METHOD ARTICLE

REVISED High-resolution sweep metagenomics using fast probabilistic inference [version 2; peer review: 2 approved]

Tommi Mäklin[1], Teemu Kallonen[2,3], Sophia David[4], Christine J. Boinett[5,6], Ben Pascoe[7], Guillaume Méric[7], David M. Aanensen[3,8,9], Edward J. Feil[7], Stephen Baker[5,6], Julian Parkhill[3], Samuel K. Sheppard[7], Jukka Corander[1-3], Antti Honkela[1,10,11]

[1]Helsinki Institute for Information Technology HIIT, Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland
[2]Department of Biostatistics, University of Oslo, Oslo, Norway
[3]Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire, UK
[4]Centre for Genomic Pathogen Surveillance, Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire, UK
[5]Hospital for Tropical Diseases, Wellcome Trust Major Overseas Programme, Oxford University Clinical Research Unit, Ho Chi Minh City, Vietnam
[6]Centre for Tropical Medicine and Global Health, Nuffield Department of Clinical Medicine, University of Oxford, Oxford, UK
[7]The Milner Centre for Evolution, Department of Biology and Biochemistry, University of Bath, Bath, UK
[8]Department of Infectious Disease Epidemiology, Imperial College London, London, UK
[9]Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, Nuffield Department of Medicine, University of Oxford, Oxford, UK
[10]Department of Public Health, University of Helsinki, Helsinki, Finland
[11]Helsinki Institute for Information Technology HIIT, Department of Computer Science, University of Helsinki, Helsinki, Finland

v2 **First published:** 30 Jan 2020, **5**:14
https://doi.org/10.12688/wellcomeopenres.15639.1
Latest published: 08 Oct 2021, **5**:14
https://doi.org/10.12688/wellcomeopenres.15639.2

Abstract

Determining the composition of bacterial communities beyond the level of a genus or species is challenging because of the considerable overlap between genomes representing close relatives. Here, we present the mSWEEP pipeline for identifying and estimating the relative sequence abundances of bacterial lineages from plate sweeps of enrichment cultures. mSWEEP leverages biologically grouped sequence assembly databases, applying probabilistic modelling, and provides controls for false positive results. Using sequencing data from major pathogens, we demonstrate significant improvements in lineage quantification and detection accuracy. Our pipeline facilitates investigating cultures comprising mixtures of bacteria, and opens up a new field of plate sweep metagenomics.

Keywords

plate sweeps, bacterial strain identification, microbial communities, metagenomics, probabilistic modeling

Open Peer Review

Approval Status ✓ ✓

	1	2
version 2 (revision) 08 Oct 2021		✓ view
version 1 30 Jan 2020	✓ view	? view

1. **Fabiano L. Thompson**, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil

2. **Mahendra Mariadassou**, Université Paris-Saclay, Jouy-en-Josas, France

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding authors: Tommi Mäklin (tommi.maklin@helsinki.fi), Antti Honkela (antti.honkela@helsinki.fi)

Author roles: Mäklin T: Conceptualization, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Kallonen T**: Conceptualization, Data Curation, Investigation, Writing – Original Draft Preparation, Writing – Review & Editing; **David S**: Investigation, Writing – Original Draft Preparation, Writing – Review & Editing; **Boinett CJ**: Formal Analysis, Investigation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Pascoe B**: Investigation, Writing – Original Draft Preparation, Writing – Review & Editing; **Méric G**: Investigation, Writing – Review & Editing; **Aanensen DM**: Investigation, Resources, Writing – Review & Editing; **Feil EJ**: Funding Acquisition, Investigation, Resources, Writing – Review & Editing; **Baker S**: Funding Acquisition, Investigation, Resources, Writing – Review & Editing; **Parkhill J**: Funding Acquisition, Investigation, Resources, Writing – Review & Editing; **Sheppard SK**: Funding Acquisition, Investigation, Resources, Writing – Original Draft Preparation, Writing – Review & Editing; **Corander J**: Conceptualization, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **Honkela A**: Conceptualization, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by the Academy of Finland (grants no. 259440 and 310261; to TM and AH) as well as the Flagship programme (Finnish Center for Artificial Intelligence FCAI; to JC and AH). TK, JC, DA and EJF are supported by the JPI-AMR consortium SpARK (MR/R00241X/1). JC was funded by the ERC (grant no. 742158). TK was funded by the Norwegian Research Council JPIAMR (grant no. 144501). SB is a Sir Henry Dale Fellow, jointly funded by the Wellcome Trust and the Royal Society [100087]. Sequencing of the Vietnamese E. coli samples was supported by the Wellcome Trust [098051]. Computational resources were provided by the 'Finnish Grid and Cloud Infrastructure' (persistent identifier urn:nbn:fi:research-infras-2016072533).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Mäklin T, Kallonen T, David S *et al.* **High-resolution sweep metagenomics using fast probabilistic inference [version 2; peer review: 2 approved]** Wellcome Open Research 2021, **5**:14 https://doi.org/10.12688/wellcomeopenres.15639.2

First published: 30 Jan 2020, **5**:14 https://doi.org/10.12688/wellcomeopenres.15639.1

REVISED **Amendments from Version 1**

We have revised our manuscript based on the feedback provided by the two reviewers. Notably, we have added a new synthetic experiment assessing the performance of mSWEEP in a setting with many *Escherichia coli* lineages simultaneously present at wildly varying coverages ranging from 50× to 0.10×. The results from this assessment are included as Figure 5 in the revised manuscript. Additionally, we have expanded the Discussion section to cover a limitation of our method related to the presence of novel or uncharacterized lineages in the samples, and also made several changes and additions throughout the manuscript related to the minimum sequencing depth required to use mSWEEP. Finally, we would like to thank both reviewers for their time and apt comments that have enabled us to improve the quality of our manuscript.

Any further responses from the reviewers can be found at the end of the article

Introduction

High-throughput sequencing technologies have enabled researchers to study bacterial populations in unprecedented detail using whole-genome sequencing of pure individual bacterial colonies. Sequencing of individual isolates has provided insights into antimicrobial resistance and the complex ecology of the spread of antimicrobial resistant variants globally. The application of community profiling metagenomics, in which the 16S rRNA gene is sequenced from complex multi-species samples, can provide information about the composition and dynamics of highly diverse bacterial populations. However, the resolution of this approach is limited due to insufficient nucleotide variation[1] and profiling beyond the level of genus/species is generally not possible. Whole-genome shotgun metagenomics delivers a much higher resolution than 16s rRNA sequencing[2] but widespread application is hindered by the cost associated with sequencing a sample to a sufficient depth to capture the diverse set of organisms and strain-level variation that may be present in the sample[3].

Current methods for taxonomic profiling of bacteria from sequencing data[4] typically perform well only up to the species-level[5] or focus on analysing predetermined single nucleotide variants (SNVs) and/or marker genes to capture the variation contained in a mixed colony of closely related strains[6–8]. Sequencing isolated colonies offers means to ignore this variation but only focusing on pure colonies is insufficient for many potential applications[9,10]. Furthermore, whilst the SNV-based approach has been successful in studies of the history of the human population, focusing solely on SNVs inadequately captures the greater variability and different modalities of variation in bacterial genomes. Conversely, solely gene-based approaches can capture some of this while potentially losing finer detail. Therefore, we aimed to strike a balance between these two approaches by making use of a complete genome reference database.

Here, we have developed the mSWEEP pipeline, which is designed to make efficient use of large collections of reference genomes that are available for numerous important human pathogens and other culturable bacterial species. mSWEEP combines clustering of the reference genomes into biologically relevant groups, fast pseudoalignment of reads to the references, fast and accurate probabilistic inference of the cluster abundances and a method for controlling false positive detections. Similar methods taking advantage of pseudoalignment either with[11,12] or without[13] the application of probabilistic modelling have been developed but we show that our combination of clustering with large reference collections vastly increases the accuracy of obtained estimates.

Although applicable to any scenario where reference genomes for the sequenced bacteria are available, mSWEEP specifically enables a new kind of high-resolution analysis in plate sweep metagenomics, where a mixture of colonies is harvested from an enrichment culture by sweeping the whole plate in contrast to isolating a single colony. Plate sweep experiments fall between whole-genome sequencing of single colonies and culture-independent metagenomics by analysing the entire complexity of a community from a specific growth medium. Since the potential species are restricted in advance by the growth medium, plate sweeps offer a cost-effective way to obtain high-depth sequencing data from only the target organisms of interest and reduce potential sources for bias when comparing enrichment cultures from different timepoints. As illustrated in our experiments, this setting is ideal for analysing samples representing populations of pathogenic bacteria, where the infecting species of primary interest have generally been previously encountered and sequenced frequently. By leveraging on existing high-resolution genomic pictures of pathogen populations, mSWEEP provides means to address a range of novel biological questions related to within-host variation, transmission and the effect of ecological factors on the microbial diversity present in samples.

Results

Lineage identification

Abundance estimation with mSWEEP is performed in two phases: *reference preparation*, performed once for a given reference collection, and *analysis* of samples (Figure 1). Reference preparation consists of defining a reference sequence database and grouping the sequences according to biological criteria such as sequence type (ST), clonal complexes (CC), or by using a clustering algorithm for bacterial genomes. Grouping related reference sequences is essential in enabling identification of the taxonomic origin of each read[11] and enables abundance estimation when the sequencing reads originate from a sequence having no exact match in the reference database but which is represented by sequences from closely related organisms within the same group (typically bacterial lineage). Consequently, accuracy of the abundance estimates provided by mSWEEP is reliant on an extensive reference database and a biologically meaningful grouping.

We constructed *detection thresholds* for the groups during the reference preparation from the reads used to assemble the reference sequences (Figure 1). We performed repeated *in silico* experiments in each reference group, where we randomly chose one reference sequence from the group, removed it from the reference set, resampled from the sequencing reads used to assemble the removed sequence, and estimated abundances from the resampled reads with mSWEEP. This process was repeated within the group for a predetermined number of iterations, and then repeated in all other groups within the reference

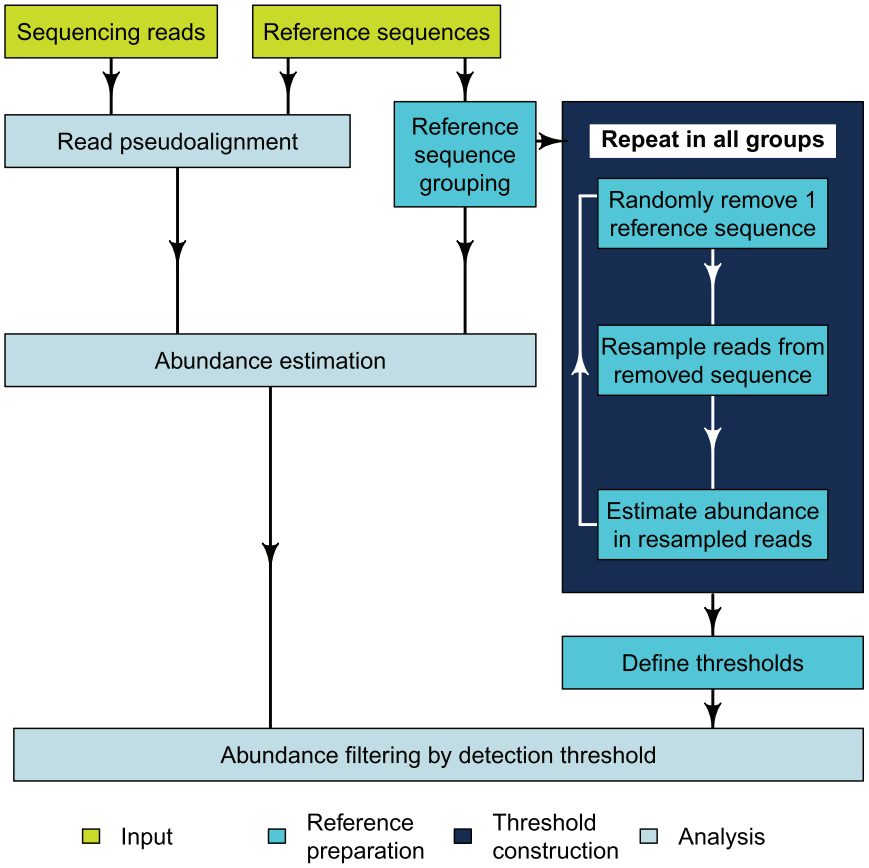

Figure 1. Flowchart of the mSWEEP pipeline describing a typical workflow for relative abundance estimation. The input part refers to the input data, reference preparation to the operations that need to be performed once per set of reference sequences, and analysis contains the steps run for every sample.

set. The detection threshold for a given group was determined by first examining abundance estimates for the given group from the repeated experiments where a different group was the true source (meaning estimates for the given group should ideally be zero), and determining from those estimates a source-specific cutoff point where only a preset number of estimates exceed the cutoff. After determining the source-specific cutoffs in all other groups, the detection threshold of the given group is obtained by taking the maximum of the source-specific cutoffs. The detection thresholds are used to filter the relative abundance estimates by setting the estimates below the cutoff point to zero. Our approach also provides a statistical confidence score for estimates exceeding the detection thresholds with the confidence determined by the number of estimates from the resampled reads allowed to exceed the source-specific cutoffs.

The first phase of analysis is pseudoaligning[14] sequencing reads to the reference sequences. Pseudoalignment produces binary *compatibility vectors* indicating which reference sequences a read pseudoaligns to. Based on the pseudoalignment count to each reference group, we defined the likelihood of a read originating from each of the groups. We assumed that 1) if multiple groups have the same total number of reference sequences, the group with a higher fraction of pseudoalignments is more likely the source for the read, and 2) the likelihood of the read to originate from a group is not dependant on the number of reference sequences in the group. Basing the likelihood on the pseudoalignment counts defines an extension of a probabilistic model that has previously been applied in RNA-sequencing[15,16] and to bacterial data[11]. The extended model utilizes multiple reference sequences from each group as opposed to the previous attempts that rely on selecting a single, best-representative sequence from each of the groups[11] . Our model obtained the relative abundances of the reference groups by considering the generating process for a sample as a pooling of sequencing reads originating from the reference groups according to some unknown proportions, corresponding to a statistical mixture model. We fit the model and inferred the mixing proportions using variational inference[16].

Assigning single-colony isolates to lineage

We compared the performance of the mSWEEP pipeline (using kallisto[14] version 0.45 for pseudoalignment and mSWEEP software version 1.1.0 for abundance estimation) against two existing methods capable of identification beyond the species-level based on leveraging reference sequence collections: metakallisto[13] (version 0.45) and the BIB pipeline[11] (commit hash 2999540). We additionally attempted to compare mSWEEP with ditasic[12] (commit hash 90fee24b), but the comparison proved infeasible

due to ditasic's quadratic scaling in the number of reference sequences — based on running the indexing step in ditasic for one day, indexing the reference data would have taken roughly 90 days and over 30 terabytes of disk space. The main differences between the chosen methods are that metakallisto attempts to identify individual strains based on all available sequences, BIB uses grouped reference sequences with a single representative sequence from each group to assign abundances to the groups, and mSWEEP identifies the presence of lineages by using grouped reference sequences with all the available sequence representatives.

As the reference data, we used bacterial sequence assemblies from four studies[17–20] augmented by single representative sequences from 27 species; a total of 3815 reference sequences. We grouped the sequences in either clonal complexes based on multilocus sequence typing[21], lineages identified with the BAPS clustering algorithm[22], or on the species-level. We removed 504 sequences from all groups represented by more than one sequence to create a dataset where the true group is known but the true sequence is not available to the method pipelines being compared (Table 1). In addition to the test data described in Table 1, we referred to a study sequencing 77 *K. pneumoniae, K. variicola*, and *K. quasipneumonae* isolates from Thailand[23] to assess the accuracy of all methods when the reference sequences and the test samples were not obtained from the same source.

mSWEEP significantly outperformed BIB and metakallisto in cases measuring accuracy of abundance estimates in the true group (Figure 2; $p < 10^{-9}$, in all comparisons, Wilcoxon signed-rank test; median error in all estimates for mSWEEP was 0.00003, for BIB 0.23, and for metakallisto 0.54). When measured by highest estimates in the incorrect groups, mSWEEP outperformed the other two methods in all cases except the *S. epidermidis* 11-group clustering and the *K. pneumoniae* out-of-reference samples (Figure 2; $p < 0.0012$, Wilcoxon signed-rank test; median error in all estimates for mSWEEP was 0.000002, for BIB 0.05, and for metakallisto 0.01). In these two latter cases, mSWEEP and metakallisto performed similarly (Figure 2; $p > 0.10$ when testing for the difference in accuracy in either direction, Wilcoxon signed-rank test). Since metakallisto attempts to identify strains rather than lineages, the observed behaviour is likely a result of the majority of the abundance estimates being spread across strains belonging to the true lineage. We also examined the performance of a modified version of metakallisto, where the estimates for individual sequences within the lineages are pooled together by summing them up. However, this modification did not increase the performance of metakallisto enough to realistically compete with the results from mSWEEP (Extended Data Figure S1[24]).

We additionally compared mSWEEP and BIB by measuring accuracy in classification based on assigning the samples to the lineage with the highest abundance estimate. With this criterion, both methods correctly identified the true clonal complex in all 100 *C. jejuni* and *C. coli* isolates, and in all 81 *S. epidermidis* isolates when using the 3-cluster grouping. In the 11-cluster *S. epidermidis* grouping, mSWEEP correctly identified the true lineage in 78 and BIB in 80 samples. In the 188 *E. coli* and 129 *K. pneumoniae* isolates, mSWEEP identified the lineage correctly in 187 and 126 samples, while BIB correctly identified 184 and 117. The *K. pneumoniae* and *E. coli* isolates that were misidentified by mSWEEP likely contain a sequence type that is missing from the reference, or are mixtures of *K. pneumoniae* and *E. coli* lineages (Extended Data Figures S1a and S1b[24]). Out of the last 61 out-of-reference *K. pneumoniae* samples, mSWEEP identified the true origin in all 61 isolates and BIB in 53.

The least accurate estimates for all methods (measured by the true positives and highest true negatives) were obtained for the 81 *S. epidermidis* isolates when using the second level of the hierarchical BAPS clustering with 11 groups (Figure 2), where none of the three methods reached the level of accuracy observed

Table 1. Reference data used to perform the analyses and to evaluate the performance of mSWEEP, metakallisto, and BIB. Clonal complexes are defined as either single-locus variants from the central sequence type (*Campylobacter jejuni, Campylobacter coli)* or double-locus variants (*Klebsiella pneumoniae* and *Escherichia coli)*. The *Staphylococcus epidermidis* lineages were identified in the original study with the BAPS clustering algorithm.

Grouping	Species	Sequences	Test sequences	Groups	Test groups
Clonal complex	*Campylobacter coli*	120	27	1	1
	Campylobacter jejuni	462	73	13	11
	Escherichia coli	1509	188	132	54
	Klebsiella pneumoniae	1351	129	79	39
Species	*Klebsiella quasipneumoniae*	9	3	1	1
	Klebsiella variicola	12	3	1	1
	Staphylococcus aureus	181		1	
Lineage	*Staphylococcus epidermidis*	143	81	3	3
Species	Multiple species with single sequences	28		28	
	total	3815	504	259	110

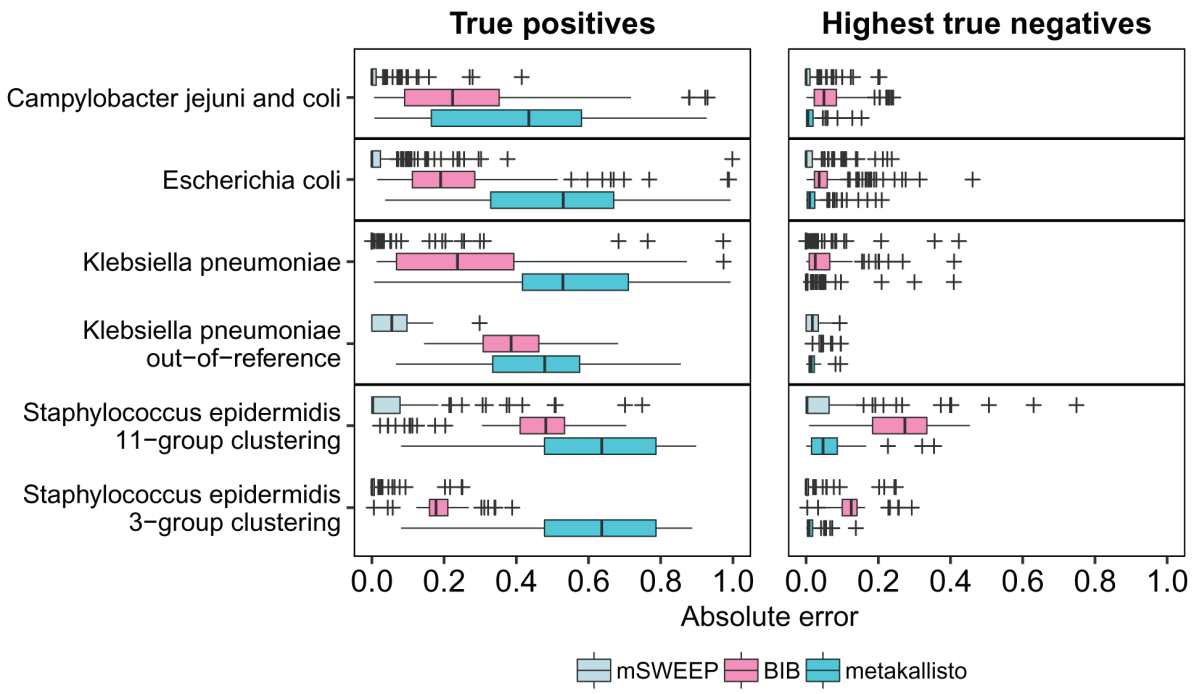

Figure 2. Error of abundance estimates in single-colony isolates (lower is better). True positives refer to the relative abundance estimates in the true lineage (mSWEEP and BIB) or the highest estimate for a strain within the true lineage (metakallisto). Highest true negatives refer to the highest estimate in the incorrect lineages. The absolute error is the difference from an abundance of one (True positives) or from zero (Highest true negatives).

in the other cases. These inaccuracies are explained by the comparably small reference for the *S. epidermidis* population (Table 1), which does not exhibit a clear cluster structure (Extended Data Figure S3a[24]) beyond the coarsest BAPS clustering into three groups. The lack of structure causes the abundance estimates to spread across the new groups defined within each of the three top-level clusters (Extended Data Figure S2c[24]).

We examined the grouping of the reference sequences by producing t-SNE[25] plots of *31*-mer distances estimated with mash[26] (version 2.0) between the reference sequences including the test isolates (Figure 3; Extended Data Figures S2a-c[24]). The *C. jejuni* and *C. coli* reference conforms to the clonal complex grouping while the *S. epidermidis* population only conforms to the coarsest 3-group BAPS clustering. The t-SNE plots correctly place the assemblies into the true groups but the method does not preserve the distances between the points or the clusters[25] and is on its own unsuited to analysing mixed isolate data.

Processing the 504 single-colony isolates with mSWEEP took an average of 23 minutes and 50 seconds per sample, metakallisto an average of 24 minutes and 42 seconds, and BIB an average of 143 minutes and 46 seconds per sample using the same reference data. mSWEEP used a maximum of 79.5 GB RAM (maximum of 24.6 GB counting only the abundance estimation step), metakallisto 108.1 GB, and BIB 31.5 GB. Resource usage was obtained by running each sample separately with a total of eight processor cores available. Reads were pseudoaligned with kallisto[14] (version 0.45) against the test reference of 3311 sequences (obtained from the 3815 reference sequences in Table 1 by removing the 504 single-colony isolates) in 259 groups (mSWEEP and metakallisto), or aligned with bowtie2[27] (version 2.3.5.1) against a reference consisting of randomly selected representative sequences from each of the 259 groups (BIB).

Quantifying synthetic mixtures of single-colony reads

We investigated the performance of mSWEEP in quantifying samples containing multiple lineages of bacteria from the same species by synthetically mixing reads from the single-colony samples. Each mixture sample was set to contain a total of one million reads from three single-colony samples from three lineages, with randomly assigned proportions from the set (0.20, 0.30, 0.50). We used a balanced incomplete block design to ensure that all lineages appear in at least 13 mixture samples, and each single-colony isolate appears at least once, producing 161 *C. jejuni* and *C. coli,* 477 *E. coli*, 584 *K. pneumoniae*, and 100 *S. epidermidis* synthetic mixture samples in total.

Compared to abundance estimates from the single-colony samples, estimates obtained from the synthetic mixture samples show that the presence of sequencing reads from multiple lineages in a synthetic mixture results in an error distribution resembling the one observed in the single-colony samples (Figure 4; Extended Data Figure S4[24]). Estimates from the synthetic *S. epidermidis* mixture samples using the 11-group split produce an error distribution that differs from the single-colony error distribution more than that observed with the other groupings.

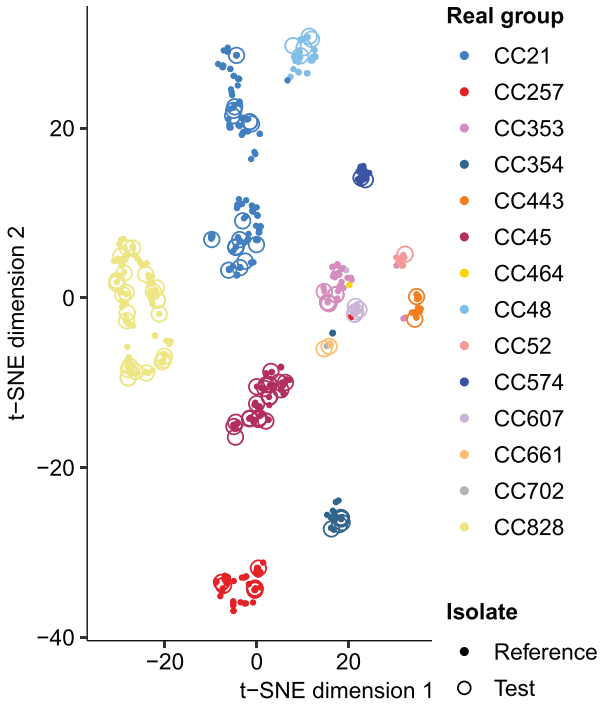

Figure 3. *C. jejuni* and *C. coli* reference 31-mer embedding. t-SNE embedding of the 31-mer distances between the reference isolates shows that the reference population conforms relatively well to the clonal complex grouping. The test cases, indicated by circles, are all correctly identified by mSWEEP and t-SNE also places them within or near the true source group.

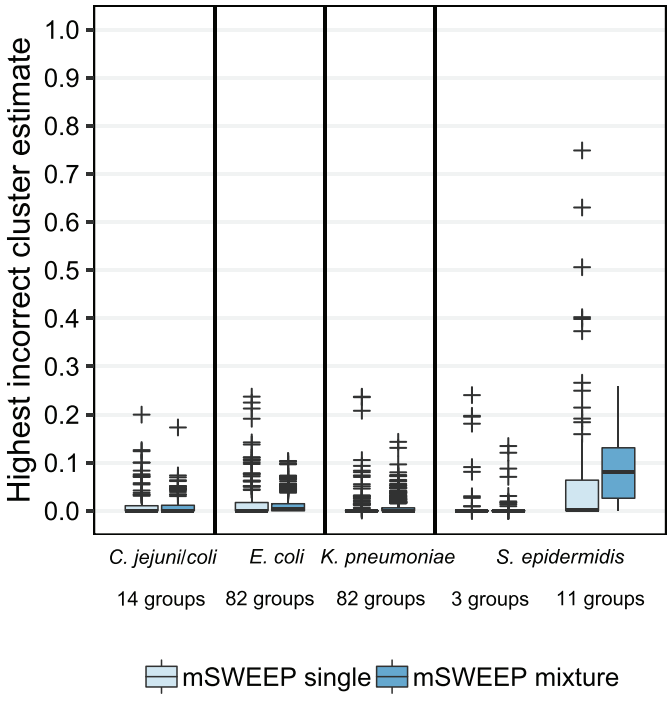

Figure 4. False positives in single-colony samples versus synthetic mixtures. Abundance estimates from synthetic mixtures of three lineages do not result in higher number of false positive estimates when compared to estimates from the single-colony samples, as measured by the largest estimate for a lineage that does not contribute any sequencing reads. The only exception is the *S. epidermidis* 11-cluster case which is not accurately identified in neither the synthetic mixtures nor the single-colony samples.

Comparing the empirical distributions of the errors from the synthetic mixtures and the single-colony isolates (Extended Data Figures S3 and S4[24]) shows that for estimates exceeding a threshold of 0.016, the accuracy of estimates from the mixture samples stochastically dominates the accuracy observed in the single-colony samples, except in the *S. epidermidis* 11-cluster case where stochastic dominance is observed only above a threshold of 0.17. Stochastic dominance establishes a partial ordering between two random variables and, in this case, implies that estimates from the mixture samples are more accurate (in a probabilistic sense) than estimates from the single-colony samples when the estimates are large enough. In the *S. epidermidis* 11-cluster case we do not establish the mixture estimates as more accurate since the distribution (Extended Data Figure S4[24]) and the observed threshold differ considerably from the other cases.

The results indicate that above this relatively low background noise level of 0.016, quantifying mixture samples is not expected to produce more false positive results than would be obtained from single-colony samples. In the synthetic mixtures, the observed background noise level corresponds to sequencing depths of around 0.30x (*E. coli* and *K. pneumoniae*) and 0.65x (*S. epidermidis*), which provides the bare minimum sequencing depth required to distinguish between the lineages of each species in samples with similar read lengths and sequencing depth. This justifies simplifying the problem of determining the detection thresholds accompanying mSWEEP, which provide a threshold for reliable detection of the reference groups in mixture samples, to determining the thresholds based on the single-colony isolates. Due to the requirement that the abundance estimates must be large enough for this assumption to hold, we incorporate the threshold observed in comparing the estimates into the detection thresholds by using it as the minimum threshold regardless of the results from the resampling procedure.

We also evaluated the performance of mSWEEP with synthetic mixtures containing more complex strain compositions. Namely, we produced 87 synthetic mixture samples each consisting of 10 *E. coli* strains from 10 different lineages mixed together at varying relative abundances (0.50, 0.25, 0.125, 0.0625, 0.0312, 0.0156, 0.0078, 0.0039, 0.0020, 0.001). The total number of reads in each sample was set to correspond to sequencing a single typical *E. coli* genome at 100× sequencing depth (around 5 million 100bp reads), resulting in sequencing depths ranging from 50× to 0.10× for the 10 different strains. Overall the design is intended to mimic performing plate sweeps with a similar amount of sequencing resources that would be available for the same number of colony picks.

The boxplot displays the relative error in the relative abundance estimates from mSWEEP compared against the true values. Error of >0% (horizontal axis) denotes estimates from mSWEEP exceeding the true value, while error of <0% denotes estimates lower than the true value. The dashed gray line corresponds to 0% error. The rows (vertical axis) separate the estimates by their approximate sequencing depth, with each sample contributing one value (estimate for one lineage) to each row.

The results from the complex synthetic mixtures (Figure 5) show that mSWEEP accurately recovers the relative abundances for the dominant lineages (50× to 12.50× sequencing depths) and

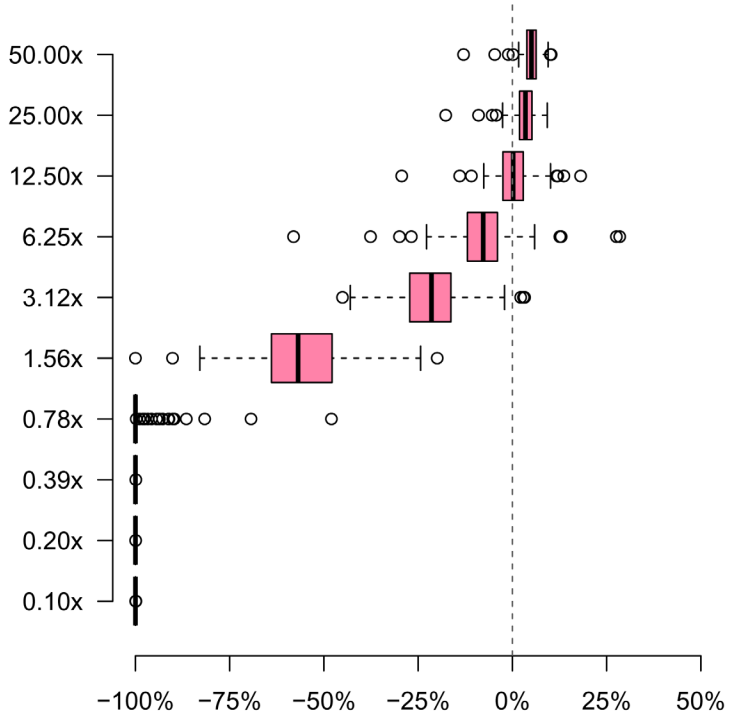

Figure 5. Relative error in 87 complex synthetic mixtures comprising 10 E.coli lineages at varying sequencing depths.

identifies the mid-range lineages (6.25× to 1.56×) correctly but underestimates their relative abundance. As a result, the relative abundances of the two most dominant lineages are slightly overestimated because, taken together, the relative abundance estimates must sum up to 1. As for the lineages with < 1× sequencing depth, their relative abundances are not recovered at all, likely because the differences between the *E. coli* lineages are not pronounced enough to separate the strains at this sequencing depth. Coincidentally, the observed accuracy cut-off point roughly corresponds to the previously established background noise level of 0.016, or approximately 1.6× sequencing depth.

Results from plate sweeps

We applied the mSWEEP pipeline to three datasets containing multiple lineages of the same species: 116 samples from *C. coli* and *C. jejuni*, 96 paired samples from *E. coli*, and 179 samples from *K. pneumoniae*. The *E. coli* samples were obtained from MacConkey plate sweeps from a cohort study of 48 Vietnamese children during a diarrhoeal episode (48 samples), and when healthy (48 samples), purposefully expecting multiple lineages in each sweep. Conversely, the *C. coli*/*C. jejuni* and *K. pneumoniae* datasets were presumed pure cultures but flagged during downstream analysis as mixed. In all three experiments, we applied the detection threshold procedure (described in more detail in the Methods section) to filter the resulting abundance estimates. We used two thresholds, corresponding to confidence scores of 0.99 and 0.90, from now on referred to as filtering by 0.99 or 0.90 confidence thresholds.

Population structure of commensal Escherichia coli from Vietnamese children

The most abundant sequence type complex identified in over half the samples (diarrhoeal and control samples) was CC10 (Figure 6; Extended Data Table S1[28]). Notably, 95% of the samples (46/48 Diarrheal and 45/48 Healthy) harboured multiple antimicrobial resistance genes (identified using the ARIBA software[29]) that belonged to three or more classes of drugs (Extended Data Figure S6[24]), which we defined as multi-drug resistance[30]. One sample was found to contain the plasmid associated resistance gene MCR-1, which confers resistance to colistin, a last line antimicrobial drug[31]. We found no significant difference in the antimicrobial resistance gene profile between the healthy and diarrhoeal samples (Two tailed, Fisher's exact test p=0.5). Extended Data Table S2[32] details how many samples harboured antimicrobial resistance genes in each antimicrobial drug class; the full antimicrobial resistance gene data can be found in Extended Data Table S3[33].

We additionally examined differences between the community composition in the healthy and diarrhoeal samples based on both the distribution of the relative abundance estimates (alpha diversity), and changes in the identified strains (beta diversity). The alpha diversity, measured by Shannon entropy (Extended Data Figure S7[24]), showed no significant differences between the two paired samples ($p > 0.90$, Wilcoxon signed-rank test; median Shannon entropy in the diarrhoeal samples was 0.60, and in the healthy samples 0.59). However, we found significant shifts in lineage composition (see Figure 6) when comparing the beta diversity, measured by Bray-Curtis dissimilarity, between the two samples ($p < 0.005$, multivariate-ANOVA). Tests were performed on relative abundance estimates filtered by both 0.99 and 0.90 confidence thresholds.

Co-occurrence patterns in Campylobacter lineages

The network diagram (Figure 7a) shows ST-clonal complex (CC) (nodes) of the isolate genomes with the thickness of edges representing the number of times that isolates from these CCs

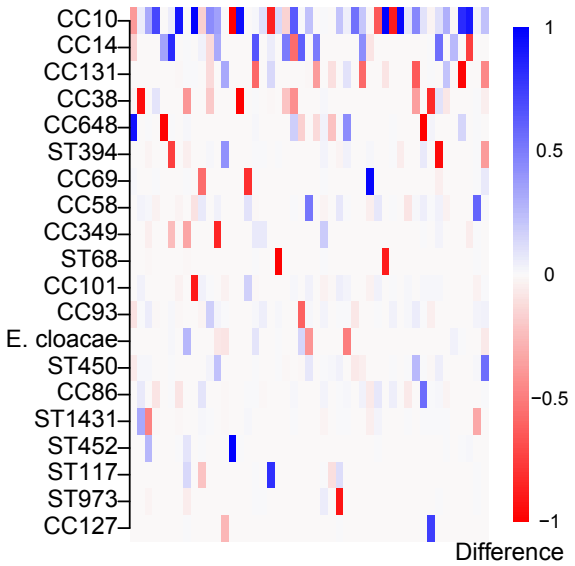

Figure 6. Difference in *Escherichia coli* clonal complex (CC) and sequence type (ST) lineage abundances during and after diarrhoea. The plot displays the differences in unfiltered relative abundance estimates before and after treatment in 20 most common (defined by the sum of relative abundances; blue denotes increase, red decrease) *E. coli* reference lineages or other species across all 47 paired samples represented in the columns.

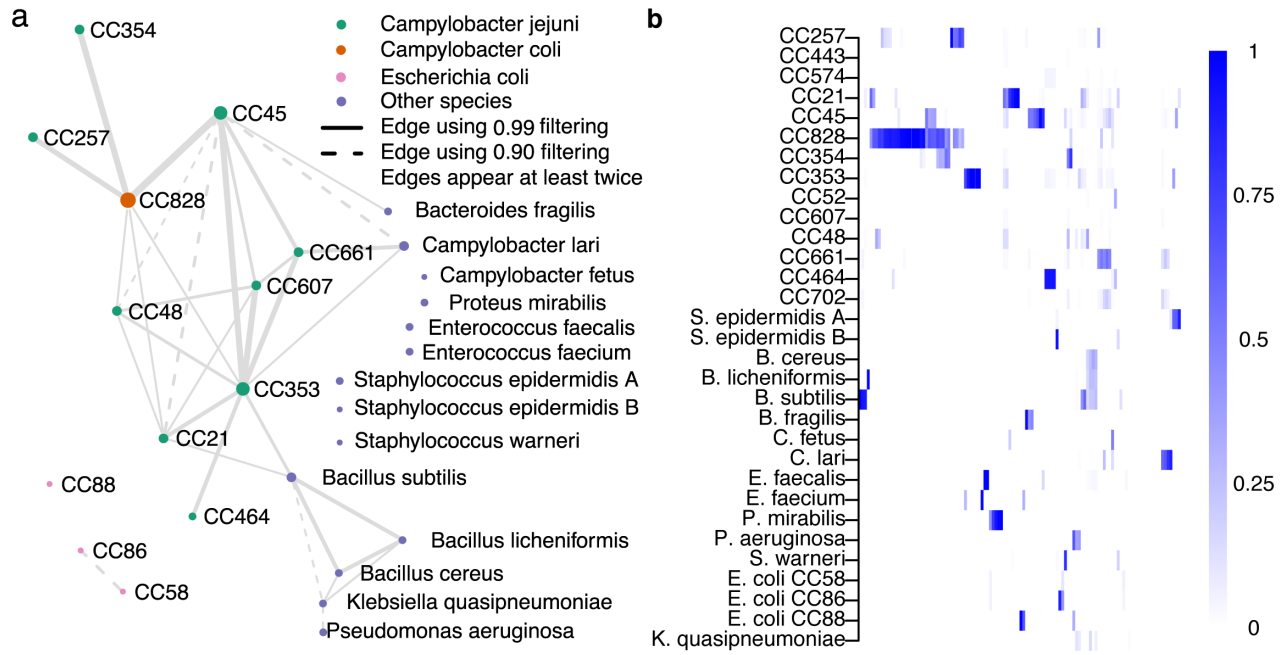

Figure 7. *C. jejuni* and *C. coli* clonal complex (CC) coexistence in 116 samples. The coexistence network in panel **a** was constructed from relative abundance estimates filtered by detection thresholds constructed using a confidence score of either 0.90 (dashed edges) or 0.99 (solid edges). An edge between two groups represents coexistence in at least two samples with the chosen threshold. Edge size is proportional to the number of times the joined nodes were observed together, and node size to the total times the group was detected. Panel **b** visualizes the unfiltered relative abundance estimates in the same reference groups (rows) as in panel a, across 116 samples (columns).

are found together in a single plate sweep sample, and the size of the node the total number of observations. The overall amount of co-occurrence between CCs (Figure 7b) provides basic information about the frequency that CCs are found together in natural populations. *C. jejuni* CCs 45, 661, 607, 353, 48, and *C. coli* CC828 are all found in samples with 4 or more other CCs and there is evidence that isolates from some CC's cohabit with other species including *Campylobacter lari* and *Bacillus subtilis*. While the sample set in this study was deliberately selected to include mixed isolate samples, quantifying the co-occurrence of species and lineages can provide information about different ecologies or lineage interactions, particularly when CCs are known to have varied sources, such as different hosts.

There is some preliminary evidence that common clinical lineages CC45 and CC21 are rarely found together in a single sample (plate sweep) while other lineages, such as the chicken associated CC353, are frequently isolated from samples containing multiple strains. From an evolutionary perspective, it is unlikely that closely related strains can stably occupy identical niches because competition would be expected to lead one to prevail. The results demonstrate co-occurrence of strains within individual host animals and multi-strain infections in humans and provide information about the complex ecology of co-occurring interacting species that leads to the observed community structure in a given sample.

Multi-drug resistant Klebsiella pneumoniae coexist with other lineages

The coexistence network (Figure 8) and the sample-lineage heatmap (Extended Data Figure S8[24]) for the 179 human clinical samples of *K. pneumoniae* demonstrates common co-occurrence of *K. pneumoniae* with a wide variety of *E. coli* lineages, as well as occasional co-occurrence with *Acinetobacter baumanii* and other species. Since both *E. coli* and *A. baumanii* grow on the media used for culture of *K. pneumoniae*, frequent co-occurrence in samples selected for their diversity is expected. Clonal complexes of *K. pneumoniae* centered on sequence types associated with high levels of multi-drug resistance (e.g. ST258, ST147 and ST101) were frequently observed co-existing with a variety of other *K. pneumoniae* lineages as well as with each other, and with other important Gram-negative pathogens. Developing a deeper understanding of these community structures and interactions will be critical for monitoring horizontal transfer of antimicrobial resistance genes between taxa.

Discussion

Metagenomics using high-throughput sequencing has become a common approach when investigating the bacterial composition of different environments or changes introduced by intervention, such as in the human gut microbiome. In most epidemiological applications, the relevant target organisms are culturable using established media which offers a clear advantage to obtaining

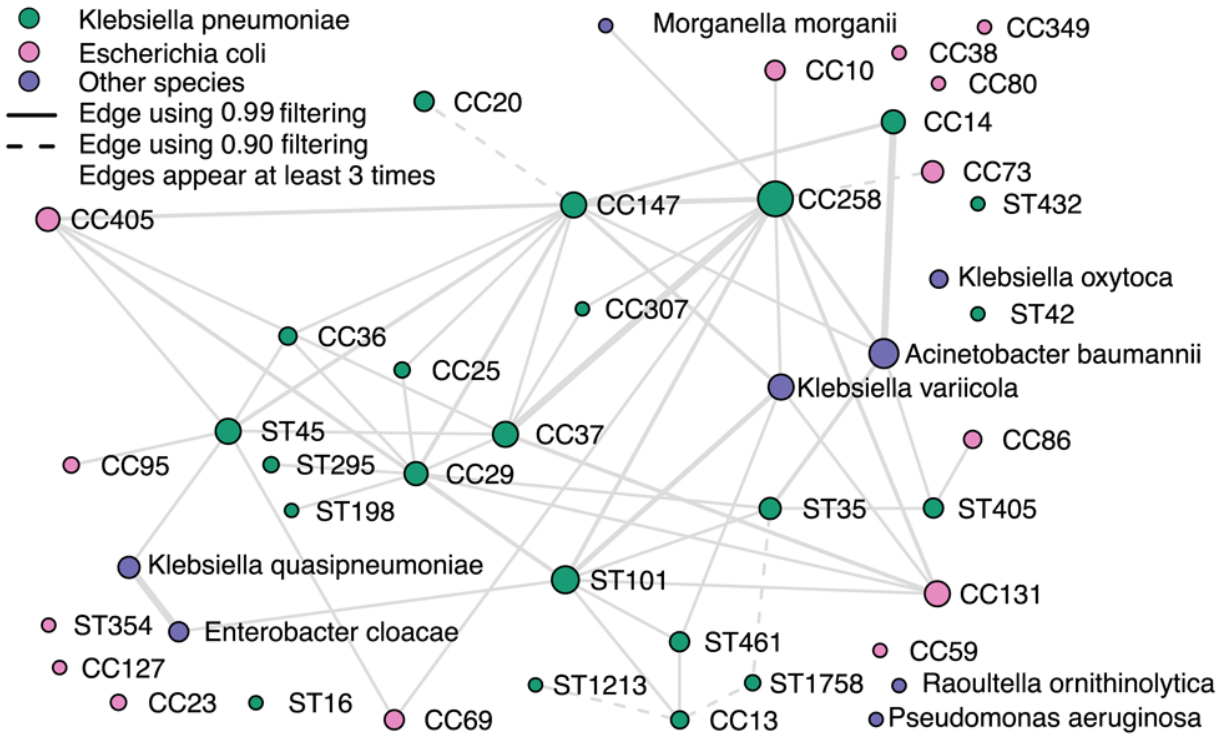

Figure 8. *K. pneumoniae* lineage coexistence in 179 mixture samples. The coexistence network was constructed from relative abundance estimates remaining after filtering by the detection thresholds. Visible edges denote coexistence in at least three samples. Dashed edges represent coexistence when using detection thresholds corresponding to the 0.90 confidence score, and solid edges using the 0.99 detection threshold. Node sizes are proportional to the number of times the lineage was observed; edge sizes to the number of times coexistence was established.

high sequencing depths in a cost-effective manner. We developed the mSWEEP pipeline to enable high-resolution inference of the lineages present in plate sweeps of enrichment cultures. mSWEEP can be used to infer the detailed population structure of a single species, or the diverse populations of bacteria typically encountered in clinical and public health settings where standard culturing media is routinely used to isolate epidemiologically relevant organisms. This pipeline also estimates the relative abundance of lineages and provides means to construct reliability cut-offs with accompanying confidence scores. Since the cut-offs are constructed before analysing any sequencing data by synthetically mimicking the properties of the in vitro data, the cut-offs and confidence scores can be used to assess the necessary sequencing depth to identify rare and low abundance lineages with confidence as well as the total number of reads required. mSWEEP was designed to have a minimal execution time using the latest advances in RNA-seq analysis and its maximum memory footprint is determined by the pseudoalignment algorithm. We demonstrated significant improvements in accuracy over the previous state-of-the-art method in our experiments.

mSWEEP provides considerable power for improving our understanding of infection by recovering a true representation of bacteria in a complex sample. For example, genotyping studies have shown that *C. jejuni* and *C. coli* colonizing the primary host (birds and mammals) form clusters of related isolates that are host-associated[34], which can be used to identify the reservoir for human infection[35]. However, multiple organisms can be isolated from the same sources[36,37]. The co-occurrence of different organisms could be a snapshot in time of a wider process of lineage succession[38] in which the resident microbiota might resist new colonizations or be displaced by recently acquired bacteria[39,40]. Further, we suggest we may be able to infer complex interactions between organisms that occupy different microniches[41] and are not in direct competition[42,43] by analysing their co-occurrence. Therefore, this approach provides a means to investigate the nature of polymicrobial infections to improve our understanding of the spread of a specific organism between hosts and transmission to humans in addition to enabling characterization of physical and temporal variation in the distribution of lineages among multi-strain samples.

Because of limitations in the initial culture and DNA isolation processes, we can only infer relative (not absolute) abundances and spike-in methods[44] must be used if an estimate of the absolute abundances is desired. However, only inferring relative abundances is not a significant limitation as the absolute abundances of target organisms are also subject to large biological and technical variation[45]. Memory requirements for large reference collections or simultaneous analysis of multiple samples necessitate a dedicated computer cluster to run the analysis pipeline, but even for very large reference collections the resource

usage is still at the level available at most bioinformatics centres. Alternatively, the reference sequences can be modified to include only the directly relevant species, which makes mSWEEP widely applicable to biologists; or traditional alignment algorithms employed to trade decreased memory usage for increased runtime.

As with any method intended to identify sequence variation, the target species need to be relatively well known to allow building of sufficiently informative reference databases. Similarly, to allow for sensible and easily interpretable inferences, the biological clustering of the reference database should be based on well-established biological entities, such as multilocus sequence types (STs) or clonal complexes (CCs) which are frequently employed as labels of lineages. These limitations suggest that mSWEEP is most relevant for extensively studied bacteria, such as pathogens, and has only limited applications when working with samples containing high numbers of novel or uncharacterized species. Some of these limitations may be overcome by using gene flow network based approaches to perform the biological clustering, better capturing the population structure in species strongly shaped by horizontal gene transfer or ecological barriers[46,47], but further study in this area is needed. However, following the introduction of mSWEEP, a separate computational tool has been developed to assess the suitability of the reference sequences to specific samples and to identify cases where mSWEEP struggles due to the presence of reads originating from novel genome sequences (https://github.com/harry-thorpe/demix_check). Consequently mSWEEP can also be used to estimate the quality of the reference collection and to discard contaminated or mis-identified genomes.

Strain identification from metagenomic data has been recently suggested by the StrainPhlAn method[48]. mSWEEP, and similar methods, are complementary to StrainPhlAn as these methods analyse similar data but from different directions. mSWEEP assigns strains present in the sample to biologically established genetically separated lineages and estimates the relative abundance of these, whereas StrainPhlAn infers SNPs and phylogenetic relations within the whole sample. Given the flexibility and generality of the mSWEEP approach, we anticipate this pipeline will pave the way for numerous novel applications of plate sweep metagenomics in many fields of microbiology.

Conclusions

mSWEEP represents a novel means to quantify the composition of bacterial communities beyond the resolution offered by bacterial identification methods based on 16S ribosomal RNA gene sequencing or whole-genome shotgun metagenomics. We have demonstrated significant improvements in accuracy over similar methods, and novel co-existence analyses using plate sweeps of enrichment cultures of the human pathogens *Campylobacter jejuni*, *Campylobacter coli*, *Escherichia coli, Klebsiella pneumoniae* and *Staphylococcus epidermidis*. We expect that mSWEEP will find use in similar studies of bacterial pathogens, where high-resolution inference is required, ample reference collections for the species of interest are available, and the plate sweep metagenomic approach can be applied in-depth at a fraction of the current cost of single-colony sequencing.

Methods

Reference construction

The reference sequences (Table 1, Extended Data Table S4[49]) are the genomic assemblies of a number of strains or species that represent the organisms of interest in a sample. We used a collection of assemblies from four studies[17–20] augmented with the genomes of a representative strain from 27 species that were identified in the real mixture data by MetaPhlAn[50].

Grouping the reference sequences

We used multilocus sequence typing (MLST) of the *C. coli, C. jejuni, E. coli* and *K. pneumoniae* reference sequences to group them into clonal complexes defined by the allelic profile of a central sequence type, and all other sequence types that vary in at most a single MLST locus (*C. coli* and *C. jejuni*) or in at most two loci (*E. coli* and *K. pneumoniae*). The *K. pneumoniae* reference contained sequences belonging to *K. variicola*, *K. quasipneumoniae*, and *K. quasivariicola* which we assigned to three groups defined by the three species. We similarly treated the 181 *S. aureus* contained in the *S. epidermidis* study as a single group, and split the 143 *S. epidermidis* sequences using the first and second levels of the hierarchical clustering produced by the hierBAPS[22] software (version 6.0). The complete grouping is provided in Extended Data Table S4[49].

Pseudoalignment

We used kallisto[14] (version 0.45) with default settings to perform pseudoalignment. Pseudoalignment produces binary compatibility vectors which indicate whether the read pseudoaligns to a reference sequence or not. In our model, we sum the pseudoalignment counts within each reference group and thus consider the observations of N reads $r_n = (r_{n,1}, \ldots, r_{n,k}, \ldots, r_{n,K})$, $n = 1, \ldots, N$, $k = 1, \ldots, K$ as containing only the information about the number of pseudoalignments $r_{n,k}$ within each of the K groups.

Abundance estimation model

We assume that the reads r_n are conditionally independent given the mixing proportions of the groups $\theta = (\theta_1, \ldots, \theta_K)$, and augment the model with the latent indicator variables $I = I_1, \ldots, I_N$ which denote the true source group of each read. The joint distribution of the collection of reads $R = r_1, \ldots, r_N$, the indicator variables $I = I_1, \ldots, I_N$ for the source group, and the mixing proportions of the groups $\theta = (\theta_1, \ldots, \theta_K)$ is defined as

$$p(R, I, \theta) = p(\theta) \prod_{n=1}^{N} \prod_{k=1}^{K} p\left(r_n \middle| I_n = k\right) p\left(I_n = k \middle| \theta\right). \quad (1)$$

The formulation in Equation (1) corresponds to a standard mixture model with observations r_n, categorically distributed latent variables I_n, event probability parameters θ, and the likelihood $p(r_n \mid I_n = k)$ of observing the full pseudoalignment count vector r_n given that the group k is the true source.

Likelihood

The likelihood $p(r_n \mid I_n = k)$ needs to be defined carefully in order to satisfy the goals of invariance to group identity and size, and monotonicity with increasing pseudoalignment counts

within a group. Given the vector r_n, we define the likelihood $p(r_n \mid I_n = k)$ of observing the whole vector r_n assuming that k is the true group in three parts depending on the number of reference sequences M_k in the group k and the pseudoalignment count $r_{n,k}$ in group k as

$$p\left(r_n \middle| I_n = k\right) = 0.01, \text{when } r_{n,k} = 0, \text{and} \tag{2}$$

$$p\left(r_n \middle| I_n = k\right) = 0.99, \text{when } M_k = 1 \text{ and } r_{n,k} > 0, \text{or} \tag{3}$$

$$p(r_n \mid I_n = k) = \frac{f\left(r_{n,k}, k\right)}{Z\left(r_n\right)} 0.99, \text{when } M_k > 1 \text{ and } r_{n,k} > 0, \tag{4}$$

with

$$\begin{gathered} f\left(r_{n,k}, k\right) \propto \binom{M_k}{r_{n,k}} \frac{B\left(\alpha_k + r_{n,k}, M_k - r_{n,k} + \beta_k\right)}{B\left(\alpha_k + M_k, \beta_k\right)}, \text{ and} \\ Z\left(r_n\right) = \prod_{k=1}^{K} \prod_{j=1}^{M_k} \frac{\alpha_k + r_{n,k} + (j-1)}{\alpha_k + \beta_k + 2M_k + (j-1)} = \prod_{k=1}^{K} \frac{\Gamma\left(\alpha_k + r_{n,k} + M_k\right)\Gamma\left(\alpha_k + \beta_k + 2M_k\right)}{\Gamma\left(\alpha_k + r_{n,k}\right)\Gamma\left(\alpha_k + \beta_k + 3M_k\right)} \end{gathered} \tag{5}$$

where $B(a, b)$ is the beta function and $a_k, \beta_k > 0$ are hyperparameters for the group k.

Equation (2) and Equation (3) zero inflate the model by an amount roughly corresponding to the error rate in both the sequencing data and the reference sequences. The denominator $Z(r_n)$ in Equation (4) and Equation (5) is a normalizing constant for the likelihood and arises from normalizing $f(r_{n,k}, k)$ over r_n. The derivation follows from using the identity $B(a+1,b) = B(a,b)\frac{a}{a+b}$ for the beta function to express each $f(r_{n,k}, k)$, $k = \{1, ..., K : M_k > 1\}$ as a product of the probability mass function of a beta-binomial distribution with parameters $(\alpha_k + M_k, \beta_k)$, and 2) the term $\prod_{j=1}^{M_k} \frac{\alpha_k + r_{nk} + (j-1)}{\alpha_k + \beta_k + 2M_k + (j-1)}$ which leads to the form that $f(r_{n,k}, k)$ has in Equation (5). Then, $Z(r_n)$ is obtained by considering normalizing over the full vector $f(r_n) = (f(r_{n,1}, 1), ..., f(r_{n,k}, k), ..., f(r_{n,K}, K))$, where $k = \{1,..., K : M_k > 1\}$.

The formulation of $f(r_{n,k}, k)$ in Equation (4) and Equation (5) intuitively arises when the probability mass function of a beta-binomial random variable with hyperparameters (α_k, β_k) is multiplied by the factor $\frac{B(\alpha_k, \beta_k)}{B(\alpha_k + M_k, \beta_k)}$. This factor is the inverse of the value of the probability mass function when the observed value is equal to the total number of sequences M_k in the group, meaning in our context a read which is compatible with all reference sequences in a group. Formulating the likelihood in this manner (Equation (4) and Equation (5)) causes groups where all sequences in the group are compatible with the read to have equal likelihoods. Compared to a model assuming independence between the groups, this formulation reduces the effect of the likelihoods being flattened in groups with large numbers of assigned reference sequences when compared to small groups, which is necessary to compare groups that differ greatly in size.

Reads with identical pseudoalignment count vectors r_n have the same likelihood and can be assigned into equivalence classes defined by the count of compatible sequences in each group. This enables a computational optimization where the likelihoods need only be calculated for the observed equivalence classes and then multiplied by the total number of times each equivalence class was observed.

Model hyperparameters

Instead of using the parametrization (α_k, β_k) in Equation (5), we use a reparameterization where

$$\pi_k = \frac{\alpha_k}{\alpha_k + \beta_k}, \phi_k = \frac{1}{\alpha_k + \beta_k}. \tag{6}$$

In this parametrization, the first parameter $\pi_k \in (0, 1)$ corresponds to the mean of the beta distribution that has been compounded with a binomial distribution to obtain the beta-binomial-derived component in Equation (5), and the second parameter $\phi_k > 0$ represents a measure of variation in the success probability of each observation[51].

We constrain the mean success rate π_k in Equation (6) to $\pi_k \in (0.5, 1)$, which produces beta-binomial distributions with an increasing probability mass function[52] in the number of compatible sequences $r_{n,k}$, which leads to the definition in Equation (5) having the same property. Increasing probability mass functions fulfil our requirement for the likelihood that of two equally sized groups with different number of compatible reference sequences, the one with more compatible sequences is always a better candidate for being the true source. The values of the parameters (π_k, ϕ_k) are set to $\pi_k = 0.65, \phi_k^{-1} = 1 - \pi_k + 0.01|M_k|^{-1}$ to capture the variance in the alignment count distributions and perform well across the set of experiments presented.

Inference

We perform inference over the mixing proportions θ of the different groups using fast collapsed variational inference[53]. The method collapses the mixing proportions θ and uses natural gradients to optimise an approximation to the posterior distribution over the indicator variables I_n, assuming the distribution factorises over θ and I_n. The same variational Bayesian method was also used in BitSeqVB[16] to obtain transcript expression levels and has been applied to estimate mixing proportions in bacterial sequencing data in BIB[11] using a different likelihood. The prior distribution on the mixing proportions θ is set to Dirichlet $(\delta\alpha, ..., \delta\alpha)$ with $\delta\alpha = 1$. The same prior was also used by BIB. Since reads originating from the same equivalence class have the same likelihood, variational inference will yield identical posterior inferences for them. This allows us to perform the inference on the smaller number of equivalence classes rather than all reads, leading to faster inference.

Detection thresholds

Detection thresholds define a means to quantify reliable identification of the reference groups through constructing a minimum relative abundance threshold on the groups. Abundance estimates that fall under the threshold are considered unreliable and set to zero. To obtain the detection thresholds (Figure 1), we generate 100 samples from each reference group within a species by resampling one million sequencing reads from a randomly chosen reference sequence for that group, roughly matching the

number of reads in our study samples, from the reads used to assemble the chosen sequence. Only reads corresponding to one sequence are used in each new sample. Reads included in the new samples are sampled with replacement with each read having the same probability of being included. The sequences used for resampling were chosen at random such that the number of reference sequences from each group corresponds to the square root of the total size of the group. Each group is represented at least once, except for groups which contain only one reference sequence where we apply the maximum detection threshold observed for other groups of the same species. Similarly, species that are represented in the reference by a single group were not resampled from, and the detection threshold was instead fixed at 0.05. After resampling, the new samples are put through pseudoalignment and mSWEEP abundance estimation without including the sequences used in resampling as pseudoalignment reference sequences.

In defining the detection thresholds, the relative abundance estimates obtained from the resampled sequencing reads are represented by $\widehat{\theta}_{n,i,j}$, where $n = 1, ..., N$ (in our examples we chose $N = 100$) indicates the new samples resampled from the reference group $i = 1, ..., K$. The third index $j = 1, ..., K$ denotes the reference group that the abundance estimate was obtained for. We first define source-specific thresholds $q_{i,j}$ that give a threshold on the reference groups j assuming that the true group i in the sample is known. The source-specific threshold $q_{i,j}$ on group $j \neq i$ is defined by ordering the relative abundance estimates for the cluster j, $\widehat{\theta}_{n,i,j}$, in an ascending order in n, and determining the cutoff point $q_{i,j}$ where $m,m \in \{1, ..., N\}$, relative abundance estimates fall below the cutoff. Using the source-specific thresholds $q_{i,j}$, we define the detection threshold q_i on group i as $q_i = max\left\{max\left\{j : q_{i,j}\right\}, \epsilon\right\}$, where ϵ is the constant minimum threshold for a specific grouping of the sequences within a species that is observed when comparing the empirical cumulative distribution functions in Extended Data Figure S6[24]. We recommend that ϵ be determined for new species by a synthetic mixing procedure similar to what was used to compare the accuracy of mixture estimates to their single-colony counterparts in Figure 4.

Based on the selected value of m, we may further define a statistical confidence score for the $M = N - m + 1$ remaining abundance estimates that exceed the detection threshold q_i as

$$1 - \frac{(N-m)+1}{N+1} = \frac{m}{N+1}, \qquad (7)$$

which corresponds roughly to the fraction of resampled samples that exceed the threshold obtained with the selected value of m. Using values of m closer to the number of samples N in constructing the detection thresholds will result in stricter thresholds and thus higher confidence in the abundance estimates that exceed the threshold. The results reported in our experiments include thresholds constructed with $m = 100$ and $m = 90$, corresponding to confidence scores (Equation 7) of approximately 0.99 and 0.90, respectively.

Implementation

The mSWEEP software provides a C++ implementation of the abundance estimation part of the pipeline described in this manuscript. After pseudoaligning the sequencing reads, mSWEEP can be called from the command line to construct the abundance estimation model and infer the relative abundances of the reference lineages as described above in the Abundance estimation model, Likelihood, Model hyperparameters, and Inference sections. The mSWEEP pipeline consists of running both the pseudoalignment and the mSWEEP abundance estimation software.

Operation

Precompiled binaries and the source code for the mSWEEP software are available in GitHub. Compiling mSWEEP from source requires a compiler with full support for the C++11 standard (for example clang version 3.3 or later, or GCC version 4.8.1 or later) and the build process utility CMake (version 2.8.12 or later). The mSWEEP software itself does not have additional external dependencies. Prospective users of the mSWEEP pipeline will also need to install kallisto[14] for pseudoalignment, and possibly construct a set of scripts tailored to their reference data should they wish to take advantage of the detection threshold approach. The GitHub repository includes usage information, a general pipeline for abundance estimation with mSWEEP, and a toy dataset for an example workflow.

A typical workflow with mSWEEP begins by indexing the set of reference sequences (reference.fasta below) for pseudoalignment (here using kallisto version 0.45)

```
kallisto -i kallisto_index reference.fasta
```

The reference sequences need to be indexed only once and the same index can be used multiple times.

Analysing sequencing data (below the paired-end reads in two files: reads_1.fastq.gz and reads_2.fastq.gz) is done by first using kallisto to pseudoalign the reads

```
kallisto pseudo -i kallisto_index -o pseudoalignments reads_1.fastq.gz reads_2.fastq.gz
```

and then running the mSWEEP software (here using version 1.1.0) to produce the relative_abundances.txt file containing the relative abundance estimates.

```
mSWEEP -f pseudoalignments -i lineages.txt -o relative_abundances.txt
```

The lineages.txt file defines the reference lineages, with each line in the file containing the name of the lineage the corresponding sequence in the reference.fasta file has been assigned to. Entries in the lineages.txt file must be in the same order as the sequences are in the reference.fasta file.

E. coli plate sweeps from Vietnamese children

In Ho Chi Minh City, 750 children were recruited into a diarrhoeal cohort study and followed for 2 years. Stool samples were collected at routine sampling points and when the children had an episode of diarrhoea. All stool samples were cultured to identify pathogens and onto MacConkey plates to isolate *E. coli* and other *Enterobacteriaceae*. The MacConkey plates were scraped and stored in 20% glycerol at -80°C. The frozen plate sweeps from 48 diarrhoea episodes, paired with 48 asymptomatic samples (96 in total), were revived on MacConkey media; plates were scraped and total genomic DNA was extracted using the Wizard genomic DNA purification kit (Promega, USA). The extracted DNA was sequenced using the Illumina HiSeq platform using the method described elsewhere[54]. Antimicrobial resistance genes were detected using the ARIBA software[27]. The raw sequence data can be found in the ENA under the accession numbers detailed in Extended Data Table S5.

Ethics approval and consent to participate

Ethical approval was required for the cohort study contributing the *E. coli* organisms. Approvals were provided by the Oxford University Tropical Research Ethics Committee (OxTREC approval 1058–13) as well as from the local partners (Institutional Review Board at the Hospital for Tropical Diseases and Hung Vuong Hospital). Written informed consent was obtained from the parent or guardian of all children for participation in the study.

Data availability

Underlying data

Figshare: mSWEEP_reference_v1-0-0.tgz, https://doi.org/10.6084/m9.figshare.8222636.v2[55]. This project contains the reference sequences and the grouping used in producing the results.

Accession numbers for the reference data can be found in Extended Data Table S4[49]. Accession numbers for the 96 Vietnamese *E. coli* plate sweeps are available in Extended Data Table S5[56]. Accession numbers for the *K. pneumoniae* mixture samples and 39 *Campylobacter* mixture samples are available in Extended Data Table S6[57]. The remaining 77 *Campylobacter* mixture samples have been submitted to Figshare:

- campylobacter_mixtures_1.tgz, https://doi.org/10.6084/m9.figshare.6445136.v1[58].
- campylobacter_mixtures_2.tgz, https://doi.org/10.6084/m9.figshare.6445190.v1[59].

The synthetic *E. coli* mixture samples, and associated metadata, that were the basis for the results presented in Figure 5 are available have been submitted to Zenodo:

- mSWEEP_revision_mixture_samples_v1-0-0.tar, https://doi.org/10.5281/zenodo.553571360.

Extended data

All *extended data* files have been submitted to Figshare under the mSWEEP project (https://figshare.com/projects/mSWEEP/64172).

Figshare: Extended Data Figures S1–S8. Additional figures supporting claims in the manuscript, https://doi.org/10.6084/m9.figshare.11379648.v3[24].

Figshare: Extended Data Table S1. Table of the most common clonal complexes and sequence types identified in the Vietnamese *E. coli* samples, https://doi.org/10.6084/m9.figshare.11379705.v1[28].

Figshare: Extended Data Table S2. Antimicrobial classes found in the Vietnamese *E. coli* samples; separated by diarrheal and healthy samples, https://doi.org/10.6084/m9.figshare.11379753.v1[32].

Figshare: Extended Data Table S3. Full antimicrobial resistance gene data in the Vietnamese *E. coli* samples, as identified by ARIBA, https://doi.org/10.6084/m9.figshare.11379756.v1[33].

Figshare: Extended Data Table S4. Description, accession numbers, and source studies for the reference sequence data used with mSWEEP, https://doi.org/10.6084/m9.figshare.11379762.v1[49].

Figshare: Extended Data Table S5. Accession numbers, status, and names of the Vietnamese *E. coli* samples, https://doi.org/10.6084/m9.figshare.11379771.v1[56].

Figshare: Extended Data Table S6. Accession numbers for the *Klebsiella* and *Campylobacter* mixture samples available in public repositories, https://doi.org/10.6084/m9.figshare.11379777.v1[57].

Software availability

Source code and precompiled binaries (generic Linux and macOS) for the mSWEEP software: https://github.com/PROBIC/mSWEEP

Archived source code as at time of publication: https://doi.org/10.5281/zenodo.3585009[61]

License: MIT

Acknowledgements

A previous version of this work is available from: https://doi.org/10.1101/332544.

References

1. Ellegaard KM, Engel P: **Beyond 16S rRNA Community Profiling: Intra-Species Diversity in the Gut Microbiota.** *Front Microbiol.* 2016; **7**: 1475.
PubMed Abstract | Publisher Full Text | Free Full Text

2. Quince C, Walker AW, Simpson JT, *et al.*: **Shotgun metagenomics, from sampling to analysis.** *Nat Biotechnol.* 2017; **35**(9): 833–844.
PubMed Abstract | Publisher Full Text

3. Yang X, Noyes NR, Doster E, *et al.*: **Use of Metagenomic Shotgun Sequencing Technology To Detect Foodborne Pathogens within the Microbiome of the Beef Production Chain.** *Appl Environ Microbiol.* 2016; **82**(8): 2433–2443.
PubMed Abstract | Publisher Full Text | Free Full Text

4. Ye SH, Siddle KJ, Park DJ, *et al.*: **Benchmarking Metagenomics Tools for Taxonomic Classification.** *Cell.* 2019; **178**(4): 779–794.
PubMed Abstract | Publisher Full Text | Free Full Text

5. Sczyrba A, Hofmann P, Belmann P, *et al.*: **Critical Assessment of Metagenome Interpretation-a benchmark of metagenomics software.** *Nat Methods.* 2017; **14**(11): 1063–1071.
PubMed Abstract | Publisher Full Text | Free Full Text

6. Greenblum S, Carr R, Borenstein E: **Extensive strain-level copy-number variation across human gut microbiome species.** *Cell.* 2015; **160**(4): 583–594.
PubMed Abstract | Publisher Full Text | Free Full Text

7. Joseph SJ, Li B, Ghonasgi T, *et al.*: **Direct amplification, sequencing and profiling of *Chlamydia trachomatis* strains in single and mixed infection clinical samples.** *PLoS One.* 2014; **9**(6): e99290.
PubMed Abstract | Publisher Full Text | Free Full Text

8. Nayfach S, Rodriguez-Mueller B, Garud N, *et al.*: **An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography.** *Genome Res.* 2016; **26**(11): 1612–1625.
PubMed Abstract | Publisher Full Text | Free Full Text

9. Paterson GK, Harrison EM, Murray GGR, *et al.*: **Capturing the cloud of diversity reveals complexity and heterogeneity of MRSA carriage, infection and transmission.** *Nat Commun.* 2015; **6**: 6560.
PubMed Abstract | Publisher Full Text | Free Full Text

10. Worby CJ, Lipsitch M, Hanage WP: **Within-host bacterial diversity hinders accurate reconstruction of transmission networks from genomic distance data.** *PLoS Comput Biol.* 2014; **10**(3): e1003549.
PubMed Abstract | Publisher Full Text | Free Full Text

11. Sankar A, Malone B, Bayliss SC, *et al.*: **Bayesian identification of bacterial strains from sequencing data.** *Microb Genom.* 2016; **2**(8): e000075.
PubMed Abstract | Publisher Full Text | Free Full Text

12. Fischer M, Strauch B, Renard BY: **Abundance estimation and differential testing on strain level in metagenomics data.** *Bioinformatics.* 2017; **33**(14): i124–i132.
PubMed Abstract | Publisher Full Text | Free Full Text

13. Schaeffer L, Pimentel H, Bray N, *et al.*: **Pseudoalignment for metagenomic read assignment.** *Bioinformatics.* 2017; **33**(14): 2082–2088.
PubMed Abstract | Publisher Full Text | Free Full Text

14. Bray NL, Pimentel H, Melsted P, *et al.*: **Near-optimal probabilistic RNA-seq quantification.** *Nat Biotechnol.* 2016; **34**(5): 525–527.
PubMed Abstract | Publisher Full Text

15. Glaus P, Honkela A, Rattray M: **Identifying differentially expressed transcripts from RNA-seq data with biological variation.** *Bioinformatics.* 2012; **28**(13): 1721–1728.
PubMed Abstract | Publisher Full Text | Free Full Text

16. Hensman J, Papastamoulis P, Glaus P, *et al.*: **Fast and accurate approximate inference of transcript expression from RNA-seq data.** *Bioinformatics.* 2015; **31**(24): 3881–3889.
PubMed Abstract | Publisher Full Text | Free Full Text

17. Kallonen T, Brodrick HJ, Harris SR, *et al.*: **Systematic longitudinal survey of invasive *Escherichia coli* in England demonstrates a stable population structure only transiently disturbed by the emergence of ST131.** *Genome Res.* 2017; **27**(8): 1437–1449.
PubMed Abstract | Publisher Full Text | Free Full Text

18. Long SW, Olsen RJ, Eagar TN, *et al.*: **Population Genomic Analysis of 1,777 Extended-Spectrum Beta-Lactamase-Producing *Klebsiella pneumoniae* Isolates, Houston, Texas: Unexpected Abundance of Clonal Group 307.** *mBio.* 2017; **8**(3): e00489-17.
PubMed Abstract | Publisher Full Text | Free Full Text

19. Meric G, Miragaia M, de Been M, *et al.*: **Ecological Overlap and Horizontal Gene Transfer in *Staphylococcus aureus* and *Staphylococcus epidermidis*.** *Genome Biol Evol.* 2015; **7**(5): 1313–1328.
PubMed Abstract | Publisher Full Text | Free Full Text

20. Yahara K, Meric G, Taylor AJ, *et al.*: **Genome-wide association of functional traits linked with *Campylobacter jejuni* survival from farm to fork.** *Environ Microbiol.* 2017; **19**(1): 361–380.
PubMed Abstract | Publisher Full Text

21. Maiden MC, Bygraves JA, Feil E, *et al.*: **Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms.** *Proc Natl Acad Sci U S A.* 1998; **95**(6): 3140–3145.
PubMed Abstract | Publisher Full Text | Free Full Text

22. Cheng L, Connor TR, Sirén J, *et al.*: **Hierarchical and spatially explicit clustering of DNA sequences with BAPS software.** *Mol Biol Evol.* 2013; **30**(5): 1224–1228.
PubMed Abstract | Publisher Full Text | Free Full Text

23. Runcharoen C, Moradigaravand D, Blane B, *et al.*: **Whole genome sequencing reveals high-resolution epidemiological links between clinical and environmental *Klebsiella pneumoniae*.** *Genome Med.* 2017; **9**(1): 6.
PubMed Abstract | Publisher Full Text | Free Full Text

24. Mäklin T, Kallonen T, David S, *et al.*: **Extended Data Figures S1-S8.** *figshare.* Figure. 2019.
http://www.doi.org/10.6084/m9.figshare.11379648.v3

25. Maaten Lvd, Hinton G: **Visualizing data using t-SNE.** *J Mach Learn Res.* 2008; **9**: 2579–2605.
Reference Source

26. Ondov BD, Treangen TJ, Melsted P, *et al.*: **Mash: fast genome and metagenome distance estimation using MinHash.** *Genome Biol.* 2016; **17**(1): 132.
PubMed Abstract | Publisher Full Text | Free Full Text

27. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2.** *Nat Methods.* 2012; **9**(4): 357–359.
PubMed Abstract | Publisher Full Text | Free Full Text

28. Mäklin T, Kallonen T, David S, *et al.*: **Extended Data Table S1.** *figshare.* Dataset. 2019.
http://www.doi.org/10.6084/m9.figshare.11379705.v2

29. Hunt M, Mather AE, Sánchez-Busó L, *et al.*: **ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads.** *Microb Genom.* 2017; **3**(10): e000131.
PubMed Abstract | Publisher Full Text | Free Full Text

30. Magiorakos AP, Srinivasan A, Carey RB, *et al.*: **Multidrug-resistant, extensively drug-resistant and pandrug-resistant bacteria: an international expert proposal for interim standard definitions for acquired resistance.** *Clin Microbiol Infect.* 2012; **18**(3): 268–281.
PubMed Abstract | Publisher Full Text

31. Liu YY, Wang Y, Walsh TR, *et al.*: **Emergence of plasmid-mediated colistin resistance mechanism MCR-1 in animals and human beings in China: a microbiological and molecular biological study.** *Lancet Infect Dis.* 2016; **16**(2): 161–168.
PubMed Abstract | Publisher Full Text

32. Mäklin T, Kallonen T, David S, *et al.*: **Extended Data Table S2.** *figshare.* Dataset. 2019.
http://www.doi.org/10.6084/m9.figshare.11379753.v2

33. Mäklin T, Kallonen T, David S, *et al.*: **Extended Data Table S3.** *figshare.* Dataset. 2019.
http://www.doi.org/10.6084/m9.figshare.11379756.v2

34. Sheppard SK, Colles FM, McCarthy ND, *et al.*: **Niche segregation and genetic structure of *Campylobacter jejuni* populations from wild and agricultural host species.** *Mol Ecol.* 2011; **20**(16): 3484–3490.
PubMed Abstract | Publisher Full Text | Free Full Text

35. Sheppard SK, Dallas JF, Strachan NJ, *et al.*: ***Campylobacter* genotyping to determine the source of human infection.** *Clin Infect Dis.* 2009; **48**(8): 1072–1078.
PubMed Abstract | Publisher Full Text | Free Full Text

36. Colles FM, McCarthy ND, Layton R, *et al.*: **The prevalence of *Campylobacter* amongst a free-range broiler breeder flock was primarily affected by flock age.** *PLoS One.* 2011; **6**(12): e22825.
PubMed Abstract | Publisher Full Text | Free Full Text

37. Sproston EL, Ogden ID, MacRae M, *et al.*: **Temporal variation and host association in the *Campylobacter* population in a longitudinal ruminant farm study.** *Appl Environ Microbiol.* 2011; **77**(18): 6579–6586.
PubMed Abstract | Publisher Full Text | Free Full Text

38. Lu J, Idris U, Harmon B, *et al.*: **Diversity and succession of the intestinal bacterial community of the maturing broiler chicken.** *Appl Environ Microbiol.* 2003; **69**(11): 6816–6824.
PubMed Abstract | Publisher Full Text | Free Full Text

39. Buffie CG, Pamer EG: **Microbiota-mediated colonization resistance against intestinal pathogens.** *Nat Rev Immunol.* 2013; **13**(11): 790–801.
PubMed Abstract | Publisher Full Text | Free Full Text

40. Nowrouzian FL, Wold AE, Adlerberth I: ***Escherichia coli* strains belonging to phylogenetic group B2 have superior capacity to persist in the intestinal microflora of infants.** *J Infect Dis.* 2005; **191**(7): 1078–1083.
PubMed Abstract | Publisher Full Text

41. Hayashi H, Takahashi R, Nishi T, *et al.*: **Molecular analysis of jejunal, ileal, caecal and recto-sigmoidal human colonic microbiota using 16S rRNA gene libraries and terminal restriction fragment length polymorphism.** *J Med Microbiol.* 2005; **54**(Pt 11): 1093–1101.
PubMed Abstract | Publisher Full Text

42. Johns BE, Purdy KJ, Tucker NP, *et al.*: **Phenotypic and Genotypic Characteristics of Small Colony Variants and Their Role in Chronic Infection.** *Microbiol Insights.* 2015; **8**: 15–23.
PubMed Abstract | Publisher Full Text | Free Full Text

43. von Bronk B, Schaffer SA, Götz A, *et al.*: **Effects of stochasticity and division of**

labor in toxin production on two-strain bacterial competition in *Escherichia coli*. *PLoS Biol.* 2017; **15**(5): e2001457.
PubMed Abstract | Publisher Full Text | Free Full Text

44. Stämmler F, Gläsner J, Hiergeist A, *et al.*: **Adjusting microbiome profiles for differences in microbial load by spike-in bacteria.** *Microbiome.* 2016; **4**(1): 28.
PubMed Abstract | Publisher Full Text | Free Full Text

45. Costea PI, Zeller G, Sunagawa S, *et al.*: **Towards standards for human fecal sample processing in metagenomic studies.** *Nat Biotechnol.* 2017; **35**(11): 1069–1076.
PubMed Abstract | Publisher Full Text

46. Arevalo P, VanInsberghe D, Elsherbini J, *et al.*: **A Reverse Ecology Approach Based on a Biological Definition of Microbial Populations.** *Cell.* 2019; **178**(4): 820–834.e14.
PubMed Abstract | Publisher Full Text

47. Bobay LM, Ellis BSH, Ochman H: **ConSpeciFix: classifying prokaryotic species based on gene flow.** *Bioinformatics.* 2018; **34**(21): 3738–3740.
PubMed Abstract | Publisher Full Text | Free Full Text

48. Truong DT, Tett A, Pasolli E, *et al.*: **Microbial strain-level population structure and genetic diversity from metagenomes.** *Genome Res.* 2017; **27**(4): 626–638.
PubMed Abstract | Publisher Full Text | Free Full Text

49. Mäklin T, Kallonen T, David S, *et al.*: **Extended Data Table S4.** *figshare.* Dataset. 2019.
http://www.doi.org/10.6084/m9.figshare.11379762.v2

50. Segata N, Waldron L, Ballarini A, *et al.*: **Metagenomic microbial community profiling using unique clade-specific marker genes.** *Nat Methods.* 2012; **9**(8): 811–814.
PubMed Abstract | Publisher Full Text | Free Full Text

51. Griffiths DA: **Maximum likelihood estimation for the beta-binomial distribution and an application to the household distribution of the total number of cases of a disease.** *Biometrics.* 1973; **29**(4): 637–648.
PubMed Abstract | Publisher Full Text

52. Berg S: **Condorcet's jury theorem, dependency among jurors.** *Social Choice and Welfare.* 1993; **10**: 87–95.
Publisher Full Text

53. Hensman J, Rattray M, Lawrence ND: **Fast Variational Inference in the Conjugate Exponential Family.** In *Advances in Neural Information Processing Systems* 25. 2012.
Reference Source

54. Quail MA, Otto TD, Gu Y, *et al.*: **Optimal enzymes for amplifying sequencing libraries.** *Nat Methods.* 2011; **9**(1): 10–11.
PubMed Abstract | Publisher Full Text

55. Mäklin T, Kallonen T, David S, *et al.*: **mSWEEP_reference_v1-0-0.tgz.** *figshare.* Dataset. 2019.
http://www.doi.org/10.6084/m9.figshare.8222636.v2

56. Mäklin T, Kallonen T, David S, *et al.*: **Extended Data Table S5.** *figshare.* Dataset. 2019.
http://www.doi.org/10.6084/m9.figshare.11379771.v2

57. Mäklin T, Kallonen T, David S, *et al.*: **Extended Data Table S6.** *figshare.* Dataset. 2019.
http://www.doi.org/10.6084/m9.figshare.11379777.v2

58. Mäklin T, Kallonen T, David S, *et al.*: **campylobacter_mixtures_1.tgz.** *figshare.* Dataset. 2018.
http://www.doi.org/10.6084/m9.figshare.6445136

59. Mäklin T, Kallonen T, David S, *et al.*: **campylobacter_mixtures_2.tgz.** *figshare.* Dataset. 2018.
http://www.doi.org/10.6084/m9.figshare.6445190

60. Mäklin T: **mSWEEP_revision_mixture_samples_v1-0-0.tar.** *zenodo.* Dataset. 2021.
http://www.doi.org/10.5281/zenodo.5535713

61. Mäklin T, Honkela A: **PROBIC/mSWEEP: v1.1.0 (17 December 2018).** (Version v1.1.0) *Zenodo.* 2019.
http://www.doi.org/10.5281/zenodo.3585009

Open Peer Review

Current Peer Review Status: ✔ ✔

Version 2

Reviewer Report 22 October 2021

https://doi.org/10.21956/wellcomeopenres.19077.r46349

✔ **Mahendra Mariadassou**
INRAE, MaIAGE, Université Paris-Saclay, Jouy-en-Josas, France

The authors answered all comments I raised and I have no further ones.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Statistics, Metagenomics.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 05 March 2021

https://doi.org/10.21956/wellcomeopenres.17135.r39845

? **Mahendra Mariadassou**
INRAE, MaIAGE, Université Paris-Saclay, Jouy-en-Josas, France

The article introduces a new method for typing of metagenomics data that goes beyond the species level (whenever) and down to the lineage or strain level. The approach does not apply to WGS datasets but to datasets sampled from plate sweeps, where the complexity of the community has been massively reduced by growth medium selection and the focus in on strain resolution.

The work is well written and easy to understand but some limitations should be pointed out:

- The work is most relevant for well studied bacteria (like pathogens) for which there are good genome reference databases.

- The comparison with metakallisto is slightly unfair, as pointed out by the authors, since it tries to tackle the more complex job of strain (rather than lineage) identification. A fairer comparison would be to pool the strains into the same group as mSWEEP and assess how it performs for those groups. This is especially important as the groups used in mSWEEP are well separated (see Fig. 3 t-SNE plot) and thus the task of assigning single-colony isolates to those groups is much easier than assigning them to strains.

- The synthetic mixtures are not very adverse to the methods: 3 samples from 3 lineages with abundances in the mix higher than 20% is a very simple mix. The Vietnamese example already correspond to more complex mixes (~10 lineages / sample) and it would have been nice to stress mSWEEP under those conditions.

- The noise level of 0.016 used as detection threshold would prevent the detection / quantification of rare lineages and means that whole plates can't be mixed if they encompass many lineages with abundances spanning several orders of magnitude. This appears to be the case for the *K. pneumionae* dataset.

How does the method compare to other strain identification tools (like DUDes[1]) that are not specific to plate sweeps? I expect the reduction in complexity induced by plate sweeps to benefit mSWEEP but it would be useful to prove it. Finally, how does the method behave when the single-colony isolate from the sweep is very far from all lineages / groups in the database? Does mSWEEP then fail with an informative message / warning?

References

1. Piro V, Lindner M, Renard B: DUDes: a top-down taxonomic profiler for metagenomics. *Bioinformatics*. 2016; **32** (15): 2272-2280 Publisher Full Text

Is the rationale for developing the new method (or application) clearly explained?

Yes

Is the description of the method technically sound?

Yes

Are sufficient details provided to allow replication of the method development and its use by others?

Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Statistics, Metagenomics.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 29 Sep 2021

Tommi Mäklin, University of Helsinki, Finland

We have thoroughly utilized the excellent feedback provided by both reviewers and made adjustments to our manuscript based on their recommendations and suggestions. Notably, we have added a new synthetic experiment assessing the performance of mSWEEP under conditions more resembling those found in the Vietnamese E. coli sequencing data as suggested by Mahendra Mariadassou. Included below is a point-by-point response to the concerns raised by both reviewers. We would also like to thank both reviewers for their time and outstanding feedback that has enabled us to improve the quality of our manuscript.

Review 2 — Mahendra Mariadassou

Mahendra Mariadassou:
The work is most relevant for well studied bacteria (like pathogens) for which there are good genome reference databases.

Author reply:
We have made it more clear in the revised manuscript that the intended application of our method is for extensively studied bacteria that have readily available reference sequences by including the following sentence in the Discussion section of our manuscript:
As with any method intended to identify sequence variation, the target species need to be relatively well known to allow building of sufficiently informative reference databases. This means that mSWEEP is most relevant for extensively studied bacteria, such as pathogens, and has only limited applications when working with samples containing high numbers of novel or uncharacterized species.

Thank you for pointing out the potential for a misunderstanding.

Mahendra Mariadassou:
The comparison with metakallisto is slightly unfair, as pointed out by the authors, since it tries to tackle the more complex job of strain (rather than lineage) identification. A fairer comparison would be to pool the strains into the same group as mSWEEP and assess how it performs for those groups. This is especially important as the groups used in mSWEEP are well separated (see Fig. 3 t-SNE plot) and thus the task of assigning single-colony isolates to those groups is much easier than assigning them to strains.

Author reply:
We agree that the comparison is slightly unfair since the methods are designed to solve different, yet related, problems. In the revised manuscript we have addressed this issue by adding a supplementary figure (Extended Data Figure 1 in the new manuscript), where the estimates from metakallisto for the sequences within the lineages have been pooled together by summing them up. While pooling improves the performance of metakallisto, the method still does not perform as well as mSWEEP, demonstrating that the probabilistic model leveraged by mSWEEP is necessary for resolving lineage-level differences, and that the conclusions in our manuscript remain largely the same.

Mahendra Mariadassou:
The synthetic mixtures are not very adverse to the methods: 3 samples from 3 lineages with abundances in the mix higher than 20% is a very simple mix. The Vietnamese example already correspond to more complex mixes (~10 lineages / sample) and it would have been nice to stress mSWEEP under those conditions.

Author reply:
We thank the reviewer for this suggestion and have accordingly added a new set of experiments, where 10 different *E. coli* lineages per sample are mixed at varying sequencing depths, resulting in coverages ranging between 50x to 0.10x. Although the new experiments are more complex, the accuracy of the results (Figure 5 in the revised manuscript) remain largely the same and, in fact, support the conclusion derived from the existing set of synthetic mixtures that mSWEEP performs well up to around the 0.016 relative abundance (corresponding to about 1.6x coverage), which was the background noise level we previously established.

Mahendra Mariadassou:
The noise level of 0.016 used as detection threshold would prevent the detection / quantification of rare lineages and means that whole plates can't be mixed if they encompass many lineages with abundances spanning several orders of magnitude. This appears to be the case for the K. pneumionae dataset.

Author reply:
The noise level of 0.016 is specific to the sequencing depth that was used to generate the synthetic mixtures. Since each mixture contains 1 000 000 reads that are 100 bases long, a relative abundance of under 0.016 would correspond to less than 16 000 reads and <0.65x sequencing depth (the exact value depending on the organism) in the species that we investigated. Considering that the differences between the lineages can be quite minimal, identification at these sequencing depths will naturally be less reliable and depend heavily upon the quality and nature of the reference sequences. Due to these factors, some sort of thresholding is, in our opinion, necessary.

Presumably the sequencing depth could be increased in order to better identify rare lineages but our data unfortunately did not provide grounds to evaluate mSWEEP in this kind of setting. Nevertheless, there is some preliminary support for this argument in the revised manuscript, where the complex synthetic mixtures contained approximately 5 times more reads than the experiments the noise level was derived from, and consequently we

were able to recover lineages at 0.0156 relative abundance (1.56x sequencing depth) and even a few of the lineages at 0.0078 relative abundance (0.78x sequencing depth). However, in general we do not expect mSWEEP to be able to separate lineages with < 1x sequencing depth unless they are far from each other in terms of genetic distance.

In any case, we thank the reviewer for pointing this issue out and have clarified the implications of the noise level in the manuscript with the following changes to the section where the noise level is mentioned:
The results indicate that above this relatively low background noise level of 0.016, quantifying mixture samples is not expected to produce more false positive results than would be obtained from single-colony samples. In the synthetic mixtures, the observed background noise level corresponds to sequencing depths of around 0.30x (E. coli and K. pneumoniae) and 0.65x (S. epidermidis), which provides the bare minimum sequencing depth required to distinguish between the lineages of each species in samples with similar read lengths and sequencing depth.

Mahendra Mariadassou:
How does the method compare to other strain identification tools (like DUDes) that are not specific to plate sweeps? I expect the reduction in complexity induced by plate sweeps to benefit mSWEEP but it would be useful to prove it.

Author reply:
We thank you for pointing out other relevant strain identification tools. However, we would like to point out that our tool is designed to be used with large, bespoke reference databases with the goal of leveraging detailed information about strain-level variation in identification. This approach necessitates replacing the traditional alignment methods, such as bowtie2 used by for example DUDes and many other strain identification tools, with pseudoalignment to achieve practical runtimes.

Additionally, DUDes specifically utilizes a version of the NCBI taxonomy standard that is no longer in use, which restricts the applications of the method. The author points out this issue, as well as the scaling issue in bowtie2 alignment, in their doctoral dissertation (http://dx.doi.org/10.17169/refubium-1123), published after the DUDes paper. To our knowledge, and based on the DUDes GitHub repository, DUDes has not been updated to address these issues nor, in fact, been maintained after its initial publication. In our opinion a comparison with mSWEEP is thus unwarranted and would likely add little of value to the comparisons already presented in our manuscript.

Moreover, other strain identification tools that make use of curated databases, such as subsets of the NCBI RefSeq database, are unsuited to analysing strain and lineage-level variation in clinical sequencing data due to the fact that for many of the strains the correct — both in-time and geographically — reference sequence is simply not necessarily included in the database. These types of databases are designed to provide a general overview of representative high-quality sequences for the species rather than something representative of the possibly highly localized variation, which renders their application beyond species-level identification conditional on the samples containing only distantly related strains. Although some methods do offer means to use customized reference databases, they are typically designed to be run with the precompiled databases and may not extend well when

stretched beyond their intended purposes to the types of application that we have designed mSWEEP for.

Mahendra Mariadassou:
Finally, how does the method behave when the single-colony isolate from the sweep is very far from all lineages / groups in the database? Does mSWEEP then fail with an informative message / warning?

Author reply:
While this is truly a limitation of our method that we did not directly tackle, as also pointed out by Fabiano L. Thompson in the review, there fortunately has been further work in solving this exact limitation. In particular, a computational tool (https://github.com/harry-thorpe/demix_check) has been developed to check whether the contents of a mixed sample correspond to a related reference sequence, or sequences from the same lineage, in mSWEEP's reference database, and produce results that can be used to verify the fact. For further explanation, we would like to refer to our answer to Fabiano L. Thompson to the similar question about the presence of a novel bacteria in the analysed sample. We have rewritten a part of our discussion section to address this concern, with the relevant paragraph now reading as:
As with any method intended to identify sequence variation, the target species need to be relatively well known to allow building of sufficiently informative reference databases. Similarly, to allow for sensible and easily interpretable inferences, the biological clustering of the reference database should be based on well-established biological entities, such as multi-locus sequence types (STs) or clonal complexes (CCs) which are frequently employed as labels of lineages. These limitations suggest that mSWEEP is most relevant for extensively studied bacteria, such as pathogens, and has only limited applications when working with samples containing high numbers of novel or uncharacterized species. However, following the introduction of mSWEEP, a separate computational tool has been developed to assess the suitability of the reference sequences to specific samples and to identify cases where mSWEEP struggles due to the presence of reads originating from novel genome sequences (https://github.com/harry-thorpe/demix_check). Consequently mSWEEP can also be used to estimate the quality of the reference collection and to discard contaminated or mis-identified genomes.

Competing Interests: No competing interests were disclosed.

Reviewer Report 16 March 2020

https://doi.org/10.21956/wellcomeopenres.17135.r37996

Fabiano L. Thompson

Laboratory of Microbiology, Institute of Biology, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil

Maklin *et al*. propose a new tool for the study of lineages and strains obtained from metagenomics. First of all, congratulations for this excellent study.

Indeed, bioinformatics tools to tap into the microbial diversity beyond genus and species level using metagenomics are limited. And particularly for those approaches involving an initial culture/plate medium step. The authors have closed this gap with mSWEEP but they have not taken advantage of this to mention in their title that the novel tool addresses strain/lineage diversity). Mathematical modelling, including maximum likelihood and binomial distances were used and formulas were shown. The author studied samples from a group of infected children to validate their system. They used MacConkey plates to obtain bacterial cells for metagenomics. Reference data comprised Campylobacter, Escherichia, Klebsiella, Staphylococcus. False positives were estimated and in general were below 0.3 (Fig 4; y axis %?). mSWEEP output is depicted in the frame of clonal complexes (CC; Figs 3, 7) which is useful in the context of previous studies on population structure of pathogens.

Some minor remarks to be incorporated in a revised version:

- Seven formulas were provided. However, the elements of each of these formulas were not explained. Please include an explanation in the new version.

- The Hiseq sequencing coverage (and read numbers) were not informed. How many ilumina reads does one need to use mSWEEP with confidence?

- Also another limitation is the need to select a reference genome for comparison. The sample may contain a novel genome, a novel pathogen. How does the system handle such possible situation?

- It is not clear why certain CC, eg. *Escherichia coli* CC405 and CC95 are connected to *Klebsiella pneumoniae* CC36, CC45 (Fig 7). This figure seems rather mixed in taxonomic terms. Is this a result of co-occurrence?

- It seems to me that relevant references need to be considered in the revised final version as they address similar questions. Arevalo *et al.* (2019)[1] and Bobay *et al*. (2018)[2].

References

1. Arevalo P, VanInsberghe D, Elsherbini J, Gore J, et al.: A Reverse Ecology Approach Based on a Biological Definition of Microbial Populations.*Cell*. 2019; **178** (4): 820-834.e14 PubMed Abstract | Publisher Full Text
2. Bobay L, Ellis B, Ochman H: ConSpeciFix: classifying prokaryotic species based on gene flow. *Bioinformatics*. 2018; **34** (21): 3738-3740 Publisher Full Text

Is the rationale for developing the new method (or application) clearly explained?

Yes

Is the description of the method technically sound?

Yes

Are sufficient details provided to allow replication of the method development and its use by others?
Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Microbiology.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 29 Sep 2021

Tommi Mäklin, University of Helsinki, Finland

We have thoroughly utilized the excellent feedback provided by both reviewers and made adjustments to our manuscript based on their recommendations and suggestions. Notably, we have added a new synthetic experiment assessing the performance of mSWEEP under conditions more resembling those found in the Vietnamese E. coli sequencing data as suggested by Mahendra Mariadassou. Included below is a point-by-point response to the concerns raised by both reviewers. We would also like to thank both reviewers for their time and outstanding feedback that has enabled us to improve the quality of our manuscript.

Review 1 — Fabiano L. Thompson

Fabiano L. Thompson:
Seven formulas were provided. However, the elements of each of these formulas were not explained. Please include an explanation in the new version.

Author reply:
We thank the reviewer for pointing out the issue of missing explanations and have revised the manuscript so that all elements are explained close to the formulas where they are used.

Fabiano L. Thompson:
The Hiseq sequencing coverage (and read numbers) were not informed. How many ilumina reads does one need to use mSWEEP with confidence?

Author reply:
While mSWEEP does not by itself require a specific number of Illumina reads to work, some practical guidelines can be derived based on the detection threshold analysis presented in the paper. In our analyses related to Extended Data Figures S3 and S4, we found that the minimum relative abundance required to accurately distinguish between strains of the same species in the same sample ("background noise level" in our manuscript) was 0.016 in synthetic samples containing 1 000 000 reads that were 100bp long. This translates to a minimum requirement of at least 16 000 reads from a specific lineage, corresponding to a coverage of about 0.30x on an average *E. coli* or *K. pneumoniae* genome and 0.65x on an *S. epidermidis* genome. Further down in the manuscript we define the detection thresholds to be either the background noise level or the value obtained from the threshold construction process, whichever of the two is higher, which results in lineage-specific thresholds that could be analysed in a similar way as the background noise level.

Unfortunately, the results from this type of analysis are specific to the samples and reference sequences in question and do not generalize to different types of data. However, by replicating this type of analysis on one's own prospective set of reference sequences, similar coverage and read number requirements could be derived before sequencing any mixed samples, providing some guidelines as to how many reads are required for mSWEEP to work accurately. The detection threshold analysis does additionally provide confidence scores for the abundance estimates exceeding the thresholds, which can be used to assess the reliability of the results.

We have expanded our Discussion section to include a remark on the possibility of using the detection thresholds to assess the required sequencing coverage by including the following statement:
This pipeline also estimates the relative abundance of lineages and provides means to construct reliability cut-offs with accompanying confidence scores. Since the cut-offs are constructed before analysing any sequencing data by synthetically mimicking the properties of the in vitro data, the cut-offs and confidence scores can be used to assess the necessary sequencing depth to identify rare and low abundance lineages with confidence as well as the total number of reads required.

We further thank the reviewer for providing us the opportunity to better our manuscript by explaining this feature of our pipeline in more detail.

Fabiano L. Thompson:
Also another limitation is the need to select a reference genome for comparison. The sample may contain a novel genome, a novel pathogen. How does the system handle such possible situation?

Author reply:
As pointed out by the reviewer and also in Mahendra Mariadassou's review, the absence of a reference sequence for a novel bacterial strain or species is absolutely a limitation of our method. In cases where there is no related genome at all in the reference genomes, a very low percentage of the reads will pseudoalign to the reference sequences at all, interrupting the analysis. In cases where the novel genome is related to the reference genomes (either a new lineage of an existing species or a closely related novel species), our method will

naturally fail to identify the correct lineage. Based on our observations, the relative abundance estimates resulting from this scenario will be spread between the most closely related genomes that are contained in the reference which can, in some cases, produce results that appear deceptively similar to those from a real mixed sample.

While our method does not address this concern — other than indirectly in some specific cases through the detection threshold analysis described in the manuscript — there have been excellent further developments of our method in solving this exact problem after the manuscript was published. In particular, a supplementary computational method (https://github.com/harry-thorpe/demix_check) has been developed for assessing whether the reference sequences contain representative sequences for the strains present in a (mixed) sample that has been processed by mSWEEP. We believe this tool neatly solves this particular limitation of our method and provides researchers an additional tool to properly judge whether the results of their analyses are correct or contain errors related to the presence of novel strains. We have rewritten a part of our discussion section to take into account the limitations pointed out by the reviewer and include our suggested solution. The relevant paragraph has been changed to the following:
As with any method intended to identify sequence variation, the target species need to be relatively well known to allow building of sufficiently informative reference databases. Similarly, to allow for sensible and easily interpretable inferences, the biological clustering of the reference database should be based on well-established biological entities, such as multi-locus sequence types (STs) or clonal complexes (CCs) which are frequently employed as labels of lineages. These limitations suggest that mSWEEP is most relevant for extensively studied bacteria, such as pathogens, and has only limited applications when working with samples containing high numbers of novel or uncharacterized species. However, following the introduction of mSWEEP, computational tools have been developed to assess the suitability of the reference sequences to specific samples and to identify cases where mSWEEP struggles due to the presence of reads originating from novel genome sequences (https://github.com/harry-thorpe/demix_check). Consequently mSWEEP can also be used to estimate the quality of the reference collection and to discard contaminated or mis-identified genomes.

Fabiano L. Thompson:
It is not clear why certain CC, eg. Escherichia coli CC405 and CC95 are connected to Klebsiella pneumoniae CC36, CC45 (Fig 7). This figure seems rather mixed in taxonomic terms. Is this a result of co-occurrence?

Author reply:
Indeed, the connections between the *E. coli* and *K. pneumoniae* clonal complexes are a result of co-occurrence in the samples. These samples were specifically selected for our study because they had been identified as contaminated or containing mixed species or strains during the culture step in another study aimed at isolating *K. pneumoniae* strains. As a result, many of the samples presented in our study are taxonomically diverse and contain several different species, or the strains of, that also grow on the types of plates that were used to culture *K. pneumoniae*. We thank the reviewer for pointing out the confusing parts in our text and have clarified the reasons for the co-occurrence in the text with the addition of the following sentence:
Since both E. coli and A. baumanii grow on the media used for culture of K. pneumoniae, frequent

co-occurrence in samples selected for their diversity is expected.

Fabiano L. Thompson:
It seems to me that relevant references need to be considered in the revised final version as they address similar questions. Arevalo et al. (2019) and Bobay et al. (2018).

Author reply:
We thank you for referring us to these highly relevant studies that we missed in our literature review and have addressed the potential benefits of combining gene flow based approaches with mSWEEP analyses in the discussion section of our revised version of the manuscript.

Competing Interests: No competing interests were disclosed.