High-resolution sweep metagenomics using fast probabilistic inference [version 1; peer review: awaiting peer review]

Tommi Mäklin1, Teemu Kallonen2,3, Sophia David4, Christine J. Boinett5,6, Ben Pascoe7, Guillaume Méric8, David M. Aanensen3,8,9, Edward J. Feil7, Stephen Baker5,6, Julian Parkhill3, Samuel K. Sheppard7, Jukka Corander1,10,11, Antti Honkela1,10,11

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

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
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, 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.

Copyright: © 2020 Mäklin T et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Mäklin T, Kallonen T, David S et al. High-resolution sweep metagenomics using fast probabilistic inference [version 1; peer review: awaiting peer review] Wellcome Open Research 2020, 5:14 (https://doi.org/10.12688/wellcomeopenres.15639.1)

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

Current methods for taxonomic profiling of bacteria from sequencing data typically perform well only up to the species-level 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. Sequencing isolated colonies offers means to ignore this variation but only focusing on pure colonies is insufficient for many potential applications. 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, and a method for controlling false positive detections. Similar methods taking advantage of pseudoalignment either with or without 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 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 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. This approach 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 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 and to bacterial data. 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. 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.

Assigning single-colony isolates to lineage
We compared the performance of the mSWEEP pipeline (using kallisto 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 (version 0.45) and the BIB pipeline (commit hash 2999540). We additionally attempted to compare mSWEEP with ditasic (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 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 \(^5\), lineages identified with the BAPS clustering algorithm \(^5\), 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 \(K.\) quasipneumoniae isolates from Thailand \(^5\) 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 metaWallisto in cases measuring accuracy of abundance estimates in the true group (Figure 2; \(p < 10^{-4}\), in all comparisons, Wilcoxon signed-rank test; median error in all estimates for mSweep was 0.00003, for BIB 0.05, and for metaWallisto 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 metaWallisto 0.01). In these two latter cases, mSweep and metaWallisto performed similarly (Figure 2; \(p > 0.10\) when testing for the difference in accuracy in either direction, Wilcoxon signed-rank test). Since metaWallisto 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 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 \(^5\)). 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 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 S2a \(^5\)) 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 S1c \(^5\)).

We examined the grouping of the reference sequences by producing t-SNE plots of 31-mer distances estimated with mash \(^5\) (version 2.0) between the reference sequences including the test isolates (Figure 3; Extended Data Figures S2a-c \(^5\)). 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

<table>
<thead>
<tr>
<th>Grouping</th>
<th>Species</th>
<th>Sequences</th>
<th>Test sequences</th>
<th>Groups</th>
<th>Test groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clonal complex</td>
<td>Campylobacter coli</td>
<td>120</td>
<td>27</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Campylobacter jejuni</td>
<td>462</td>
<td>73</td>
<td>13</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>Escherichia coli</td>
<td>1509</td>
<td>188</td>
<td>132</td>
<td>54</td>
</tr>
<tr>
<td></td>
<td>Klebsiella pneumoniae</td>
<td>1351</td>
<td>129</td>
<td>79</td>
<td>39</td>
</tr>
<tr>
<td>Species</td>
<td>Klebsiella quasipneumonia</td>
<td>9</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Klebsiella variicola</td>
<td>12</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Staphylococcus aureus</td>
<td>181</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Lineage</td>
<td>Staphylococcus epidermidis</td>
<td>143</td>
<td>81</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Species</td>
<td>Multiple species with single sequences</td>
<td>28</td>
<td></td>
<td>28</td>
<td></td>
</tr>
<tr>
<td>total</td>
<td></td>
<td>3815</td>
<td>504</td>
<td>259</td>
<td>110</td>
</tr>
</tbody>
</table>
**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 (mSWEET 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).

**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 mSWEET and t-SNE also places them within or near the true source group.
correctly place the assemblies into the true groups but the method does not preserve the distances between the points or the clusters 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 (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 (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 S3). 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.

Comparing the empirical distributions of the errors from the synthetic mixtures and the single-colony isolates (Extended Data Figures S3 and S4) 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

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.
we do not establish the mixture estimates as more accurate since the distribution (Extended Data Figure S3) 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. 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.

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 5; Extended Data Table S1). Notably, 95% of the samples (46/48 Diarrheal and 45/48 Healthy) harboured multiple antimicrobial resistance genes (identified using the ARIBA software) that belonged to three or more classes of drugs (Extended Data Figure S5), which we defined as multi-drug resistance. One sample was found to contain the plasmid associated resistance gene MCR-1, which confers resistance to colistin, a last line antimicrobial drug. 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 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.

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).

---

**Figure 5.** 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.
The alpha diversity, measured by Shannon entropy (Extended Data Figure S6), 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 5) 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 6a) 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 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 6b) 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 7) and the sample-lineage heatmap (Extended Data Figure S7) 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 baumannii* and other species. Both *E. coli* and *A. baumannii* grow on the media used for culture of *K. pneumoniae*. Clonal complexes of *K. pneumoniae* centered on sequence types associated with high levels of multi-drug resistance (e.g. ST258, ST147 and ST101) were.

---

**Figure 6.** *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).
frequently observed co-existing with a variety of other K. pneumo-
niae lineages as well as with each other, and with other important
Gram-negative pathogens. Developing a deeper understand-
ing of these community structures and interactions will be criti-
cal 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
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 encoun-
tered in clinical and public health settings where standard cultur-
ing media is routinely used to isolate epidemiologically relevant
organisms. This pipeline also estimates the relative abundance
of lineages and reliability cut-offs. 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 under-
standing of infection by recovering a true representation of
bacteria in a complex sample. For example, genotyping stud-
ies have shown that C. jejuni and C. coli colonizing the primary
host (birds and mammals) form clusters of related isolates that are
host-associated35, which can be used to identify the reservoir for
human infection36. However, multiple organisms can be isolated
from the same sources37,38. The co-occurrence of different organ-
isms could be a snapshot in time of a wider process of lineage
succession39 in which the resident microbiota might resist new
colonizations or be displaced by recently acquired bacteria40,41.
Further, we suggest we may be able to infer complex interactions
between organisms that occupy different microniches42 and are
not in direct competition43,44 by analysing their co-occurrence.
Therefore, this approach provides a means to investigate the
nature of polymicrobial infections to improve our understand-
ing of the spread of a specific organism between hosts and transmis-
sion to humans in addition to enabling characterization of physi-
cal and temporal variation in the distribution of lineages among
multi-strain samples.

Because of limitations in the initial culture and DNA iso-
lization processes, we can only infer relative (not absolute)

Figure 7. 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.
abundances and spike-in methods 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. 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 multi-locus sequence types (STs) or clonal complexes (CCs) which are frequently employed as labels of lineages. As a by-product, 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, 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) 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 augmented with the genomes of a representative strain from 27 species that were identified in the real mixture data by MetaPhlAn.

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 software (version 6.0). The complete grouping is provided in Extended Data Table S4.

Pseudoalignment
We used kallisto (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 the reads $r_n = (r_{n,1}, \ldots, r_{n,k})$, $n = 1, \ldots, N$, containing only the information about the number of pseudoalignments $r_{nk}$ 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 $I = I_1, \ldots, I_K$ 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 $I = I_1, \ldots, I_K$, for the source group, and the mixing proportions of the groups $\theta = (\theta_1, \ldots, \theta_K)$ is defined as

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

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

Likelihood
The likelihood $p(r_n | I_n = k)$ 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 $r_n$ whose components $r_{n,k}$ denote the pseudoalignment count in group $k$, we define the likelihood $p(r_n | I_n = k)$ of observing the whole vector $r_n$ assuming that $k$ is the true group in three parts

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

$$p(r_n | I_n = k) = 0.99, \text{ when } M_k = 1 \text{ and } r_{n,k} > 0, \text{ or } \quad (3)$$
\[ p(r_k | I_k = k) = \frac{f(r_k, k)}{Z(r_k)} \cdot 0.99, \text{ when } M_k > 1 \text{ and } r_{k,i} > 0, \] (4)

with

\[ f(r_k, k) = \frac{M_k}{Z(r_k)} \frac{B(\alpha + r_k, M_k - r_k + \beta_k)}{B(\alpha + M_k, \beta_k)} \]

and

\[ Z(r_k) = \prod_{k=1}^{\infty} \frac{\alpha + r_k + (j-1)}{\alpha + \beta_k + 2M_k + (j-1)} \frac{\Gamma(\alpha + r_k + M_k) \Gamma(\alpha + \beta_k + 3M_k)}{\Gamma(\alpha + r_k) \Gamma(\alpha + \beta_k + 3M_k)} \] (5)

where \( B(a, b) \) is the beta function and \( M_k \) is the number of reference sequences in 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 \( Z(r_k) \) in Equation (4) and Equation (5) arises from deriving the normalizing constant for normalizing \( f(r_k, k) \) over \( r_k \). The derivation follows from using the identity \( B(a+1,b) = B(a,b) \frac{a}{a+b} \) to express each \( f(r_{k,i}, k) \), \( k \in \{1, \ldots, K : M_k > 1\} \) as a product of the probability mass function of a beta-binomial distribution with parameters \((\alpha_k, M_k, \beta_k)\), and the term \( \prod_{k=1}^{\infty} \frac{\alpha + r_k + (j-1)}{\alpha + \beta_k + 2M_k + (j-1)} \), which leads to the form \( f(r_{k,i}, k) \) has in Equation (5). Then, \( Z(r_{k,i}) \) is obtained by considering normalizing over the full vector \( f(r_{k,i}, k) \), \( k = \{1, \ldots, K : M_k > 1\} \).

The formulation of \( f(r_{k,i}, k) \) in Equation (4) and Equation (5) intuitively arises when the probability mass function of a beta-binomial random variable with hyperparameters \((\alpha_k, \beta_k)\) is multiplied by the factor \( \frac{B(\alpha_r, \beta_r)}{B(\alpha_r + M_k, \beta_r)} \). 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 \( r_k \) 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 \((\alpha_k, \beta_k)\) in Equation (5), we use a reparameterization where

\[ \pi_k = \frac{\alpha_k}{\alpha_k + \beta_k}, \phi_k = \frac{1}{\alpha_k + \beta_k}. \] (6)

The first parameter \( \pi_k \) 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 \( \phi_k \) represents a measure of variation in the success probability of each observation.\(^{[6]}\)

We constrain the mean success rate \( \pi_k \) in Equation (6) to \( \pi_k \in (0.5, 1) \), which produces beta-binomial distributions with an increasing probability mass function\(^{[7]}\) in the number of compatible sequences \( r_{k,i} \), which leads to the definition in Equation (5) having the same property. Increasing probability mass functions fulfill 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 \((\pi_k, \phi_k)\) are set to \( \pi_k = 0.65, \phi_k^2 = 1 - \pi_k + 0.01 \) 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 \( \theta \) of the different groups using fast collapsed variational inference.\(^{[3]}\) The method collapses the mixing proportions \( \theta \) and uses natural gradients to optimise an approximation to the posterior distribution over the indicator variables \( I_k \), assuming the distribution factors over \( \theta \) and \( I_k \). The same variational Bayesian method was also used in BitSeqVB\(^{[4]}\) to obtain transcript expression levels and has been applied to estimate mixing proportions in bacterial sequencing data in BIB\(^{[5]}\) using a different likelihood. The prior distribution on the mixing proportions \( \theta \) is set to Dirichlet \((\alpha, ..., \alpha)\) with \( \alpha = 1 \). 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 \( \theta_{i,j} \), where \( n = 1, \ldots, N \) (in our examples we chose \( N = 100 \)) indicates the new samples resampled from the reference group \( i = 1, \ldots, K \). The third index \( j = 1, \ldots, K \) 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 \( j \), \( \tilde{\theta}_{i,j} \), in an ascending order in \( n \), and determining the cutoff point \( q_{ij} \) where \( m,m \in [1, \ldots, N] \), 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

\[
q_i = \max \left\{ \max \left\{ j : q_{ij} \right\} \right\},
\]

where \( \epsilon \) 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 S5. We recommend that \( \epsilon \) 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 define a statistical confidence score for the abundance estimates exceeding the detection threshold \( q_i \) as

\[
1 - \frac{(N-m)+1}{N+1} = \frac{m}{N+1},
\]

which corresponds roughly to the fraction of resampled samples that exceed the threshold corresponding to the value of \( m \). Using a value of \( m \) closer to the number of samples \( N \) in constructing the detection thresholds result in stricter thresholds and thus more confidence in the abundance estimates exceeding the threshold. 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)

```bash
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

```bash
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.

```bash
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. Antimicrobial resistance genes were detected using the ARIBA software. 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.v24. This project contains the reference sequences and the grouping used in producing the results.

Data are available under the terms of the Creative Commons Zero “No rights reserved” data waiver (CC0 1.0 Public domain dedication).

Accession numbers for the reference data can be found in Extended Data Table S44. Accession numbers for the 96 Vietnamese E. coli plate sweeps are available in Extended Data Table S54. Accession numbers for the K. pneumoniae mixture samples and 39 Campylobacter mixture samples are available in Extended Data Table S64. The remaining 77 Campylobacter mixture samples have been submitted to Figshare:

- campylobacter_mixtures_1.tgz, https://doi.org/10.6084/m9.figshare.6445136.v15
- campylobacter_mixtures_2.tgz, https://doi.org/10.6084/m9.figshare.6445190.v16

Data are available under the terms of the Creative Commons Attribution 4.0 International license (CC-BY 4.0).

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

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

References


