Title: Precision Identification of Diverse Bloodstream Pathogens from the Gut Microbiome

Authors: Fiona B. Tamburini, BS[1]*; Tessa M. Andermann, MD, MPH[2]*; Ekaterina Tkatchenko, BS[3]; Fiona Senchyna[4], BS; Niaz Banaei[2,4], MD; Ami S. Bhatt, MD, PhD[1,3]**

1 Department of Genetics, Stanford University

2 Department of Medicine, Division of Infectious Diseases, Stanford University

3 Department of Medicine, Division of Hematology, Stanford University

4 Clinical Microbiology Laboratory, Stanford University Medical Center

* These authors contributed equally

**To whom correspondence should be addressed: asbhatt@stanford.edu

Abstract

Bloodstream infection is the most common infectious complication in hematopoietic cell transplantation recipients. Culture-based sequencing approaches have been previously applied to interrogate the intestine as a reservoir of these bloodstream pathogens. To evaluate the genomic concordance of bloodstream pathogens and bacterial strains within the intestinal microbiome using whole genome sequencing (WGS), we developed StrainSifter, a bioinformatic pipeline to compare nucleotide variation between bacterial isolate strains and stool metagenomes. We applied StrainSifter to bloodstream isolates and stool metagenome samples from a cohort of 30 hematopoietic stem cell transplant recipients with bloodstream infections at a single academic center. StrainSifter is designed to 1) identify single nucleotide variants (SNVs) between isolate and metagenomic short reads aligned to assembled draft genomes using stringent alignment, coverage, and variant frequency criteria, and 2) output both phylogenetic trees and SNV counts for clinical and epidemiologic strain comparison. Using StrainSifter, we identified enteric BSI isolates - including *Enterococcus faecium, Klebsiella pneumoniae*, *Escherichia coli*, and *Streptococcus mitis* - that were highly concordant with those in the gut microbiota. We also identified highly concordant

strains of *Staphylococcus epidermidis* and *Pseudomonas aeruginosa* between the bloodstream and gut microbiome, an unexpected finding for typically non-enteric organisms. In contrast to previous studies demonstrating intestinal domination by pathogenic bacteria preceding BSI, we found significant heterogeneity in relative pathogen abundance prior to infection. These findings demonstrate not only the utility of StrainSifter in strain matching between isolate and metagenomic sequencing data, but also provide a more precise and comprehensive investigation of the intestine as a reservoir of diverse pathogens capable of causing bloodstream infections in hematopoietic stem cell transplant patients.

Introduction

Microbial pathogenicity and transmission depend in part on strain-level variability in both gene content and at the single-nucleotide variant (SNV) level, as different strains of the same organism can vary widely in their ability to cause disease (Snitkin et al. 2011; Lieberman et al. 2013). Whole genome sequencing (WGS) has facilitated the exploration of strain-level determinants of virulence and has enabled the precise tracking of pathogen transmission (Snitkin et al. 2011; Snitkin et al. 2012). Methods for strain tracking and comparison most commonly rely on either sequencing a defined number of genes or identifying single nucleotide variants (SNVs) across the whole genome. For example, whole genome sequencing of cultured *Klebsiella pneumoniae* stool isolates in ICU patients demonstrated the role of *K. pneumoniae* colonization in systemic infections (Gorrie et al. 2017; Martin et al. 2016; Nielsen et al. 2017) and genome-wide SNV comparison is routinely employed for epidemiologic strain tracking (Salipante et al. 2015; Snitkin et al. 2012; Harris et al. 2013; Chen et al. 2017; Kovanen et al. 2014). While strain comparisons have primarily been performed on bacterial isolates, newer computational tools (MIDAS, StrainPhlAn, and metaSNV) profile strain variation between metagenomic samples over time without the bias inherent in culturing isolates (Costea et al. 2017; Nayfach et al. 2016; Truong et al. 2017).

Despite the increasing use of WGS for strain comparison, bioinformatic tools have not been developed for

51 the specific purpose of comparing disease-causing bacterial isolates to metagenomic samples at the
52 nucleotide level within the same patient. It is desirable to achieve the highest resolution of comparison
53 across the whole genome of an organism in order to precisely compare clinical isolates to strains within a
54 sampled metagenome. Additionally, most tools depend on the availability of reference genomes and are
55 thus often restricted to studying well characterized bacterial pathogens, limiting the types of organisms
56 that are available for strain comparisons. To date, no reference-free tool exists with the stringent filtering
57 steps necessary to identify genomic variants specific to a single strain between genomes and
58 metagenomes. We sought to compare strain-level identity between bloodstream infection isolates and
59 stool metagenomes using samples from immunocompromised patients at elevated risk of systemic
60 infections, with the goal of better understanding the origins of bloodstream infection in this patient
61 population.
62
63 Hematopoietic stem cell transplant (HCT) recipients are at high risk for bloodstream infections, which
64 occur in 20-60% of patients primarily during periods of neutropenia and mucositis (Blennow et al. 2014;
65 Gudiol et al. 2014; Mikulska et al. 2009). Prior to engraftment of donor stem cells, HCT recipients are
66 particularly susceptible to translocation of enteric organisms from the intestinal microbial reservoir into
67 the bloodstream. During periods of mucositis, primary bloodstream infections with enteric Gram-negative
68 and Gram-positive organisms are thought to arise from mucosal barrier injury (See et al. 2017). In
69 contrast, non-enteric commensal and environmental bacteria access the bloodstream through other routes
70 including intravenous lines and sites where skin epithelial integrity has become compromised. Existing
71 methods for strain comparison have been helpful in identifying the origins of bloodstream infections in
72 HCT patients using culture-based pulsed-field gel electrophoresis (PFGE), or PCR-based multi-locus
73 sequence typing (MLST) (Tancrede and Andremont 1985; Samet et al. 2013). Although these methods
74 are rapid, affordable, and standardized across many organisms, they have several limitations: 1) they
75 require isolation and culture of candidate pathogens from the stool or other suspected pathogen reservoirs,
76 a step that can introduce selection bias, and 2) they often rely on detailed characterization of a small

77 number of genes and therefore lack the discriminatory power required for robust, genome-wide strain
78 comparison (W. Li, Raoult, and Fournier 2009). No studies to date have investigated the origins of
79 bloodstream infections in HCT or other immunocompromised patients at the level of single nucleotide
80 variation across entire genomes.
81
82 In this work, we present StrainSifter, a straightforward bioinformatics pipeline based on whole genome
83 sequencing to comprehensively and precisely compare single nucleotide variants between isolate
84 genomes and metagenomes without bias introduced by culturing. Additionally, unlike most previous
85 strain comparison bioinformatic tools (MIDAS and StrainPhlAn), variant calling within our pipeline does
86 not depend on the availability of a reference genome database, a particularly relevant consideration in
87 immunocompromised patients who develop infections with lesser known and poorly studied opportunistic
88 pathogens. We have applied this tool to track bacterial strain translocation from the gut to the bloodstream
89 in a unique cohort of immunocompromised patients at high risk for translocation and bloodstream
90 infection, specifically patients with hematologic malignancies who have undergone hematopoietic stem
91 cell transplantation. We find that StrainSifter is able to identify strains of both intestinal mucosa-
92 associated (*Enterococcus faecium*, *Escherichia coli, Klebsiella pneumoniae, Enterobacter cloacae, and*
93 *Streptococcus mitis*) and typically non-mucosal bacteria (*Staphylococcus epidermidis* and *Pseudomonas*
94 *aeruginosa*) within the gut microbiome as being concordant with the same strain cultured from the blood
95 of patients with bloodstream infections. Intriguingly, this suggests that both enteric and classically non-
96 enteric pathogens can colonize and even thrive within the intestinal reservoir of immunocompromised
97 hosts.
98
99 **Methods**
100 **Cohort selection**
101 A retrospective cohort study, approved by the institutional review board under the IRB protocol #8903
102 (PI: Dr. David Miklos, co-PI: Dr. Ami Bhatt), was performed at Stanford Hospital. Informed consent was

103 obtained from all subjects prior to specimen collection. At the time of cohort identification (July 2017), a
104 stool biospecimen collection containing 964 stool samples from 402 patients was available for
105 investigation. This collection consisted of convenience samples collected from autologous and allogeneic
106 stem cell transplantation patients at Stanford University Hospital between October 5 2015 and June 9
107 2017. Patients were included in this study if a stool sample had been collected within 30 days prior to an
108 episode of bloodstream infection (BSI) for which a blood isolate was also available. From this final
109 cohort, we sequenced all stool samples in our collection within 60 days prior to and 30 days after BSI.
110

111 **Bloodstream isolate identification**

112 Bloodstream isolates from HCT patients who received medical care at Stanford University Hospital were
113 obtained from the Stanford Hospital Clinical Microbiology Laboratory. All isolates considered typical
114 bloodstream pathogens by National Healthcare Safety Network (NHSN) guidelines are stored in a
115 glycerol suspension at -80°C for 12 months ("Centers for Disease Control and Prevention" 2018). Blood
116 culture isolates considered to be skin-associated bacteria (including viridans group *Streptococcus spp*. and
117 coagulase-negative *Staphylococcus spp.*) were saved if they were recovered in two or more blood culture
118 sets as per NHSN criteria ("Centers for Disease Control and Prevention" 2018). Isolates were identified
119 by standard biochemical testing and matrix-assisted laser desorption and ionization time-of-flight mass
120 spectrometry (MALDI-TOF) (Bruker Daltonics, Bremen, Germany).
121

122 **Sample processing**

123 Bloodstream bacterial isolates were plated on brain heart infusion agar with 10% horse blood. DNA was
124 extracted from isolates using the Gentra Puregene Yeast/Bact. Kit per manufacturer's instructions. Stool
125 samples were collected and stored at 4°C for up to 24 hours prior to homogenization, aliquoting and
126 storage at 80°C. DNA was extracted from stool using the QIAamp DNA Stool Mini Kit (QIAGEN®) per
127 manufacturer's instructions, with an initial bead-beating step prior to extraction using the Mini-
128 Beadbeater-16 (BioSpec Products) and 1 mm diameter Zirconia/Silica beads (BioSpec Products). Bead-

129 beating consisted of 7 rounds of alternating 30 second bead-beating bursts followed by 30 seconds
130 cooling on ice. DNA concentration for all samples was measured using Qubit® Fluorometric Quantitation
131 (Life Technologies). DNA sequencing libraries from both isolates and stool were prepared using the
132 Nextera XT DNA Library Prep Kit (Illumina®) with isolates and stool microbiota libraries prepared at
133 separate times following DNAse treatment of all lab surfaces to avoid cross-contamination. Library
134 concentration was measured using Qubit® Fluorometric Quantitation (Life Technologies) and library
135 quality and size distributions were analyzed with the Bioanalyzer 2100 (Agilent). Prepared libraries were
136 multiplexed and subjected to 100 base pair, paired-end sequencing on the HiSeq 4000 platform
137 (Illumina®).
138
139 **Shotgun sequencing analysis**
140 *Preprocessing*
141 Sequencing data were demultiplexed by unique barcodes (bcl2fastq v2.20.0.422, Illumina®). Read
142 quality was assessed using FastQC v0.11.4. Reads were deduplicated to remove PCR and optical
143 duplicates using SuperDeduper v1.4 with the start location in the read at 5 bp (-s 5) and minimum length
144 of 50 bp (-l 50). Deduplicated reads were trimmed using TrimGalore v0.4.4, a wrapper for CutAdapt
145 v1.16, with a minimum quality score of 30 for trimming (-q 30), minimum read length of 50 (--length 50)
146 and the "--nextera" flag to remove Illumina Nextera adapter sequences. Draft genomes of bacterial
147 isolates were assembled using SPAdes v3.8.0 (Bankevich et al. 2012) with default parameters. Summary
148 statistics for each BSI assembly were generated using 'basic_assembly_stats.py' from GAEMR v1.0.1
149 ("GAEMR" n.d.).
150
151 *Taxonomic classification*
152 Bloodstream isolates were taxonomically classified by aligning draft genome contigs to the BLAST nt
153 database and identifying the most common top hit for each contig. Gut metagenomic reads were
154 taxonomically classified via the One Codex platform, a web-based tool for assigning read-level

classifications based on unique *k*-mer signatures relative to a curated reference database [2017 database] (Minot, Krumm, and Greenfield 2015).

Phylogenetic tree building and variant identification with the StrainSifter pipeline

StrainSifter is a pipeline deployed via Snakemake with modules for variant calling and phylogenetic tree building, available at GitHub (https://github.com/bhattlab/strainsifter). StrainSifter accepts as input an assembled bacterial draft genome and two or more short read data sets (isolate or metagenomic), and can report a phylogenetic tree of input samples as well as pairwise SNV counts. To build the phylogenetic trees reported in this manuscript, an NCBI RefSeq reference genome for each infectious species (based on clinical laboratory classifications) was provided to StrainSifter. For the variant counting reported herein, BSI draft genomes were supplied to StrainSifter. For both analyses, all stool and BSI short read datasets were provided as input. For both phylogeny and SNV-counting modules, preprocessed short reads are first aligned to the reference genome using the Burrows-Wheeler aligner (BWA) v0.7.10 (Li and Durbin 2009). Alignments are filtered to include only high-confidence alignments with mapping quality of at least 60 using the "view" tool from the SAMtools suite (v1.7)(Li et al. 2009) (samtools view -b -q 60), and further filtered using BamTools "filter" (v2.4.0) to include only reads with the desired number or fewer mismatches (i.e. for five or fewer mismatches: bamtools filter -tag 'NM:<6') (Barnett et al. 2011) For phylogenetic tree construction, reads with 5 or fewer mismatches were included; for determining strain single-nucleotide variants, reads were limited to 1 or fewer mismatches. Per-base coverage is calculated from each resulting BAM file using bedtools genomecov (v2.26.0) and processed with custom python scripts to identify samples meeting a minimum average coverage of 5X across at least 50% of the genome. Only samples meeting the coverage requirement are continued through the pipeline. Pileup files are created from BAM files using SAMtools "mpileup", and are analyzed using custom python scripts to identify bases occurring with at least 0.8 frequency at positions covered 5X or greater. Only bases with a minimum phred score of >=20 are considered. Consensus sequences for each sample are created, wherein bases that cannot be confidently determined given the described parameters are called as "N".

181

182 To create a phylogenetic tree, core positions are identified on a per-species basis, where core positions are
183 defined as positions in the reference genome where a base could be confidently called for all samples
184 analyzed. To generate phylogenetic trees, core positions with variants in at least one sample are identified
185 and concatenated into fasta files. Fasta files are aligned using MUSCLE v3.8.31 (Edgar 2004) and a
186 maximum-likelihood phylogenetic tree is computed using FastTree v2.1.7 (Price, Dehal, and Arkin 2010).
187 Phylogenetic trees are visualized in R using the ape, phangorn, and ggtree packages. Pairwise SNVs are
188 determined from the consensus sequences using a custom python script.

189

190 *Synthetic multilocus sequence typing*

191 Metagenomic short reads were assembled using metaSPAdes v3.8.0 (Nurk et al. 2017). Multilocus
192 sequence typing (MLST) schemes and sequences were downloaded from the PubMLST database
193 ("PubMLST" n.d.). MLST gene sequences were aligned to metagenome assemblies, and the top hit for
194 each alignment was chosen based on E-value, percent identity, and alignment length. Only MLST
195 sequences that were present in the metagenomic assembly with 100% identity across the entire length of
196 the sequence were reported. MLST types generated by our in-house analysis were confirmed with the
197 SRST2 synthetic multilocus sequence typing tool (Inouye et al. 2014).

198

199 *Plots*

200 Plots were generated using the R programming language using the ggplot2, reshape2, and dplyr packages.

201

202 **<u>Results</u>**

203 **Sample collection and cohort**

204 During the study period, we identified 74 patients within our stool biospecimen collection with a
205 bloodstream infection (77 BSI episodes). Of these, we identified 30 patients (32 bloodstream isolates)
206 with at least 1 collected stool sample within 30 days prior to their bacterial BSI. From our final cohort of

207 30, we sequenced all stool samples collected between 60 days prior to and 30 days after the date of BSI,
208 83 samples in total. Patients had a median of 2 stool samples (range 1-8), with the sampling distribution
209 presented in Figure 1. Stools were collected at a median of 9 days prior to BSI (range -58 to +31).
210 Clinical characteristics of the 30 patients are listed in Table 1 (see Table S1 for individual patient data).
211 All patients in the cohort were on at least one antibiotic within 30 days prior to BSI, with
212 fluoroquinolones being the most commonly administered antibiotic.
213
214 **Table 1.**

Baseline characteristics		**N (%)**
Age (years)		
	<30	4 (13%)
	≤30-39	4 (13%)
	≤40-49	6 (20%)
	≤50-59	3 (10%)
	≤60	11 (37%)
Sex (%male)		17 (57%)
Underlying diagnosis		
	AML	7 (23%)
	ALL	6 (20%)
	CML	2 (7%)
	Lymphoma	7 (23%)
	MDS/Myelofibrosis	6 (20%)
	Other	2 (7%)
Conditioning regimen		
	Myeloablative	16 (53%)
	Reduced intensity	9 (30%)
	Non-myeloablative	5 (17%)

Transplant source		
	Peripheral blood	20 (67%)
	Bone marrow	8 (27%)
	Double umbilical cord blood	2 (7%)
Type of donor		
Autologous		3 (10%)
Allogeneic		27 (90%)
	Matched related donor	10 (37%)
	Matched unrelated donor	14 (52%)
	Mismatched unrelated donor	3 (11%)
TPN within 30 days		15 (50%)
Antibiotics within 30 days*		30 (100%)
	Fluoroquinolones	26 (87%)
	Beta-lactam	14 (47%)
	Carbapenem	6 (20%)
	Vancomycin (IV)	12 (40%)
Bacteremia species		
Gram-positive	*Staphylococcus aureus*	8 (25%)
	MSSA	5 (16%)
	MRSA	3 (9%)
	Coagulase-negative Staphylococcus spp.	4 (13%)
	Streptococcus mitis	7 (22%)
	Enterococcus faecium	2 (6%)
	Rothia mucilaginosa	2 (6%)
Gram-negative	*Escherichia coli*	4 (13%)
	Klebsiella spp.	3 (9%)
	Klebsiella pneumoniae	2 (6%)

Klebsiella oxytoca	1 (3%)
Enterobacter cloacae	1 (3%)
Pseudomonas aeruginosa	1 (3%)

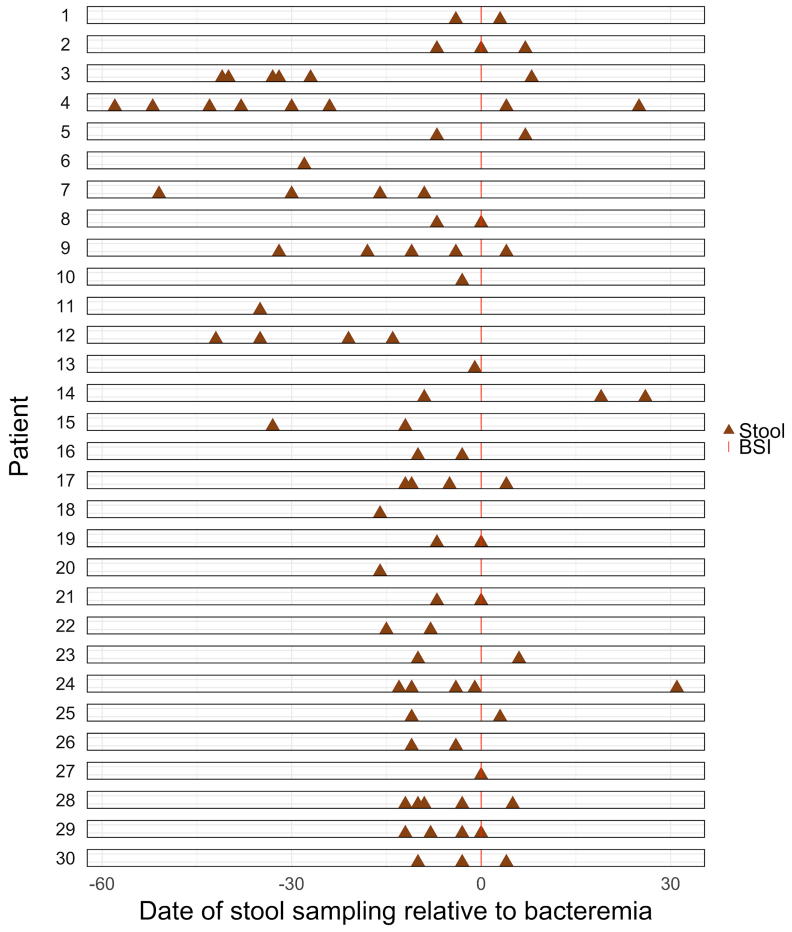
List of abbreviations: AML (acute myeloid leukemia); ALL (acute lymphoblastic leukemia); CML (chronic myeloid leukemia).

***Antibiotics are not exclusive and will not add up to 100%**

221 **Figure 1. Stool sampling**

222 Timeline of stool sampling relative to BSI and hematopoietic cell transplant for patients included in the

223 study.



224

225

226 **Infectious taxa are present in the gut microbiome near the time of infection**

227 We sequenced an average of 68,704,214 reads per stool sample (20,298,792-148,162,892) and 27,999,060

228 reads per isolate (10,652,310-61,058,906) (Table S2). Isolate reads were assembled into draft genomes

229 (Table S3).

230

231 MALDI-TOF is unable to discriminate between some closely related taxa (e.g. *Streptococcus* spp.), so we

232 sought to confirm the clinical laboratory BSI classifications by aligning draft genome contigs to the

233 BLAST nt database. Of 32 BSIs, we identified 3 for which WGS and MALDI-TOF results disagreed.
234 Two *Streptococcus* isolates that had been classified as *Streptococcus mitis* by MALDI-TOF were
235 identified as *Streptococcus oralis,* and one as *Streptococcus pseudopneumoniae* by WGS (Table S4).
236
237 To determine community composition of pre- and post-BSI stool samples, we taxonomically classified
238 gut metagenomic reads via the One Codex platform. Considering species level read classifications, we
239 observe the target species in the gut at a threshold of 0.1% or greater relative abundance for 19 of 83
240 stools (23%) or 20 of 95 stool-BSI pairs (21%) taking into account episodes of polymicrobial BSI; one
241 patient developed a BSI with two bacterial species, both of which were present in the stool above
242 threshold level (Table 2; full taxonomic classifications in table S5). Read-level taxonomic classifiers are
243 often unable to discriminate between closely related species, so we also compared BSIs and stool read
244 classifications at the genus level. We observe the target genus in the gut community for 27% of stools, or
245 23% of stool-BSI pairs (Table S5, S6).
246
247 **Table 2.**

Patient	Day relative to BSI	BSI organism	Relative abundance (%)
1	-4	Staphylococcus aureus	0
1	3	Staphylococcus aureus	0
2	-7	Enterobacter cloacae	0
2	0	Enterobacter cloacae	0.74
2	7	Enterobacter cloacae	0
2	-7	Klebsiella pneumoniae	0
2	0	Klebsiella pneumoniae	3.47
2	7	Klebsiella pneumoniae	0.01
3	-27	Escherichia coli	87.15
3	-32	Escherichia coli	97.25
3	-33	Escherichia coli	96.4
3	-40	Escherichia coli	0.05

3	-41	Escherichia coli	0.43
3	8	Escherichia coli	0.01
3	-27	Staphylococcus aureus	0
3	-32	Staphylococcus aureus	0
3	-33	Staphylococcus aureus	0
3	-40	Staphylococcus aureus	0
3	-41	Staphylococcus aureus	0
3	8	Staphylococcus aureus	0
4	-24	Staphylococcus aureus	0
4	-30	Staphylococcus aureus	0
4	-38	Staphylococcus aureus	0
4	-43	Staphylococcus aureus	0
4	-52	Staphylococcus aureus	0
4	-58	Staphylococcus aureus	0
4	25	Staphylococcus aureus	0
4	4	Staphylococcus aureus	0
5	-7	Staphylococcus aureus	0
5	7	Staphylococcus aureus	0
6	-28	Klebsiella oxytoca	0
7	-16	Escherichia coli	0.79
7	-30	Escherichia coli	0.22
7	-51	Escherichia coli	2.05
7	-9	Escherichia coli	0.38
8	-7	Rothia mucilaginosa	0
8	0	Rothia mucilaginosa	0
9	-11	Escherichia coli	0.31
9	-18	Escherichia coli	0
9	-32	Escherichia coli	0
9	-4	Escherichia coli	0.03
9	4	Escherichia coli	0
10	-3	Staphylococcus aureus	0

11	-35	Escherichia coli	0.04
12	-14	Staphylococcus aureus	0
12	-21	Staphylococcus aureus	0
12	-35	Staphylococcus aureus	0.65
12	-42	Staphylococcus aureus	0.09
13	-1	Staphylococcus epidermidis	2.51
14	-9	Klebsiella pneumoniae	0
14	19	Klebsiella pneumoniae	0
14	26	Klebsiella pneumoniae	0
15	-12	Staphylococcus epidermidis	0
15	-12	Staphylococcus epidermidis	0
15	-33	Staphylococcus epidermidis	0
15	-33	Staphylococcus epidermidis	0
16	-10	Streptococcus mitis	0
16	-3	Streptococcus mitis	0
17	-11	Streptococcus mitis	0
17	-12	Streptococcus mitis	0
17	-5	Streptococcus mitis	0
17	4	Streptococcus mitis	0.03
18	-16	Staphylococcus epidermidis	0
18	-16	Staphylococcus epidermidis	0
19	-7	Pseudomonas aeruginosa	0
19	0	Pseudomonas aeruginosa	0.83
20	-16	Staphylococcus epidermidis	0
21	-7	Staphylococcus aureus	0.01
21	0	Staphylococcus aureus	0.25
22	-15	Streptococcus mitis	0
22	-8	Streptococcus mitis	2.33
23	-10	Enterococcus faecium	11.61
23	6	Enterococcus faecium	0
24	-1	Rothia mucilaginosa	0.04

24	-11	Rothia mucilaginosa	0
24	-13	Rothia mucilaginosa	0
24	-4	Rothia mucilaginosa	0
24	31	Rothia mucilaginosa	0
25	-11	Enterococcus faecium	0.36
25	3	Enterococcus faecium	94.09
26	-11	Streptococcus mitis	0
26	-4	Streptococcus mitis	0
27	0	Staphylococcus aureus	0.51
28	-10	Streptococcus mitis	0
28	-12	Streptococcus mitis	0.04
28	-3	Streptococcus mitis	0
28	-9	Streptococcus mitis	0.02
28	5	Streptococcus mitis	0
29	-12	Streptococcus mitis	0
29	-3	Streptococcus mitis	0
29	-8	Streptococcus mitis	0.01
29	0	Streptococcus mitis	0.03
30	-10	Streptococcus mitis	0
30	-3	Streptococcus mitis	0
30	4	Streptococcus mitis	0

Gut domination with an infectious pathogen does not always precede bloodstream infection

We sought to determine whether bloodstream infection is preceded by gut microbiota domination by a pathogen as has been previously reported (Ubeda et al. 2010; Taur et al. 2012). We defined domination as 30% species relative abundance or greater of an organism in the gut, as previously described (Taur et al. 2012). Of the 19 stools in which the BSI organism was detected, only 4 stool samples (21%) were dominated by the BSI pathogen. Stool samples from patient 3 were dominated by *Escherichia coli* (>85% relative abundance) 27, 32, and 33 days prior to the development of an *E. coli* BSI (Figure 2A; table 2);

257 *Enterococcus faecium* dominated the intestinal microbiome (94% relative abundance) in patient 25 three
258 days after *E. faecium* BSI (Figure 2B; table 2).
259
260 Several patients experienced BSI with organisms present at relatively lower abundance in the gut prior to
261 bloodstream infection: *Staphylococcus epidermidis* at approximately 2.5% relative abundance in the gut
262 of patient 13 one day prior to their *S. epidermidis* infection, *Pseudomonas aeruginosa* at low abundance
263 (0.83%) in the gut of patient 19 on the day of infection, and *Klebsiella pneumoniae* at 3.5% relative
264 abundance in the gut of patient 2, 9 days prior to BSI (Figure 2C,D; table 2).
265
266 We also observe gut domination by non-BSI-causing pathogens. Patient 14 experienced a *K. pneumoniae*
267 BSI, yet stool samples at two timepoints were dominated by other pathogens: *E. coli* at 67% relative
268 abundance 9 days prior to BSI and *E. faecium* at 85% relative abundance 19 days after BSI (Table 2).
269
270

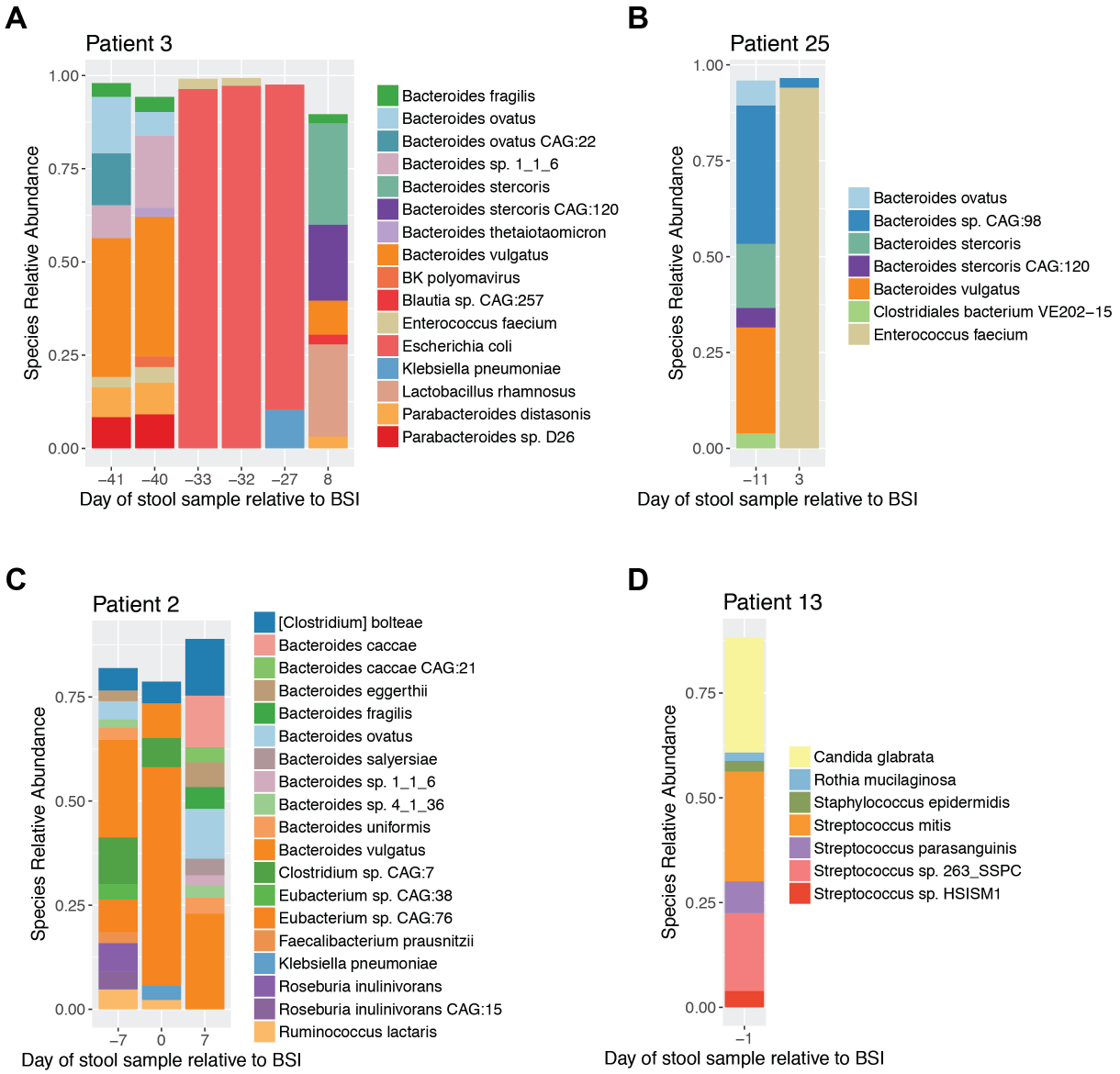
271 **Figure 2. Gut domination by a pathogen does not always precede bloodstream infection.** Domination

272 by *E. coli* (A) and Enterococcus faecium (B) occurs prior to bacteremia, while *K. pneumoniae* (C) and *S.*

273 *epidermidis* (D) are present in the gut microbiota of their respective patients prior to BSI but not at high

274 abundance. Plots show species present at 2% relative abundance or greater (host reads excluded).

275

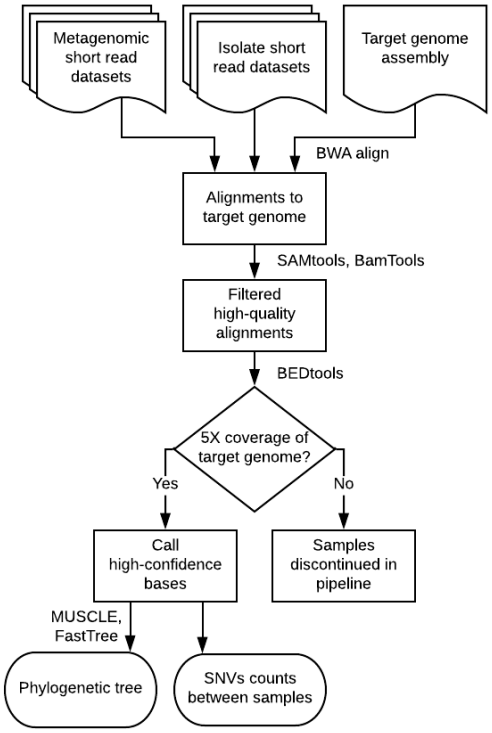


276

277 **Phylogenetic analysis indicates close relatedness of gut and BSI strains**

278 We developed the StrainSifter pipeline to perform nucleotide-level comparisons between genomes and

279 metagenomes. StrainSifter is deployed as a Snakemake workflow, and the pipeline steps are described in
280 Figure 3. Briefly, the inputs to StrainSifter are a draft genome (the "target" genome) and one or more
281 other short read datasets (isolate genomes or metagenomes). StrainSifter detects whether the target
282 organism is present with sufficient abundance in the short read datasets and outputs phylogenetic trees
283 and SNV counts between samples. Despite an abundance of research studies focused on strain-level
284 comparisons of microbes, there is still no clear definition of what constitutes a "strain," and it is unclear
285 whether genomes that differ by a few nucleotides should be considered distinct strains. To investigate the
286 overall relatedness of strains of each BSI species in our metagenomes and isolates, we first took a
287 comparative phylogenomics approach by applying StrainSifter to assess inter-patient strain relatedness.
288 Rather than compare strains from the same patient to publicly-available reference genomes, we instead
289 compared the phylogenetic relatedness of a patient's gut and BSI strains to strains from the other patient
290 samples in our collection. We reasoned that strains from the same geographic location are more likely to
291 be closely phylogenetically related than strains that are both spatially and temporally distant, so
292 comparing inter-patient strain relatedness in the context of the other strains in our sample collection
293 should more precisely indicate strains or lineages unique to a patient.
294
295 **Figure 3. StrainSifter pipeline**
296 Flowchart showing input, analysis steps, and output of the StrainSifter pipeline.



For each BSI species, we used StrainSifter to identify metagenomic samples with at least 5X coverage of the NCBI representative genome for each target organism, found core SNVs present in each sample, and inferred maximum-likelihood phylogenetic trees from multiple sequence alignments of concatenated core SNVs. Reference genomes were selected based on the clinical laboratory MALDI-TOF classification of BSI isolates. Phylogenetic trees demonstrate a higher level of intra-patient versus inter-patient genetic relatedness among colonizing and infecting organisms (Figure 4). BSI and intestinal strains of typically enteric species such as *Enterococcus faecium* (patient 25), *Klebsiella pneumoniae* (patient 2), *E. coli* (patient 3, patient 7), and *S. mitis* (patient 22) are closely phylogenetically related. Unexpectedly, we observe close intra-patient strain relatedness between gut and BSI for typically non-enteric taxa including *S. epidermidis* strains (patient 13) and *Pseudomonas aeruginosa* (patient 19, not pictured). Phylogenetic trees for *P. aeruginosa* and *Enterobacter cloacae* are not shown as these species were not observed in any other samples.

Single-nucleotide variants reveal strain identity

312 We next used StrainSifter to compare single-nucleotide variants (SNVs) between strains in our sample
313 collection. We executed the StrainSifter pipeline for each draft genome, with all stool short read datasets
314 as input. StrainSifter identified samples with sufficient coverage (at least 5X coverage across 50% or
315 more of the genome) of the target organism in the gut metagenome and determined the number of variants
316 shared between each pair of input samples that contained the target strain.
317
318 We observe 1 SNV between the BSI and gut content of Patient 13 who developed a *S. epidermidis* BSI
319 (Table 3). This indicates that the bloodstream strain is concordant to the strain found in the gut 1 day
320 prior. Patient 3 developed an *E. coli* bloodstream infection and we observe zero SNVs between the BSI
321 strain and all samples 33, 32, and 27 days prior to the BSI, indicating that the identical E. coli BSI strain
322 colonized the gut over a month before the onset of infection. Further, we observe zero discriminating
323 SNVs between identical strains of *Pseudomonas aeruginosa* in both the blood and stool specimens.
324 Notably, all stool samples containing the BSI-causing organism at sufficient coverage to profile with
325 StrainSifter were more closely related to their respective BSI isolate based on SNV distance than to any
326 sample from other patients. None of the 30 patients included in our study had sufficient *Staphylococcus*
327 *aureus* in their stool samples to profile with StrainSifter, indicating that this organism infrequently
328 colonizes the gut of HCT patients, even at low abundance.
329
330 **Table 3.**

BSI organism	Sample 1	Sample 2	SNVs	Bases compared	SNVs per base compared	Total bases	Percentage of bases compared
Enterobacter cloacae	Patient 2 BSI	Patient 2 stool day 0	5	4499921	1.11E-06	4552154	98.85%
Enterococcus faecium	Patient 25 BSI	Patient 25 stool day 3	2	2931442	6.82E-07	2935264	99.87%
Escherichia coli	Patient 3 BSI	Patient 3 stool day -33	0	5032224	0.00E+00	5056158	99.53%
Escherichia coli	Patient 3 BSI	Patient 3 stool day -32	0	5032230	0	5056158	99.53%

Escherichia coli	Patient 3 BSI	Patient 3 stool day -27	0	5030454	0	5056158	99.49%
Escherichia coli	Patient 7 BSI	Patient 7 stool day -51	259	4388254	5.90E-05	5267890	83.30%
Klebsiella pneumoniae	Patient 2 BSI	Patient 2 stool day 0	4	5379270	7.44E-07	5390737	99.79%
Klebsiella pneumoniae	Patient 14 BSI	Patient 14 stool day -9	72	3924126	1.83E-05	5407718	72.57%
Pseudomonas aeruginosa	Patient 19 BSI	Patient 19 stool day 0	0	6830248	0	6882743	99.24%
Staphylococcus epidermidis	Patient 13 BSI	Patient 13 stool day -1	1	2592411	3.86E-07	2604609	99.53%
Streptococcus mitis	Patient 22 BSI	Patient 22 stool day -8	28	1345797	2.08E-05	2084062	64.58%

331

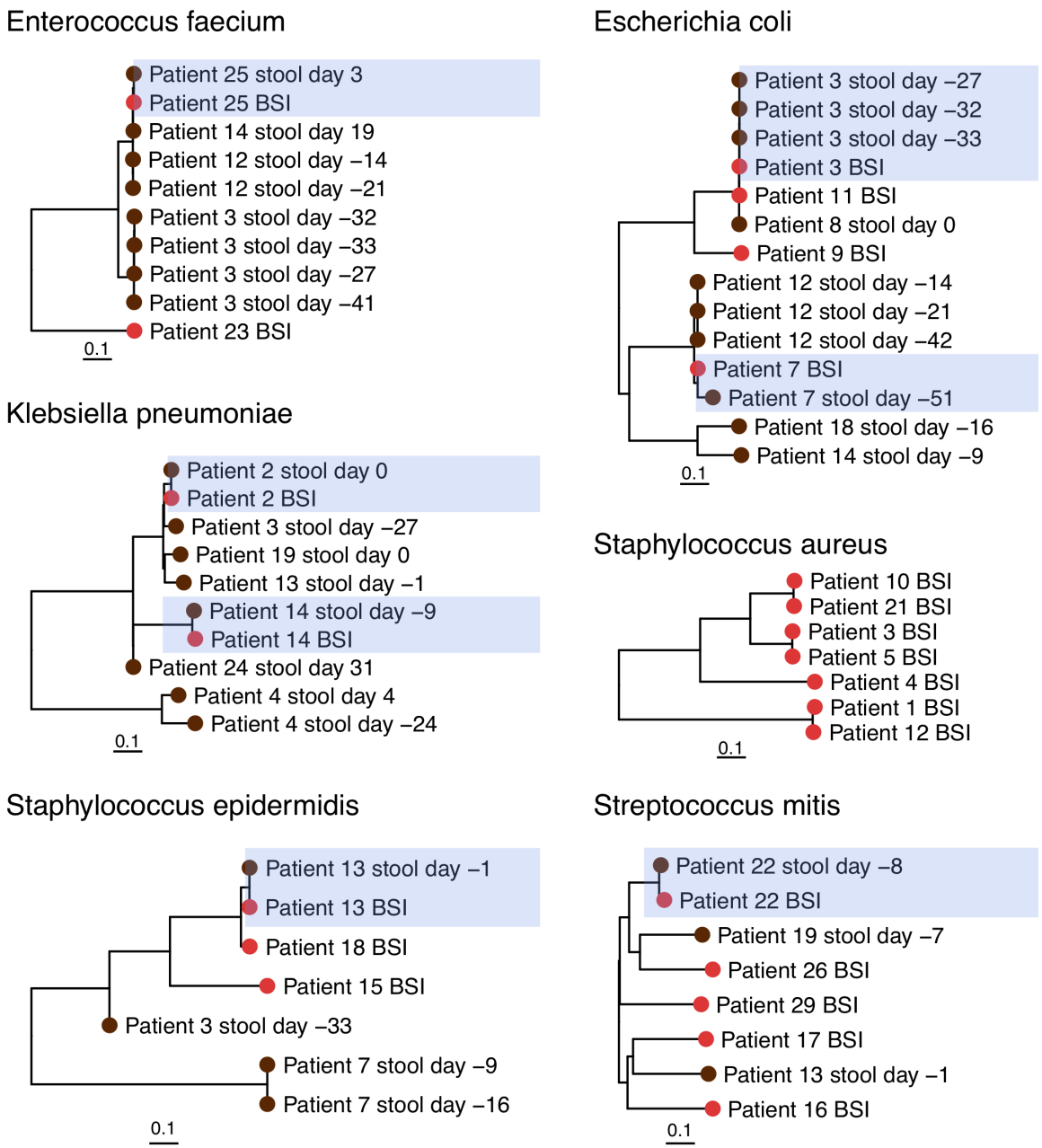
332

333 **Figure 4. Intra-patient samples are more closely related than inter-patient samples.** Branch tip

334 colors indicate stool and BSI samples from the same patient. Samples from the same patient are more

335 closely phylogenetically related than other samples.

336



337

338

339 ***In silico* multilocus sequence typing is concordant with StrainSifter results**

340 To compare whole-genome sequencing-based approaches to traditional strain typing methods, we
341 performed *in silico* multilocus sequence typing (MLST) on BSI genomes and assembled stool
342 metagenomes (Table 3). In many instances, MLST types for gut metagenome strains could not be
343 resolved or could only be partially detected, likely due to low abundance or absence of the target
344 organism in the gut community. MLST data is concordant with StrainSifter SNV profiling in only 4
345 infections for which MLST schemes exist for the bloodstream isolate of interest (Table S7). MLST types
346 could not be determined for either of patient 2's BSI isolates (*K. pneumoniae* and *E. cloacae*) because the
347 sequences at several loci in each genome were not present in the respective MLST databases. We were
348 unable to type *Rothia mucilaginosa* and *Streptococcus mitis*, as standard MLST schemes do not exist for
349 these organisms.
350
351 **Discussion**
352
353 **StrainSifter is a useful pipeline for strain detection within metagenomes**
354 High resolution strain comparisons are required to identify the origins of bloodstream infections using
355 metagenomic samples. While many bioinformatic tools exist for strain comparison between clinical
356 isolates or SNV profiling between metagenomes, no pipeline exists for the specific purpose of detecting
357 the presence of a single strain in a metagenome. In this study, we have developed StrainSifter, a
358 straightforward and facile pipeline for identifying phylogenetic relationships between isolate and
359 metagenomic WGS samples.
360
361 In our study, StrainSifter was able to precisely and comprehensively compare gut metagenomic content to
362 the genomes of diverse bloodstream isolates using high resolution WGS. Although techniques such as *in*
363 *silico* MLST are able to compare genomes to metagenomes, MLST loci are only 450 to 500 base pairs on
364 average, and standard 7-gene MLST schemes therefore analyze less than 4 kilobases of DNA - on the
365 order of 0.1% of the average bacterial genome. By examining genome-wide variants, StrainSifter greatly

366 increases the confidence of strain comparisons.

367

368 Using StrainSifter, we found that the enteric organisms *E. faecium, E.coli*, *K. pneumoniae*, *E. cloacae*,
369 and *S. mitis* were concordant between blood and stool microbiota samples. This confirms what has
370 previously been reported in the literature (Tancrede and Andremont 1985; Samet et al. 2013). Previously
371 published studies have indicated that gut domination with a pathogen is a risk factor for subsequent
372 bloodstream infection (Taur et al. 2012; Ubeda et al. 2010; "Centers for Disease Control and Prevention"
373 2018). While *E. coli* and *E. faecium* domination events were observed preceding BSI, we find *K.*
374 *pneumoniae*, *E. cloacae,* and *S. epidermidis* in relatively low abundance in stool samples on the day of
375 BSI and one day prior.

376

377 Interestingly, we also find that typically non-enteric organisms *S. epidermidis* and *P. aeruginosa* were not
378 only present in the gut, but strains within the gut microbiota and bloodstream were concordant. Gut
379 colonization by *S. epidermidis* and *P. aeruginosa* is unexpected given current National Health and Safety
380 guidelines delimiting enteric versus non-enteric organisms (Taur et al. 2012; Ubeda et al. 2010; "Centers
381 for Disease Control and Prevention" 2018; Advani et al. 2017). Yet, colonization by atypical organisms
382 may not be surprising given the degree of disruption of the intestinal microbiota by antibiotics,
383 chemotherapy, and radiation. While *P. aeruginosa* intestinal colonization has been previously
384 documented (Nesher et al. 2014), intestinal colonization with coagulase-negative *Staphylococcus* is not
385 well described. In our own study, multiple stool samples were determined to have more than 80% relative
386 abundance of *S. epidermidis*. In addition, we observe the same colonizing strain of coagulase-negative
387 *Staphylococcus* over time in the same individual (patient 7), arguing against only transient membership
388 within the intestinal microbiome.

389

390 Previous research using existing strain typing and culture-based strategies has only tentatively
391 demonstrated evidence for gut translocation of coagulase-negative *Staphylococcus spp.* and *P. aeruginosa*

392 in immunocompromised patients (Wade et al. 1982; Rotstein et al. 1988; Nesher et al. 2014; MacFie et al.
393 1999; Costa et al. 2006). Here our straightforward approach allows for greater precision and confidence in
394 strain comparisons between the gut microbiota and infecting organism in the bloodstream. Given that our
395 analysis focused entirely on the stool microbiota, we cannot rule out the possibility of the same strain of
396 coagulase-negative *Staphylococcus* colonizing the entire body surface--including the gut, nares, skin, or
397 central line--from which the infection may have originated instead. Previous research has suggested,
398 however, that the majority of patients with coagulase-negative *Staphylococcus* bloodstream infection are
399 colonized with a multitude of strains that differ between the intestine and the skin (Herwaldt et al. 1992).
400
401 Determining whether two strains are related based on the number of SNVs that differ between their
402 genomes is not straightforward. While approximate bacterial mutation rates due to genetic drift have been
403 established, most estimates are derived from experiments in culture media. As gut and bloodstream
404 environments impose many selective pressures beyond genetic drift, such estimates are likely not relevant
405 in the clinical context. For instance, it is possible that bacterial pathogens that originate from the intestine
406 and translocate into the bloodstream may acquire single nucleotide variants over time due to selective
407 pressures. For this reason, it is difficult to define a threshold number of SNVs beyond which two strains
408 are not considered to be the same. Regardless of the absolute number of SNVs between strains or across
409 all strains analyzed with StrainSifter, strains from the same patient consistently share fewer SNVs
410 compared to those from other patients.
411
412 Although StrainSifter can precisely identify shared variants between genomes and metagenomes, it is
413 limited to profiling only the dominant strain of a given organism in a community. However, it has been
414 shown that gut metagenomes frequently contain only one predominant strain of each species, so
415 StrainSifter is likely to function optimally in most situations (Truong et al. 2017). As is the case with
416 existing tools for metagenomic strain comparison, StrainSifter is unable to analyze low-abundance strains
417 since adequate coverage (5X) is required for confident variant calling.

418

419 **Future directions**

420 StrainSifter can be used to precisely identify strains within relatively biomass-rich metagenomic samples
421 without relying on external reference genome databases in which representation of opportunistic
422 pathogens may be lacking. Compared to stool, samples from the nares, skin, or the environment typically
423 have far less microbial biomass, limiting the number of sequencing reads that can be obtained. In order to
424 provide the most precise epidemiologic information about the origins of bloodstream and other infections,
425 StrainSifter will need to be tested in low microbial biomass samples in a larger prospective study of
426 immunocompromised patients. Given the high resolution of WGS-based strain comparisons, we may
427 indeed discover instances of organisms thought to originate from non-enteric sites instead colonizing and
428 translocating from the intestine in severely immunocompromised patients, as was potentially
429 demonstrated by *S. epidermidis* and *P. aeruginosa* in our study. Given the growing body of research on
430 therapies to improve gut microbiota diversity and bolster colonization resistance against pathogens, more
431 precisely identifying the origins of bloodstream infections in the stem cell transplant population may
432 influence how hospitals and healthcare providers can most effectively work to prevent infections
433 (Galloway-Peña, Jenq, and Shelburne 2017). Finally, once StrainSifter is able to establish the origins of
434 an infection, the wealth of information offered by WGS can be interrogated to identify potential
435 functional differences, including in gene content and metabolic pathways, between strains colonizing the
436 gut, nares, or skin that result in disseminated infections compared to those that do not.

437

438 In conclusion, we have shown that StrainSifter offers a straightforward bioinformatic pipeline to precisely
439 compare bacterial isolates to metagenomic samples, circumventing the need for external reference
440 genome databases or culture-based methods of strain comparison. We successfully applied StrainSifter to
441 blood culture isolates and stool samples from hematopoietic stem cell transplant patients and found
442 classically non-enteric bacteria in the gut microbiome that were concordant with bloodstream isolates. We
443 anticipate that StrainSifter will be an easy-to-use tool in the context of hospital or infectious outbreak

epidemiologic studies, and that characterization of the microbiome dynamics that occur prior to infection may help us precisely identify potential reservoirs of pathogens.

Acknowledgements

We thank Joyce Kang for her assistance with stool sample processing, as well the other members of the Bhatt Lab for providing feedback on the study design, bioinformatics pipeline, and manuscript revisions. We would like to thank Nick Greenfield and the One Codex team for help with using their platform. We appreciate Drs. Matthew Kelly, Chris Severyn, and Doyle Ward for their feedback and ideas on the manuscript. We would especially like to thank the patients and nurses on the Blood and Marrow Transplant service for their enthusiastic participation in this project. This work was supported in part by National Science Foundation Graduate Research Fellowship (FBT), the National Institutes of Health, National Center for Advancing Translational Science, Clinical and Translational Science Award KL2 TR000083 and UL1 TR001085 (TMA). ASB was funded in part by the National Cancer Institute NIH K08 award, #CA184420, Damon Runyon Clinical Investigator Award, and Amy Strelzer Manasevit Award. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

461 **References**

462 Advani, Sonali, Rachael Lee, Mariann Schmitz, Martha Long, and Bernard Camins. 2017. "2015 NHSN
463 CAUTI Definition Change and Its Impact on CLABSI Rates at an Academic Medical Center." *Open*
464 *Forum Infectious Diseases* 4 (suppl_1): S180–S180.

465 Bankevich, Anton, Sergey Nurk, Dmitry Antipov, Alexey A. Gurevich, Mikhail Dvorkin, Alexander S.
466 Kulikov, Valery M. Lesin, et al. 2012. "SPAdes: A New Genome Assembly Algorithm and Its
467 Applications to Single-Cell Sequencing." *Journal of Computational Biology: A Journal of*
468 *Computational Molecular Cell Biology* 19 (5): 455–77.

469 Barnett, Derek W., Erik K. Garrison, Aaron R. Quinlan, Michael P. Strömberg, and Gabor T. Marth.
470 2011. "BamTools: A C++ API and Toolkit for Analyzing and Managing BAM Files."
471 *Bioinformatics* 27 (12): 1691–92.

472 Blennow, O., P. Ljungman, E. Sparrelid, J. Mattsson, and M. Remberger. 2014. "Incidence, Risk Factors,
473 and Outcome of Bloodstream Infections during the Pre-Engraftment Phase in 521 Allogeneic
474 Hematopoietic Stem Cell Transplantations." *Transplant Infectious Disease: An Official Journal of*
475 *the Transplantation Society* 16 (1): 106–14.

476 "Centers for Disease Control and Prevention." 2018. Bloodstream Infection Event (Central Line-
477 Associated Bloodstream Infection and Non-Central Line-Associated Bloodstream Infection). The
478 Centers for Disease Control and Prevention Web Site. February 2018.
479 http://www.cdc.gov/nhsn/PDFs/pscManual/4PSC_CLABS current.pdf.

480 Chen, Yi, Yan Luo, Heather Carleton, Ruth Timme, David Melka, Tim Muruvanda, Charles Wang, et al.
481 2017. "Whole Genome and Core Genome Multilocus Sequence Typing and Single Nucleotide
482 Polymorphism Analyses of Listeria Monocytogenes Associated with an Outbreak Linked to Cheese,
483 United States, 2013." *Applied and Environmental Microbiology* 83 (15): 1–15.

484 Costa, Silvia F., Antonio A. Barone, Marisa H. Miceli, Inneke M. van der Heijden, Robson E. Soares,
485 Anna S. Levin, and Elias J. Anaissie. 2006. "Colonization and Molecular Epidemiology of

486 Coagulase-Negative Staphylococcal Bacteremia in Cancer Patients: A Pilot Study." *American*
487 *Journal of Infection Control* 34 (1): 36–40.
488 Costea, Paul Igor, Robin Munch, Luis Pedro Coelho, Lucas Paoli, Shinichi Sunagawa, and Peer Bork.
489 2017. "metaSNV: A Tool for Metagenomic Strain Level Analysis." *PloS One* 12 (7): 1–9.
490 Edgar, Robert C. 2004. "MUSCLE: Multiple Sequence Alignment with High Accuracy and High
491 Throughput." *Nucleic Acids Research* 32 (5): 1792–97.
492 "GAEMR." n.d. Accessed April 19, 2018. https://www.broadinstitute.org/software/gaemr/.
493 Galloway-Peña, Jessica R., Robert R. Jenq, and Samuel A. Shelburne. 2017. "Can Consideration of the
494 Microbiome Improve Antimicrobial Utilization and Treatment Outcomes in the Oncology Patient?"
495 *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* 23
496 (13): 3263–68.
497 Gorrie, Claire L., Mirjana Mirčeta, Ryan R. Wick, David J. Edwards, Nicholas R. Thomson, Richard A.
498 Strugnell, Nigel F. Pratt, et al. 2017. "Gastrointestinal Carriage Is a Major Reservoir of Klebsiella
499 Pneumoniae Infection in Intensive Care Patients." *Clinical Infectious Diseases: An Official*
500 *Publication of the Infectious Diseases Society of America* 65 (2): 1–8.
501 Gudiol, C., C. Garcia-Vidal, M. Arnan, I. Sánchez-Ortega, B. Patiño, R. Duarte, and J. Carratalà. 2014.
502 "Etiology, Clinical Features and Outcomes of Pre-Engraftment and Post-Engraftment Bloodstream
503 Infection in Hematopoietic SCT Recipients." *Bone Marrow Transplantation* 49 (6): 824–30.
504 Harris, Simon R., Edward J. P. Cartwright, M. Estée Török, Matthew T. G. Holden, Nicholas M. Brown,
505 Amanda L. Ogilvy-Stuart, Matthew J. Ellington, et al. 2013. "Whole-Genome Sequencing for
506 Analysis of an Outbreak of Meticillin-Resistant Staphylococcus Aureus: A Descriptive Study." *The*
507 *Lancet Infectious Diseases* 13 (2): 130–36.
508 Herwaldt, L. A., R. J. Hollis, L. D. Boyken, and M. A. Pfaller. 1992. "Molecular Epidemiology of
509 Coagulase-Negative Staphylococci Isolated from Immunocompromised Patients." *Infection Control*
510 *and Hospital Epidemiology: The Official Journal of the Society of Hospital Epidemiologists of*
511 *America* 13 (2): 86–92.

512 Inouye, Michael, Harriet Dashnow, Lesley-Ann Raven, Mark B. Schultz, Bernard J. Pope, Takehiro
513 Tomita, Justin Zobel, and Kathryn E. Holt. 2014. "SRST2: Rapid Genomic Surveillance for Public
514 Health and Hospital Microbiology Labs." *Genome Medicine* 6 (11): 90.
515 Kovanen, Sara M., Rauni I. Kivistö, Mirko Rossi, Thomas Schott, Ulla-Maija Kärkkäinen, Tamara
516 Tuuminen, Jaakko Uksila, Hilpi Rautelin, and Marja-Liisa Hänninen. 2014. "Multilocus Sequence
517 Typing (MLST) and Whole-Genome MLST of Campylobacter Jejuni Isolates from Human
518 Infections in Three Districts during a Seasonal Peak in Finland." *Journal of Clinical Microbiology*
519 52 (12): 4147–54.
520 Lieberman, Tami D., Kelly B. Flett, Idan Yelin, Thomas R. Martin, Alexander J. McAdam, Gregory P.
521 Priebe, and Roy Kishony. 2013. "Genetic Variation of a Bacterial Pathogen within Individuals with
522 Cystic Fibrosis Provides a Record of Selective Pressures." *Nature Genetics* 46 (1): 82–87.
523 Li, H., and R. Durbin. 2009. "Fast and Accurate Short Read Alignment with Burrows–Wheeler
524 Transform." *Bioinformatics* 25 (14): 1754–60.
525 Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo
526 Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. "The
527 Sequence Alignment/Map Format and SAMtools." *Bioinformatics* 25 (16): 2078–79.
528 Li, Wenjun, Didier Raoult, and Pierre-Edouard Fournier. 2009. "Bacterial Strain Typing in the Genomic
529 Era." *FEMS Microbiology Reviews* 33 (5): 892–916.
530 MacFie, J., C. O'Boyle, C. J. Mitchell, P. M. Buckley, D. Johnstone, and P. Sudworth. 1999. "Gut Origin
531 of Sepsis: A Prospective Study Investigating Associations between Bacterial Translocation, Gastric
532 Microflora, and Septic Morbidity." *Gut* 45 (2): 223–28.
533 Martin, Rebekah M., Jie Cao, Sylvain Brisse, Virginie Passet, Weisheng Wu, Lili Zhao, Preeti N. Malani,
534 Krishna Rao, and Michael A. Bachman. 2016. "Molecular Epidemiology of Colonizing and Infecting
535 Isolates of Klebsiella Pneumoniae." *mSphere* 1 (5): 1–12.
536 Mikulska, Malgorzata, Valerio Del Bono, Anna Maria Raiola, Barbara Bruno, Francesca Gualandi,
537 Domenico Occhini, Carmen di Grazia, Francesco Frassoni, Andrea Bacigalupo, and Claudio Viscoli.

538 2009. “Blood Stream Infections in Allogeneic Hematopoietic Stem Cell Transplant Recipients:
539 Reemergence of Gram-Negative Rods and Increasing Antibiotic Resistance.” *Biology of Blood and*
540 *Marrow Transplantation: Journal of the American Society for Blood and Marrow Transplantation*
541 15 (1): 47–53.
542 Minot, Samuel S., Niklas Krumm, and Nicholas B. Greenfield. 2015. “One Codex: A Sensitive and
543 Accurate Data Platform for Genomic Microbial Identification.” *bioRxiv*, 027607.
544 Nayfach, Stephen, Beltran Rodriguez-Mueller, Nandita Garud, and Katherine S. Pollard. 2016. “An
545 Integrated Metagenomics Pipeline for Strain Profiling Reveals Novel Patterns of Bacterial
546 Transmission and Biogeography.” *Genome Research* 26 (11): 1612–25.
547 Nesher, L., K. V. I. Rolston, D. P. Shah, J. T. Tarrand, V. Mulanovich, E. J. Ariza-Heredia, and R. F.
548 Chemaly. 2014. “Fecal Colonization and Infection with Pseudomonas aeruginosa in Recipients of
549 Allogeneic Hematopoietic Stem Cell Transplantation.” *Transplant Infectious Disease: An Official*
550 *Journal of the Transplantation Society* 17 (1): 33–38.
551 Nielsen, Karen Leth, Marc Stegger, Kristoffer Kiil, Paul A. Godfrey, Michael Feldgarden, Berit Lilje,
552 Paal S. Andersen, and Niels Frimodt-Møller. 2017. “Whole-Genome Comparison of Urinary
553 Pathogenic Escherichia Coli and Faecal Isolates of UTI Patients and Healthy Controls.”
554 *International Journal of Medical Microbiology: IJMM* 307 (8): 497–507.
555 Nurk, Sergey, Dmitry Meleshko, Anton Korobeynikov, and Pavel A. Pevzner. 2017. “metaSPAdes: A
556 New Versatile Metagenomic Assembler.” *Genome Research* 27 (5): 824–34.
557 Price, Morgan N., Paramvir S. Dehal, and Adam P. Arkin. 2010. “FastTree 2--Approximately Maximum-
558 Likelihood Trees for Large Alignments.” *PloS One* 5 (3). Public Library of Science: e9490.
559 “PubMLST.” n.d. Accessed April 20, 2018. https://pubmlst.org/.
560 Rotstein, C., D. Higby, K. Killion, and E. Powell. 1988. “Relationship of Surveillance Cultures to
561 Bacteremia and Fungemia in Bone Marrow Transplant Recipients with Hickman or Broviac
562 Catheters.” *Journal of Surgical Oncology* 39 (3): 154–58.
563 Salipante, Stephen J., Dhruba J. SenGupta, Lisa A. Cummings, Tyler A. Land, Daniel R. Hoogestraat,

564 and Brad T. Cookson. 2015. “Application of Whole-Genome Sequencing for Bacterial Strain Typing
565 in Molecular Epidemiology.” *Journal of Clinical Microbiology* 53 (4): 1072–79.
566 Samet, A., A. Śledzińska, B. Krawczyk, A. Hellmann, S. Nowicki, J. Kur, and B. Nowicki. 2013.
567 “Leukemia and Risk of Recurrent Escherichia Coli Bacteremia: Genotyping Implicates E. Coli
568 Translocation from the Colon to the Bloodstream.” *European Journal of Clinical Microbiology &*
569 *Infectious Diseases: Official Publication of the European Society of Clinical Microbiology* 32 (11):
570 1393–1400.
571 See, Isaac, Minn M. Soe, Lauren Epstein, Jonathan R. Edwards, Shelley S. Magill, and Nicola D.
572 Thompson. 2017. “Impact of Removing Mucosal Barrier Injury Laboratory-Confirmed Bloodstream
573 Infections from Central Line-Associated Bloodstream Infection Rates in the National Healthcare
574 Safety Network, 2014.” *American Journal of Infection Control* 45 (3): 321–23.
575 Snitkin, E. S., A. M. Zelazny, P. J. Thomas, F. Stock, D. K. Henderson, T. N. Palmore, J. A. Segre, and
576 Nisc Comparative Sequencing Program. 2012. “Tracking a Hospital Outbreak of Carbapenem-
577 Resistant Klebsiella Pneumoniae with Whole-Genome Sequencing.” *Science Translational Medicine*
578 4 (148): 148ra116–48ra116.
579 Snitkin, Evan S., Adrian M. Zelazny, Clemente I. Montero, Frida Stock, Lilia Mijares, NISC Comparative
580 Sequence Program, Patrick R. Murray, and Julie A. Segre. 2011. “Genome-Wide Recombination
581 Drives Diversification of Epidemic Strains of Acinetobacter Baumannii.” *Proceedings of the*
582 *National Academy of Sciences of the United States of America* 108 (33): 13758–63.
583 Tancrede, C. H., and A. O. Andremont. 1985. “Bacterial Translocation and Gram-Negative Bacteremia in
584 Patients with Hematological Malignancies.” *The Journal of Infectious Diseases* 152 (1): 99–103.
585 Taur, Ying, Joao B. Xavier, Lauren Lipuma, Carles Ubeda, Jenna Goldberg, Asia Gobourne, Yeon Joo
586 Lee, et al. 2012. “Intestinal Domination and the Risk of Bacteremia in Patients Undergoing
587 Allogeneic Hematopoietic Stem Cell Transplantation.” *Clinical Infectious Diseases: An Official*
588 *Publication of the Infectious Diseases Society of America* 55 (7): 905–14.
589 Truong, Duy Tin, Adrian Tett, Edoardo Pasolli, Curtis Huttenhower, and Nicola Segata. 2017. “Microbial

590 Strain-Level Population Structure & Genetic Diversity from Metagenomes." *Genome Research* 27
591 (4): 626–38.
592 Ubeda, Carles, Ying Taur, Robert R. Jenq, Michele J. Equinda, Tammy Son, Miriam Samstein, Agnes
593 Viale, et al. 2010. "Vancomycin-Resistant Enterococcus Domination of Intestinal Microbiota Is
594 Enabled by Antibiotic Treatment in Mice and Precedes Bloodstream Invasion in Humans." *The*
595 *Journal of Clinical Investigation* 120 (12): 4332–41.
596 Wade, J. C., S. C. Schimpff, K. A. Newman, and P. H. Wiernik. 1982. "Staphylococcus Epidermidis: An
597 Increasing Cause of Infection in Patients with Granulocytopenia." *Annals of Internal Medicine* 97
598 (4): 503–8.

599