PLOS One

RESEARCH ARTICLE

Doing more with less: Genomic quasi-G-primes differentiate septic from healthy patients

Congzhou M. Sha[1], **Michail Patsakis**[1,2,3], **Ioannis Mouratidis**[1,3,4], **Xiaoyuan Wei**[4,5,6], **Taejung Chung**[4,5], **Jasna Kovac**[4,5], **Ilias Georgakopoulos-Soares**[1,3,4]*

1 Department of Molecular and Precision Medicine, Institute for Personalized Medicine, The Pennsylvania State University College of Medicine, Hershey, Pennsylvania, United States of America, **2** National Technical University of Athens, School of Electrical and Computer Engineering, Athens, Greece, **3** Department of Pharmacology and Toxicology, The University of Texas at Austin College of Pharmacy, Austin, Texas, United States of America, **4** Huck Institutes of the Life Sciences, The Pennsylvania State University, University Park, Pennsylvania, United States of America, **5** Department of Food Science, The Pennsylvania State University College of Agricultural Sciences, University Park, Pennsylvania, United States of America, **6** Department of Internal Medicine, Morsani College of Medicine, University of South Florida, Tampa, Florida, United States of America

* ilias@austin.utexas.edu

OPEN ACCESS

Citation: Sha CM, Patsakis M, Mouratidis I, Wei X, Chung T, Kovac J, et al. (2026) Doing more with less: Genomic quasi-G-primes differentiate septic from healthy patients. PLoS One 21(2): e0341828. https://doi.org/10.1371/journal.pone.0341828

Editor: Padmapriya P Banada, Rutgers Biomedical and Health Sciences, UNITED STATES OF AMERICA

Received: August 26, 2025

Accepted: January 13, 2026

Published: February 6, 2026

Data availability statement: The *Staphylococcus* raw reads have been deposited in the NCBI database, under accession number PRJNA1005703. The Supplemental Data is available on Zenodo (doi: 10.5281/zenodo.14205065). We have included Jupyter

Abstract

Sepsis is a life-threatening state of disseminated infection, and treatment requires knowledge of the organism responsible. The gold standard for sepsis diagnosis is blood culture, which requires days of growth. Next-generation sequencing has been proposed as an alternative; however, existing methods may lack sensitivity. In this work, we explore the idea of genomic quasi-G-primes, which are short DNA sequences specific to a single species within a group of relevant species. We first validated the genomic quasi-G-prime classification in controlled *Staphylococcus aureus* sequencing experiments, and then applied the same approach to blood-derived sequencing data from septic and healthy patients, where genomic quasi-G-prime profiles distinguished disease states. Our method is highly space-efficient, permitting fast classification on modest hardware and enabling it to outperform existing taxonomic classification approaches in this task.

Introduction

Next-generation sequencing (NGS) is increasingly used in clinical practice, such as in tumor clinical oncology to guide cancer therapy [1], and to diagnose rare genetic diseases [2]. Multiple experimental methods are encompassed by NGS, and the term refers to the ability to sequence many short nucleotide sequences (RNA or DNA) in parallel, allowing for an unbiased view of all nucleotides present in a sample. One of the fundamental barriers to interpreting NGS data is assignment of nucleotide reads to specific species, posing a significant challenge in microbiomics and complicating the application of NGS to infectious disease.

notebooks in the Supplemental Data, which reproduce our entire analysis. For the sepsis data, we only include the final master table (publication/data/master_table_17.ser.gz) which can be used by Gprime.jl, and classified read counts, because the associated genome assemblies and patient sample sequencing data require terabytes of space. Similarly, we only report the results of Kraken 2. Gprime.jl is also available on GitHub (https://github.com/Georgakopoulos-Soares-lab/Gprime.jl).

Funding: This project was funded in part by startup funds from the Penn State College of Medicine (to IGS) and by the Huck Innovative and Transformational Seed Fund (HITS) award from the Huck Institutes of the Life Sciences at Penn State University (to IGS). Additional support was provided by the National Institute of Food and Agriculture awards PEN04853 (to JK) and Multistate Project 4666 (to JK), and by the Lester Earl and Veronica Casida Career Development Endowment (to JK). CMS is grateful for support from the Penn State College of Medicine's Medical Scientist Training Program.

Competing interests: The authors Congzhou M Sha, Michail Patsakis, Ioannis Mouratidis, and Ilias Georgakopoulos-Soares hold a patent pending for technology related to this work.

Recently, the concept of genomic quasi-primes has been proposed to aid in organism identification and classification (Figs 1A-C) [3–5]. The essential insight is that not all sequences detected in a sample are equally informative, and one should focus on the subset of short sequences which are most likely to be unique and differentiate between species. In the absence of noise in the data, one would consider only those short sequences which occur in the genome of one species and not in any other, so that detection of such a sequence in a sample suggests that the species it belongs to is also present in the sample. In other words, such a sequence is highly specific, allowing the researcher to rule in the species. These sequences are not shared between any two or more genomes, resulting in the disjoint regions shown in Fig 1C. Note that in this work, we restricted our attention to quasi-primes for a specific set of bacterial, viral, and fungal species known to cause sepsis in humans.

Because NGS is also highly sensitive [6,7], if species-specific sequences are not detected, it is unlikely that the associated species is present at clinically relevant levels, allowing the researcher to rule out the species as the major infectious cause. The combination of high specificity and high sensitivity of NGS lends itself to identifying the causative organism in sepsis. In this work, we provide a proof of concept that genomic quasi-primes provide sufficient signal-to-noise ratio to identify *Staphylococcus aureus* in sequencing experiments and to classify septic versus healthy patients in clinical settings.

Unlike polymerase chain reaction (PCR), NGS is a hypothesis-free method of capturing the genetics of a sample. PCR relies upon curation of primers, which biases the sequences which are detected. In principle, NGS detects all nucleotide sequences present in the sample without preference. Although we use NGS data in this study, we note that identifying quasi-primes may also be useful in designing PCR primers to improve the sensitivity of PCR in detecting foreign genomes.

We will review the mathematical description of quasi-primes. We refer to a sequence of length k as a kmer. For a set of genomes $G = \{G_1, G_2, \cdots, G_n\}$, let $G_{i,k}$ denote the set of all kmers in G_i. A **kmer quasi-G-prime of $G_{i,k}$** is a sequence of length k that occurs only in G_i and not in any other G_j. The addition of "G" in quasi-G-prime indicates that the quasi-primes were restricted to a small set of clinically relevant genomes, rather than in typical analyses of universal taxonomies.

Existing approaches to organism identification in sepsis rely on universal taxonomy classification (e.g., Kraken [8,9], Kaiju [10]) to classify reads, whereas clinically relevant organisms occupy a small portion of the taxonomic space. Programs such as Kraken are also highly parameter dependent and can result in spurious classifications [11]. In this work, we first show that 12mer genomic quasi-primes differentiate between closely related *Staphylococcus* spp. We then evaluate our algorithm on real-world patient data and show that by limiting our sequences to 17mer genomic quasi-primes, we achieve sensitive and specific classification of septic vs healthy patients.

Materials and methods

Overview

We first provide an overview of our methods. We defined our organisms of interest (Supplemental Data), including common causes of sepsis as well as the human

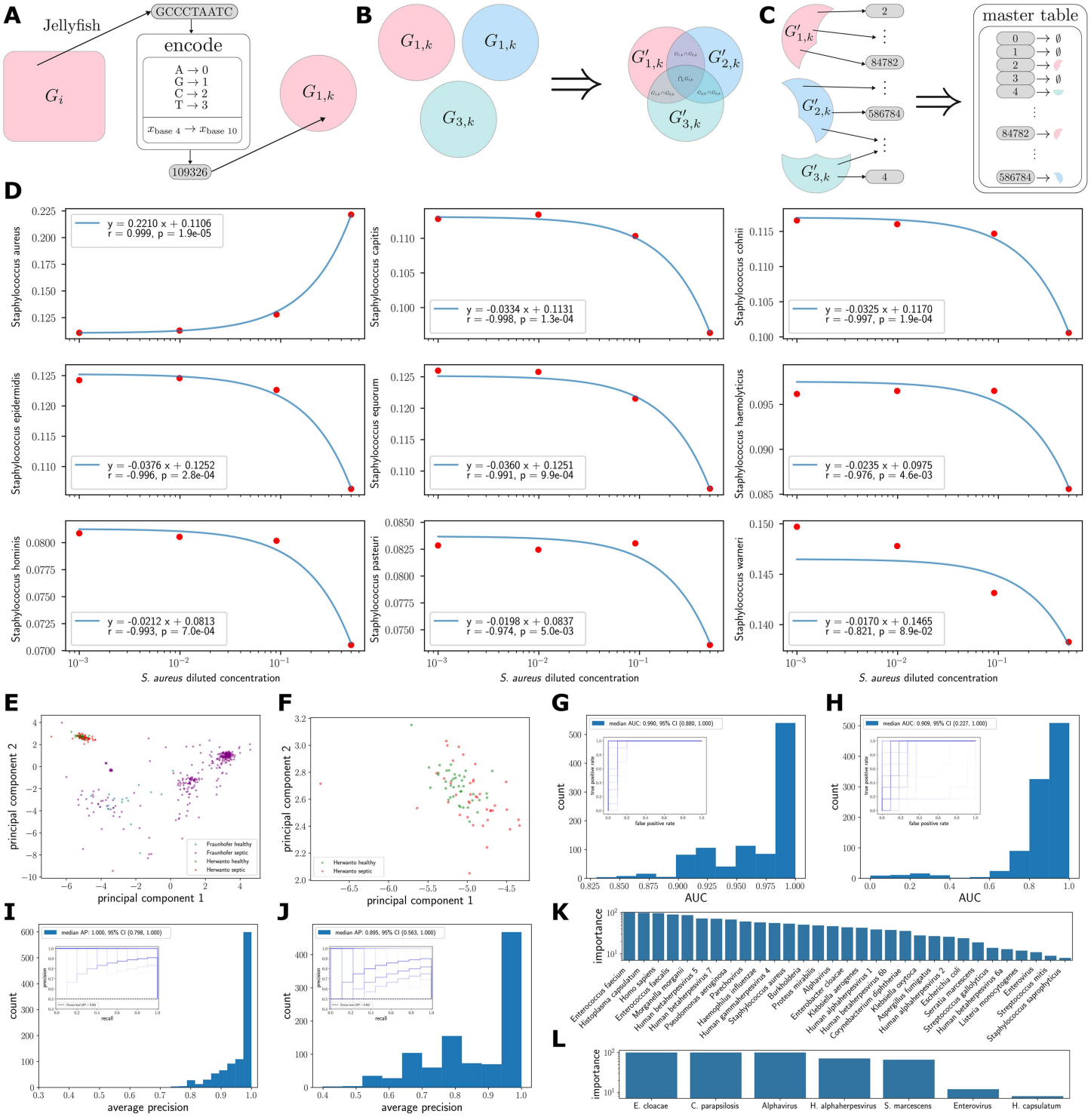

Fig 1. Graphical illustration of quasi-prime pipeline, with *in vitro* and clinical proof-of-concept. A. All kmers present in a genome $\mathbf{G}_1$ are encoded as integers and placed in a set $\mathbf{G}_{1,k}$. **B**. Once kmer sets have been constructed for all genomes of interest, the quasi-prime sets $\mathbf{G}'_{i,k}$ are determined through set operations. **C**. Now that the $\mathbf{G}'_{i,k}$ are disjoint, we construct the mapping from quasi-primes to species. **D**. Proportion of reads assigned to each species as a function of the diluted DNA concentrations of *S. aureus*, along with linear regressions and **r** and **p** values. **E and F**. Principal component analysis of normalized read counts. The Herwanto data points occupy a distinct corner (**F**) of the space. **G and H**. AUC distributions for logistic

regression of Herwanto and Fraunhofer, respectively. Underlying ROC curves are inset at low opacity to visualize the curve density. **I and J.** Average precision distributions for Herwanto and Fraunhofer, respectively. Underlying precision–recall curves are inset at low opacity to visualize the curve density. **K and L**. are the most important species in classifying septic vs healthy patients for Herwanto and Fraunhofer, respectively.

https://doi.org/10.1371/journal.pone.0341828.g001

genome [12]. We acquired the genomes from NCBI Datasets [13,14]. Representative organisms were chosen based on a list of infectious diseases found at https://emedicine.medscape.com/infectious_diseases.

To choose an appropriate kmer size, we considered a balance of sequence specificity and performance. The average number of occurrences of a kmer in a sequence is equal to the length of the genome divided by 4^k. For a single 5'-3' sequence of length L, there are $L-k+1$ contiguous subsequences of length k. Since $k \ll L$, we will assume that there are L such kmers. Since there are 4^k possible kmers and the L kmers are distributed among these, the average number of occurrences of each kmer in the sequence is $L/4^k$. If we also consider the reverse complement of the sequence, the sequence length is effectively doubled, though we ignore this fact in this estimation. To have an average of one occurrence per kmer, we must therefore have that

$$L \approx 4^k \rightarrow k \approx \log_4 L. \tag{1}$$

This argument generalizes to the case of multiple sequences and chromosomes, where L becomes the total number of bases across all nucleotides.

Using Eq (1) as a rough guide, for our proof-of-concept classification of *Staphylococcus* spp., the combined length of the staphylococcal genomes was 23,039,377 base pairs. Since $\log_4 23,039,377 \approx 12$, we chose to classify 12mers for this first analysis. For the sepsis analysis, the human genome alone was over 3 billion bases in length. Since $\log_4 3 \times 10^9 \approx 16$ and the remainder of the genomes which we included in sepsis classification contributed a billion or so more bases, we chose to classify 17mers for the sepsis analysis.

Finally, we created a master table, in which each 17mer was mapped to its corresponding species (Fig 1C). Since each 17mer was encoded in base 4 and we chose fewer than $2^8 = 256$ species, this resulted in a byte vector of length 4^{17}, with each entry being the species ID encoded as a byte. The result was a dictionary that fits within 17.2 GB of RAM and is highly accessible to consumers. With this dictionary, our method classifies 2 billion 17mers in 20 minutes on a single core of a 2.2 GHz Xeon E5-2650v4, with a processing speed of several million bps / second. Additionally, we parallelized the various tasks in our pipeline, accelerating the construction of the master table and classification of samples. Our pipeline is implemented in Julia, and we call our package Gprime.jl.

Shotgun metagenomic sequencing of *Staphylococcus aureus* in a background microbiome

S. aureus ATCC 12600 (type strain, alias NCTC 8532, isolated from pleural fluid) purchased from the ATCC was selected as the target species, and eight other *Staphylococcus* species were used as the background microbiome (Table 1). These isolates were cultured on BHI agar for 16–18 hours. Using 10 µL disposable loops, four loops of overnight cultured cells were collected and transferred to tubes containing beads and 500 µL of PBS. The tubes were then horizontally vortexed, followed by centrifugation. Two hundred microliters of the supernatant were subsequently used for gDNA extraction via the QIAGEN QIAamp DNA Blood Mini Kit. The extracted DNA from these isolates was adjusted to the same concentration (e.g., ~120 ng/µL). The adjusted DNA from eight other *Staphylococcus* species were included, mixed and treated as in the background microbiome to assess the specificity of quasi-primes for the detection of *S. aureus* DNA sequences in a test sample that included *S. aureus* and the background microbiome, combined in a 1:1 ratio. *S. aureus* DNA was tenfold diluted in background microbiome DNA (0:1, 1:1, 1:10, 1:100, and 1:1000) to assess the limit of detection of Nanopore sequencing for the detection of *S. aureus* DNA in background microbiome DNA.

Table 1. *S. aureus* and 8 other *Staphylococcus* species included in the limit of detection experiment.

Isolate Number	Species	Source of isolation	Target or background microbiome	Genome Size
PS03046	*S. aureus* (ATCC12600)	Pleural fluid	Target	2.8 Mbp
PS01783	*S. epidermidis*	Ready-to-eat food product	Background microbiome	2.5 Mbp
PS01400	*S. hominis*	Raw milk	Background microbiome	2.3 Mbp
PS01871	*S. capitis*	Raw milk	Background microbiome	2.5 Mbp
PS01480	*S. equorum*	Raw milk	Background microbiome	2.8 Mbp
PS01395	*S. warneri*	Raw milk	Background microbiome	2.5 Mbp
PS01784	*S. cohnii*	Dairy powder	Background microbiome	2.7 Mbp
PS01910	*S. pasteurii*	Whey protein isolate liquid	Background microbiome	2.5 Mbp
PS01361	*S. haemolyticus*	Bovine manure	Background microbiome	2.6 Mbp

https://doi.org/10.1371/journal.pone.0341828.t001

Acquisition of human data

We used the NCBI Datasets to acquire all the sequencing data. We extracted the 17mers using Jellyfish 2.3.0 [15]. For the human genome, we used the T2T-CHM13 version 2.0 [12] assembly. We examined the metadata for each study to determine how septic vs healthy patients were encoded. For each accession number, we downloaded and parsed the XML summary of all sequences from NCBI BioSamples. The list of genomes by accession number that were used can be found in our Supplemental Data on Zenodo (publication/data/genomes_downloaded.txt).

Septic vs healthy patient datasets were also acquired from NCBI BioProjects, with accession numbers as follows: PRJEB13247 [16], PRJEB21872 [17,18], PRJEB30958 [18], and PRJNA647880 [19]. The first three accessions originated from the same institution/research group, which we refer to as Fraunhofer (18 healthy samples and 305 septic samples). The last accession is Herwanto (40 healthy and 39 septic). All three Fraunhofer accessions consisted of cell-free DNA isolated from plasma, sequenced using the Illumina HiSeq 2500 platform. Herwanto consisted of cell-free RNA isolated from peripheral blood, sequenced using the Illumina HiSeq 4000 platform.

Identification of quasi-primes and creation of the master table

First, for each genome, we identified all subsequences of length k (i.e., the kmers) in that genome. For example, let genome 1 be ATCGGC. Then the set of 3mers $G_{1,3}$ is {ATC, TCG, CGG, GGC}. If the set of 3mers for genome 2 were $G_{2,3}$ = {ATC, CCC, GGA}, then the 3mers specific to genome 1 and not included in genome 2 would be {TCG, CGG, GGC}. If we had more genomes, we would similarly exclude 3mers in genome 1 which were contained in those other genomes, until finally we have only those 3mers in genome 1 which are not contained in any other genome, which we would call $G'_{1,3}$. In set notation, we perform a set subtraction between $G_{1,3}$ and the union of all other 3mer sets $\bigcup_{j\neq i} G_{j,3}$: $G'_{1,3} = G_{1,3} \setminus \cup_{j\neq i} G_{j,3}$.

For species s with multiple genome assemblies available $\{s_n\}$ (such as for the various strains of *E. coli*), we merged the set of kmers across all such genome assemblies: $G_{i,k} = \cup_n s_n$. Once we calculate all of the $G'_{i,k}$, we create a master table M, which maps each sequence in $G'_{i,k}$ to i. We interpreted each sequence as a number in base 4, representing $A \Longleftrightarrow 0$, $G \Longleftrightarrow 1$, $C \Longleftrightarrow 2$, $T \Longleftrightarrow 3$. For example, a sequence ATCG would be encoded as $0321_4 = 0 \times 4^3 + 3 \times 4^2 + 2 \times 4^1 + 1 \times 4^0 = 57$ in base 10. The map M can thereby be represented as an array of size 4^k. If the preceding sequence (ATCG) corresponds to species 7, then the 57^{th} entry of M is set to 7.

For 17mers, M is an array of size $4^{17} = 17,179,869,184$. The data type of the entries can be optimized to the number of species. For example, a byte can represent an unsigned integer from 0 to 255. If we have at most 256 species, then Mcan be efficiently represented as an array of bytes of length 4^{17}, corresponding to exactly 4^{17} bytes ($\approx$17 gigabytes) of memory.

Classification and normalization of read counts

We performed classification of 17mers for each sample via Jellyfish [15] and the master table, resulting in a vector of read counts for each sample. To normalize the read counts, we first divided all read counts by the total number of reads. Next, for all the samples (Herwanto + Fraunhofer), we centered the normalized read counts by subtracting the mean and dividing by the standard deviation. By performing these two operations in this order, we retained the distributional information of each species' read count in the context of the sample's run characteristics before standardizing the inputs to regularize our logistic regressions.

Resampling Fraunhofer

Whereas the Herwanto dataset was appropriately balanced between healthy and septic samples (40 and 39, respectively), only 5.6% (18 of 323) of the Fraunhofer dataset were from healthy patients. Such imbalanced data easily lead to overfitting [20]. As a reasonable starting point, we assume that the prior probability that a given sample is septic is 50%, i.e., the clinician is 50% sure that the patient is septic. To do this, we combined the 18 healthy Fraunhofer samples with 18 samples chosen uniformly at random from the remaining 305 septic Fraunhofer samples. We performed this undersampling repeatedly to estimate the empirical distributions of our performance metrics.

Logistic regression

We used scikit-learn 1.5.1 [21] to perform logistic regressions (Figs 1E-F), with hard-coded random seeds for reproducibility. The Herwanto and undersampled Fraunhofer datasets were stratified by healthy vs septic status and uniformly randomly split into training and test sets at a ratio of 3:1, except as noted in the code.

Logistic regressions are used for binary classification tasks, such as classifying the reads in this work as healthy vs septic. Logistic regressions require the fitting of the weights $a_0, a_1, \cdots, a_n$ in

$$p(x) = \frac{1}{1 + e^{-a_0 + \sum_i a_i x_i}},$$

where the x_i are the input features (i.e., the normalized read counts), $p(x) = 1$ represents the positive class, $p(x) = 0$ represents the negative class, and the coefficients a_i are fit so that the binary log-loss L (or cross-entropy loss) is minimized:

$$L = -\sum_x \left[y(x) \ln p(x) + (1 - y(x)) \ln (1 - p(x)) \right],$$

where x are the inputs taken from the training set, $y(x)$ is the true label of the sample (septic vs healthy), and $p(x)$ is the output of the logistic regression from above. An additional regularization term was added to L to help mitigate overfitting:

$$L_{reg} = \frac{1}{C} \sum_i a_i^2,$$

so that the total loss to be minimized was $L_{total} = L + L_{reg}$. Finally, we also performed a hyperparameter search, resulting in a relative penalty factor of C = 100 for the L^2 logistic weight regularization and 50 training iterations.

Statistics

Linear regressions with associated significance testing were performed using SciPy [22]. Hotelling's T^2-test [23] was performed using the pingouin package [24].

Performance metrics

The logistic regression produces numbers in the range [0,1], which can be interpreted as the probability that a given input belongs to the positive class. By varying the threshold for these probabilities to be considered positive vs negative, we may characterize the classification performance of the logistic regression. The classification performance metrics we examined were the receiver operating characteristic (ROC) curve and its associated area under the curve (AUC) (Figs 1G-H), as well as the precision–recall (PR) curve and its associated average precision (AP) (Figs 1I-J). These quantities were calculated using the scikit-learn package [21].

For a given cutoff, logistic regression yields rates of true positives (TPs), false positives (FPs), true negatives (TNs), and false negatives (FNs), which are used to calculate performance metrics. The ROC is a plot of the TP as a function of the FP, and the AUC of the ROC is computed via the trapezoidal rule on the ROC curve. For the PR curve, the precision P and recall R are defined as follows:

$$P = \frac{TP}{TP + FP},$$

$$R = \frac{TP}{TP + FN}.$$

The AP is defined as:

$$AP = \sum_{n} (R_n - R_{n-1}) P_n,$$

where (R_n, P_n) are the recall and precision, respectively, calculated at the thresholds. Note that this definition differs from the trapezoidal rule for the AUC curve.

To estimate the median AUC, median AP, and 95% confidence intervals, we performed bootstrapping by repeatedly resampling Herwanto and Fraunhofer into training and test sets (Figs 1G-J). The distributions, medians, and 95% confidence intervals in Figs 1G-J were calculated from 1000 bootstrap samples, whereas only the first 100 of the corresponding AUC/PR curves were plotted for ease of visualization.

To evaluate the importance of each species to the logistic regression, we trained 100 logistic regression models with random, stratified splits of Herwanto and Fraunhofer and counted the number of times a given feature had nonzero permutation feature importance (Figs 1K-L). All calculations are shown in the included Jupyter notebooks [25] (Supplemental Data).

Results

First, to validate quasi-prime read assignment, we used shotgun metagenomic sequencing of DNA from *Staphylococcus aureus* combined at varying concentrations with eight other *Staphylococcus* species. Using reference assemblies for these species we determined their 12mer quasi-primes. As noted in the Methods, we chose a length of 12 for this proof-of-concept analysis to match the number of possible 12mers to the total size of the staphylococcal genomes, so that each 12mer corresponds to an average of a single species. We classified the reads for each sample and plotted the read proportion as a function of the initial *Staphylococcus aureus* concentration (Fig 1D). The concentrations were logarithmically spaced, and thus, the linear regressions appeared exponential. There was excellent agreement between the classifications and *Staphylococcus aureus* concentrations ($p = 1.9e\text{-}5$). These results demonstrate the ability of our approach for distinguishing bacterial strains in multi-culture experiments.

Second, we used human data from NCBI Datasets which labeled septic versus healthy patients, as described in the Methods section. Blood culture data were unavailable. We performed principal component analysis on the normalized

read count data, plotting Herwanto (cell-free RNA) versus Fraunhofer (cell-free DNA) (Fig 1E). We found that the Herwanto data (Fig 1F) were clearly distinct from the Fraunhofer data (Hotelling's $T^2 = 6080.3$, $p < 10^{-10}$). Therefore, subsequent analyses of the Herwanto and Fraunhofer datasets were performed independently.

For simplicity and interpretability, we used logistic regressions to predict whether the patient would be septic or healthy in both datasets. For Herwanto, we performed repeated random splitting of the points between the training and test sets. Owing to the severe data imbalance in Fraunhofer, we performed repeated undersampling of that dataset (Methods). We plotted the distribution of AUCs and the associated receiver operating characteristic (ROC) curves (Figs 1G-H). We found that the median AUCs for the datasets were high (approximately 0.9 for both), indicating high sensitivity and specificity. We then examined the distribution of average precision and the associated precision–recall curves (Figs 1I-J); the median average precision was similarly high (Herwanto: 0.877, Fraunhofer 1.0).

We examined the most important factors (Methods) for each logistic model (Figs 1K-L), which revealed that the *Homo sapiens* count was near the top in importance for both datasets. The Fraunhofer dataset emphasized other common sepsis pathogens, such as *Pseudomonas*, *Enterococcus*, and *Haemophilus*.

Finally, we performed additional analyses using Kraken 2, which are included as Supporting Information (Supplemental Methods, S1-S3 Figs). We find that the full (~60–70 GB) database achieved similar performance to our method when the output was restricted to the species we tested, whereas the truncated (~16 GB) and custom-built Kraken 2 databases did not produce sufficient read assignments to classify septic vs healthy patients.

Discussion

A major predictive factor of sepsis was the number of reads classified as human, presumably because a lower proportion of reads classified as human indicates increased reads classified as exogenous. Common etiologies of sepsis, including *Pseudomonas*, *Enterococcus*, and *Haemophilus*, were identified [16–19,26].

Due to the small number of taxa that can thrive in the human body, we may limit the size of the dictionary of quasi-G-primes significantly compared with that of universal taxonomic classifiers such as Kraken 2 with its standard databases. Furthermore, organisms belonging to similar taxa are often treated with the same set of standard broad-spectrum antibiotics. For example, Gram-negative organisms are often treated with third- or fourth-generation cephalosporins, while Gram-positive organisms and methicillin-resistant *Staphylococcus aureus* are treated with vancomycin, and thus ceftriaxone/cefepime + vancomycin is a common treatment of choice [27,28]. A potential diagnostic benefit of our method's sensitivity is that it may become possible to rule out significant bacteremia via next-generation sequencing and thereby improve antimicrobial stewardship for patients. However, further validation with blood cultures is necessary to assess the clinical accuracy of our method in identifying the causative organism in sepsis. The major weakness of this work and of the literature is the paucity of blood-sequenced sepsis cases with known organisms due to the poor sensitivity of the gold standard of blood culture for diagnosis; therefore, we were limited solely to sepsis/healthy classification.

There is a need for improved normalization of experimental techniques between research groups and NGS techniques, as indicated by the heterogeneity seen in the principal component analysis. The cell-free DNA sequencing of the Fraunhofer dataset potential provides a less biased view of the microbiome population, compared to the cell-free RNA sequencing of the Herwanto dataset. However, we were able to predict septic versus healthy patients in both populations in separate logistic regressions, indicating that there is sufficient signal in both types of sequencing for classification. Our importance analysis was limited due to the low number of healthy controls in Fraunhofer. In the Supporting Information (S1 File, S1 Fig, S2 Fig, S3 Fig), we demonstrate that our method is more memory efficient, sensitive, and accurate than Kraken 2.

There are many strains of specific bacterial species, which can significantly affect the availability of quasi-G-primes. For example, in this work, we performed a union on all strains of *Staphylococcus aureus*, which may be inappropriate for

methicillin-sensitive vs methicillin-resistant strains, where the clinical question is whether to use MRSA-targeted therapies. It may therefore be beneficial to focus on plasmid-associated kmers which confer specific antibiotic resistance in such cases.

In this work and in our other recent publications [3–5], there is a focus on quasi-primes because they are specific to a single species. With quasi-G-primes, one requires $|G|-1$ kmers to differentiate among all $|G|$ species in a dataset with perfect specificity. In principle, only $\lceil \log_2 |G| \rceil$ specially selected kmers are needed for 100% specificity; in the case that the first kmer divides G into two halves, A and B, the second kmer perfectly divides A and B into four quarters, and so on. In practice, sequencing is a stochastic and noisy process; thus, many more quasi-G-primes were used in this work to increase its sensitivity. Furthermore, our current selection of species was informed by clinical relevance, and in our future work we plan to encompass all available species to further refine our quasi-G-prime lists.

In future work specific to sepsis, it may also be beneficial to group bacterial taxa or strains by the class of antibiotics which are generally most effective and search for kmers that differentiate between those groups of bacteria, enabling the narrowing of antibiotics in the hospital setting. There is a spectrum of specificity to kmers, from quasi-G-primes specific to a single organism to the gestalt genomic fingerprints considered in microbiomics, all of which may be useful in the treatment of sepsis.

The simple idea of genomic quasi-primes enables highly sensitive and specific determination of species, especially when the plausible taxonomies are limited. Future applications of quasi-primes in personalized medicine may even use the patient's own genome in place of the generic human genome as a control, to increase detection of foreign genomic material. Our method is effective and has potential for practical application in biology and medicine. By varying the definition and grouping of taxa, future work may achieve different classifications, which may inform clinical practice.

Supporting information

S1 Fig. Full Kraken 2 database analysis. Applying the full Kraken 2 database (~60–70 GB RAM) to the same benchmarking tasks as provided in the main text. **A-I** correspond to subfigures **1D-L** in the main text. The results are largely similar to that of our method, at the cost of RAM and additionally some lower accuracy and specificity for the Herwanto dataset.
(SVG)

S2 Fig. Truncated Kraken 2 database analysis. Applying the truncated Kraken 2 database (~16 GB RAM) to the same benchmarking tasks as provided in the main text. **A** shows the benchmarking of this database against cultured *Staphylococcus* spp. When restricted to the list of sepsis-relevant genomes, this truncated database did not generate any counts for the sepsis species in septic patients, and we were unable to plot those results.
(SVG)

S3 Fig. Custom Kraken 2 database analysis. Applying the custom Kraken 2 databases to the same benchmarking tasks as provided in the main text. **A-I** correspond to subfigures **1D-L** in the main text. In **A**, the database was built only upon the *Staphylococcus* species which were cultured. Notably, Kraken 2 did not pick up any counts for *Staphylococcus warneri*. The sepsis classification task was similarly difficult for the custom database (**B-I**), with few septic samples demonstrating any read counts.
(SVG)

S1 File. Supplemental Methods and Discussion.
(DOCX)

Acknowledgments

Portions of this work were performed while MP, IM, IGS, XW, and TC were affiliated with the Pennsylvania State University.

Author contributions

Conceptualization: Congzhou M Sha, Michail Patsakis, Ioannis Mouratidis, Ilias Georgakopoulos-Soares.

Data curation: Congzhou M Sha, Jasna Kovac.

Formal analysis: Congzhou M Sha.

Investigation: Congzhou M Sha.

Methodology: Congzhou M Sha, Ioannis Mouratidis, Ilias Georgakopoulos-Soares.

Project administration: Ilias Georgakopoulos-Soares.

Resources: Xiaoyuan Wei, Taejung Chung, Jasna Kovac, Ilias Georgakopoulos-Soares.

Software: Congzhou M Sha.

Supervision: Jasna Kovac, Ilias Georgakopoulos-Soares.

Validation: Congzhou M Sha, Ilias Georgakopoulos-Soares.

Visualization: Congzhou M Sha.

Writing – original draft: Congzhou M Sha.

Writing – review & editing: Congzhou M Sha, Michail Patsakis, Ioannis Mouratidis, Xiaoyuan Wei, Taejung Chung, Jasna Kovac, Ilias Georgakopoulos-Soares.

References

1. Ghoreyshi N, Heidari R, Farhadi A, Chamanara M, Farahani N, Vahidi M, et al. Next-generation sequencing in cancer diagnosis and treatment: clinical applications and future directions. Discov Oncol. 2025;16(1):578. https://doi.org/10.1007/s12672-025-01816-9 PMID: 40253661
2. Schuler BA, Nelson ET, Koziura M, Cogan JD, Hamid R, Phillips JA 3rd. Lessons learned: next-generation sequencing applied to undiagnosed genetic diseases. J Clin Invest. 2022;132(7):e154942. https://doi.org/10.1172/JCI154942 PMID: 35362483
3. Mouratidis I, Chan CSY, Chantzi N, Tsiatsianis GC, Hemberg M, Ahituv N, et al. Quasi-prime peptides: identification of the shortest peptide sequences unique to a species. NAR Genom Bioinform. 2023;5(2):lqad039. https://doi.org/10.1093/nargab/lqad039 PMID: 37101657
4. Mouratidis I, Konnaris MA, Chantzi N, Chan CSY, Patsakis M, Provatas K, et al. Identification of the shortest species-specific oligonucleotide sequences. Genome Res. 2025;35(2):279–95. https://doi.org/10.1101/gr.280070.124 PMID: 39746719
5. Bochalis E, Patsakis M, Chantzi N, Mouratidis I, Chartoumpekis DV, Georgakopoulos-Soares I. Taxonomic quasi-primes: peptides charting lineage-specific adaptations and disease-relevant loci. Protein Sci. 2025;34(9):e70241. https://doi.org/10.1002/pro.70241 PMID: 40852837
6. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014;15(3):R46. https://doi.org/10.1186/gb-2014-15-3-r46 PMID: 24580807
7. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol. 2019;20(1):257. https://doi.org/10.1186/s13059-019-1891-0 PMID: 31779668
8. Menzel P, Ng KL, Krogh A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nat Commun. 2016;7:11257. https://doi.org/10.1038/ncomms11257 PMID: 27071849
9. Wright RJ, Comeau AM, Langille MGI. From defaults to databases: parameter and database choice dramatically impact the performance of metagenomic taxonomic classification tools. Microb Genom. 2023;9(3):000949. https://doi.org/10.1099/mgen.0.000949 PMID: 36867161
10. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, et al. The complete sequence of a human genome. Science. 2022;376(6588):44–53. https://doi.org/10.1126/science.abj6987 PMID: 35357919
11. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44(D1):D733-45. https://doi.org/10.1093/nar/gkv1189 PMID: 26553804
12. Sayers EW, Cavanaugh M, Clark K, Pruitt KD, Sherry ST, Yankie L, et al. GenBank 2024 Update. Nucleic Acids Res. 2024;52(D1):D134–7. https://doi.org/10.1093/nar/gkad903 PMID: 37889039

13. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 2011;27(6):764–70. https://doi.org/10.1093/bioinformatics/btr011 PMID: 21217122
14. 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(1):73. https://doi.org/10.1186/s13073-016-0326-8 PMID: 27368373
15. 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(8):1796. https://doi.org/10.3390/ijms18081796 PMID: 28820494
16. 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(5):e394–402. https://doi.org/10.1097/CCM.0000000000003658 PMID: 30720537
17. Herwanto V, Tang B, Wang Y, Shojaei M, Nalos M, Shetty A, et al. Blood transcriptome analysis of patients with uncomplicated bacterial infection and sepsis. BMC Res Notes. 2021;14(1):76. https://doi.org/10.1186/s13104-021-05488-w PMID: 33640018
18. Johnson JM, Khoshgoftaar TM. Survey on deep learning with class imbalance. J Big Data. 2019;6(1). https://doi.org/10.1186/s40537-019-0192-5
19. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research. 2011;12: 2825–2830. Available from: http://dl.acm.org/citation.cfm?id=2078195%5Cnhttp://arxiv.org/abs/1201.0490
20. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020;17(3):261–72. https://doi.org/10.1038/s41592-019-0686-2 PMID: 32015543
21. Hotelling H. The Generalization of Student's Ratio. Ann Math Statist. 1931;2(3):360–78. https://doi.org/10.1214/aoms/1177732979
22. Vallat R. Pingouin: statistics in Python. JOSS. 2018;3(31):1026. https://doi.org/10.21105/joss.01026
23. Kluyver T, Ragan-Kelley B, Pérez F, Granger B, Bussonnier M, Frederic J, et al. Jupyter Notebooks – a publishing format for reproducible computational workflows. Positioning and Power in Academic Publishing: Players, Agents and Agendas. IOS Press. 2016. https://doi.org/10.3233/978-1-61499-649-1-87
24. Evaluation and management of suspected sepsis and septic shock in adults - UpToDate. [cited 5 Nov 2025]. Available: https://www.uptodate.com/contents/evaluation-and-management-of-suspected-sepsis-and-septic-shock-in-adults?search=organisms%20in%20sepsis&source=search_result&selectedTitle=3~150&usage_type=default&display_rank=3
25. Tunkel AR, Hartman BJ, Kaplan SL, Kaufman BA, Roos KL, Scheld WM, et al. Practice guidelines for the management of bacterial meningitis. Clin Infect Dis. 2004;39(9):1267–84. https://doi.org/10.1086/425368 PMID: 15494903
26. Lieberthal AS, Carroll AE, Chonmaitree T, Ganiats TG, Hoberman A, Jackson MA. Practice guidelines for the management of bacterial meningitis. Clinical Infectious Diseases. 2012;53.
27. Chen M, Zhao H. Next-generation sequencing in liquid biopsy: cancer screening and early detection. Hum Genom. 2019;13(1). https://doi.org/10.1186/s40246-019-0220-8
28. Short NJ, Kantarjian H, Ravandi F, Konopleva M, Jain N, Kanagal-Shamanna R, et al. High-sensitivity next-generation sequencing MRD assessment in ALL identifies patients at very low risk of relapse. Blood Adv. 2022;6(13):4006–14. https://doi.org/10.1182/bloodadvances.2022007378