High-resolution sweep metagenomics using ultrafast read mapping and inference

Tommi Mäklin[1], Teemu Kallonen[2,3], Sophia David[4], Ben Pascoe[5], Guillaume Méric[5], David M. Aanensen[3,6,7], Edward J. Feil[5], Samuel K. Sheppard[5], Jukka Corander[1,2,3], and Antti Honkela[1,8]

1 Helsinki Institute for Information Technology, Department of Mathematics and Statistics, University of Helsinki, Finland
2 Department of Biostatistics, University of Oslo, Norway
3 Wellcome Trust Sanger Institute, Hinxton, Cambridge, United Kingdom
4 Centre for Genomic Pathogen Surveillance, Wellcome Genome Campus, Hinxton, Cambridge, United Kingdom
5 The Milner Center for Evolution, Department of Biology and Biochemistry, Bath University, Bath, United Kingdom
6 Department of Infectious Disease Epidemiology, Imperial College London, London, United Kingdom
7 Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, Nuffield Department of Medicine, University of Oxford, Oxford, United Kingdom
8 Department of Public Health, University of Helsinki, Finland

Abstract

Traditional 16S ribosomal RNA sequencing and whole-genome shotgun metagenomics can determine the composition of bacterial communities on genus level and species level but high-resolution inference on the strain level is challenging due to close relatedness between strain genomes. We present the mSWEEP pipeline for identifying and estimating relative abundances of bacterial strains from plate sweeps of enrichment cultures. mSWEEP uses a database of biologically grouped sequence assemblies as a reference and achieves ultra-fast mapping and accurate inference using pseudoalignment, Bayesian probabilistic modeling, and a control for false positive results. We use sequencing data from the major human pathogens *Campylobacter jejuni*, *Campylobacter coli*, *Klebsiella pneumoniae* and *Staphylococcus epidermidis* to demonstrate that mSWEEP significantly outperforms previous state-of-the-art in strain quantification and detection accuracy. The introduction of mSWEEP opens up a new field of plate sweep metagenomics and facilitates investigation of bacterial cultures composed of mixtures of organisms at differing levels of variation.

Introduction

High-throughput sequencing technologies have enabled researchers to study bacterial populations in unprecedented detail using whole-genome sequencing (WGS) of single pure bacterial colonies. Sequencing of single isolates has revealed complex ecology behind antibiotic resistance and the spread of antibiotic resistant

isolates across the globe. The application of community profiling metagenomics, in which the 16S rRNA gene is sequenced from multi-species samples, can provide information about the composition and dynamics of highly diverse bacterial populations, but the resolution is limited because assignment beyond the level of genus is often not possible due to insufficient nucleotide variation to distinguish species or strains[1]. Whole-genome shotgun metagenomics delivers much higher resolution[2] but widespread application is hindered by costs associated with sequencing the sample to sufficient depth to capture the typically diverse set of organisms present in a sample[3].

We address these challenges by leveraging a solution from a combination of targeted enrichment of the organisms of interest with powerful probabilistic algorithms. Our solution places itself between single colony WGS and culture-independent metagenomics, reaching unprecedented level of biological resolution. Our method, called mSWEEP, identifies and quantifies the presence of bacterial strains from short read sequencing data obtained from plate sweeps of any type of enrichment cultures. These are made by harvesting a mixture of colonies from a plate culture by sweeping the whole plate in contrast to picking a single colony. The generality of the method stems from its applicability to DNA sequenced en massé from arbitrary culturing medium, which may target biological variation ranging from family level to species and strain levels using desired markers. mSWEEP uses a database of biologically grouped sequence assemblies as a reference to which sequencing reads are pseudoaligned[4] and an efficient Bayesian inference engine[5,6] to estimate the relative group abundances. mSWEEP provides statistical confidence scores for group and co-occurrence detections, enabling reliable detection of co-existing strains in the plate sweep samples, which provides the 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

Strain identification method overview

Abundance estimation with mSWEEP is performed in two phases: *reference preparation* and *analysis* (Figure 1). Reference preparation consist of defining the reference database and grouping them according to biological criteria such as

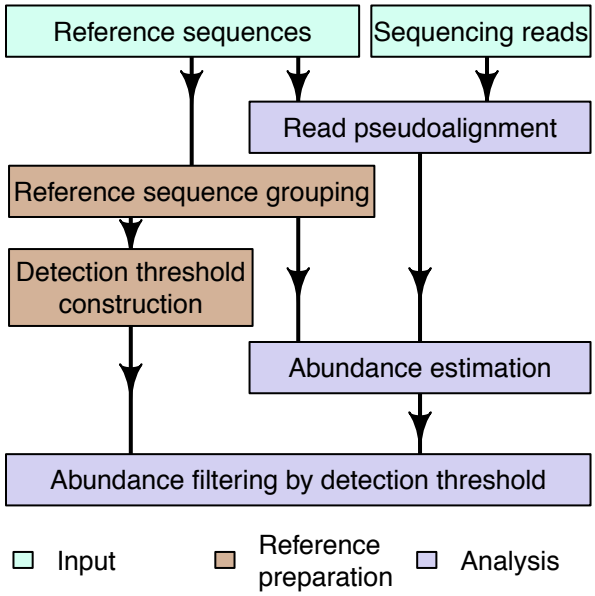

Figure 1 Flowchart of the mSWEEP pipeline describing a typical workflow for relative abundance estimation. The input portion refers to the input data, reference preparation to the operations that need to be performed once per each set of reference sequences, and analysis contains the steps to run for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

sequence types (ST), clonal complexes, 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[6], and accuracy of the mSWEEP abundance estimates is consequently reliant on an expansive reference database and an accurate grouping.

Reference preparation also includes constructing *detection thresholds* on the groups by resampling sequencing reads from the reads used to assemble the reference sequences. We randomly select references from each group within a species, and resample reads from one reference at a time to produce new samples containing reads belonging to only one group. The detection threshold is determined by examining errors observed in the new samples with a known source and maximizing over all possible sources. By determining the minimum relative abundance estimate required for a group to be considered reliably identified, the detection thresholds help avoid incorrectly calling the presence of a group and provide a statistical confidence score for the estimates exceeding the threshold, corresponding to a level of error deemed acceptable in abundance estimates from the resampled sequencing reads. The three steps in reference preparation need to be performed once for a set of reference sequences.

Pseudoalignment[4] in the analysis phase uses the reference sequences to produce binary *compatibility vectors* indicating which reference sequences a read

pseudoaligns with. The compatibility vectors are the basis for estimating the relative abundances of the groups.

Based on the compatibility vectors and the grouping, we define a likelihood for each read to have originated from each of the groups. The likelihood is based on the number of pseudoalignments to each group. We assume that if multiple groups have the same total number of reference sequences in them, then groups with a higher fraction of pseudoalignments are always better candidates for having originated the read. By basing the likelihood on the number of pseudoalignments to each group, we define an extension of the Bayesian model applied to RNA-sequencing[5, 7] and bacterial data[6] that uses multiple reference sequences from each group. Representing the groups with multiple sequences better captures the variation within the groups and enables abundance estimation for species that are difficult to identify with existing methods.

We obtain the relative abundances of the groups by considering a sample as the product of mixing sequencing reads from the groups according to some unknown mixing proportions, corresponding to a mixture model formulation. The model is fitted using a variational Bayesian approach[5] which estimates the mixing proportions (relative abundances) of the groups.

Assigning single-strain isolates to source

We compared the performance of mSWEEP against the existing method for estimating the relative abundances of bacterial groups, BIB[6]. We used 2279 assemblies from three studies[8-10] as the reference set sequences for mSWEEP. mSWEEP outperformed BIB in all five sets of test isolates containing single strains of bacteria.

The reference contained 462 *C. jejuni* and 120 *C. coli* assemblies[10] grouped by sequence type complex (ST complex), 181 *S. aureus* assemblies[8] considered a single group, 143 *S. epidermidis* assemblies[8] grouped by BAPS clustering[11] into a 3-cluster and a 11-cluster grouping, and 1373 *K. pneumoniae* assemblies[9] grouped in ST complexes defined by a central sequence type and its single locus variants. We removed 135 *K. pneumoniae* isolates and one *S. epidermidis*, *C. jejuni*, or *C. coli* isolate at a time, for a total of 81, 73, and 27 test isolates from each species, to obtain within-reference test samples where the true group of origin is known.

To assess accuracy when the test samples have not been sequenced in the same location or do not originate from the same part of the world as the samples used to assemble the reference sequences, we used 61 out-of-reference *K. pneumoniae* isolates from Thailand[12] where the true ST complex is still contained in our reference. Since the *K. pneumoniae* reference sequences originate from Houston, Texas, the isolates from Thailand should be different from the 135 within-reference test samples.

mSWEEP significantly outperformed BIB in all examined cases (Figure 2; $p < 10^{-6}$, in all comparisons, Wilcoxon signed-rank test) when measured by accuracy of the abundance estimates in the true source group and the highest incorrect estimate in each isolate. Both mSWEEP and BIB correctly identified the true ST complex in all 100 *C. jejuni* and *C. coli* isolates and the correct BAPS cluster using the 3-cluster grouping for the 81 *S. epidermidis* isolates. Compared to the abundance estimates from mSWEEP, BIB exhibited considerable uncertainty in both cases.

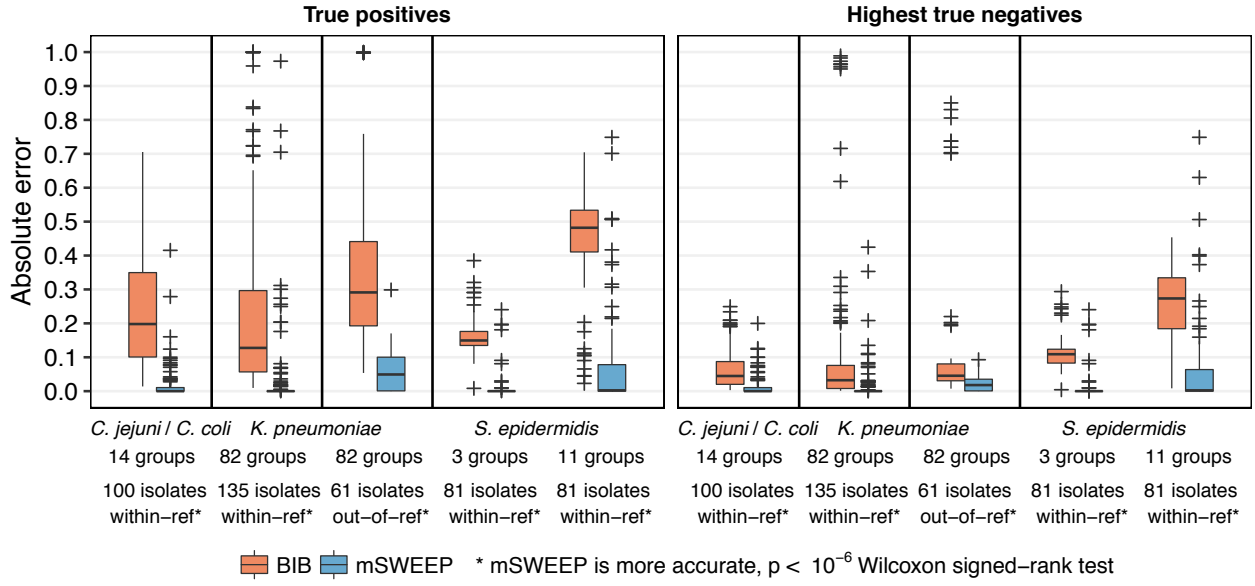

Figure 2 Error of abundance estimates in single-strain isolates (lower is better). True positives represent the relative abundance estimates in the group that is the true source for the isolate. Highest true negatives contain the highest estimate in an incorrect group for the isolates. The absolute error is deviation from an abundance of one (True positives) or zero (Highest true negatives.) mSWEEP significantly outperformed BIB in all examined cases when measured by either type of error. The three outliers in the *K. pneumoniae* isolates are cases where the supposed single-strain isolate in fact contained a mixture or a novel sequence type.

The 135 *K. pneumoniae* within-reference isolates were identified with a similar accuracy with mSWEEP identifying the true group in 132 isolates and BIB in 123. Compared to mSWEEP, the abundance estimates from BIB again contained significant uncertainties. The three *K. pneumoniae* isolates misidentified by mSWEEP were found to contain a novel sequence type or mixtures of *K. pneumoniae* groups and *E. coli* (Supplementary Figure 1).

The least accurate estimates for both BIB and mSWEEP were obtained when the 81 *S. epidermidis* isolates are split into 11 BAPS clusters (Figure 2) with mSWEEP identifying the true group in 78 and BIB in 80 isolates. Neither of the methods reached the level of accuracy observed in the other cases. The inaccuracies are explained by the reference *S. epidermidis* population not exhibiting a clear cluster structure (Supplementary figure 2a) beyond the initial BAPS clustering into three groups, causing the abundance estimates to spread across the new groups defined within the three-cluster split (Supplementary figure 3). The observed behaviour emphasises the need for a reference set that both adequately captures wide variation within the species and has a meaningful grouping.

Abundance estimates from the 61 out-of-reference *K. pneumoniae* isolates resulted in a loss in accuracy for both methods when compared to the 135 within-reference isolates. mSWEEP identified the true origin in all 61 out-of-reference isolates and BIB in 53. The accuracy of mSWEEP and BIB in true positive estimates fell compared to the within-reference isolates but the the errors in true negative cases remained relatively small.

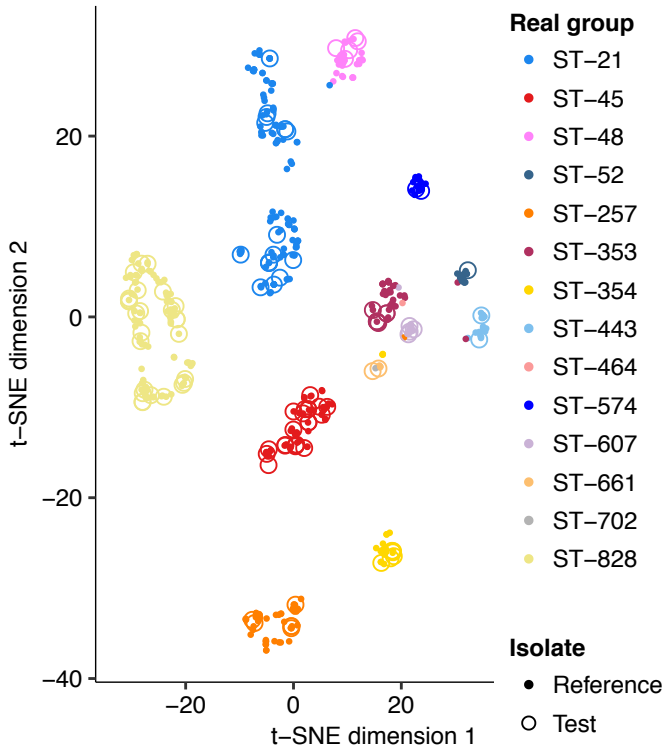

Figure 3 C. jejuni and C. coli reference 31-mer embedding. t-SNE embedding of the 31-mer distances between the reference isolates shows that the reference population conforms relatively well to the defined ST complex grouping. The test cases, indicated by circles, are all correctly identified by mSWEEP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

We examined the grouping of the reference sequences by producing t-SNE plots of *31*-mer distances between the reference sequences including the test isolates (Figure 3, Supplementary figures 2a and 2b). The *C. jejuni* and *C. coli* reference conforms to the ST complex grouping while the *S. epidermidis* population only conforms to the first 3-group BAPS clustering. The t-SNE plots correctly place the assemblies to the true groups but the method does not preserve the distances between the points or the clusters[13] and is unsuited to analyzing mixed isolates.

Processing the 377 single-strain isolates with mSWEEP took an average of 2 minutes and 21 seconds per sample. BIB took over ten times longer with an average of 30 minutes and 42 seconds per sample using the same reference data. Both timings were obtained with eight processor cores, with mSWEEP having 2143 reference sequences in 100 groups and BIB one from each group.

Quantifying synthetic strain mixtures

We investigated the performance of mSWEEP in quantifying mixture samples by constructing 161 *C. jejuni* and *C. coli,* 584 *K. pneumoniae*, and 100 *S. epidermidis* synthetic mixture samples from the single-strain reads used to compare mSWEEP and BIB. Groups included in each mixture sample were determined according to a balanced incomplete block design, ensuring that all groups appear at least 13 times

and each isolate appears at least once. Each sample was set to contain a total of one million reads from three single-strain isolates of a single species with randomly assigned proportions. We examine the distribution of the highest false positive estimates in each mixture sample and compare it to the results from the single-strain isolates.

Abundance estimates obtained from the synthetic mixture samples using the *C. jejuni* and *C. coli* ST complex grouping, the *K. pneumoniae* ST complex grouping, and the 3 BAPS-clusters *S. epidermidis* grouping show that the presence of sequencing reads from multiple groups in a synthetic mixture sample results in an error distribution resembling what was observed in the single-strain isolates (Figure 4, Supplementary Figure 4). Estimates from the synthetic *S. epidermidis* mixture samples using the 11 BAPS-clusters grouping produce an error distribution that differs from the single-strain error distribution more than what was observed with the other groupings.

Comparing the empirical distributions of relative abundance estimation errors from the synthetic mixtures and the single-strain isolates (Supplementary Figures 4 and 5) 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-strain samples, except in the *S. epidermidis* 11-cluster case where stochastic dominance is observed only above a threshold of 0.17. Stochastic

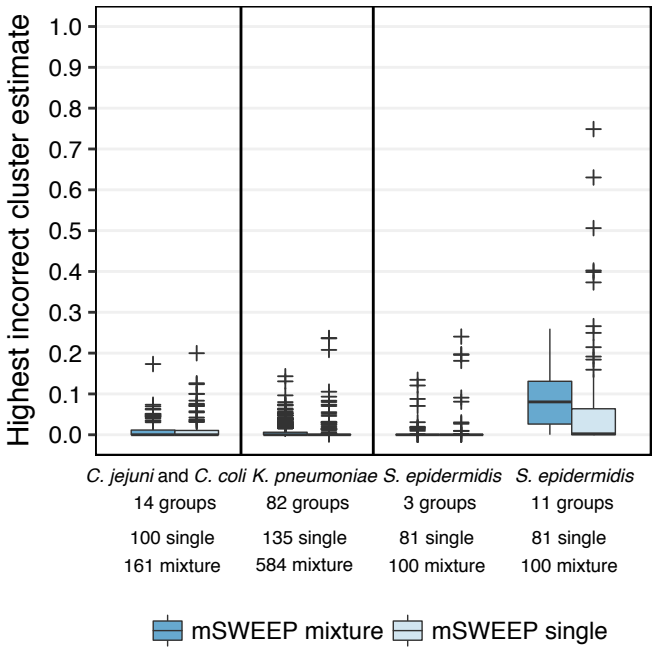

Figure 4 False positives in single-strain samples versus synthetic mixtures. Abundance estimates from synthetic mixtures containing sequencing reads from three groups of a species do not result in higher false positive estimates than what is obtained from isolates from a single group when measured by the largest estimate for a group that does not contribute any sequencing data to the sample. The only different case is *S. epidermidis* using the 11-cluster grouping, which cannot be accurately identified even in the single-strain case.

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-strain 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 (Supplementary Figure 4) 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 what would be obtained from single-strain samples, allowing us to simplify the problem of determining detection thresholds for the groups in mixture samples to determining them from the single-strain 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.

Mixture data from Campylobacter jejuni and Klebsiella pneumoniae isolates

We applied the mSWEEP pipeline to two datasets containing sequencing reads from 116 *C. jejuni* and 179 *K. pneumoniae* mixture samples. Both datasets were originally supposed to contain pure cultures but were flagged in assembly as potential mixtures. Reference data from the previous experiments was expanded with 1509 *E. coli* assemblies[14] and 27 single representative sequences from additional species that were identified in the mixture samples by MetaPhlAn[15]. We constructed detection thresholds on the reference groups using the aforementioned procedure and filtered the relative abundance estimates from the mixture samples by them.

The network diagram (Figure 5a) 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 5b) provides basic information about the frequency that CC's are found together in nature populations. *C. jejuni* CCs 45, 661, 607, 353, 48, and *C. coli* CC828 are all found in samples with 4 or more other CCs

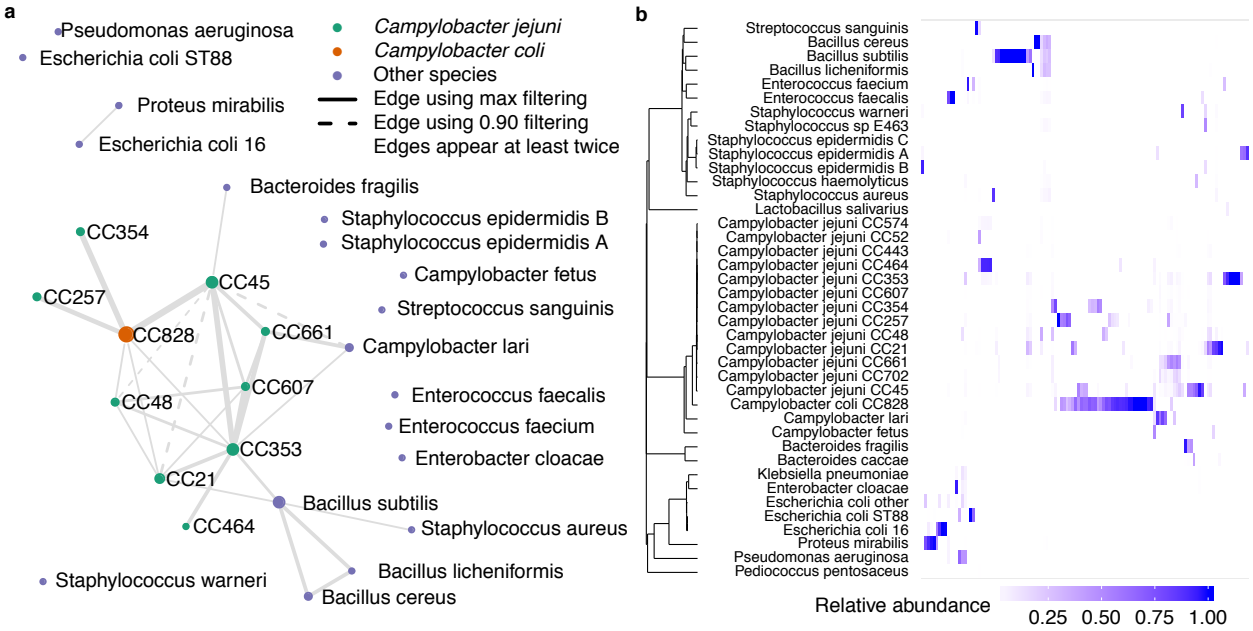

Figure 5 *C. jejuni* clonal complex coexistence in 116 mixture samples. Coexistence network was constructed from the relative abundances that has been filtered by the detection thresholds. The displayed edges (panel a) show groups that were observed to coexist at least two times. Edge and node sizes are proportional to times observed. Dashed edges are present only when using the less strict 0.90 confidence threshold, solid edges when using the 0.90 or the max threshold. Panel b contains a visualization of the unfiltered relative abundance estimates in all reference groups. The tree was constructed by first calculating average *31*-mer distance from sequences assigned to a group to all other sequences, then averaging the resulting distances to the other groups, and last hierarchically clustering the results with the average linkage method.

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 strains can provide information about different ecologies or strain interactions, particularly when CCs are known to have varied sources, such as different hosts.

There is some preliminary evidence that common clinical strains 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.

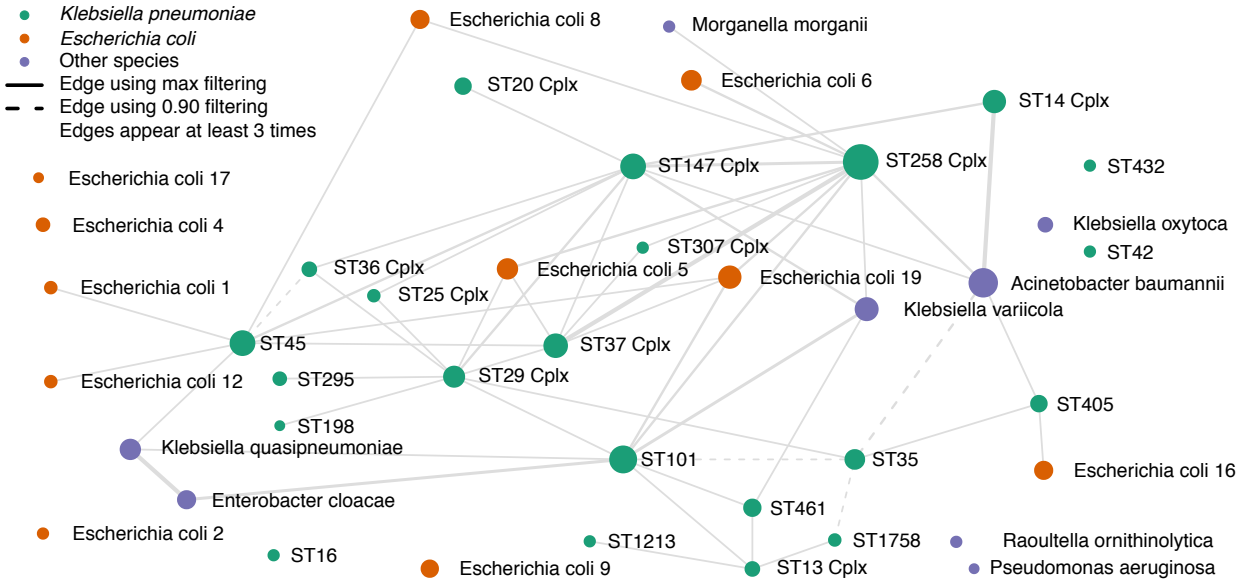

Figure 6 *K. pneumoniae* strain group coexistence in 179 mixture samples. Coexistence network was constructed from the relative abundance estimates remaining after filtering by detection threshold. Edges shown are observed at least three times. Edge and node size is proportional to number of times observed. Dashed edges represent coexistence that is observed only when using the 0.90 confidence detection thresholds.

The coexistence network for the 179 human clinical samples of *K. pneumoniae* (Figure 6, Supplementary Figure 6) demonstrates common co-occurrence of *K. pneumoniae* with a wide variety of *E. coli* strain groups, as well as occasional co-occurrence with *Acinetobacter baumanii* and other species. Both *E. coli* and *A. baumanii* grow on the media used for culture of *K. pneumoniae*. Clonal complexes of *K. pneumoniae* 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* strains 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 anti-microbial resistance genes between taxa.

Discussion

Metagenomics using high-throughput sequencing has become common practice when investigating bacterial composition in different environments or changes introduced by interventions to e.g. the human gut microbiome. In most epidemiological applications the relevant target organisms are culturable using established media which offers a clear advantage to obtain high sequencing depths in a cost-effective manner. To enable high-resolution inference about the strains

present in plate sweeps of enrichment cultures, we have developed the mSWEEP pipeline. mSWEEP can be used to infer high-resolution population structure of single species or diverse populations of bacteria. Our pipeline also gives estimates of the relative abundance of strains and estimate reliability cut-offs. mSWEEP has been designed to have minimal execution time using the latest advances in RNA-seq analysis and having minimal memory footprint so that most typical analyses can be run on a regular laptop, and we have demonstrated significant improvements in accuracy over previous state-of-the-art method.

mSWEEP has considerable power for improving our understanding of infection by recovering a true representation of bacteria in a sample. Genotyping studies have shown that *C. jejuni* and *C. coli* strains colonizing the primary host (birds and mammals) form clusters of related isolates that are host-associated[16] which can be used to identify the reservoir for human infection[17]. However, multiple strains can be isolated from the same sources[18, 19]. The co-occurrence of multiple strains could be a snapshot in time of a wider process of lineage succession[20] in which the resident microbiota might resist new colonizations or be displaced by incoming bacteria[21, 22] or indicate complex interactions between strains that occupy different microniches[23] and are not in direct competition[24, 25]. mSWEEP provides means to investigate the nature of polymicrobial infections which could improve understanding of the spread of a strain 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 abundances, not absolute. However, this is not a significant limitation as even the absolute abundances of target organisms are subject to large biological and technical variation. Large reference collections of over 1 000 genomes or simultaneous analysis of multiple samples increase the memory requirement and may require a dedicated computer cluster to run, but these are still at the level available at most bioinformatics centres, which makes the method widely applicable by biologists. As with any method targeting to identify sequence variation, the target species need to be relatively well known to allow building 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 strains. 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[26]. mSWEEP is complementary to StrainPhlAn as the two methods analyse similar data from different directions. mSWEEP assigns strains present in the sample to biologically established genetically separated clades or clusters and estimates the relative abundance of these, whereas StrainPhlAn infers SNPs and phylogenetic relations of the whole sample. Given the flexibility and generality of the mSWEEP approach, we anticipate that it paves way for numerous novel applications of plate sweep metagenomics in many fields of microbiology.

Online Methods

Reference construction

The reference sequences (Supplementary Table 1) are the genomic assemblies of a number of strains or species that represent the organisms of interest in a sample. We use a collection of assemblies from various studies[8-10, 14].

Grouping the reference sequences

We grouped the *C. jejuni* and *K. pneumoniae* reference sequences into clonal complexes defined by a central sequence type and its single locus variants. The *S. epidermidis* and *E. coli* sequences were clustered using the BAPS software[11] (version 6.0). We split the *S. epidermidis* and *E. coli* populations according to the first clustering level produced by BAPS.

Pseudoalignment

We used kallisto[4] (version 0.44) with default settings to perform pseudoalignment. Pseudoalignment produces binary compatibility vectors which summarize the observations of the reads $r_n = (r_{n,1}, \ldots, r_{n,k})$ as numbers of compatible sequences (observed pseudoalignments) in each group $r_{n,k}$ for a read r_n.

Abundance estimation model

The reads r_n are modelled by K-dimensional count vectors containing the count of pseudoalignments to reference sequences in each of the K groups. 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

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

The formulation in Equation (1) corresponds to a mixture model with observations r_n, categorically distributed latent variables I_n and their parameters θ.

Likelihood

Given a pseudoalignment count vector r_n, whose components $r_{n,k}$ denote the count of observed pseudoalignments to reference sequences in each group k, and the total number of reference sequences in the group, $|k|$, we define the likelihood $p(r_n \mid I_n = k)$ of the read that produced the count vector r_n originating from the k:th group as

$$p(r_n \mid I_n = k) \propto \binom{|k|}{n} \frac{B(\alpha_k + r_{n,k}, |k| - r_{n,k} + \beta_k)}{B(\alpha_k, \beta_k)} \frac{B(\alpha_k, \beta_k)}{B(\alpha_k + |k|, \beta_k)}, \quad (2)$$

where $B(\alpha, \beta)$ is the beta function, when the group k contains at least two reference sequences $|k| > 1$ and at least one reference sequence in group k is compatible with the read r_n, $0 < r_{nk} \leq |k|$. When the group contains only one reference sequence and that sequence is compatible with the read r_n, the likelihood is set to $p(r_n | I_n = k) \propto 1$. In both cases, when the group contains no compatible sequences, regardless of the total number, the likelihood is set to $p(r_n \mid I_n = k) \propto 0.01/0.99$.

Equation (2) is based on the beta-binomial model, but the likelihoods are renormalized by the factor $\frac{B(\alpha_k, \beta_k)}{B(\alpha_k + |k|, \beta_k)}$, which represents the likelihood of a read given compatibility with all sequences in the group. The renormalization causes groups where $r_{n,k} = |k|$ in at least two groups, that is all reference sequences in at least two groups are compatible with the read, to have equal likelihoods and also reduces the effect of the likelihoods being flattened in groups with large numbers of assigned reference sequences when compared to small groups, allowing the model to better compare group that differ largely in size.

Reads with identical pseudoalignment count vectors r_n have the same likelihoods and can be assigned into equivalence classes defined by the count of compatible sequences in each group. The likelihoods need only be calculated for the observed equivalence classes.

Model hyperparameters

Instead of using the parametrization (α_k, β_k) of Equation (2), we reparametrize the likelihood as

$$\pi_k = \frac{\alpha_k}{\alpha_k + \beta_k}, \phi_k = \frac{1}{\alpha_k + \beta_k} \quad (3)$$

where the first parameter π_k corresponds to the mean of the beta distribution compounded with a binomial distribution to obtain the beta-binomial distribution part in Equation (2), and the second parameter ϕ_k represents a measure of variation in the success probability of each observation[27].

We constrain the mean success rate π_k to $\pi_k \in (0.5, 1)$, producing only distributions with an increasing probability mass function[28], fulfilling our assumption 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 $\pi_k = 0.65$, and $\phi_k^{-1} = 1 - \pi_k + 0.01|k|^{-1}$.

Inference

Obtaining the posterior distribution over the mixing proportions θ of the different groups is done by fitting an approximate posterior distribution over the indicator variables I using variational inference. The same variational Bayesian method is also used in BitSeqVB[5] to obtain transcript expression levels and has been applied to estimate mixing proportions in bacterial sequencing data in BIB[6]. The prior distribution on the mixing proportions θ is set to be Dirichlet$(\alpha, \ldots, \alpha)$ 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 the minimum relative abundance estimate in each group to be considered a reliable identification. Relative abundance estimates that fall under the corresponding detection threshold are set to zero. To obtain detection thresholds on the groups within a species, we generate 100 samples from each group by resampling one million sequencing reads from the reads used to assemble the reference sequences. Each sample contains resampled sequencing reads from only one reference sequence. The reads are sampled with replacement and each read has the same probability of being included. Reference sequences used in the resampling are 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 and each group is represented at least once. The resampled sequencing reads are put through mSWEEP abundance estimation. Reference sequences used in resampling are not included in the set of reference sequences used when performing estimation on the resampled sequencing reads. Species in the reference represented by a single sequence were not resampled from and the detection threshold fixed at 0.05.

The relative abundance estimates obtained from the resampled sequencing reads are represented by $\hat{\theta}_{\mathrm{i,j,k}}$, where $i = 1,\ldots,100$ indicates samples that were sampled from the reference group $j = 1,\ldots,K$, and $k = 1,\ldots,K$ denotes the reference group that the abundance estimate was observed for. To obtain the detection thresholds, we first define source-specific thresholds $q_{j,k}$ that give a threshold on the reference groups k assuming that the true group j in the sample is known. The source-specific threshold $q_{j,k}$ on group $k \neq j$ is defined by ordering the relative abundance estimates for the cluster k, $\hat{\theta}_{\mathrm{i,j,k}}$, where $i = 1,\ldots,100$, in an ascending order and determining the cutoff point where $100p$, $p \in (0,1)$, relative abundance estimates fall below the cutoff. Using the source-specific thresholds $q_{j,k}$, we define the detection threshold q_j on group j as $q_j = max\{ max\{ k : q_{j,k}\}, \epsilon\}$, where ϵ is the constant minimum threshold for the species that was observed when comparing the empirical cumulative distribution functions in Supplementary Figure 5. 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-strain counterparts.

The quantile p, used to determine the cutoff points for the source-specific thresholds provides a statistical confidence score for the abundance estimates that exceed the detection thresholds q_j. Values of p closer to one result in stricter detection thresholds and more confidence in the remaining abundance estimates.

Software and data availability

mSWEEP software is available in GitHub: https://github.com/PROBIC/mSWEEP. Accession numbers for the reference data can be found in Supplementary Table 1. Accession numbers for the *K. pneumoniae* mixture samples and 39 *Campylobacter* mixture samples are available in Supplementary Table 2. The remaining 77 *Campylobacter* mixture samples have been submitted to figshare under DOIs 10.6084/m9.figshare.6445136 and 10.6084/m9.figshare.6445190.

Acknowledgements

TM and AH were supported by the Academy of Finland grants no. 259440 and 310261. TK, JC, DA and EF are supported by the JPI-AMR consortium SpARK (MR/R00241X/1). JC was funded by the ERC grant no. 742158. T.K. was funded by the Norwegian Research Council JPIAMR grant no. 144501.

Author Contributions

T.M., J.C., and A.H. developed the model and designed the method comparison experiments. T.M. implemented the model and ran the analyses. T.K. built the K. pneumoniae and E. coli references. S.D., G.M., D.M.A., E.J.F., and S.K.S. collected the mixture data and interpreted the results from it. T.M., T.K., S.D., S.K.S., J.C., and A.H. wrote the paper. J.C. and A.H. conceived the study. All authors discussed the results and commented on the manuscript.

References

1. Ellegaard, K.M. & Engel, P. Beyond 16S rRNA Community Profiling: Intra-Species Diversity in the Gut Microbiota. *Front Microbiol* **7**, 1475 (2016).
2. Quince, C., Walker, A.W., Simpson, J.T., Loman, N.J. & Segata, N. Shotgun metagenomics, from sampling to analysis. *Nature biotechnology* **35**, 833-844 (2017).
3. Yang, X. et al. Use of Metagenomic Shotgun Sequencing Technology To Detect Foodborne Pathogens within the Microbiome of the Beef Production Chain. *Appl Environ Microbiol* **82**, 2433-2443 (2016).

4. Bray, N.L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34**, 525-527 (2016).
5. Hensman, J., Papastamoulis, P., Glaus, P., Honkela, A. & Rattray, M. Fast and accurate approximate inference of transcript expression from RNA-seq data. *Bioinformatics* **31**, 3881-3889 (2015).
6. Sankar, A. et al. Bayesian identification of bacterial strains from sequencing data. *Microb Genom* **2**, e000075 (2016).
7. Glaus, P., Honkela, A. & Rattray, M. Identifying differentially expressed transcripts from RNA-seq data with biological variation. *Bioinformatics* **28**, 1721-1728 (2012).
8. Meric, G. et al. Ecological Overlap and Horizontal Gene Transfer in Staphylococcus aureus and Staphylococcus epidermidis. *Genome Biol Evol* **7**, 1313-1328 (2015).
9. Long, S.W. 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* **8** (2017).
10. Yahara, K. et al. Genome-wide association of functional traits linked with Campylobacter jejuni survival from farm to fork. *Environ Microbiol* **19**, 361-380 (2017).
11. Cheng, L., Connor, T.R., Siren, J., Aanensen, D.M. & Corander, J. Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. *Mol Biol Evol* **30**, 1224-1228 (2013).
12. Runcharoen, C. et al. Whole genome sequencing reveals high-resolution epidemiological links between clinical and environmental Klebsiella pneumoniae. *Genome Med* **9**, 6 (2017).
13. Maaten, L.v.d. & Hinton, G. Visualizing data using t-SNE. *Journal of machine learning research* **9**, 2579-2605 (2008).
14. Kallonen, T. 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).
15. Segata, N. et al. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat Methods* **9**, 811-814 (2012).
16. Sheppard, S.K. et al. Niche segregation and genetic structure of Campylobacter jejuni populations from wild and agricultural host species. *Mol Ecol* **20**, 3484-3490 (2011).
17. Sheppard, S.K. et al. Campylobacter genotyping to determine the source of human infection. *Clin Infect Dis* **48**, 1072-1078 (2009).

18. Sproston, E.L. et al. Temporal variation and host association in the Campylobacter population in a longitudinal ruminant farm study. *Applied and environmental microbiology* **77**, 6579-6586 (2011).
19. Colles, F.M., McCarthy, N.D., Layton, R. & Maiden, M.C.J. The prevalence of Campylobacter amongst a free-range broiler breeder flock was primarily affected by flock age. *PLoS One* **6**, e22825 (2011).
20. Lu, J. et al. Diversity and succession of the intestinal bacterial community of the maturing broiler chicken. *Applied and environmental microbiology* **69**, 6816-6824 (2003).
21. Buffie, C.G. & Pamer, E.G. Microbiota-mediated colonization resistance against intestinal pathogens. *Nat Rev Immunol* **13**, 790-801 (2013).
22. Nowrouzian, F.L., Wold, A.E. & Adlerberth, I. Escherichia coli strains belonging to phylogenetic group B2 have superior capacity to persist in the intestinal microflora of infants. *J Infect Dis* **191**, 1078-1083 (2005).
23. Hayashi, H., Takahashi, R., Nishi, T., Sakamoto, M. & Benno, Y. 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* **54**, 1093-1101 (2005).
24. Johns, B.E., Purdy, K.J., Tucker, N.P. & Maddocks, S.E. Phenotypic and Genotypic Characteristics of Small Colony Variants and Their Role in Chronic Infection. *Microbiol Insights* **8**, 15-23 (2015).
25. von Bronk, B., Schaffer, S.A., Gotz, A. & Opitz, M. Effects of stochasticity and division of labor in toxin production on two-strain bacterial competition in Escherichia coli. *PLoS Biol* **15**, e2001457 (2017).
26. Truong, D.T., Tett, A., Pasolli, E., Huttenhower, C. & Segata, N. Microbial strain-level population structure and genetic diversity from metagenomes. *Genome research* **27**, 626-638 (2017).
27. Griffiths, D.A. Maximum Likelihood Estimation for Beta-Binomial Distribution and an Application to Household Distribution of Total Number of Cases of a Disease. *Biometrics* **29**, 637-648 (1973).
28. Berg, S. Condorcets Jury Theorem, Dependency among Jurors. *Soc Choice Welfare* **10**, 87-95 (1993).