecipitation sequencing
CTCF	CCCTC-binding factor
DBD	Deoxyribonucleic acid binding domain
dd-cfDNA	Donor-derived cell-free DNA
DNA	Deoxyribonucleic acid
DNase-Seq	DNase I hypersensitive sites sequencing
FOXA1	Forkhead box A1
GTF2A	General transcription factor IIA
GTF2B	General transcription factor IIB
H3K27ac	Histone H3 lysine 27 acetylation
H3K4	Histone H3 lysine 4
ICU	Intensive care unit
ML	Machine learning
NETosis	Neutrophil extracellular traps
NF-κB	Nuclear factor kappa-light-chain-enhancer of activated B cells
NFR	Nucleosome-free region
NGS	Next-generation sequencing
NIPT	Non-invasive prenatal testing
PCR	Polymerase chain reaction
RNA	Ribonucleic acid
SATB1	SATB homeobox 1
SOX2	SRY-box transcription factor 2
TF	Transcription factors
TSS	Transcription start site

References

Alcaide, M., Cheung, M., Hillman, J., Rassekh, S. R., Deyell, R. J., Batist, G., Karsan, A., Wyatt, A. W., Johnson, N., Scott, D. W., and Morin, R. D. (2020). Evaluating the quantity, quality and size distribution of cell-free dna by multiplex droplet digital pcr. *Scientific reports*, 10(1):12564.

Andrews, S. (2010). Fastqc: a quality control tool for high throughput sequence data.

Bailey, T. L., Johnson, J., Grant, C. E., and Noble, W. S. (2015). The meme suite. *Nucleic acids research*, 43(W1):W39–49.

Barral, A. and Zaret, K. S. (2024). Pioneer factors: roles and their regulation in development. *Trends in genetics : TIG*, 40(2):134–148.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bryant, J., Carter, R. J., Keira Cheetham, R., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. A., Humphray, S. J., Irving, L. J., Karbelashvili, M. S., Kirk, S. M., Li, H., Liu, X., Maisinger, K. S., Murray, L. J., Obradovic, B., Ost, T., Parkinson, M. L., Pratt, M. R., Rasolonjatovo, I. M. J., Reed, M. T., Rigatti, R., Rodighiero, C., Ross, M. T., Sabot, A., Sankar, S. V., Scally, A., Schroth, G. P., Smith, M. E., Smith, V. P., Spiridou, A., Torrance, P. E., Tzonev, S. S., Vermaas, E. H., Walter, K., Wu, X., Zhang, L., Alam, M. D., Anastasi, C., Aniebo, I. C., Bailey, D. M. D., Bancarz, I. R., Banerjee, S., Barbour, S. G., Baybayan, P. A., Benoit, V. A., Benson, K. F., Bevis, C., Black, P. J., Boodhun, A., Brennan, J. S., Bridgham, J. A., Brown, R. C., Brown, A. A., Buermann, D. H., Bundu, A. A., Burrows, J. C., Carter, N. P., Castillo, N., Chiara E Catenazzi, M., Chang, S., Neil Cooley, R., Crake, N. R., Dada, O. O., Diakoumakos, K. D., Dominguez-Fernandez, B., Earnshaw, D. J., Egbujor, U. C., Elmore, D. W., Etchin, S. S., Ewan, M. R., Fedurco, M., Fraser, L. J., Fuentes Fajardo, K. V., Scott Furey, W., George, D., Gietzen, K. J., Goddard, C. P., Golda, G. S., Granieri, P. A., Green, D. E., Gustafson, D. L., Hansen, N. F., Harnish, K., Haudenschild, C. D., Heyer, N. I., Hims, M. M., Ho, J. T., Horgan, A. M., Hoschler, K., Hurwitz, S., Ivanov, D. V., Johnson, M. Q., James, T., Huw Jones, T. A., Kang, G.-D., Kerelska, T. H., Kersey, A. D., Khrebtukova, I., Kindwall, A. P., Kingsbury, Z., Kokko-Gonzales, P. I., Kumar, A., Laurent, M. A., Lawley, C. T., Lee, S. E., Lee, X., Liao, A. K., Loch, J. A., Lok, M., Luo, S., Mammen, R. M., Martin, J. W., McCauley, P. G., McNitt, P., Mehta, P., Moon, K. W., Mullens, J. W., Newington, T., Ning, Z., Ling Ng, B., Novo, S. M., O'Neill, M. J., Osborne, M. A., Osnowski, A., Ostadan, O., Paraschos, L. L., Pickering, L., Pike, A. C., Pike, A. C., Chris Pinkard, D., Pliskin, D. P., Podhasky, J., Quijano, V. J., Raczy, C., Rae, V. H., Rawlings, S. R., Chiva Rodriguez, A., Roe, P. M., Rogers, J., Rogert Bacigalupo, M. C., Romanov, N.,

Romieu, A., Roth, R. K., Rourke, N. J., Ruediger, S. T., Rusman, E., Sanches-Kuiper, R. M., Schenker, M. R., Seoane, J. M., Shaw, R. J., Shiver, M. K., Short, S. W., Sizto, N. L., Sluis, J. P., Smith, M. A., Ernest Sohna Sohna, J., Spence, E. J., Stevens, K., Sutton, N., Szajkowski, L., Tregidgo, C. L., Turcatti, G., Vandevondele, S., Verhovsky, Y., Virk, S. M., Wakelin, S., Walcott, G. C., Wang, J., Worsley, G. J., Yan, J., Yau, L., Zuerlein, M., Rogers, J., Mullikin, J. C., Hurles, M. E., McCooke, N. J., West, J. S., Oaks, F. L., Lundberg, P. L., Klenerman, D., Durbin, R., and Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59.

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., and Yang, S. (2019). Analytical and clinical validation of a microbial cell-free dna sequencing test for infectious disease. *Nature microbiology*, 4(4):663–674.

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics (Oxford, England)*, 30(15):2114–2120.

Boyle, A. P., Davis, S., Shulha, H. P., Meltzer, P., Margulies, E. H., Weng, Z., Furey, T. S., and Crawford, G. E. (2008). High-resolution mapping and characterization of open chromatin across the genome. *Cell*, 132(2):311–322.

Brenner, T., Skarabis, A., Stevens, P., Axnick, J., Haug, P., Grumaz, S., Bruckner, T., Luntz, S., Witzke, O., Pletz, M. W., Ruprecht, T. M., Marschall, U., Altin, S., Greiner, W., and Berger, M. M. (2021). Optimization of sepsis therapy based on patient-specific digital precision diagnostics using next generation sequencing (digisep-trial)-study protocol for a randomized, controlled, interventional, open-label, multicenter trial. *Trials*, 22(1):714.

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., and Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, dna-binding proteins and nucleosome position. *Nature methods*, 10(12):1213–1218.

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. *Scientific reports*, 6:27859.

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. *Proceedings of the National Academy of Sciences of the United States of America*, 116(37):18738–18744.

Cheng, A. P., Cheng, M. P., Gu, W., Sesing Lenz, J., Hsu, E., Schurr, E., Bourque, G., Bourgey, M., Ritz, J., Marty, F. M., Chiu, C. Y., Vinh, D. C., and de Vlaminck, I. (2021). Cell-free dna tissues of origin by methylation profiling reveals significant cell, tissue, and organ-specific injury related to covid-19 severity. *Med (New York, N.Y.)*, 2(4):411–422.e5.

Drucker, E. and Krapfenbauer, K. (2013). Pitfalls and limitations in translation from biomarker discovery to clinical utility in predictive and personalised medicine. *The EPMA journal*, 4(1):7.

Duncan, C. F., Youngstein, T., Kirrane, M. D., and Lonsdale, D. O. (2021). Diagnostic challenges in sepsis. *Current infectious disease reports*, 23(12):22.

Eun, H.-M. (1996). Rna polymerases. In *Enzymology Primer for Recombinant DNA Technology*, pages 491–565. Elsevier.

Evans, L., Rhodes, A., Alhazzani, W., Antonelli, M., Coopersmith, C. M., French, C., Machado, F. R., Mcintyre, L., Ostermann, M., Prescott, H. C., Schorr, C., Simpson, S., Wiersinga, W. J., Alshamsi, F., Angus, D. C., Arabi, Y., Azevedo, L., Beale, R., Beilman, G., Belley-Cote, E., Burry, L., Cecconi, M., Centofanti, J., Coz Yataco, A., de Waele, J., Dellinger, R. P., Doi, K., Du, B., Estenssoro, E., Ferrer, R., Gomersall, C., Hodgson, C., Møller, M. H., Iwashyna, T., Jacob, S., Kleinpell, R., Klompas, M., Koh, Y., Kumar, A., Kwizera, A., Lobo, S., Masur, H., McGloughlin, S., Mehta, S., Mehta, Y., Mer, M., Nunnally, M., Oczkowski, S., Osborn, T., Papathanassoglou, E., Perner, A., Puskarich, M., Roberts, J., Schweickert, W., Seckel, M., Sevransky, J., Sprung, C. L., Welte, T., Zimmerman, J., and Levy, M. (2021). Surviving sepsis campaign: international guidelines for management of sepsis and septic shock 2021. *Intensive care medicine*, 47(11):1181–1247.

Fan, H. C., Blumenfeld, Y. J., Chitkara, U., Hudgins, L., and Quake, S. R. (2010). Analysis of the size distributions of fetal and maternal cell-free dna by paired-end sequencing. *Clinical chemistry*, 56(8):1279–1286.

Fleischmann-Struzek, C., Mellhammar, L., Rose, N., Cassini, A., Rudd, K. E., Schlattmann, P., Allegranzi, B., and Reinhart, K. (2020). Incidence and mortality of hospital- and icu-treated sepsis: results from an updated and expanded systematic review and meta-analysis. *Intensive care medicine*, 46(8):1552–1562.

Graw, J. (2015). *Genetik*. Springer Berlin Heidelberg, Berlin, Heidelberg.

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. *Critical care medicine*, 47(5):e394–e402.

Holwerda, S. J. B. and de Laat, W. (2013). Ctcf: the protein, the binding partners, the binding sites and their chromatin loops. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 368(1620):20120369.

Hudecova, I., Smith, C. G., Hänsel-Hertsch, R., Chilamakuri, C. S., Morris, J. A., Vijayaraghavan, A., Heider, K., Chandrananda, D., Cooper, W. N., Gale, D., Garcia-Corbacho, J., Pacey, S., Baird, R. D., Rosenfeld, N., and Mouliere, F. (2022). Characteristics, origin, and potential for cancer diagnostics of ultrashort plasma cell-free dna. *Genome research*, 32(2):215–227.

Ignatiadis, M., Sledge, G. W., and Jeffrey, S. S. (2021). Liquid biopsy enters the clinic - implementation issues and future challenges. *Nature reviews. Clinical oncology*, 18(5):297–312.

Isbel, L., Grand, R. S., and Schübeler, D. (2022). Generating specificity in genome regulation through transcription factor sensitivity to chromatin. *Nature reviews. Genetics*, 23(12):728–740.

Jamshidi, A., Liu, M. C., Klein, E. A., Venn, O., Hubbell, E., Beausang, J. F., Gross, S., Melton, C., Fields, A. P., Liu, Q., Zhang, N., Fung, E. T., Kurtzman, K. N., Amini, H., Betts, C., Civello, D., Freese, P., Calef, R., Davydov, K., Fayzullina, S., Hou, C., Jiang, R., Jung, B., Tang, S., Demas, V., Newman, J., Sakarya, O., Scott, E., Shenoy, A., Shojaee, S., Steffen, K. K., Nicula, V., Chien, T. C., Bagaria, S., Hunkapiller, N., Desai, M., Dong, Z., Richards, D. A., Yeatman, T. J., Cohn, A. L., Thiel, D. D., Berry, D. A., Tummala, M. K., McIntyre, K., Sekeres, M. A., Bryce, A., Aravanis, A. M., Seiden, M. V., and Swanton, C. (2022). Evaluation of cell-free dna approaches for multi-cancer early detection. *Cancer cell*, 40(12):1537–1549.e12.

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. *Clinical chemistry*, 68(9):1184–1195.

Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-dna interactions. *Science (New York, N.Y.)*, 316(5830):1497–1502.

Kalantar, K. L., Neyton, L., Abdelghany, M., Mick, E., Jauregui, A., Caldera, S., Serpa, P. H., Ghale, R., Albright, J., Sarma, A., Tsitsiklis, A., Leligdowicz, A., Christenson, S. A., Liu, K., Kangelaris, K. N., Hendrickson, C., Sinha, P., Gomez, A., Neff, N., Pisco, A., Doernberg, S. B., Derisi, J. L., Matthay, M. A., Calfee, C. S., and Langelier, C. R. (2022). Integrated host-microbe plasma metagenomics for sepsis diagnosis in a prospective cohort of critically ill adults. *Nature microbiology*, 7(11):1805–1816.

Knight, S. R., Thorne, A., and Lo Faro, M. L. (2019). Donor-specific cell-free dna as a biomarker in solid organ transplantation. a systematic review. *Transplantation*, 103(2):273–283.

Koboldt, D. C., Ding, L., Mardis, E. R., and Wilson, R. K. (2010). Challenges of sequencing

human genomes. *Briefings in bioinformatics*, 11(5):484–498.

Kustanovich, A., Schwartz, R., Peretz, T., and Grinshpun, A. (2019). Life and death of circulating cell-free dna. *Cancer biology & therapy*, 20(8):1057–1067.

Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T. R., and Weirauch, M. T. (2018). The human transcription factors. *Cell*, 172(4):650–665.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics (Oxford, England)*, 25(14):1754–1760.

Li, Y., Guo, J., Yang, H., Li, H., Shen, Y., and Zhang, D. (2021). Comparison of culture-negative and culture-positive sepsis or septic shock: a systematic review and meta-analysis. *Critical care (London, England)*, 25(1):167.

Li, Y., Ma, M., Xu, X., Li, Q., and Ji, C. (2022). Value of digital pcr in the early diagnosis of sepsis: A systematic review and meta-analysis. *Journal of critical care*, 72:154138.

Li, Z., Schulz, M. H., Look, T., Begemann, M., Zenke, M., and Costa, I. G. (2019). Identification of transcription factor binding sites using atac-seq. *Genome biology*, 20(1):45.

Liu, Y., Fu, L., Kaufmann, K., Chen, D., and Chen, M. (2019). A practical guide for dnase-seq data analysis: from data management to common applications. *Briefings in bioinformatics*, 20(5):1865–1877.

Lo, Y. M., Corbetta, N., Chamberlain, P. F., Rai, V., Sargent, I. L., Redman, C. W., and Wainscoat, J. S. (1997). Presence of fetal dna in maternal plasma and serum. *Lancet (London, England)*, 350(9076):485–487.

Lo, Y. M., Tein, M. S., Pang, C. C., Yeung, C. K., Tong, K. L., and Hjelm, N. M. (1998). Presence of donor-specific dna in plasma of kidney and liver-transplant recipients. *Lancet (London, England)*, 351(9112):1329–1330.

Lo, Y. M., Zhang, J., Leung, T. N., Lau, T. K., Chang, A. M., and Hjelm, N. M. (1999). Rapid clearance of fetal dna from maternal plasma. *American journal of human genetics*, 64(1):218–224.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):550.

Loy, C., Ahmann, L., de Vlaminck, I., and Gu, W. (2024). Liquid biopsy based on cell-free dna and rna. *Annual review of biomedical engineering*, 26(1):169–195.

Loyfer, N., Magenheim, J., Peretz, A., Cann, G., Bredno, J., Klochendler, A., Fox-Fisher, I., Shabi-Porat, S., Hecht, M., Pelet, T., Moss, J., Drawshy, Z., Amini, H., Moradi, P., Nagaraju, S., Bauman, D., Shveiky, D., Porat, S., Dior, U., Rivkin, G., Or, O., Hirshoren, N., Carmon, E., Pikarsky, A., Khalaileh, A., Zamir, G., Grinbaum, R., Abu Gazala,

M., Mizrahi, I., Shussman, N., Korach, A., Wald, O., Izhar, U., Erez, E., Yutkin, V., Samet, Y., Rotnemer Golinkin, D., Spalding, K. L., Druid, H., Arner, P., Shapiro, A. M. J., Grompe, M., Aravanis, A., Venn, O., Jamshidi, A., Shemer, R., Dor, Y., Glaser, B., and Kaplan, T. (2023). A dna methylation atlas of normal human cell types. *Nature*, 613(7943):355–364.

Mandel, P. and Metais, P. (1948). Les acides nucléiques du plasma sanguin chez l'homme. *Comptes rendus des seances de la Societe de biologie et de ses filiales*, 142(3-4):241–243.

Moding, E. J., Nabet, B. Y., Alizadeh, A. A., and Diehn, M. (2021). Detecting liquid remnants of solid tumors: Circulating tumor dna minimal residual disease. *Cancer discovery*, 11(12):2968–2986.

Moser, T., Kühberger, S., Lazzeri, I., Vlachos, G., and Heitzer, E. (2023). Bridging biological cfdna features and machine learning approaches. *Trends in genetics : TIG*, 39(4):285–307.

Mouliere, F. (2022). A hitchhiker's guide to cell-free dna biology. *Neuro-oncology advances*, 4(Suppl 2):ii6–ii14.

Müller, J., Hartwig, C., Sonntag, M., Bitzer, L., Adelmann, C., Vainshtein, Y., Glanz, K., Decker, S. O., Brenner, T., Weber, G. F., von Haeseler, A., and Sohn, K. (2024). A novel approach for in vivo dna footprinting using short double-stranded cell-free dna from plasma. *Genome research*, 34(8):1185–1195.

Nakato, R. and Sakata, T. (2021). Methods for chip-seq analysis: A practical workflow and advanced applications. *Methods (San Diego, Calif.)*, 187:44–53.

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., and Wapner, R. J. (2015). Cell-free dna analysis for noninvasive examination of trisomy. *The New England journal of medicine*, 372(17):1589–1597.

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., and Nolte, F. S. (2023). Plasma microbial cell-free dna sequencing from over 15,000 patients identified a broad spectrum of pathogens. *Journal of clinical microbiology*, 61(8):e0185522.

Rhee, C., Dantes, R., Epstein, L., Murphy, D. J., Seymour, C. W., Iwashyna, T. J., Kadri, S. S., Angus, D. C., Danner, R. L., Fiore, A. E., Jernigan, J. A., Martin, G. S., Septimus, E., Warren, D. K., Karcz, A., Chan, C., Menchaca, J. T., Wang, R., Gruber, S., and Klompas, M. (2017). Incidence and trends of sepsis in us hospitals using clinical vs claims data, 2009-2014. *JAMA*, 318(13):1241–1249.

Samura, O. (2020). Update on noninvasive prenatal testing: A review based on current worldwide research. *The journal of obstetrics and gynaecology research*, 46(8):1246–1254.

Scheer, C. S., Fuchs, C., Gründling, M., Vollmer, M., Bast, J., Bohnert, J. A., Zimmermann, K., Hahnenkamp, K., Rehberg, S., and Kuhn, S.-O. (2019). Impact of antibiotic administration on blood culture positivity at the beginning of sepsis: a prospective clinical cohort study. *Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases*, 25(3):326–331.

Sedlazeck, F. J., Rescheneder, P., and von Haeseler, A. (2013). Nextgenmap: fast and accurate read mapping in highly polymorphic genomes. *Bioinformatics (Oxford, England)*, 29(21):2790–2791.

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., and Angus, D. C. (2016). The third international consensus definitions for sepsis and septic shock (sepsis-3). *JAMA*, 315(8):801–810.

Snyder, M. W., Kircher, M., Hill, A. J., Daza, R. M., and Shendure, J. (2016). Cell-free dna comprises an in vivo nucleosome footprint that informs its tissues-of-origin. *Cell*, 164(1-2):57–68.

Snyder, T. M., Khush, K. K., Valantine, H. A., and Quake, S. R. (2011). Universal noninvasive detection of solid organ transplant rejection. *Proceedings of the National Academy of Sciences of the United States of America*, 108(15):6229–6234.

Sonntag, M., Mueller, J., Brenner, T., Decker, S. O., von Haeseler, A., Sohn, K., Anaesthesia, G., and (GAIN-CARE), I. C. T. G. (2025). Human in vivo footprints from blood plasma samples for improved diagnostics in septic patients. *medRxiv*.

Ulz, P., Perakis, S., Zhou, Q., Moser, T., Belic, J., Lazzeri, I., Wölfler, A., Zebisch, A., Gerger, A., Pristauz, G., Petru, E., White, B., Roberts, C. E. S., John, J. S., Schimek, M. G., Geigl, J. B., Bauernhofer, T., Sill, H., Bock, C., Heitzer, E., and Speicher, M. R. (2019). Inference of transcription factor binding from cell-free dna enables tumor subtype prediction and early detection. *Nature communications*, 10(1):4666.

Vaddavalli, P. L. and Schumacher, B. (2022). The p53 network: cellular and systemic dna damage responses in cancer and aging. *Trends in genetics : TIG*, 38(6):598–612.

Vincent, J.-L., Brealey, D., Libert, N., Abidi, N. E., O'Dwyer, M., Zacharowski, K., Mikaszewska-Sokolewicz, M., Schrenzel, J., Simon, F., Wilks, M., Picard-Maureau, M., Chalfin, D. B., Ecker, D. J., Sampath, R., and Singer, M. (2015). Rapid diagnosis of infection in the critically ill, a multicenter study of molecular detection in bloodstream infections, pneumonia, and sterile site infections. *Critical care medicine*,

43(11):2283–2291.

Vincent, J.-L., Rello, J., Marshall, J., Silva, E., Anzueto, A., Martin, C. D., Moreno, R., Lipman, J., Gomersall, C., Sakr, Y., and Reinhart, K. (2009). International study of the prevalence and outcomes of infection in intensive care units. *JAMA*, 302(21):2323–2329.

Wingender, E., Schoeps, T., Haubrock, M., Krull, M., and Dönitz, J. (2018). Tfclass: expanding the classification of human transcription factors to their mammalian orthologs. *Nucleic acids research*, 46(D1):D343–D347.

Xiao, Y., Bi, M., Guo, H., and Li, M. (2022). Multi-omics approaches for biomarker discovery in early ovarian cancer diagnosis. *EBioMedicine*, 79:104001.

Yan, F., Powell, D. R., Curtis, D. J., and Wong, N. C. (2020). From reads to insight: a hitchhiker's guide to atac-seq data analysis. *Genome biology*, 21(1):22.

Yuwono, N. L., Warton, K., and Ford, C. E. (2021). The influence of biological and lifestyle factors on circulating cell-free dna in blood plasma. *eLife*, 10.

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., and Liu, X. S. (2008). Model-based analysis of chip-seq (macs). *Genome biology*, 9(9):R137.

Zhong, P., Bai, L., Hong, M., Ouyang, J., Wang, R., Zhang, X., and Chen, P. (2024). A comprehensive review on circulating cfrna in plasma: Implications for disease diagnosis and beyond. *Diagnostics (Basel, Switzerland)*, 14(10).