The circulating phageome reflects bacterial infections

[#]Haddock, NL[1], [#]Barkal, LJ[2], Ram-Mohan, N[3], Kaber, G[4], Chiu, CY[5], Bhatt, AS[6,7], Yang, S[3], *Bollyky, PL[4]

[#]These authors have contributed equally to this manuscript.
*Corresponding author

Affiliations

1. Immunology Program, School of Medicine, Stanford University, Stanford, California, 94305
2. Division of Pulmonary, Allergy, and Critical Care Medicine, School of Medicine, Stanford University, Stanford, California, 94305
3. Department of Emergency Medicine, School of Medicine, Stanford University, Stanford, California, 94305
4. Division of Infectious Diseases and Geographic Medicine, School of Medicine, Stanford University, Stanford, California, 94305
5. Department of Medicine, Division of Infectious Diseases, University of California San Francisco, San Francisco, CA, USA
6. Division of Hematology, School of Medicine, Stanford University, Stanford, California, 94305
7. Department of Genetics, Stanford University, Stanford, California, 94305

Abstract

Bacteriophage, viruses that infect bacteria, are abundant in the human body but the relationship between the phageome and bacterial population dynamics is unclear. Because bacteriophage are often highly specific to bacterial host strains and species, we asked whether bacteriophage present in cell-free DNA (cfDNA) reflect bacterial infections in sepsis and other infections. To address this, we generate a computational pipeline for identifying bacteriophage sequences in cfDNA. In three large, independent cohorts of infected patients and asymptomatic controls, we demonstrate that all individuals have a circulating phageome. Moreover, infection associates with overrepresentation of pathogen-specific phage, allowing for identification of bacterial pathogens. We further show that phage can identify pathovariant Escherichia coli infections and distinguish between closely-related pathogenic bacteria such as Staphylococcus aureus and frequent contaminants such as coagulase-negative Staphylococcus. Phage DNA may have utility in identifying bacterial pathogens in infection and studying bacterial dynamics.

Abstract word count: 143

Introduction

Bacterial culture methods developed over a century ago remain the standard of care for identifying bacterial pathogens in sepsis and other settings. Unfortunately, these methods are often imprecise and can be confounded by both false positives and negatives. Improving identification and subsequent study of bacterial pathogens is a critical medical need[1].

Next generation sequencing (NGS) of cell-free DNA (cfDNA) holds promise for filling this diagnostic gap. cfDNA is made up of DNA fragments, typically 50-200 bp long, found in circulation[2]. cfDNA is primarily human in origin with a small but rich microbial compartment[2–4]. With high-throughput sequencing, cfDNA lends itself to non-invasive testing and has transformed diagnostics in perinatal testing, autoimmune disorders, cancer staging, and transplant rejection[5–8]. There have been efforts to develop cfDNA infection diagnostics in sepsis[9–12] and other settings[13,14].

Retrospective cohort studies this approach failed to identify the pathogen in many sepsis cases, and these approaches do not establish thresholds distinguishing colonization from infection[15,16]. Other molecular approaches such as PCR-based diagnostics yield similarly mixed results[17] and have limited widespread adoption. There is an urgent need for improvement to methods of studying bacterial dynamics underlying infection to facilitate diagnostic improvements.

Bacteriophage (phage), or viruses that infect bacteria, may provide insight into the bacterial population dynamics underlying sepsis and opportunistic infections. Though abundant in the human body, they have historically been challenging to identify in samples and have been termed microbial “dark matter”[18,19].

Nonetheless, aspects of phage biology make them uniquely well-suited for interrogating bacterial infections. Many phage have narrow host ranges limited to a few closely related bacteria, though information is sparse given the vast diversity of phage[20–22]. Composition of phages may thereby reflect bacterial populations at the species and strain level. Additionally, some phage translocate across epithelia and into the bloodstream[23–25]. Despite structural and genetic diversity, most phage within humans are non-enveloped DNA viruses[26,27] and can be sequenced using existing NGS methods and detected in cfDNA[3,28].

Targeted analysis of phage in metagenomic data has historically been limited by lack of available phage sequences[18,29–31]. Additionally, many metagenomic annotation software packages require metagenome assembly prior to use – frequently impossible in low microbial biomass specimens. Fortunately, continuous rapid increases in submitted genomes to NCBI’s GenBank has facilitated improved identification of these phage as more of these “dark matter” genomes become known.

Here we describe an approach to confidently identifying and interpreting DNA phage sequences in cfDNA. We apply this pipeline to samples - bodily fluids and plasma - collected from three independent human patient cohorts, each comprised of infected patients as well as healthy controls. Two of these cohorts involve previously published metagenomic and culture data[32,33] while the third was generated here. Using these samples, we reveal that at the site of infection, phageomes reflect infection through pathogen-specific phage overrepresentation. We demonstrate that infection associates with pathogen specific overrepresentation in the circulating phageome, allowing for identification of bacterial pathogens. Furthermore, we show that phage can distinguish between closely related pathogenic and commensal bacterial species where bacterial cfDNA may fall short. Our data suggest that circulating phage sequences in plasma provide a non-invasive approach to studying bacterial dynamics in infection.

Results

Local infections are associated with pathogen-specific phage overrepresentation

To assess whether phage sequences present in conventional NGS data can be used to identify bacterial pathogens, we first analyzed a publicly available dataset of 92 infected bodily fluid samples (Bioproject PRJNA558701) generated via Illumina sequencing[32]. These included samples from infected wounds, joints, urine, serum, and other sites as well as culture negative controls. Causative infectious agents were identified by culture or 16S rRNA sequencing, including *Escherichia coli*, *Streptococcus spp.*, *Staphylococcus spp.*, and *Klebsiella spp*. This dataset provides an excellent setting to investigate the relationships between pathogen and bacteriophage at the site of infection.

To identify phages within the metagenomic sequencing data, we applied a phage annotation pipeline (Fig. 1A). Raw data was quality controlled and trimmed using standard bioinformatics tools, FASTQC[34] and Trimmomatic[35]. Host reads were subtracted by mapping to the human reference genome GRCh38 via Bowtie2. A BLAST search utilizing the full NCBI Nucleotide database revealed that on average, 9.17% of identifiable non-human reads were from bacteriophage, indicating that at the average site of infection bacteriophage comprise a considerable proportion of free DNA (Fig. 1B, Fig. S1). A first-pass BLAST search[36,37] was performed against a custom phage database constructed from the NCBI Nucleotide database. Reads with significant hits were subjected to a secondary and more stringent human sequence removal, in which all reads with significant BLAST hits to any human sequence deposited in the NCBI nuccore were removed and annotations were made again using the custom phage BLAST search.

To link phage with bacterial host(s), we created a phage dictionary, compiled using virushostDB[38], naming convention of bacteriophages with clearly identified host genera, and NCBI nucleotide source host fields for bacteriophage sequence entries. This dictionary includes taxonomic classifications for both the phage and for the bacterial host, if known (Supplemental file 1). Using aggressive human read depletion, we find fewer human reads interfering with phage calls (Fig. S2A), with an average of 25.18% of unique phage hits depleted. We also find fewer phage without known bacterial hosts (Fig. S2B). Subsequent annotations reflect a diverse pool of bacteriophage across many families (Fig. 1C). In short, the average site of infection has a robust, diverse phageome. Additional human depletion steps and a curated bacteriophage dictionary allow for cleaner phageome annotations which can be meaningfully analyzed with respect to infection etiology.

We find that bacterial infections have notable pathogen-specific enrichment of phage, such that *E. coli* phage are overrepresented in *E. coli* infection. Similar trends are seen for infections by *Streptococcus spp.*, *Staphylococcus aureus*, and *Klebsiella spp.* (Fig. 1D). Ecological diversity of phage, as calculated by the Shannon Diversity Index (SDI), a measure of entropy often applied in ecology to quantify species richness and evenness[39], is disrupted on a per-infection basis (Fig. 1E-H), with higher levels of specific phage diversity in patients with infection by the corresponding host but not in controls or those with other infection etiology.

These findings demonstrate that infection of serum and other bodily fluids is associated with overrepresented pathogen-specific phage in proportion and ecological diversity.

The circulating phageome is mostly unaltered by sepsis but reflects infection etiology

We next sought to determine if there is a circulating phageome in healthy subjects and whether it is disrupted in patients with blood culture positive infections. We analyzed plasma from 62 patients with sepsis seen in the Stanford University Emergency Department as well as 10 asymptomatic controls (Fig. 2A). The sepsis samples represent twenty different bacterial infections as confirmed by positive blood cultures (Table S1) and were analyzed using the workflow depicted in Fig. 2A. Briefly, cfDNA was collected from the samples, sequenced on an Illumina platform to an average depth of 18.67 million reads, and run through the phage annotation and analysis pipeline detailed above for the dataset in Figure 1.

We find that plasma cfDNA from both asymptomatic controls and septic patients is comprised of mostly human reads (Fig. 2B), consistent with previous studies[2–4]. Non-human reads, despite making up a relatively small proportion of all cfDNA reads, represent a rich microbial community (Fig 2C). Bacterial hits include many genera. In asymptomatic individuals the greatest single genus represented *Propionibacterium* hits, a known skin commensal[40] (Fig. S4). Less than 1% of reads correspond to mammalian viruses, while considerably more reads map to bacteriophage: 8.80% and 9.62% of non-human reads in asymptomatic and septic patient samples, respectively (Fig. 2C, Fig S3). There is a broad variety of phage families represented (Fig. 2D), though the tailed phage families *Myoviridae* and *Siphoviridae* predominate. There is no statistical difference in phage family representation within the circulating phageomes of asymptomatic controls and septic patients (Fig. 2D). SDI of circulating phage is not significantly different between the groups (Fig. 2E), though septic patients did have a wider range in phage SDI. These findings indicate that there is no broad "Sepsis Phageome" signature.

Of the identified phage in both the asymptomatic and sepsis groups that have known bacterial hosts in the phage dictionary, most correspond to *Pseudomonas*, *Escherichia, Klebsiella, Propionibacterium*, and *Streptococcus* genera (Fig. 2F). Bacterial host proportions are not significantly different across asymptomatic and sepsis groups (Fig. 2F, Fig. S5A, Table S2), though there is inter-individual variation (Fig S5B-C). Many phage have as yet unknown bacterial hosts, accounting for 47.2% and 37.4% of phage in the asymptomatic and septic patient groups, respectively (Fig 2F).

As sepsis is commonly caused by a single bacterial strain[41], we investigated circulating phage on a per-infection basis. We find that patients with sepsis due to *E. coli, Streptococcus spp.*, *Staphylococcus spp.*, and *Klebsiella spp.* infections have a corresponding overrepresentation in phage proportion (Fig 2G) as well as the SDI of the pathogen specific phage population often also increases in sepsis. This is the case in sepsis caused by *E. coli* (Fig 2H) as well as for *Streptococcus* (Fig 2I)*, Staphylococcus* (Fig 2J), and with *Klebsiella* (Fig 2K).

Taken together, the content of non-human cfDNA, including the bacterial, eukaryotic, and viral DNA as well as the bacteriophage families, diversity, and annotated bacterial hosts, are broadly similar in health and in sepsis. However, septic patients with positive blood cultures have disrupted phageomes corresponding to the bacterial cause of their sepsis.

The circulating phageome reflects individual pathogens in another large study.

We then sought to validate our findings in a larger, more clinically diverse group of patients. The SepSeq study, a cohort study of 348 patients, investigated the ability of bacterial cfDNA to diagnose the cause of sepsis in patients triggering an emergency department sepsis alert with a wide array of infectious etiologies for which data are publicly available[33].

There are four patient populations in the SepSeq study: patients with 1) sepsis and positive blood cultures, 2) sepsis and negative blood cultures but documented infection elsewhere, 3) Systemic Inflammatory Response Syndrome (SIRS) who clinically appear septic but ultimately have no infectious etiology identified, and 4) asymptomatic controls (Fig 3A). The original study reported overall promising diagnostic statistics using bacterial cfDNA, but with notable gaps: under-sensitivity in samples with negative initial blood cultures, and oversensitivity to chronic infections and commensals. We sought to determine first, if the phageome trends we observed in our other dataset are consistent in the SepSeq study in patients with bacterial infections and second, if phage present in cfDNA might offer additional useful information in the identification or study of pathogenic bacterial dynamics.

Analyzing the SepSeq data with our annotation pipeline (Fig 3A), we again find that circulating phageomes are remarkably similar in health and sepsis. While overall phage SDI higher in patients

with sepsis and positive blood cultures (Fig 3B), proportions of bacterial hosts represented by the identified phage (Fig 3C) are largely stable. We again find the most common bacterial hosts to be *Enterobacter spp.*, *Escherichia spp.*, *Staphylococcus spp.*, and *Streptococcus spp.*, even in healthy patients (Fig 3C). Phage specific to the known cause of infection are again overrepresented in those phageomes (Fig 3D).

Within *E. coli* phage specifically, there is a significant increase in phage diversity between patients with *E. coli* infection and those without positive cultures, either who grew another organism or none at all (SIRS) (Fig 3E). Notably, there was no difference in *E. coli* phage diversity in *E. coli* infections between those patients who had positive blood cultures, or negative blood cultures but positive site, (e.g. urine), cultures. The intermediate nature of site-culture positive infections is also reflected in *Streptococcus* infections where only blood culture positive infections have significantly higher *Streptococcus* phage diversity (Fig 3F). The trend for *Staphylococcus* infections is similar to that of *E. coli* infections; patients with SIRS or other infections have low *Staphylococcus* phage diversity while true *Staphylococcus* infections have greater phage diversity, and there is no significant difference between blood culture positive or negative groups (Fig 3G). Lastly, *Klebsiella* phage diversity is significantly increased in *Klebsiella* infections, and to a lesser extent in other causes of sepsis (Fig 3H).

Within the SepSeq study, there is once again a healthy circulating phageome and no “Sepsis Phageome”. Diversity of phage specific to the bacterial genus causing infection, however, is consistently overrepresented in the context of bloodstream infection. Notably, specific phage diversity is also often increased in infections that cause sepsis without positive blood cultures. This suggests plasma cfDNA sequencing can provide a noninvasive alternative to site of infection fluid sequencing towards the study of bacterial infections.

E. coli phage reflect host strain characteristics

E. coli infections are common causes of sepsis, including in our cohorts, and *E. coli* phage are among the best studied. Here, *E. coli* phage are sporadic in asymptomatic patients, more common in patients with SIRS or non-*E. coli* causes of sepsis, and ubiquitous in patients with *E. coli* sepsis (Fig 4A). There is notable enrichment for *E. coli* phage from *Siphoviridae, Podoviridae*, *Myoviridae* phage families in *E. coli* sepsis while *Autographiviridae* phage remain at stably low levels despite infection status (Fig 4B-E). There is also slight but significant enrichment for *E. coli* phage from *Siphoviridae, Podoviridae*, *Myoviridae* phage families in patients with SIRS or non-*E. coli* sepsis (Fig 4B-E).

There is incredible diversity of *E. coli* phage with more than 500 unique *E. coli* phage identified across the 36 *E. coli* sepsis patients. Not all are present in every sample, and in fact the majority of phage are present in only one or two patients. However, there is a subset of moderately represented *E. coli* phage present in many but not all patients with *E. coli* sepsis. Most strikingly, there is a group of broadly represented, or ubiquitous, *E. coli* phage that are present in nearly all patients with *E. coli* sepsis but not the other patients or controls (Fig 4F). These groups of *E. coli* phage have notable variation at the phage family level. Within the group of rare phage, there are predominantly *Myoviridae* phage. *Podoviridae* phage predominate in the moderately represented group of phage, though there is still a sizable proportion of *Myoviridae* phage. And in the broadly represented group of *E. coli* phage, there is clear enrichment for *Siphoviridae* phage (Fig 4G).

E. coli phage are well studied and 65.56% of these phage have a clear bacterial host strain noted in the literature (Supplemental File 2). Using this rich literature captured in our phage dictionary, we annotate bacterial host down to the level of the bacterial strain for the *E. coli* phage in each prevalence subset. We find that lab strains of *E. coli*, most derived initially from patient samples, are common hosts across each phage subset. Conversely, phages associated with toxin-producing strains of *E. coli* (STEC, ETEC) are overrepresented in the broadly represented phages (Fig. 4H) while across all *E. coli* Sepsis *E. coli* phages this is only true of STEC associated phages (Fig. S6).

There is a corresponding decrease in environmental host *E. coli* bacterial strains (sewage, manure, or water). Additionally, we find that in the infection site dataset from Figure 1, including direct sequencing of *E. coli* infected fluids that host bacterial strain ecology mirrors what we find in *E. coli* sepsis plasma – toxigenic pathovariants of *E. coli* are highly represented (Fig. S7). Though strain level identities of the pathogenic *E. coli* in these samples were not known, these data indicate differences in the host-strain reflection of *E. coli* phage found in *E. coli* sepsis versus other or no infection samples.

Together, these data indicate that *E. coli* phage can provide an ecological view of *E. coli* infections, both at local sites of infection and in the context of sepsis – with phages from samples of true infection being associated with toxin producing *E. coli* host pathovariants. This level of granularity will be helpful as we learn more about the dysregulation of colonizing *E. coli* strains as true infection is established, and may have utility in distinguishing true infection from simple colonization.

Phage can distinguish between Staphylococcus aureus and coagulase-negative Staphylococcus

Distinguishing between true infection and contamination or simple colonization remains a challenging clinical problem[42,43]. In the case of *Staphylococcus* infections, bacterial cfDNA can potentially underperform in distinguishing species within a bacterial genus. We wondered if bacteriophages were able to determine if the *Staphylococcus* species is *S. aureus*, the presence of which in blood culture is always considered pathogenic, versus one of the coagulase-negative *Staphylococcus* species (CoNS), the presence of which is often a contaminant.

We find that when mapping to a *S. aureus* reference genome, bacterial cfDNA is unable to distinguish between *S. aureus* and CoNS in either our cohort (Fig 5A) or the SepSeq cohort (Fig 5B). Consistent with this, the corresponding ROC plots have AUCs near 0.5 (Fig 5C). Notably, however, the populations of phage that infect *S. aureus* and CoNS are distinct and the diversity of *S. aureus*-specific phage is informative: there is a significant increase in *S. aureus* phage diversity only in *S. aureus* infections (Fig 5E). This is reflected in a ROC plot with AUC of 0.96 (Fig 5F). However, this increase did not reach significance in our smaller cohort (Fig. 5D).

These data suggest that increased diversity of *S. aureus* phage is sensitive and specific for *S. aureus* infections. This relationship can also be seen in the correlation plot where *S. aureus* phage diversity correlates strongly with *S. aureus* infection, which is not the case for CoNS phage SDI and CoNS infection (Fig 5G). Taken together, subgroups of phage, specifically *S. aureus* phage, are sensitive and specific indicators of true *S. aureus* infection compared to the genetically similar but less virulent CoNS. This suggests that bacteriophage may provide informative targets in the improvement of NGS diagnostics where gaps remain in bacterial DNA performance.

Discussion

We report here that bacteriophage sequences are present in cfDNA from serum and other infected bodily fluids and that this information can be used to identify bacterial pathogens. We demonstrate the utility of this approach in three, large independent data sets as well as in their associated asymptomatic controls, all generated using conventional NGS methods.

The data presented here indicate that all individuals have a circulating phageome. Some phage transcytose across tissue and enter systemic circulation; our data suggests that this may occur on a substantial scale. We find there is little global alteration in the phageome in bacterial sepsis, consistent with sepsis typically being due to single or small numbers of pathogens. The origins and functional significance of phage cfDNA in the human body await further investigation.

We show that for some bacteria this approach can be used to study bacteriophage ecology in infection and to distinguish between closely related bacterial species to distinguish true infections and contamination. This potential is demonstrated here in the group of patients with *E. coli* sepsis, where we show that phage ubiquitous in *E. coli* sepsis tend to be associated with host pathovariants that produce toxins. We show that *S. aureus* phage have excellent discriminatory power to distinguish *S. aureus* infection from CoNS while CoNS phage have poor correlation with CoNS culture which are often due to contaminated cultures. Future experimental studies in pre-clinical models are needed to test this hypothesis and to inform our understanding of phage cfDNA in infection.

While the increase in the diversity of phage corresponding to bacterial pathogens in sepsis appears robust (parallel increase in bacterial cfDNA during sepsis is the foundation of various diagnostic platforms[11,32,33,44,45]), its ecological meaning is unclear. Because bacteria may harbor multiple prophages[46], this increase in diversity may reflect lysis by multiple phage due to rapid bacterial expansion or stress. However, other explanations are possible. Understanding phage dynamics associated with bacterial infection will increase robustness of phageome studies in sepsis and other settings. Further studies are needed to investigate how phage gain access to the bloodstream in sepsis and how this varies by site of infection and bacterial burden.

This study has several limitations. One limitation is that two of the three cohorts in this study were collected at the same location - the Stanford Emergency Department. However, sample preparation, year of acquisition, and sequencing methods differed between the groups. Additionally, because of the breadth of sepsis etiologies, each individual type of infection had relatively low numbers of samples, making comparisons challenging. Though microbial cfDNA biomass may be higher at infection sites, we believe that the tradeoff of a noninvasive approach facilitates easier acquisition of patient samples and enables retrospective study in banked plasma samples from infected patients.

In summary, we show here that there is a circulating phageome detectable in both asymptomatic and infected individuals using NGS sequencing of cfDNA, and that phage cfDNA can help to identify bacterial infections. We believe there is great promise in using phage to noninvasively study and diagnose bacterial infections.

Methods

Sample collection

The Emergency Department Sepsis Biobank study protocol and informed consent (32851) was reviewed and approved by the University's IRB (Stanford). Samples were collected from patients presenting to the Stanford University Hospital Emergency Department who triggered a sepsis alert, which included blood sample collection at the time of enrollment as well as collection of the results from standard of care microbiological testing they received during the first seven days of their admission. Patients were eligible to enroll if they triggered a sepsis alert, were 18 years of age or older, had a temperature of >38°C or <36°C and met at least one systemic inflammatory response syndrome criteria (heart rate >90 beats per minute; respiratory rate >20 breaths per minute or a partial pressure of carbon dioxide <32 mm Hg; white blood cell count of either > 12,000 cells μl^{-1} or <4,000 cells μl^{-1} and >10% bands).

Blood samples were collected from peripheral blood draw or indwelling venous catheter into BD Vacutainer EDTA blood collection tubes (Becton Dickinson, Franklin Lakes, NJ) at the same time as the patients were undergoing blood draw for standard of care blood cultures. In the vast majority of patients, this occurred prior to antibiotic administration. Samples were then stored at 4°C for no more than 72 hours, centrifuged at 1500 rpm for 10 min to obtain plasma, and stored thereafter at -80°C. Samples were deidentified using a unique code that could be linked to the patient identifier to later

obtain gold standard microbiological testing results for each patient. Plasma samples from 62 patients with positive blood culture results were randomly selected for further phageome analyses as described below.

Asymptomatic control plasma was purchased from Innovative Research (IPLASK2E2ML Novi, Michigan).

DNA extraction, library preparation, and NGS

DNA was extracted from 1 mL of each human plasma sample using the DNeasy Blood and Tissue kit (Qiagen, 69504, Hilden, Germany) according to the manufacturer protocol and recommendations. Extracted DNA was delivered to SeqMatic (Fremont, California) for DNA library preparation, which was performed using the Nextera XT library preparation kit (Illumina, FC-131-1096, San Diego, California) and then subsequent sequencing on the NovaSeq 6000 platform (1 x 100bp) at an average of 18.67 million reads per sample.

Sequencing read quality control

Raw sequencing read quality was assessed using the FASTQC software[34]. Trimming of adapter sequences and low quality individual reads was done using Trimmomatic 0.39[35] with default settings for single end sequencing. MultiQC[47] was used to assemble a quality report across all samples post-trimming to confirm average read quality above the default trimming threshold.

Human read removal

Reads passing quality control were then mapped to human reference genome GRCh38 with Bowtie2 v2.4.4[48]. Reads that did not map were considered to be putative non-human reads and were extracted to a BAM file using Samtools v1.15 with flag 0x4 and converted to a FASTQ file[49]. Seqtk was used to convert FASTQ files to FASTA files[50]. These sequences were then subjected to an additional layer of human read removal: BLAST v2.13.0[36,37] was then used to align putative non-human reads to all human-associated sequences (those with taxids 9606, 63221, or 741158) in NCBI Nucleotide (www.ncbi.nlm.nih.gov/nuccore), retrieved on 11/28/2021. All reads with hits to human sequences with an e-value below 0.0005 were discarded and all remaining reads were considered to be non-human.

Superkingdom read annotation

Superkingdom distribution of non-human reads was done by nucleotide BLAST search against the full NCBI Nucleotide database (retrieved on 11/28/2021 as above). A negative taxid list containing primate taxids was provided to avoid any incorrect annotations that could result from residual human-origin reads not removed by the above rigorous human read depletion. Output included taxids, read ID, and hit superkingdom. The distribution of superkingdoms was assessed in R Statistical Software (v4.1.2; R Core Team 2021[51]). Reads with matches to both bacteria and bacteriophage were characterized as potential prophage.

Bacteriophage annotations

The phage sequence reference database was constructed from phage genomes retrieved from NCBI Nucleotide on 12/07/2021 (NCBI, US National Library of Medicine, retrieved as above); the search was restricted to viral sequences and used terms “phage” or “bacteriophage.” Duplicate sequences, or sequences of only ORFs, genes, proteins, regions, or modified genomes were excluded. This yielded 26,159 phage sequences, which were combined into a single FASTA file and used to generate a local phage BLAST database. Cleaned and processed reads were annotated by performing a BLAST alignment search of non-human reads against our phage reference database. BLAST parameters were adjusted to require an e-value of less than 0.0005. All reads with bacteriophage hits below the significance threshold were then subjected to a BLAST alignment

search limiting the subset of the NCBI Nucleotide database with human taxonomic IDs to identify bacteriophage matching reads with human homology. The reads with hits to human sequences were then removed, and the remaining sequences were subjected to a second BLAST search against the phage reference database with a word size of 28, use of 8 threads, an e-value cutoff of less than 0.0005, a culling limit of 10 to capture potential identities of various bacteriophages for reads covering highly conserved sequences across bacteriophages. Output included qseqid, sseqid, stitle, pident, length, and evalue. A dictionary of phage and their bacterial host(s) was created using annotations from virus-host DB [Mihara2016], from the phage NCBI taxonomy entry, by phage name, and manual curation from manuscripts associated with the phage sequence submission. Phage with only genus level host detail available were noted as such. Further characterization of host bacterial qualities was performed through literature search.

Interpretation of phage annotations

Phage sequence hits were analyzed using R. Phage accession IDs were extracted and the taxize package was used to assign corresponding taxonomic lineage information[52]. Phage diversity was calculated using the SDI [39] through the R package vegan[53].

Bacterial mapping

Bacterial reads were assessed through mapping to the reference *S. aureus* bacterial genome NC_007795.1 using Bowtie2 v2.4.4[48] with default settings. Number of mapping reads were obtained using samtools view, with flag -F 0x04 used to obtain a coverage output from the mapped SAM file output.

Statistics

Statistics were performed using Prism 9.3.1 (GraphPad Software, San Diego, California). This included unpaired, two tailed t-tests, Brown-Forsythe ANOVAs, receiver-operator curve generation, and Pearson correlations. Please see figure legends for specific statistical tests performed in each figure.

References

1. *Improving the prevention, diagnosis and clinical management of sepsis*. (2017).
2. Grabuschnig, S. *et al.* Putative Origins of Cell-Free DNA in Humans: A Review of Active and Passive Nucleic Acid Release Mechanisms. *Int. J. Mol. Sci.* **21**, 1–24 (2020).
3. Kowarsky, M. *et al.* Numerous uncharacterized and highly divergent microbes which colonize humans are revealed by circulating cell-free DNA. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 9623–9628 (2017).
4. Cheng, A. P. *et al.* Cell-free DNA profiling informs all major complications of hematopoietic cell transplantation. *Proc. Natl. Acad. Sci. U. S. A.* **119**, (2022).
5. Fan, H. C., Blumenfeld, Y. J., Chitkara, U., Hudgins, L. & Quake, S. R. Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 16266–16271 (2008).
6. De Vlaminck, I. *et al.* Circulating cell-free DNA enables noninvasive diagnosis of heart transplant rejection. *Sci. Transl. Med.* **6**, (2014).
7. Snyder, T. M., Khush, K. K., Valantine, H. A. & Quake, S. R. Universal noninvasive detection of solid organ transplant rejection. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 6229–6234 (2011).
8. Schwarzenbach, H., Hoon, D. S. B. & Pantel, K. Cell-free nucleic acids as biomarkers in cancer patients. *Nat. Rev. Cancer* **11**, 426–437 (2011).
9. Grumaz, S. *et al.* Next-generation sequencing diagnostics of bacteremia in septic patients. *Genome Med.* **8**, (2016).
10. Grumaz, C. *et al.* Rapid Next-Generation Sequencing-Based Diagnostics of Bacteremia in Septic Patients. *J. Mol. Diagn. JMD* **22**, 405–418 (2020).
11. Chen, P. *et al.* Rapid diagnosis and comprehensive bacteria profiling of sepsis based on cell-free DNA. *J. Transl. Med.* **18**, (2020).
12. Wang, L. *et al.* Plasma Microbial Cell-Free DNA Sequencing Technology for the Diagnosis of Sepsis in the ICU. *Front. Mol. Biosci.* **8**, (2021).
13. Eichenberger, E. M. *et al.* Microbial Cell-Free DNA Identifies the Causative Pathogen in Infective Endocarditis and Remains Detectable Longer Than Conventional Blood Culture in Patients with Prior Antibiotic Therapy. *Clin. Infect. Dis.* (2022) doi:10.1093/CID/CIAC426.
14. Burnham, P. *et al.* Urinary cell-free DNA is a versatile analyte for monitoring infections of the urinary tract. *Nat. Commun. 2018 91* **9**, 1–10 (2018).
15. Hogan, C. A. *et al.* Clinical Impact of Metagenomic Next-Generation Sequencing of Plasma Cell-Free DNA for the Diagnosis of Infectious Diseases: A Multicenter Retrospective Cohort Study. *Clin. Infect. Dis. Off. Publ. Infect. Dis. Soc. Am.* **72**, 239–245 (2021).
16. Cheng, H. K. *et al.* Combined use of metagenomic sequencing and host response profiling for the diagnosis of suspected sepsis. *bioRxiv* 854182–854182 (2019) doi:10.1101/854182.
17. Sinha, M. *et al.* Emerging Technologies for Molecular Diagnosis of Sepsis. *Clin. Microbiol. Rev.* **31**, (2018).
18. Hatfull, G. F. Dark Matter of the Biosphere: the Amazing World of Bacteriophage Diversity. *J. Virol.* **89**, 8107–8110 (2015).
19. Shkoporov, A. N. & Hill, C. Bacteriophages of the Human Gut: The 'Known Unknown' of the Microbiome. *Cell Host Microbe* **25**, 195–209 (2019).
20. de Jonge, P. A., Nobrega, F. L., Brouns, S. J. J. & Dutilh, B. E. Molecular and Evolutionary Determinants of Bacteriophage Host Range. *Trends Microbiol.* **27**, 51–63 (2019).
21. Flores, C. O., Meyer, J. R., Valverde, S., Farr, L. & Weitz, J. S. Statistical structure of host-phage interactions. *Proc. Natl. Acad. Sci. U. S. A.* **108**, (2011).
22. Koskella, B. & Meaden, S. Understanding Bacteriophage Specificity in Natural Microbial Communities. *Viruses 2013 Vol 5 Pages 806-823* **5**, 806–823 (2013).
23. Nguyen, S. *et al.* Bacteriophage Transcytosis Provides a Mechanism To Cross Epithelial Cell Layers. *mBio* **8**, (2017).
24. Górski, A. *et al.* Bacteriophage translocation. *FEMS Immunol. Med. Microbiol.* **46**, 313–319 (2006).

25. Navarro, F. & Muniesa, M. Phages in the Human Body. *Front. Microbiol.* **8**, 566–566 (2017).
26. Manrique, P. *et al.* Healthy human gut phageome. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 10400–10405 (2016).
27. Zhang, T. *et al.* RNA viral community in human feces: prevalence of plant pathogenic viruses. *PLoS Biol.* **4**, 0108–0118 (2006).
28. Huang, Y.-F. *et al.* Analysis of microbial sequences in plasma cell-free DNA for early-onset breast cancer patients and healthy females. *BMC Med. Genomics* **11**, 16 (2018).
29. Nayfach, S. *et al.* Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. *Nat. Microbiol.* **6**, 960–970 (2021).
30. Camarillo-Guerrero, L. F., Almeida, A., Rangel-Pineros, G., Finn, R. D. & Lawley, T. D. Massive expansion of human gut bacteriophage diversity. *Cell* **184**, 1098-1109.e9 (2021).
31. Tisza, M. J. & Buck, C. B. A catalog of tens of thousands of viruses from human metagenomes reveals hidden associations with chronic diseases. *Proc. Natl. Acad. Sci.* **118**, e2023202118 (2021).
32. Gu, W. *et al.* Rapid pathogen detection by metagenomic next-generation sequencing of infected body fluids. *Nat. Med. 2020 271* **27**, 115–124 (2020).
33. Blauwkamp, T. A. *et al.* Analytical and clinical validation of a microbial cell-free DNA sequencing test for infectious disease. *Nat. Microbiol. 2019 44* **4**, 663–674 (2019).
34. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data.
35. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinforma. Oxf. Engl.* **30**, 2114–2120 (2014).
36. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
37. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
38. Mihara, T. *et al.* Linking Virus Genomes with Host Taxonomy. *Viruses* **8**, (2016).
39. Shannon, C. E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948).
40. Kirschbaum, J. O. & Kligman, A. M. THE PATHOGENIC ROLE OF CORYNEBACTERIUM ACNES IN ACNE VULGARIS. *Arch. Dermatol.* **88**, 832–833 (1963).
41. Hotchkiss, R. S. *et al.* Sepsis and septic shock. *Nat. Rev. Dis. Primer* **2**, 16045–16045 (2016).
42. Doern, G. V. *et al.* Practical Guidance for Clinical Microbiology Laboratories: A Comprehensive Update on the Problem of Blood Culture Contamination and a Discussion of Methods for Addressing the Problem. *Clin. Microbiol. Rev.* **33**, (2019).
43. Yamamoto, K., Mezaki, K. & Ohmagari, N. Simple indictor of increased blood culture contamination rate by detection of coagulase-negative staphylococci. *Sci. Rep.* **11**, (2021).
44. Long, Y. *et al.* Diagnosis of Sepsis with Cell-free DNA by Next-Generation Sequencing Technology in ICU Patients. *Arch. Med. Res.* **47**, 365–371 (2016).
45. Barrett, S. L. R. *et al.* Cell free DNA from respiratory pathogens is detectable in the blood plasma of Cystic Fibrosis patients. *Sci. Rep.* **10**, (2020).
46. Wang, X. *et al.* Cryptic prophages help bacteria cope with adverse environments. *Nat. Commun.* **1**, (2010).
47. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinforma. Oxf. Engl.* **32**, 3047–3048 (2016).
48. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
49. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, (2021).
50. GitHub - lh3/seqtk: Toolkit for processing sequences in FASTA/Q formats.
51. R Core Team (2020). — European Environment Agency.
52. Chamberlain, S. A. & Szöcs, E. Taxize: Taxonomic search and retrieval in R. *F1000Research* **2**, (2013).
53. CRAN - Package vegan.

Acknowledgements

We would like to thank Tim Blauwkamp (Karius Inc) for his assistance providing additional metadata for the SepSeq dataset. PLB is supported by the NIH (R01 HL148184-01, R01 AI12492093, R01 DC019965), the Cystic Fibrosis Foundation, and a grant from the Emerson Collective. NLH is supported by the NSF GRFP. LJB is supported by the NIH (T32 HL129970-06). SY is supported by the NIH (R01 AI153133, R01 AI137272, R01 AI138978)

Author Contributions

N.L.H., L.J.B., N.R.M., G.K., S.Y., and P.L.B. designed the study. N.L.H., L.J.B., and N.R.M performed experiments. N.L.H., L.J.B., G.K., N.R.M., S.Y., A.S.B., C.Y.C. and P.L.B. analyzed data. N.L.H., L.J.B., and P.L.B. wrote the manuscript.

Competing Interests

ASB has consulted for biomX and is on the scientific advisory boards of ArcBio and Caribou Biosciences.

Materials & Correspondence

All commercially obtained materials used in this study have catalog numbers noted in relevant methods sections. The entire volume of human plasma samples were used as described in the methods and therefore cannot be made available.
Sequencing data with human reads removed have been deposited into NCBI SRA under bioproject PRJNA860730 and can be accessed at reviewer link :
https://dataview.ncbi.nlm.nih.gov/object/PRJNA860730?reviewer=22tn13ho0q7324nf0rlkuuli76.
Publicly available data utilized: The SepSeq study data has been previously published under bioproject PRJNA507824. The bodily fluid sequencing study data has been previously published under bioproject PRJNA558701. No new computational tools were developed as part of this study.

Tables

Table S1: Distribution of blood culture positive bacterial identifications in the sequenced sepsis cohort (N = 62).

Cronobacter sakazakii	1
Cryptococcus neoformans	1
Enterobacter aerogenes	2
Enterobacter cloacae	1
Enterococcus sp.	1
Escherichia coli	14
Escherichia coli and Klebsiella pneumoniae	1
Escherichia coli and Streptococcus sp.	1
Gram Positive cocci	1
Klebsiella pneumoniae	5
Moraxella sp.	1
Paenibacillus sp.	1
Proteus vulgaris	1
Pseudomonas aeruginosa	6
Salmonella enterica	1
Serratia marcescens	1
Staphylococcus aureus	5
Staphylococcus aureus and coagulate-negative Staphylococcus sp.	1
Staphylococcus aureus and Klebsiella pneumoniae	1
Staphylococcus sp., coagulase-negative	5
Streptococcus sp.	11
Streptococcus pneumoniae	(2)
Streptococcus mitis	(2)
Streptococcus pyogenes	(1)

Table S2: Mann-Whitney test summary of Phage SDI between Asymptomatic and Septic patient samples. All genus specific SDIs had an adjusted p value greater than 0.05.

	Discovery?	P value	Mean rank of Asymptomatic	Mean rank of Sepsis	Mean rank diff.	Mann-Whitney U	Adj. p value
Campylobacter	No	0.113756	26.75	38.07	-11.32	212.5	0.57447
Pseudomonas	No	0.497073	40.75	35.81	4.935	267.5	0.772375
Burkholderia	No	0.775651	34.75	36.78	-2.032	292.5	0.870453
Acinetobacter	No	0.175584	28.1	37.85	-9.755	226	0.591131
Escherichia	No	0.003646	19.15	39.3	-20.15	136.5	0.073656
Klebsiella	No	0.679095	33.9	36.92	-3.019	284	0.857357
Salmonella	No	0.560555	40.1	35.92	4.181	274	0.808801
Shigella	No	0.421434	32.1	37.21	-5.11	266	0.709415
Aeromonas	No	0.214896	44.2	35.26	8.942	233	0.620128

Not annotated	No	0.044566	48.8	34.52	14.28	187	0.300079
Synechococcus	No	0.844428	35.35	36.69	-1.335	298.5	0.872229
Staphylococcus	No	0.863593	35.4	36.68	-1.277	299	0.872229
Streptococcus	No	0.654721	39.1	36.08	3.019	284	0.857357
Lactococcus	No	0.774446	35.3	36.69	-1.394	298	0.870453
Propionibacterium	No	0.417956	41.05	35.77	5.284	264.5	0.709415
Mycobacterium	No	0.030548	48.9	34.5	14.4	186	0.300079
Gordonia	No	0.409273	41.1	35.76	5.342	264	0.709415
Dickeya	No	0.143066	27.45	37.96	-10.51	219.5	0.577986
Cutibacterium	No	0.301805	42.2	35.58	6.619	253	0.709415
Other	No	0.336353	42.5	35.53	6.968	250	0.709415

Figures

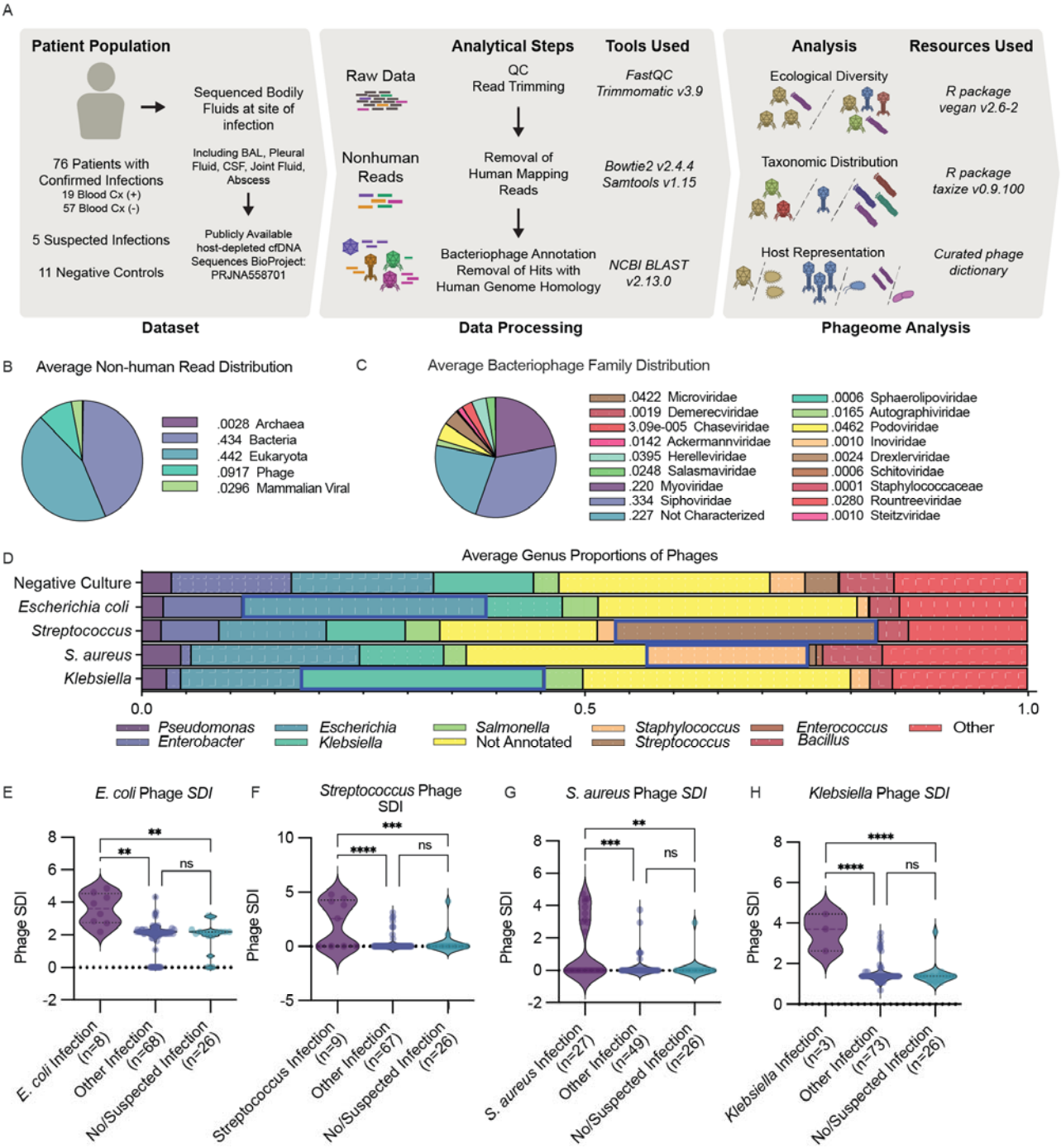


Figure 1. Local infections are associated with pathogen-specific phage overrepresentation.
A) Schematic of sample numbers, sample processing, and analytical approach. B) Distribution of non-human reads identity by Archeal, Eukaryotic, Bacterial, Mammalian Viral, and Bacteriophage categories. (Archaea mean 0.0028 SD 0.026, Eukaryotic mean 0.442 SD 0.322, Bacterial mean 0.434 SD 0.362, Mammalian Viral mean 0.0296 SD 0.067, Bacteriophage mean 0.0917 SD 0.1086) C) Average bacteriophage distribution by bacteriophage family. D) Average distribution of unique phages by bacterial host genus by infection category, with causative pathogen's own bar outlined in blue box. E-H) Infection specific Shannon DI calculated and shown as violin plots with median and quartiles shown by dashed lines for E) E. coli (One-way anova $P < 0.0001$, mean E. coli Infection Phage SDI = 3.591, mean Other Infection Phage SDI = 1.884, mean No/Suspected Infection Phage SDI =1.884. One-way anova Multiple comparisons: E. coli Infection vs other infection $P < 0.0001$, E. coli Infection vs No/Suspected Infection $P = 0.0002$, Other Infection vs No/Suspected Infection $P > 0.9999$), F) Streptococcus (One-way anova $P < 0.0001$, mean Streptococcus Infection Phage SDI = 2.187, mean Other Infection Phage SDI = 0.2422, mean No/Suspected Infection Phage SDI 0.3494. One-way anova Multiple comparisons: Streptococcus Infection vs other infection $P < 0.0001$, Streptococcus Infection vs No/Suspected Infection $P = 0.0001$, Other Infection vs No/Suspected Infection $P = 0.9251$), G) S. aureus (One-way anova $P = 0.0003$, mean S. aureus Infection Phage SDI = 1.362, mean Other Infection Phage SDI = 0.2178, mean No/Suspected Infection Phage SDI 0.1963. One-way anova Multiple comparisons: S. aureus Infection vs other infection $P = 0.0003$, S. aureus Infection vs No/Suspected Infection $P = 0.0082$, Other Infection vs No/Suspected Infection $P = 0.9979$), and H) Klebsiella Infections (One-way anova $P < 0.0001$, mean Klebsiella Infection Phage SDI = 3.606, mean Other Infection Phage SDI = 1.518, mean No/Suspected Infection Phage SDI 1.490. One-way anova Multiple comparisons: Klebsiella Infection vs other infection $P < 0.0001$, Klebsiella Infection vs No/Suspected Infection $P < 0.0001$, Other Infection vs No/Suspected Infection $P = 0.9820$)

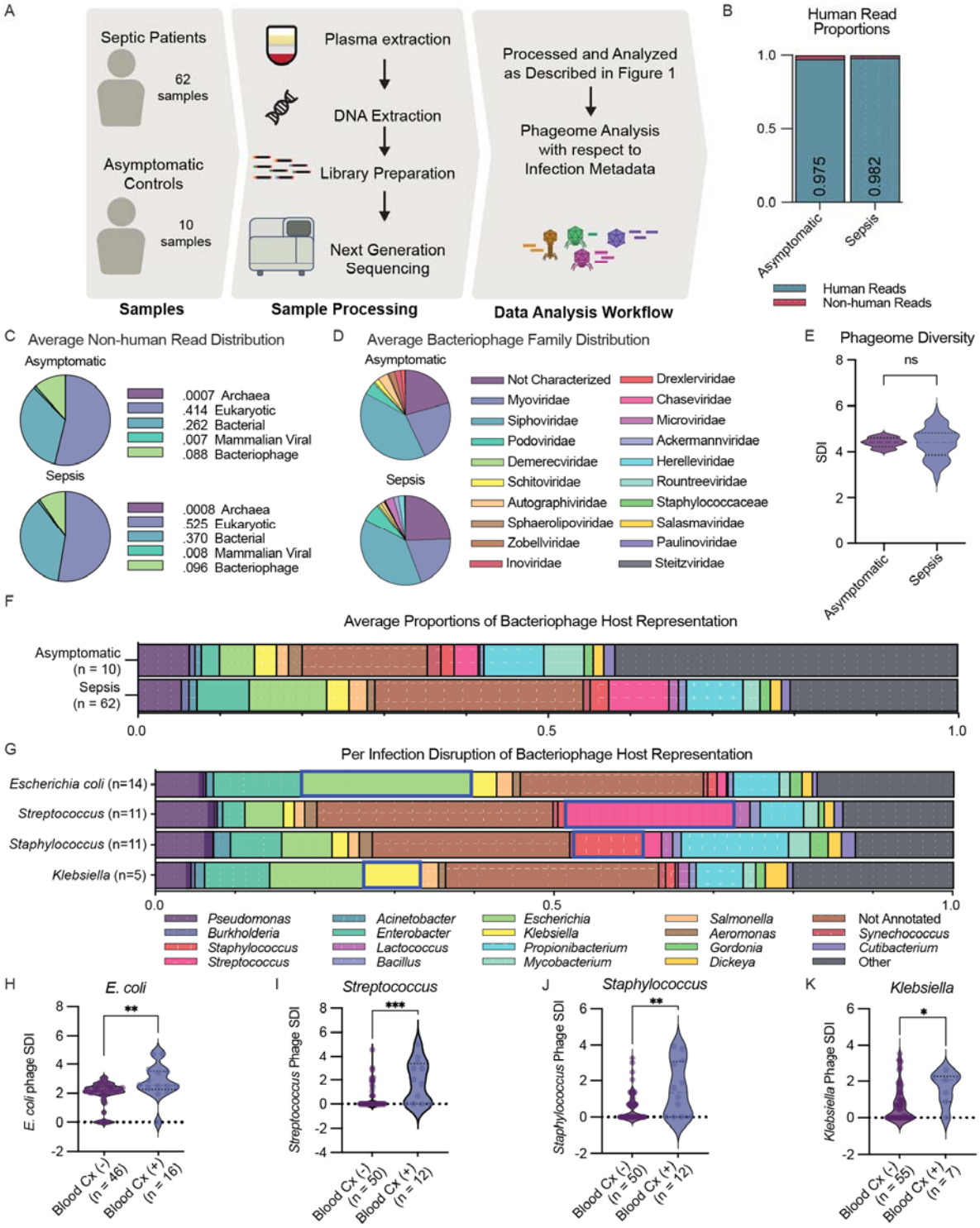


Figure 2. The circulating phageome is mostly unaltered by sepsis but reflects infection etiology.
A) Schematic of sample numbers, sample processing, and analytical approach. B) Proportions of human and non-human reads. C) Distribution of non-human reads identity by Archeal, Eukaryotic, Bacterial, Mammalian Viral, and Bacteriophage categories. Asymptomatic (Archaea mean 0.0007 SD 0.0003, Eukaryotic mean 0.420 SD 0.126, Bacterial mean 0.486 SD 0.126, Mammalian Viral mean 0.006 SD 0.002, Bacteriophage mean 0.0877 SD 0.0223) Sepsis (Archaea mean 0.0008 SD 0.0005, Eukaryotic mean 0.522 SD 0.278, Bacterial mean 0.368 SD 0.282, Mammalian Viral mean 0.008 SD 0.018, Bacteriophage mean 0.1009 SD 0.083) D) Distribution of bacteriophage distribution of bacteriophage family. E) Shannon Diversity Index in asymptomatic and septic patient phageomes is not significantly different (Asymptomatic: mean = 4.41, range = 0.0655, Sepsis: mean = 4.38, range = 2.992, P = 0.89 as determined by unpaired two-tailed T-test) shown by violin plot with median and quartiles shown by dashed lines. F) Average distribution of unique phages by bacterial host genus in asymptomatic and septic patient phageomes. In Asymptomatic samples, mean proportions and SD (Pseudomonas: mean 0.059, SD 0.035, Acinetobacter: mean 0.007 SD 0.004, Escherichia: mean 0.043 SD 0.021, Salmonella: mean 0.015 SD 0.013, Not Annotated: mean 0.16 SD 0.06, Burkholderia: mean 0.0063 SD 0.0060, Enterobacter: mean 0.0024 SD 0.034, Klebsiella: mean 0.024 SD 0.027, Aeromonas: mean 0.017 SD 0.011, Synechococcus: mean 0.013 SD 0.033, Staphylococcus: mean 0.015 SD 0.027, Lactococcus: mean 0.0019 SD 0.0045, Propionibacterium: mean 0.074 SD 0.082, Gordonia: mean 0.012, SD 0.016, Cutibacterium: mean 0.014 SD 0.015, Streptococcus: mean 0.024 SD 0.037, Bacillus: mean 0.057 SD 0.053, Mycobacterium: mean 0.048 SD 0.047, Dickeya: mean 0.013, Other Host: mean 0.42 SD 0.065). In Sepsis samples, mean proportions and SD (Pseudomonas: mean 0.057 SD 0.072, Acinetobacter: mean 0.017 SD 0.024, Escherichia: mean 0.096 SD 0.080, Salmonella: mean 0.015 SD 0.017, Not Annotated: mean 0.23 SD 0.11, Burkholderia: mean 0.0095 SD 0.012, Enterobacter: mean 0.051 SD 0.062, Klebsiella: mean 0.029 SD 0.046, Aeromonas: mean 0.014 SD 0.014, Synechococcus: mean 0.0071 SD 0.016, Staphylococcus: mean 0.023 SD 0.054, Lactococcus: 0.015 SD 0.054, Propionibacterium: mean 0.049 SD 0.074, Gordonia: mean 0.012 SD 0.021, Cutibacterium: mean .0081 SD 0.014, Streptococcus: mean 0.030 SD 0.063, Bacillus: mean 0.017 SD 0.023, Mycobacterium: mean 0.021 SD 0.040, Dickeya: 0.019 SD 0.015, Other Host: mean 0.27 SD 0.14). G) Average proportions of unique phage by bacterial host in samples with single identified pathogen, by cultured pathogen (Streptococcus, Staphylococcus, Pseudomonas, Klebsiella, and Escherichia) with bar corresponding to row's infectious pathogen outlined in blue box. H-K) Infection specific Shannon DI calculated and shown as violin plots with median and quartiles shown by dashed lines by H) E. coli infection status (mean Blood Cx (-): 2.00, mean Blood Cx (+): 2.82, P = 0.004). I) Streptococcus infection status (mean Blood Cx (-): 0.535, mean Blood Cx (+): 1.89, P = 0.0008). J) Staphylococcus infection status (mean Blood Cx (-): 0.560, mean Blood Cx (+): 1.59, P = 0.002). K) Klebsiella infection status (mean Blood Cx (-): 0.775, mean Blood Cx (+): 1.63, P = 0.038). All statistics were performed using two-tailed, unpaired t tests.

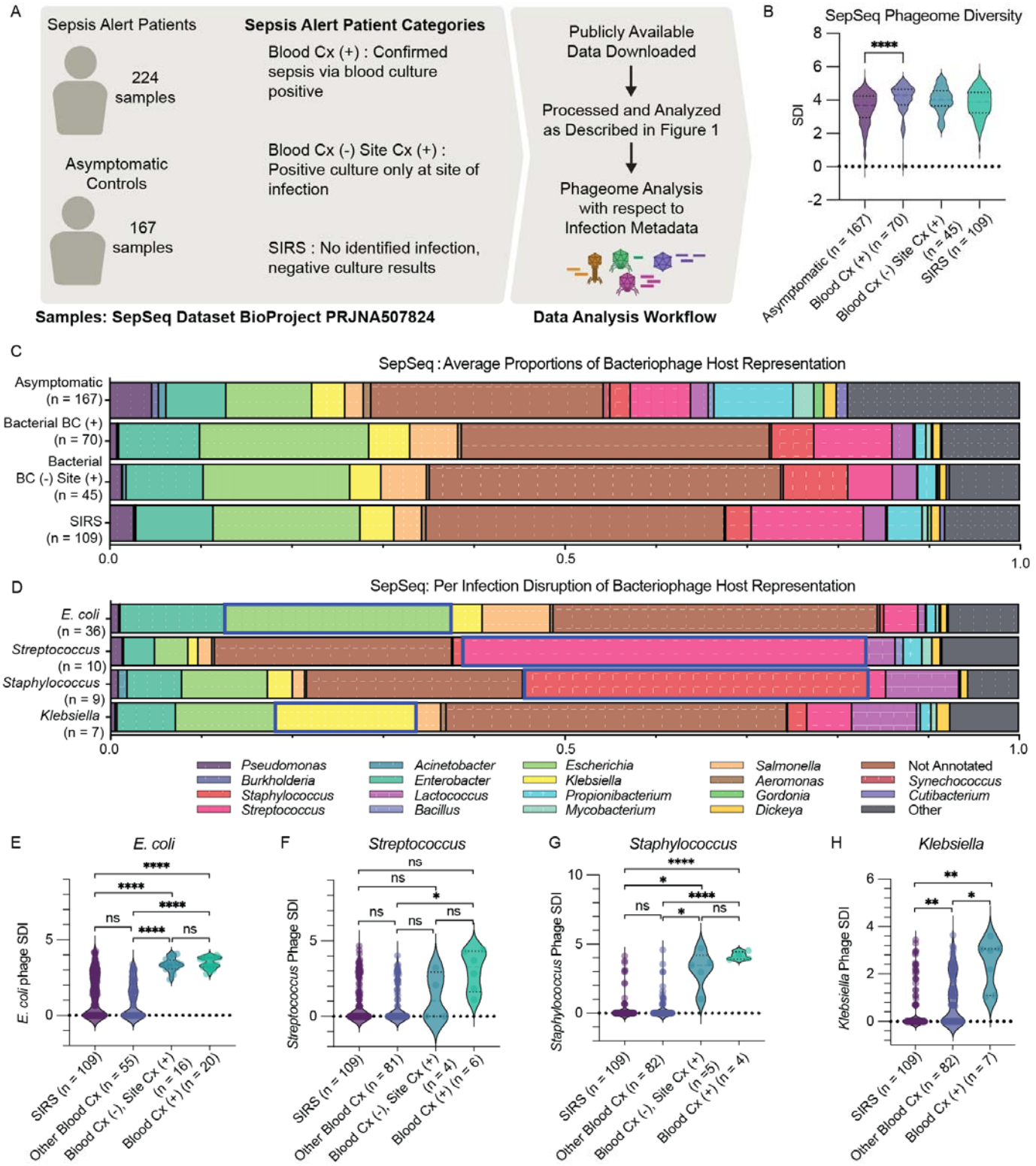


Figure 3. The circulating phageome reflects individual pathogens in another large study.
A) SepSeq Sample Schematic. B) Shannon DI of Phages by sample category, (mean Blood Cx (+): 4.05, mean Asymptomatic: 3.48, $P < 0.001$). C) Average proportions of Unique Phage by bacterial host by sample category. D) Per Infection disruption of average proportions unique phage with single identified pathogen, by cultured pathogen in Streptococcus, Staphylococcus, Pseudomonas, Klebsiella, and Escherichia infection, with bar corresponding to row's infectious pathogen outlined in blue box. (E-I) Pathogen Host Phage Shannon Diversity by infection category, all statistics performed with Brown-Forsythe ANOVA with multiple conparisons shown as violin plots with median and quartiles shown by dashed lines for E) E. coli (mean SIRS: 1.45, mean Other Blood Cx: 1.07, mean Blood Cx (-) Site Cx (+): 3.35, mean Blood Cx (+): 3.48, Brown-Forsythe ANOVA $p < 0.0001$), F) Streptococcus (mean SIRS: 1.02, mean Other Blood Cx: 0.760, mean Blood Cx (-) Site Cx (+): 1.32, mean Blood Cx (+): 3.03, Brown-Forsythe ANOVA $P = 0.017$), G) Staphylococcus (mean SIRS: 0.266, mean Other Blood Cx: 0.316, mean Blood Cx (-) Site Cx (+): 3.17, mean Blood Cx (+): 4.16, Brown-Forsythe ANOVA $P = 0.0009$), H) Klebsiella (mean SIRS: 0.510, mean Other Blood Cx: 1.00, mean Blood Cx (+): 2.43, Brown-Forsythe ANOVA $P < 0.0001$)

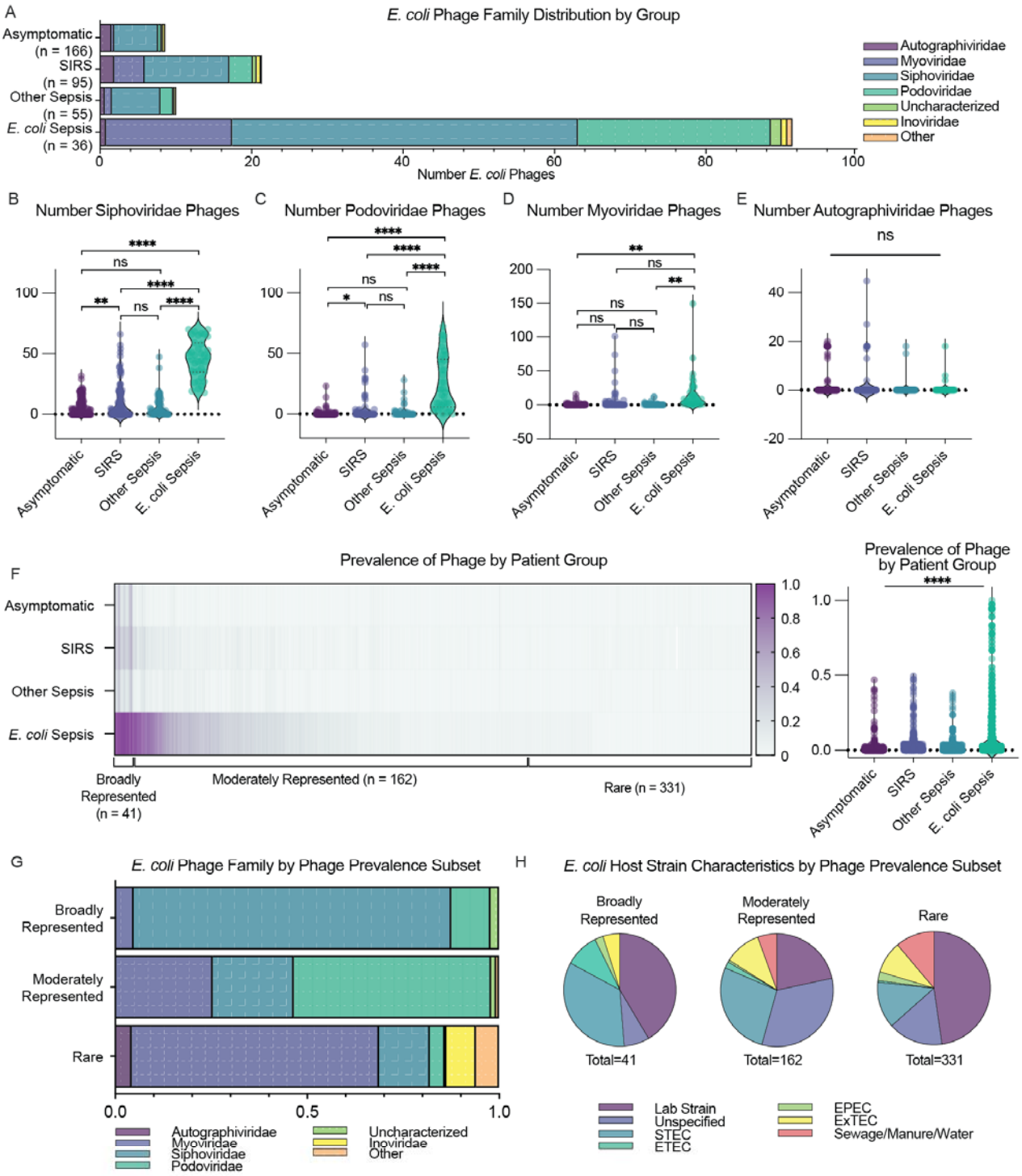


Figure 4: E. coli phage reflect host strain characteristics

A) In samples with E. coli phage detectable, E. coli bacteriophage numbers by family vary by infection status. (B-E) Number of E. coli phages by phage families B) E. coli Siphoviridae phages (mean Asymptomatic: 5.80, mean SIRS: 11.19, mean Other Sepsis: 6.40, mean E. coli Sepsis: 45.67, Brown-Forsythe ANOVA P < 0.0001), C) E. coli Podovidirae phages (mean Asymptomatic: 0.44, mean SIRS: 3.08, mean Other Sepsis: 1.66, mean E. coli Sepsis: , Brown-Forsythe ANOVA P < 0.0001), D) Number of E. coli Myoviridae phages (mean Asymptomatic: 0.36, mean SIRS: 4.02 , mean Other Sepsis: 0.98 , mean E. coli Sepsis: 16.67 , Brown-Forsythe ANOVA P < 0.0001), E) E. coli Autographiviridae phages (Brown-Forsythe ANOVA test P = 0.32), F) Some individual E. coli phages are prevalent across multiple samples, with E. coli Sepsis samples reflecting phages more highly represented across the group (mean 0.1524 representation across E. coli Sepsis E. coli phages versus 0.01231 for Asymptomatic, 0.03364 for SIRS, and 0.01631 for Other Sepsis samples). G) Stacked bar plot of phage representation across samples by group of E. coli phages subset by representation level in E. coli Sepsis. H) E. coli phage morphology by phage representation groups. All violin plots have median and quartiles shown by dashed lines

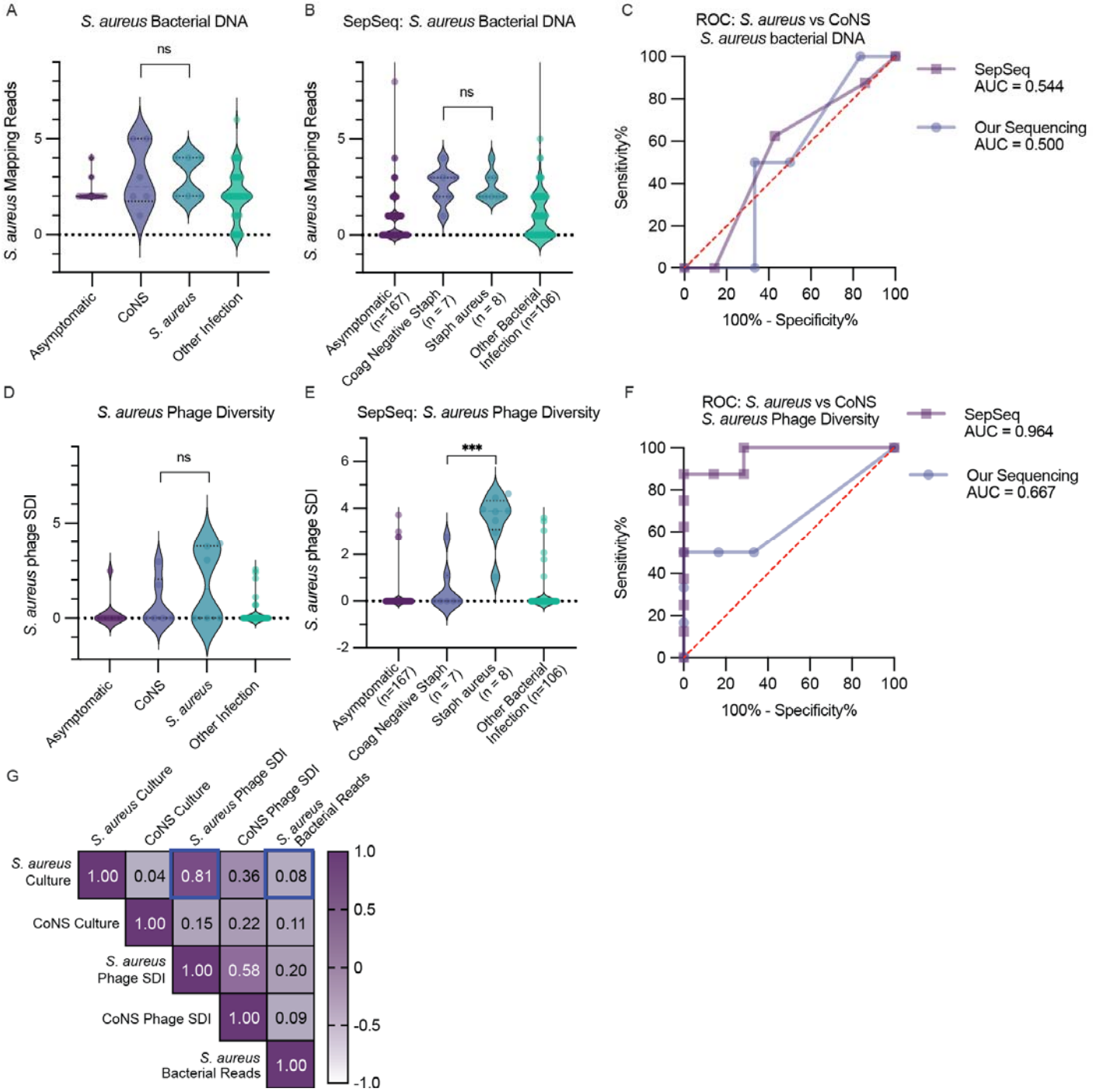
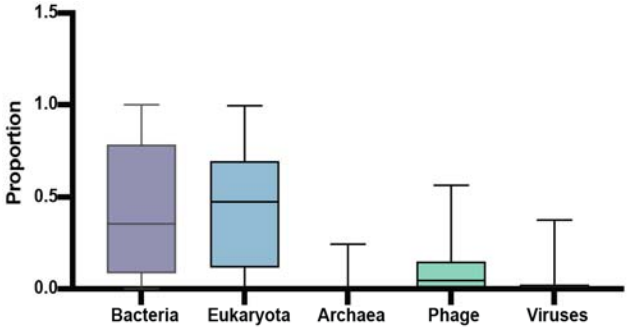


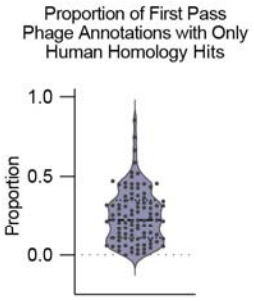
Figure 5. Phage can distinguish between Staphylococcus aureus and coagulase-negative Staphylococcus.
A) Reads mapping to S. aureus bacterial genome in our sequencing by infection category (mean CoNS: 3, mean S. aureus: 3, P > 0.99) B) Reads mapping to S. aureus bacterial genome in SepSeq by infection category (mean CoNS: 2.57, mean S. aureus: 2.38, P = 0.657), C) ROC lot for S. aureus bacterial genome in distinguishing S. aureus culture vs CoNS culture (AUC Our Sequencing 0.500, 95% Confidence Interval 0.146 to 0.854, P > 0.99, AUC SepSeq 0.544, 95% Confidence Interval 0.225 to 0.858, P = 0.772), D) S. aureus phage SDI by infection category in our sequencing (mean CoNS: 0.78, mean S. aureus: 1.78, P = 0.32), E) S. aureus SDI by infection category in SepSeq (mean CoNS: 0.553, mean S. aureus: 3.54, P = 0.0002), F) ROC plot for S. aureus phage SDI in distinguishing S. aureus culture from CoNS culture (AUC Our Sequencing 0.667, 95% Confidence Interval 0.339 to 0.995, P = 0.337, AUC SepSeq 0.964, 95% Confidence Interval 0.877 to 1, P = 0.0026), G) Correlation matrix by culture status, phage diversity, and S. aureus mapping bacterial reads in SepSeq samples (S. aureus culture and S. aureus phage SDI Pearson R: 0.81, 95% Confidence Interval 0.736 to 0.861, P = 1.65e-029, CoNS Culture and S. aureus phage SDI Pearson R: 0.15, 95% Confidence Interval -0.031 to 0.32, P = 0.10, S. aureus culture and S. aureus Bacterial Reads Pearson R: 0.08, 95% Confidence Interval -0.098 to 0.254, P = 0.37)



	Bacteria	Eukaryota	Archaea	Phage	Viruses
Minimum	0	0.0006613	0	0	0
Maximum	0.9992	0.9941	0.244	0.5645	0.3746
Range	0.9992	0.9934	0.244	0.5645	0.3746
Mean	0.434	0.4419	0.002847	0.09167	0.02957
Std. Deviation	0.3623	0.3219	0.02585	0.1086	0.06703
Std. Error of Mean	0.0384	0.03412	0.00274	0.01151	0.007105

Figure S1. Bar plots of proportions of non-human read identities by BLAST search with a table of descriptive statistics.

A



B

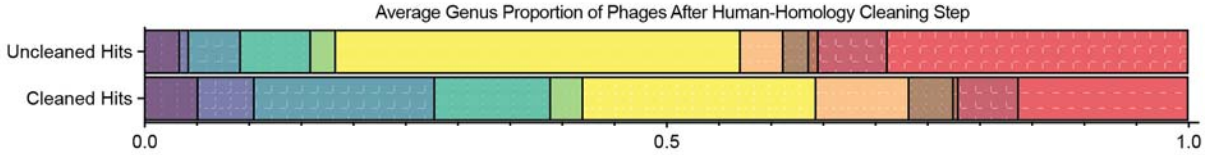
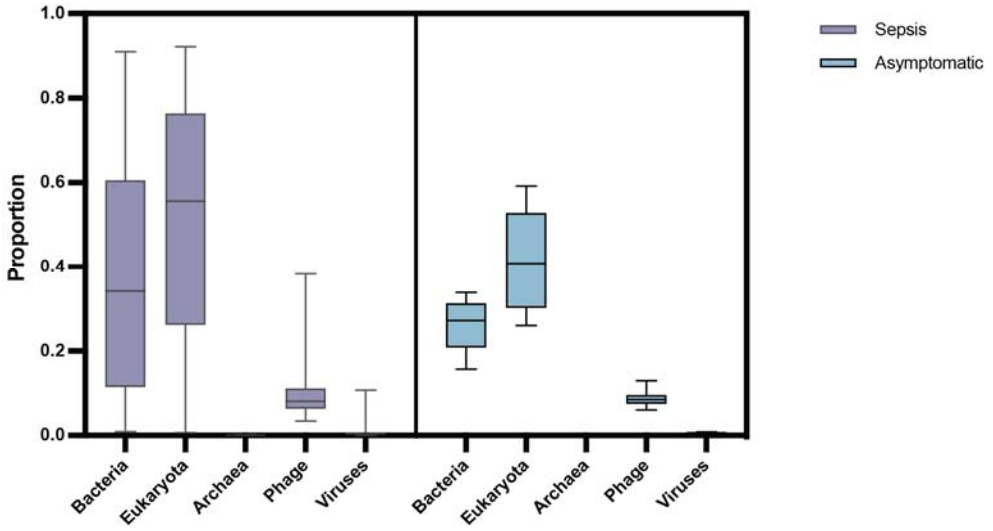


Figure S2. A secondary human sequence removal from phageome annotations removes potential false positives and impacts phageome host representation. A) Proportions of bacteriophage hits removed in secondary human sequence homolog removal step mean 0.2518 SD 0.1708, B) Average distribution of unique phages by bacterial host genus with and without secondary human sequence homology removal. Uncleaned Hits, mean proportions and SD (*Pseudomonas*: mean 0.0342 SD 0.0635, *Enterobacter* mean 0.0085 SD 0.0142, *Escherichia* mean 0.0497 SD 0.0473, *Klebsiella* mean 0.0673 SD 0.0712, *Salmonella* mean 0.0240 SD 0.0347, Not Annotated mean 0.387 SD 0.129, *Staphylococcus* mean 0.0409 SD 0.0569, *Streptococcus* mean 0.0246 SD 0.0459, *Enterococcus* mean 0.0093 SD 0.0196, *Bacillus* mean 0.0660 SD 0.0749, Other mean 0.288 SD 0.102), Cleaned Hits, mean proportions and SD (*Pseudomonas*: mean 0.0514 SD 0.0730, *Enterobacter* mean 0.0539 SD 0.142, *Escherichia* mean 0.173 SD 0.108, *Klebsiella* mean 0.111 SD 0.0743, *Salmonella* mean 0.0308 SD 0.0228, Not Annotated mean 0.223 SD 0.119, *Staphylococcus* mean 0.0890 SD 0.167, *Streptococcus* mean 0.0426 SD 0.141, *Enterococcus* mean 0.0005 SD 0.0203, *Bacillus* mean 0.0578 SD 0.0329, Other mean 0.162 SD 0.0690)



	Sepsis (n=62)					Asymptomatic (n=10)				
Category	Bacterial	Eukaryota	Archaea	Phage	Virus	Bacterial	Eukaryota	Archaea	Phage	Virus
Minimum	0.00924	0.006629	0	0.03389	0	0.1569	0.2599	0.0003482	0.06051	0.003917
Maximum	0.9098	0.9219	0.002253	0.3829	0.1069	0.3389	0.5922	0.00141	0.1295	0.008574
Range	0.9006	0.9152	0.002253	0.349	0.1069	0.182	0.3323	0.001062	0.069	0.004657
Mean	0.3697	0.5251	0.0007543	0.09615	0.00829	0.2615	0.4138	0.0007406	0.08804	0.006535
Std. Deviation	0.2653	0.2691	0.0004522	0.05743	0.01626	0.06084	0.1178	0.0003298	0.01896	0.001499
Std. Error of Mean	0.03369	0.03417	0.00005742	0.007294	0.002065	0.01924	0.03726	0.0001043	0.005997	0.000474

Figure S3. Bar plots of proportions of non-human read identities by BLAST search with a table of descriptive statistics for samples from both Septic and Asymptomatic individuals.

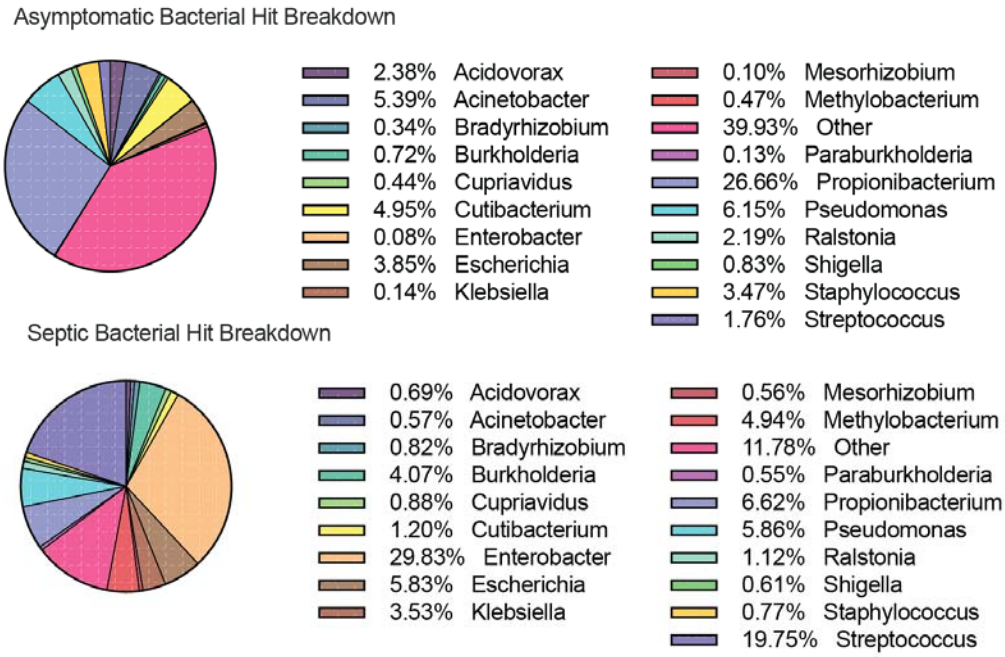


Figure S4. Average proportion of bacterial hit genus in asymptomatic and septic nonhuman cfDNA as identified by BLAST search.

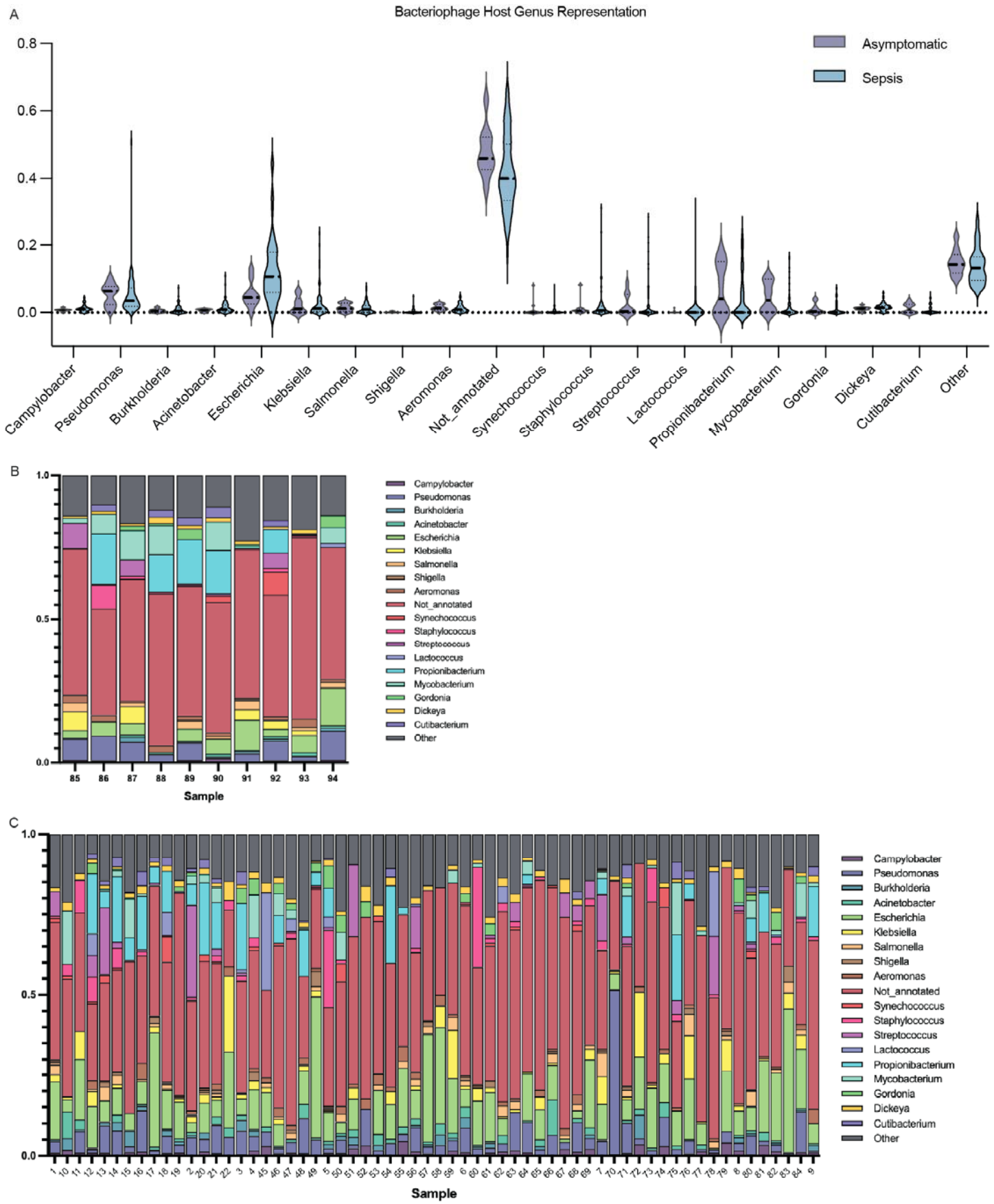


Figure S5. Phage bacterial host distribution does not change in sepsis, though individual variation remains. A) Violin plot of bacteriophage host genus proportions between Asymptomatic and Septic patient samples, associated statistics are in Table S1. B) Phage bacterial host genus proportions per asymptomatic patient. C) Phage bacterial host genus proportions per septic patient.

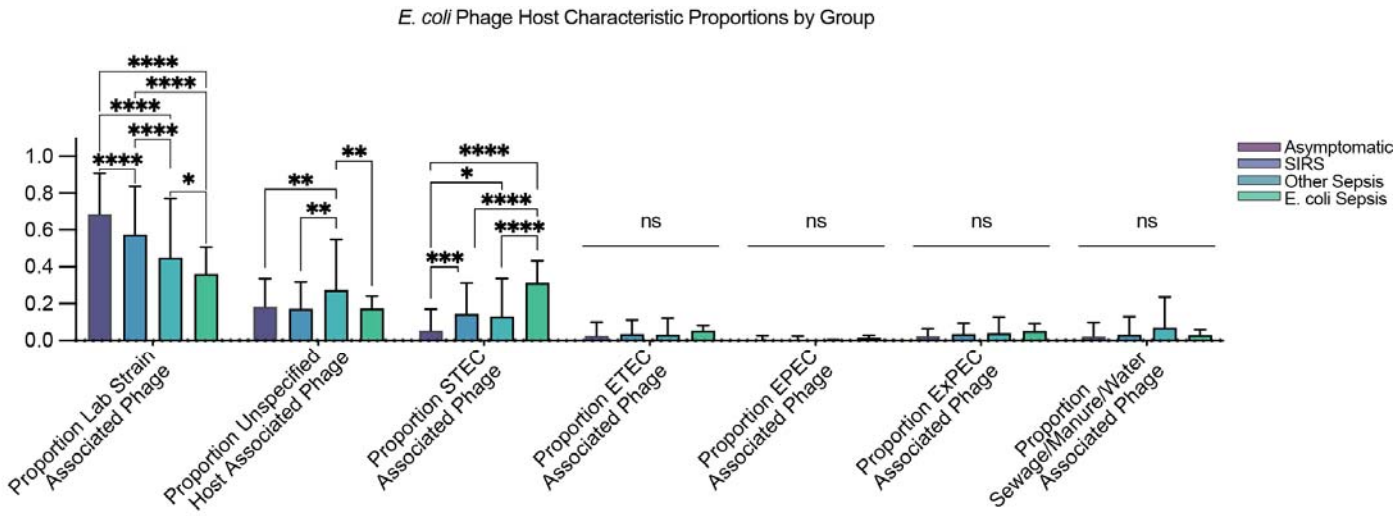


Figure S6. *E. coli* Phage Host Characteristic Proportions by Group. Differences in proportion of *E. coli* phage host characteristics from the literature are shown. Analyzed by two way ANOVA with multiple comparisons. Shown on plots are ns if no differences were found across characteristic proportion or between conditions.

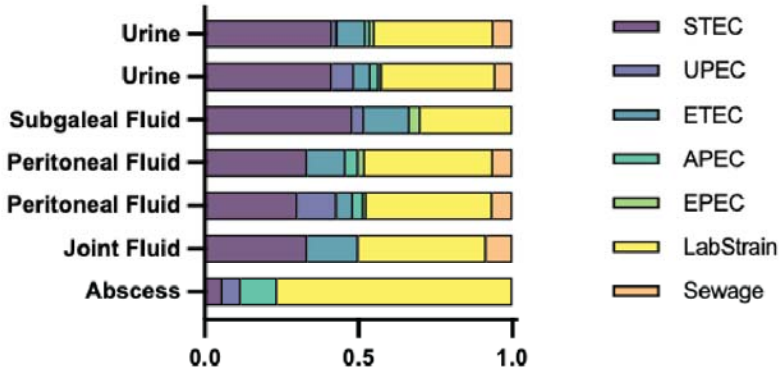


Figure S7. E. coli phage characteristics by site from re-analysis of data from Gu et al. Fraction of phage annotations corresponding to each strain type of E. coli in patients with documented E. coli infections. Each bar is one patient, the infected fluid sequenced is labeled to the left of each bar.