

Culture-enriched metagenomic sequencing enables in-depth profiling of the cystic fibrosis lung microbiota

Fiona J. Whelan¹, Barbara Waddell², Saad A. Syed³, Shahrokh Shekariz³, Harvey R. Rabin^{2,4}, Michael D. Parkins^{2,4} and Michael G. Surette^{2,3,5*}

Amplicon sequencing (for example, of the 16S rRNA gene) identifies the presence and relative abundance of microbial community members. However, metagenomic sequencing is needed to identify the genetic content and functional potential of a community. Metagenomics is challenging in samples dominated by host DNA, such as those from the skin, tissue and respiratory tract. Here, we combine advances in amplicon and metagenomic sequencing with culture-enriched molecular profiling to study the human microbiota. Using the cystic fibrosis lung as an example, we cultured an average of 82.13% of the operational taxonomic units representing 99.3% of the relative abundance identified in direct sequencing of sputum samples; importantly, culture enrichment identified 63.3% more operational taxonomic units than direct sequencing. We developed the PLate Coverage Algorithm (PLCA) to determine a representative subset of culture plates on which to conduct culture-enriched metagenomics, resulting in the recovery of greater taxonomic diversity—including of low-abundance taxa—with better metagenome-assembled genomes, longer contigs and better functional annotations when compared to culture-independent methods. The PLCA is also applied as a proof of principle to a previously published gut microbiota dataset. Culture-enriched molecular profiling can be used to better understand the role of the human microbiota in health and disease.

The field of microbiology began with the visualization of microorganisms¹ and continued once we learned to control their growth. The advent of next-generation sequencing revolutionized microbiology by allowing microbial genomics and community analysis without the requirement of culture. These technologies have expanded our understanding of the human microbiome, identifying the effects that environment, diet and host genetics have on these complex communities^{2–5}. Culture-independent studies have made crucial contributions to microbiome research. However, without bacterial culture, we lose the advantages of classical microbiology, resulting in a lack of mechanistic studies and an alternative focus on microbial ‘dysbiosis’ and microbiome diversity^{6,7}.

Although it is commonly stated that the majority of the human microbiota is unculturable, numerous studies conclusively counter this. In 1974, Finegold and others cultured ~300 species from 40 faecal specimens using both aerobic and anaerobic culture⁸. More recently, Goodman and others cultured almost half of the human gut microbiota, recovering 316 operational taxonomic units (OTUs) of the 631 identified by culture-independent techniques⁹. Lagier and others cultured 340 species of bacteria from three stool samples¹⁰. Furthermore, recent studies recovered 88% of family-level OTUs¹¹ and 95% of all OTUs identified in faecal specimens¹² (importantly, both of these studies also identified more OTUs via culture-dependent than culture-independent methods). Other human-associated communities have been profiled with culture, including urine¹³, skin¹⁴, oral¹⁵ and cystic fibrosis communities¹⁶.

Marker gene profiling (for example, 16S rRNA gene sequencing) provides a simple and rapid method to assess the taxonomic

composition of a community. Although metagenomics can assess functional capacity in addition to taxonomy, the usefulness of these data is dependent on how well short-read sequences can be assembled into contigs. The quality of assembly can be impacted by the complexity of the community, the sequencing technology and/or the proportions of host DNA contamination^{17–19}. Culture enrichment is uniquely positioned to improve metagenomic assembly by allowing the proliferation of microorganisms (thus abating host DNA contamination) on media that ‘biologically bin’ samples, thus decreasing their complexity. Coupling culture enrichment with computational approaches allows us to separate promiscuous and fastidious microorganisms to better resolve these communities. Here, we use sputum from cystic fibrosis as an example of a complex microbial community where host DNA can represent a large proportion of the community (>99%)^{20–23}. We present a culture-enriched metagenomic strategy, which overcomes these limitations and improves metagenomic results, providing a more comprehensive profile of the microbial community.

In this study, we merge culture-dependent and -independent techniques to better understand microbial communities. Culture-enriched 16S rRNA gene sequencing established that 82.13% of all OTUs—representing 99.3% of the relative abundance—in the cystic fibrosis lung microbiota are culturable. Furthermore, culture enrichment increased OTU recovery when compared to culture-independent sequencing. We introduce the PLate Coverage Algorithm (PLCA), which uses 16S rRNA gene sequencing to optimize culture-enriched metagenomics, and show that culture-enriched metagenomics improves the recovery of metagenomic-assembled

¹School of Life Sciences, University of Nottingham, Nottingham, UK. ²Department of Microbiology, Immunology and Infectious Diseases, University of Calgary, Calgary, Alberta, Canada. ³Department of Biochemistry and Biomedical Sciences, McMaster University, Hamilton, Ontario, Canada.

⁴Department of Medicine, University of Calgary, Calgary, Alberta, Canada. ⁵Department of Medicine, McMaster University, Hamilton, Ontario, Canada.

*e-mail: surette@mcmaster.ca

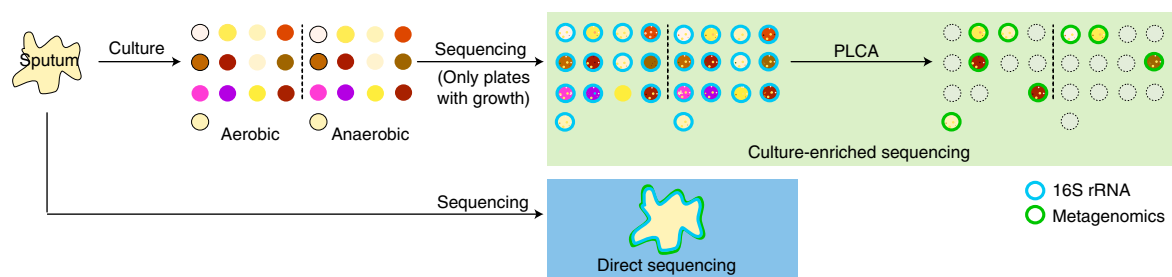


Fig. 1 | The culture-enriched metagenomic sequencing workflow. Sputum samples collected from cystic fibrosis patients were plated onto 13 selective and non-selective media and incubated either aerobically or anaerobically. 16S rRNA gene sequencing was conducted on the sputum sample (direct sequencing) as well as on each media type (culture-enriched sequencing).

genomes and produces more thorough functional annotations than direct metagenomic approaches. We identify the advantages of culture-enriched metagenomics: increased taxonomic and functional information due to a decrease in contaminating host DNA and the ability to computationally and biologically bin microbial species.

Results

In this study, we devised a strategy for culture-enriched metagenomic profiling (Fig. 1). This strategy is based on previous culture enrichment for amplicon-based sequencing conducted on the cystic fibrosis lung microbiome, which was able to culture 43 of 48 bacterial families identified in sputum¹⁶. On collection, samples were immediately plated onto 13 different media types (Extended Data Fig. 1) under aerobic and anaerobic conditions. 16S rRNA gene sequencing was performed on the sputum sample (direct sequencing) as well as on the collective organisms grown in each medium/environment pairing (culture-enriched sequencing) for a total of 26 culture-enriched samples per sputum (Fig. 1). The OTU diversity in the direct sequencing and distribution in the culture-enriched sequencing was used in conjunction with the PLCA (details below) to determine a representative subset of culture-enriched plates that adequately represented the sample. Shotgun metagenomic sequencing was performed on the original sample and the culture-enriched subset.

The majority of the cystic fibrosis lung microbiota is culturable. We first identified the culturable fraction of the cystic fibrosis lung microbial community in 20 sputum samples from 10 patients. We defined an OTU to be culturable if it contained ≥ 10 reads and was recovered from ≥ 1 culture-enriched plate at a relative abundance of $\geq 0.01\%$. Across the dataset, an average of 82.13% (range, 64.6–100%) of OTUs identified by direct sequencing were culturable; this culturable fraction represents an average of 99.3% (range, 97.8–100%) of the relative abundance in the associated direct sequencing results (Fig. 2a). When the genus-level OTU taxonomic assignments were compared to a list of species previously identified via culture enrichment²⁴, and to previous culture enrichment of the cystic fibrosis lung¹⁶ and gut¹², we identified 18 genera cultured in this study that had not previously been identified via large-scale culture enrichment (Supplementary Table 1). Of the OTUs that were never cultured across the dataset, we observed an over-representation of the Spirochaetes (7 of 7 OTUs identified in culture-independent sequencing) and the SR1 (1 of 1), and many members of the Tenericutes (7 of 22) and Saccharibacteria (2 of 3) phyla (Fig. 2b, blue ring). Together, these results indicate that most OTUs in the cystic fibrosis lung microbiota are culturable and that those OTUs that are not cultured are taxonomically restricted, and historically challenging to culture. Importantly, culture-independent methods do not distinguish DNA from viable versus non-viable organisms, so some bacteria not recovered may not be viable in these samples.

These results still hold if a more stringent definition of culturable is used (Extended Data Fig. 2a,b).

Culture enrichment increases OTU recovery. Culture-enriched 16S rRNA gene sequencing consistently recovered more OTUs than direct sequencing (Fig. 2b and Extended Data Fig. 2c–e). For example, in the direct sequencing of sample 1, 49 OTUs were recovered, 42 of which were also identified by culture enrichment (Fig. 2a); in addition to these 42 OTUs, an additional 124 OTUs were identified by culture enrichment (Extended Data Fig. 2c). This enrichment in OTU recovery did not correlate with variability in the α -diversity of the original sample (Extended Data Fig. 2e). We hypothesized that the ability to enrich may be due to the recovery of low-abundance taxa. To test this hypothesis, we resequenced a sample to a depth 24 times deeper than the original direct sequencing (972,834 versus 41,199 reads) and rarefied at decreasing depths (range, 500,000–1,000 reads). We observed that the number of OTUs recovered only by culture decreases as the sequencing depth increases (Extended Data Fig. 3). These cultured OTUs were typical members of the cystic fibrosis lung microbiota, including *Streptococcus* sp., *Prevotella* sp. and *Veillonella* sp., indicating that culture allows for the enrichment of taxa present at low abundance in the original sample.

Culture enrichment's increase in OTU recovery is dependent on medium type and oxygen availability. The variety of media and environmental conditions used is important in capturing the diversity of microbial communities. The use of both anaerobic and aerobic conditions encourages the recovery of different taxa, as is evident in the taxonomic distribution and β -diversity relationships of the 16S rRNA gene sequencing results of direct and culture-enriched sequencing (Fig. 3a,b). For example, *Veillonella* sp. and *Prevotella* sp. were recovered exclusively under anaerobic conditions; conversely, *Rothia* sp. and *Pseudomonas* sp. were obtained at greater abundances in aerobic culture (Fig. 3a). The α -diversity of each culture condition (medium + aerobic/anaerobic environment) varied with each sample (Supplementary Fig. 2) and no single condition consistently best recapitulated the originating sputum sample (Extended Data Fig. 4). Hierarchical clustering of the taxa recovered from each culture condition further indicates the importance of using both selective and non-selective culture conditions (Fig. 3c and Extended Data Figs. 5 and 6). Although some organisms, such as *Streptococcus* sp., will grow under many conditions, others, such as *Neisseria* sp., *Rothia* sp. and *Stenotrophomonas* sp. were only recovered from a subset of culture conditions. Furthermore, there were also OTU-dependent differences in growth patterns within some genera (for example, *Prevotella* OTUs; Fig. 3d and Extended Data Fig. 6). Across the dataset, anaerobic culture recovered almost half of all cultured OTUs; furthermore, we nearly doubled the number of recovered OTUs by expanding our culture conditions beyond the four typically employed in a clinical laboratory (Fig. 3e).

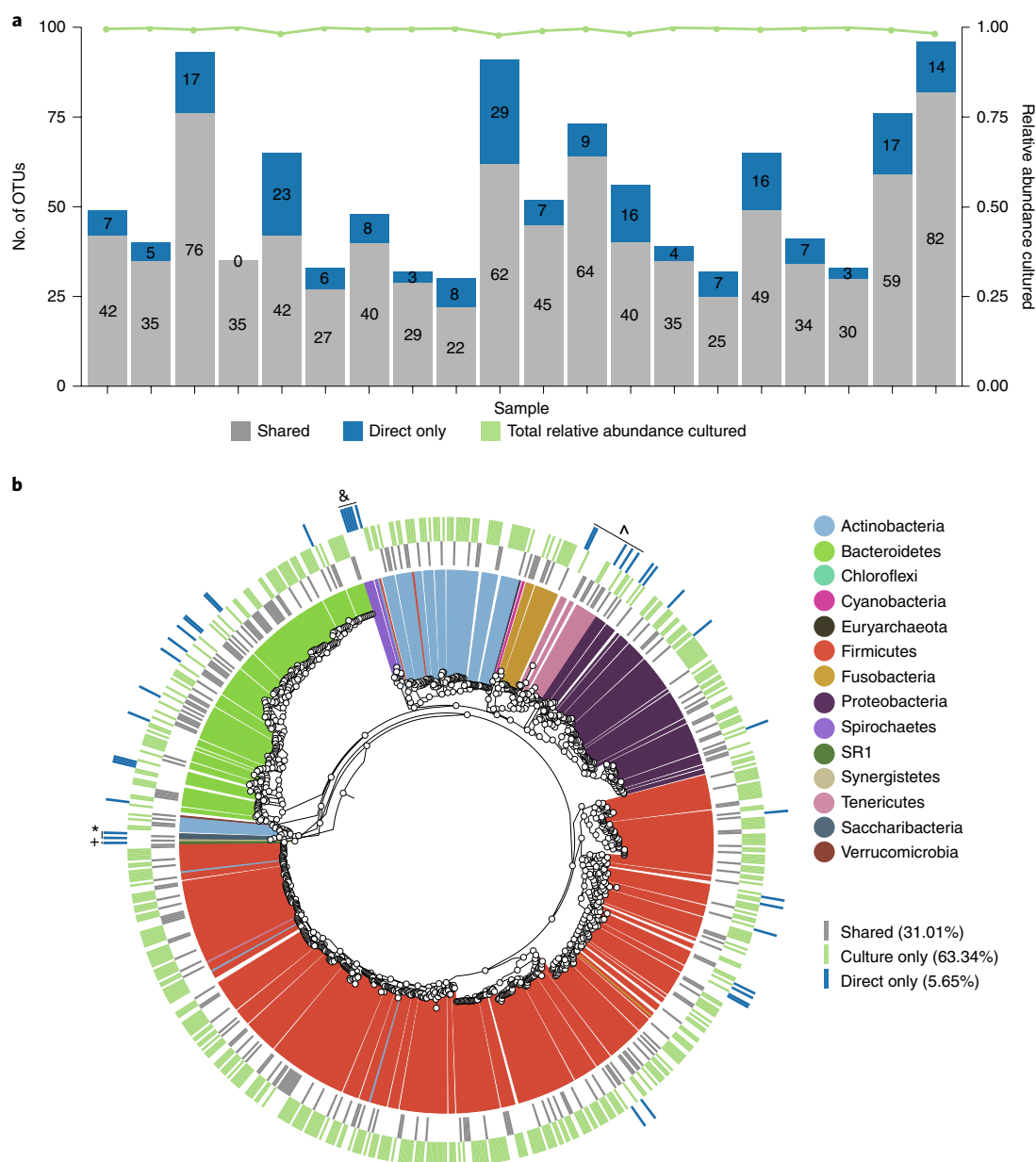


Fig. 2 | The majority of the cystic fibrosis lung microbiota is culturable. a, The majority of OTUs identified in the direct 16S rRNA gene sequencing were also recovered by culture-enriched sequencing. **b**, In fact, 63.3% of OTUs across the dataset were identified only by culture-enriched sequencing (green ring). In contrast, 5.7% of OTUs were not cultured, including many Tenericutes (*) and Saccharibacteria (*) and all Spirochaetes (&) and SR1 (+). Similar results were obtained with a more stringent relative abundance cutoff (Extended Data Fig. 1a,b).

Another advantage of culture enrichment is that it allows for the post hoc recovery of organisms of interest from frozen bacterial stocks (see Methods). As an example, *Stenotrophomonas* sp. were isolated from two bacterial stocks in which they accounted for relatively low relative abundances (1.3 and 1.5%). To facilitate recovery, the stocks were replated on a medium type specific to *Stenotrophomonas* sp. (Fig. 3f). Taxonomic assignment of isolated colonies as *Stenotrophomonas maltophilia* was confirmed via full-length 16S rRNA gene sequencing (Supplementary Table 2).

The PLCA informs culture-enriched metagenomic sequencing.

The conditions used for culture enrichment are necessarily broad; the cystic fibrosis lung microbiota, like other human-associated communities, can be composed of a wide range of organisms, from common pathogens (for example, *Pseudomonas*, *Staphylococcus* and

*Haemophilus*²⁵) to anaerobes (for example, *Prevotella*, *Fusobacterium* and *Veillonella*^{26–28}) and emerging pathogens (*Stenotrophomonas maltophilia* and *Achromobacter*²⁹). Although the lung could harbour any of these organisms, an individual's microbiota is a unique subset of these possibilities. This means that while the variety of culture conditions used is necessary to capture the diversity across individuals, not every plate is needed to enrich the microbiota of a given sample. This is also true of other human-associated communities; the gut, for example, harbours many organisms that are prevalent across the population, whereas some species can be quite specific to the individual³⁰. It is not known a priori which subset of plates would best recapitulate a given community; however, we can use 16S rRNA gene sequencing to identify the taxonomic distribution across cultured plates, and to choose the subset of plates that best recapitulates the community on which to conduct culture-enriched

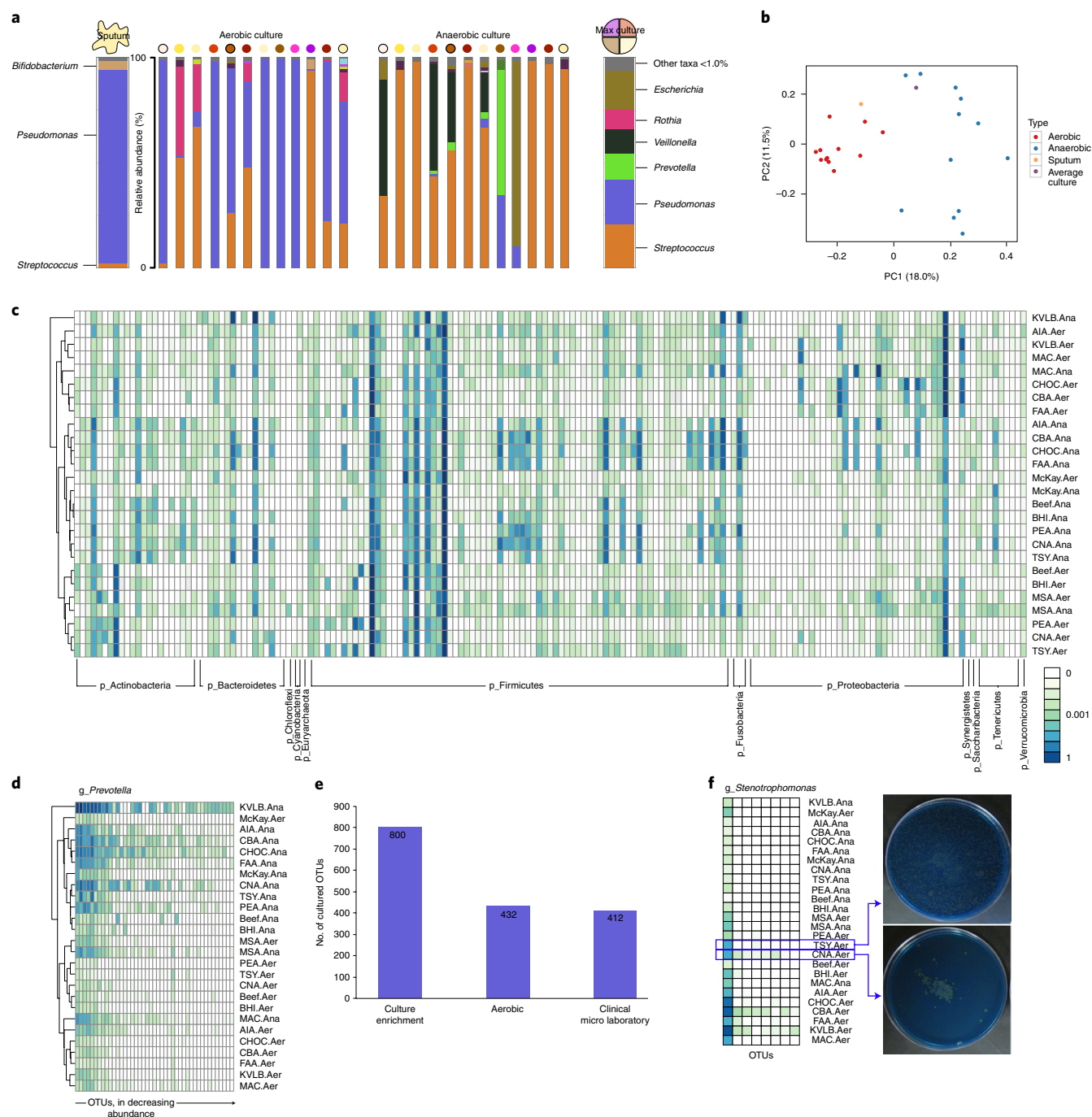


Fig. 3 | Taxonomic diversity captured across culture enrichment conditions. a, b, The use of both anaerobic and aerobic conditions allows greater organism recovery, as shown by a pair of direct and culture-enriched sequencing of a representative sample (taxonomic summaries (**a**) and unweighted Unifrac β -diversity metric ($n=26$, **b**)). In **a**, culture-enriched plates, indicated with circles as in Fig. 1, are displayed alphabetically (Extended Data Fig. 1). In **b**, the direct sequencing clusters with some aerobic samples due to the abundance of *Pseudomonas* sp., and not due to a lack of bacterial growth (Supplementary Fig. 1). **c**, A heatmap showing the maximum observed relative abundance (range 0-1) of each phylum across culturing conditions. Aerobic (Aer) and anaerobic (Ana) culture condition acronyms and recipes are provided in the Methods. Genus-level labelling is available in Extended Data Fig. 5. **d**, Within a genus, different OTUs can also have different culture preferences (Extended Data Fig. 5). **e**, The number of OTUs obtained from culture enrichment is compared to the number obtained if only aerobic culturing was used, or if culture was restricted to that of a standard clinical microbiology laboratory (CBA.Aer, MAC.Aer, MSA.Aer, CHOC.Ana). **f**, Cultured organisms can be recovered from frozen bacterial stocks. Here, *Stenotrophomonas* sp. were isolated from stocks with a relative abundance of 1.3% on CNA.Aer and 1.5% on TSY.Aer, respectively. Plates with low abundance of *Stenotrophomonas* sp. were purposefully chosen to indicate the power of this approach.

metagenomic sequencing. As such, we implemented the PLCA, which determines the minimum number of culture-enriched plates necessary to capture the taxonomic diversity of a sample with

culture-enriched metagenomics. The PLCA is not specific to a particular set of culture conditions, or particular microbiome, but instead can be used with any community that contains a culturable majority.

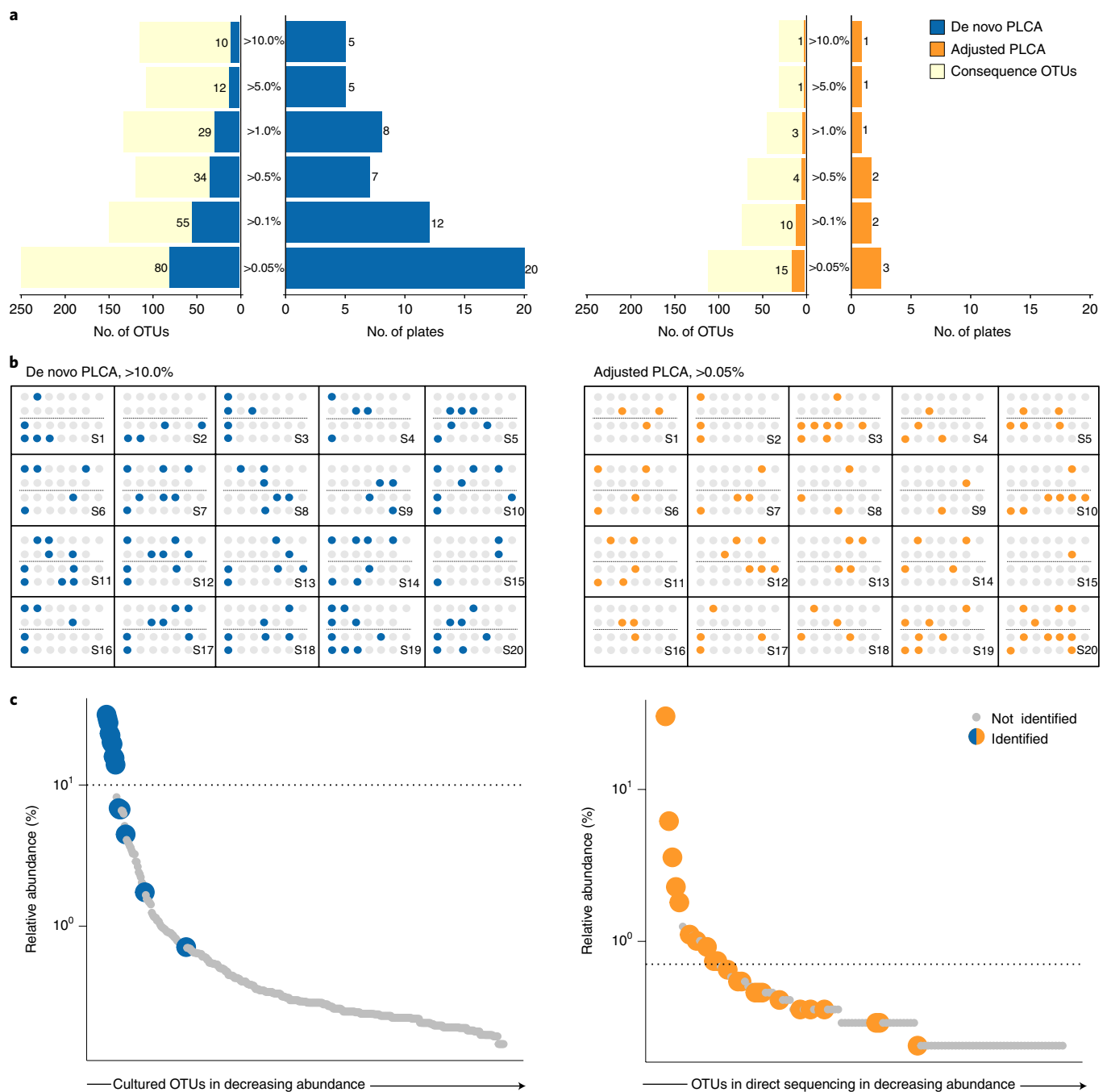


Fig. 4 | The PLCA determines an optimal plate set for culture-enriched metagenomics. **a**, The number of plates and OTUs necessary to capture the taxonomic diversity above varying thresholds of a representative sample using de novo and adjusted PLCAs. The resultant OTUs are divided into those obtained above the threshold and those captured by consequence of being present on plates within the optimal plate set. A similar output for all other samples is available (Extended Data Fig. 6). **b**, The plate subsets for the de novo and adjusted PLCAs for each sample in the dataset. Each culture condition is represented with a grey dot, which is coloured in the samples (S) in which it is part of the PLCA optimal plate set. **c**, The number of identified OTUs obtained with the de novo (blue) and adjusted PLCA (orange) when applied to sample 1 with the thresholds (dotted lines) displayed in **b**. Because the aim of the de novo PLCA is to recover the most abundant cultured organisms, the OTUs identified by culture enrichment are displayed; in contrast, the adjusted PLCA aims to recover abundant OTUs from the original sample and thus the OTUs identified in the direct sequencing are shown.

There are two versions of the PLCA. The de novo PLCA recapitulates the culture-enriched community independently of direct sequencing, whereas the adjusted PLCA focuses on the OTUs recovered from the direct sequencing results (Supplementary Fig. 3). That is to say, the user can decide between recovery of all cultured organisms (de novo PLCA) or preferential recovery of

the abundant organisms from the original community (adjusted PLCA). The use of either version is dependent on whether the user is interested in the composition of the original sample—for example, when answering clinically relevant research questions—or in questions concerning sample biodiversity. Using these algorithms at different thresholds (see Methods) across the dataset highlights

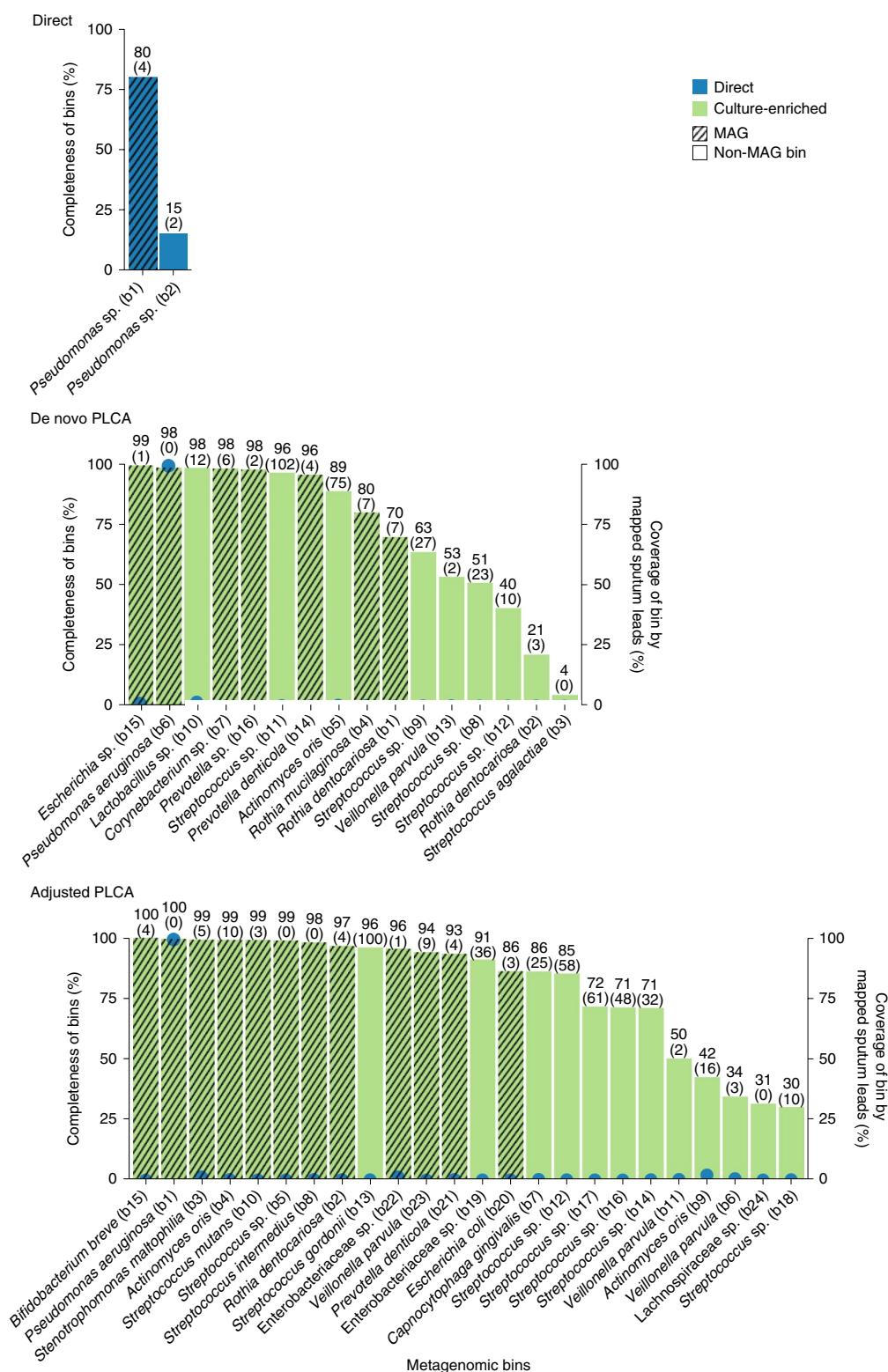


Fig. 5 | MAG and non-MAG bins resulting from culture-enriched and direct metagenomic sequencing of the first sample in the dataset. Three sets of metagenomic sequencing were performed on a representative sample in order to compare approaches. The statistics for the bins result from direct (blue) and culture-enriched (green) metagenomic sequencing. Culture-enriched metagenomics resulted in more metagenomic bins and more MAGs (hashed overlay) compared to direct sequencing. The coverage and contamination values (in parentheses) are displayed above each bin. Blue dots indicate the percent coverage of direct sequencing reads to each bin Source data. b, bin.

the uniqueness of these communities: no two samples have the same optimal plate set, and every culture condition is necessary for at least one sample, as indicated by each condition in the de novo

and adjusted outcomes being necessary for the culture-enriched metagenomic sequencing of at least one sample in the dataset (Fig. 4a,b and Extended Data Fig. 7).

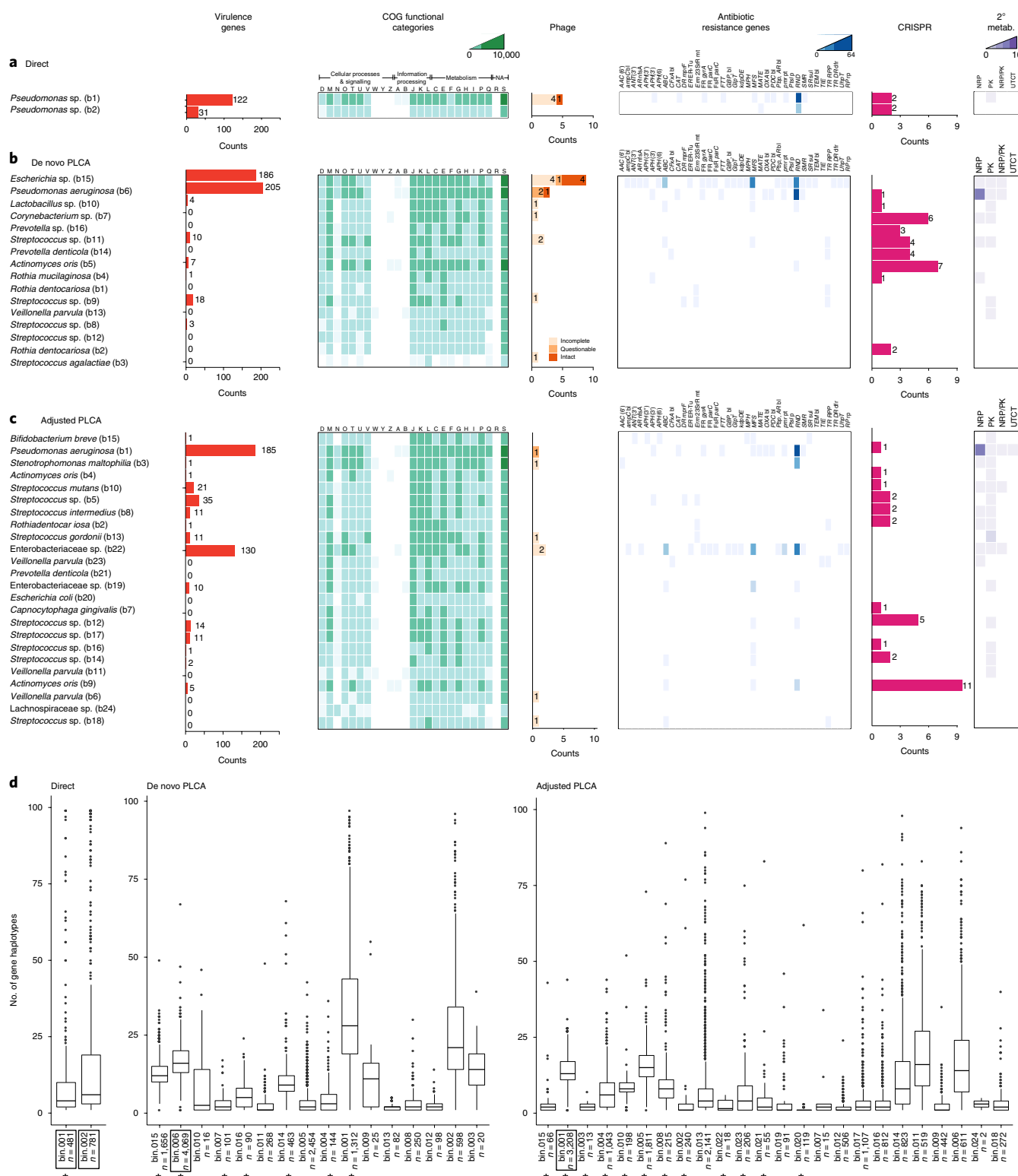


Fig. 6 | The taxonomic and functional diversity of direct and culture-enriched metagenomic sequencing. **a–c**, Bins are ordered by completeness. From left to right, estimations of virulence gene counts, the prevalence of proteins within COG functional categories, phage predictions, antibiotic resistance genes, counts of CRISPR genes and secondary metabolite predictions were estimated for direct (**a**) and culture-enriched (**b,c**) metagenomic sequencing. Predictions are split into rows based on metagenomic bin assignments (bold assignments indicate MAGs). **d**, Haplotype diversity of the open reading frames within each bin. Asterisks identify MAGs and bins are boxed if they correspond to *Pseudomonas* sp. Sample numbers are present on the x axis. Boxplots display the first and third quartiles, with a horizontal line to indicate the median, and whiskers extend to 1.5 times the interquartile range. bl, beta-lactamase; sul, sulfonamide resistant; p, phosphotransferase; pt, phosphoethanolamine transferase; dfr, dihydrofolate reductase; rp, undecaprenyl pyrophosphate related protein; 2° metab., secondary metabolites; NRP, non-ribosomal peptide; PK, polyketide; NRP/PK, non-ribosomal peptide/polyketide; UTCT, unknown thiotemplated cluster type.

Culture-enriched metagenomics provides improved taxonomic and functional resolution. To test the PLCA, we used both versions of the algorithm on a representative sample. The PLCA indicated culture-enriched metagenomics on five (de novo PLCA) and three (adjusted PLCA) culture plates (see Methods): Beef (aerobic) and AIA, KVLB, MAC and McKay (anaerobic) for de novo PLCA; TSY and McKay (aerobic) and CHOC (anaerobic) for the adjusted PLCA. Moderate taxonomic concordance was observed between the 16S rRNA gene and metagenomic sequencing, with some differences in composition and relative abundance (Extended Data Fig. 8). Following metagenomic co-assembly, binning and taxonomic assignment (see below), we compared the observed versus expected results of each algorithm. The de novo PLCA recovered 10 metagenomic bins with matching taxonomic assignments to the 10 OTUs above the PLCA threshold, in addition to a further 6 bins matching OTUs below the threshold (Fig. 4c and Supplementary Table 3). Similarly, the adjusted PLCA recovered 10 of 11 expected OTUs and matched an additional 14 OTUs below the adjusted PLCA threshold (Fig. 4c and Supplementary Table 4). When the adjusted PLCA was applied to the other 19 samples in this dataset, culture-enriched metagenomic sequencing of a collective 91 (of a possible 520) plates resulted in the recovery of 83.4% of targeted species with an additional 413 metagenomic bins recovered from other species (Extended Data Fig. 9a). We also applied the PLCA to previously published culture enrichment data from the gut microbiota¹², establishing that this algorithm is not specific to the lung microbiome or a specific set of culture conditions (Extended Data Fig. 9b).

For each of the de novo and adjusted PLCA plate sets and the direct sequencing, we co-assembled shotgun metagenomic reads into contigs, which were binned and defined as metagenome-assembled genomes (MAGs) or non-MAG bins based on sequence composition (see Methods). Culture-enriched metagenomic sequencing resulted in 7 MAGs and 9 non-MAG bins (de novo PLCA) and 12 MAGs and 12 non-MAG bins (adjusted PLCA); in contrast, 1 MAG and 1 non-MAG bin were recovered from direct metagenomic sequencing (Fig. 5). When the direct sequencing metagenomic reads were mapped onto the culture-enriched metagenomic bins, significant coverage was only observed in 1 bin in both the de novo and adjusted PLCA results (Fig. 5, blue dots). Both direct sequencing metagenomic bins were taxonomically assigned as *Pseudomonas* sp.; in contrast, the culture-enriched bins were taxonomically diverse, spanning 14 genera (Fig. 5).

The increased taxonomic diversity obtained via culture enrichment directly translates into increased functional information about this microbial community (Fig. 6). Consistently, across clusters of orthologous groups (COGs), functional categories and predictions of virulence genes, phage, antibiotic resistance, clustered regularly interspaced short palindromic repeats (CRISPRs) and secondary metabolites, culture-enriched metagenomic sequencing provides a greater diversity and number of functional identifications. For example, genes contributing to the GacS/GacA two-component system were identified in *Pseudomonas* sp. bins from both culture enrichment datasets but not in the direct sequencing of the sputum sample. Previous research has shown that strains of *P. aeruginosa* lacking this system are less able to colonize in mouse models³¹, indicating their importance for *P. aeruginosa* virulence. The detection of these genes in the culture-enriched data suggests that this system is present in the cystic fibrosis lung microbiota but is not identified in the direct metagenomic sequencing due to poor sequencing depth/assembly. Among the *Pseudomonas* sp. results, there were 10, 16 and 17 'perfect' hits against the Comprehensive Antibiotic Resistance Database (CARD) in the direct, de novo PLCA and adjusted PLCA datasets, respectively; among these, two beta-lactamases (OXA-50, PDC-1) and two repressors (*nalD*, *nfxB*) were found only in the culture-enriched sequencing. These *Pseudomonas*-specific results indicate that even when bins overlap in taxonomic assignment between

techniques, the culture enriches for functional annotation; in addition to this, the functional characterizations of all other bins (15 and 23 in the de novo and adjusted PLCA, respectively) would not have been possible from direct sequencing alone.

Previous research has established the presence of heterogeneous populations of single species in cystic fibrosis airways (for example, in *P. aeruginosa*³², *Burkholderia cepacia* complex³³ and *Stenotrophomonas maltophilia*³⁴) as well as in the human microbiome as a whole (for example, *Bifidobacterium longum* populations in the infant gut microbiome³⁵, various strains following faecal microbiota transplantation³⁶). As such, we calculated the genetic variability within each metagenomic bin by identifying haplotypes within open reading frames³⁷ (Fig. 6d). In some metagenomic bins, we identify a consistent and small number of open reading frame haplotypes, suggesting that the bin represents a single genomic population (that is, one strain). However, in most bins, we see great haplotype diversity indicative of heterogeneous populations (that is, multiple strains). As expected, the number of haplotypes per gene within these bins is diverse, indicating the known spectrum of evolutionary pressures within bacterial genomes³⁸. On average, more gene haplotypes were identified in the *Pseudomonas* sp. culture-enriched bins than in the direct sequencing; however, the direct sequencing identified a greater number of prevalent haplotype gene outliers. There was no correlation between bin completeness or MAG status and mean haplotype frequency.

Discussion

The decrease in cost and increase in massively parallelized sequencing technology has revolutionized the way that we study the human microbiota, leading to an increased understanding of these communities and how they relate to health and disease. The gut has arguably been the most well-studied human microbiome, with numerous studies linking it to various diseases (for example, inflammatory bowel disease³⁹ and irritable bowel syndrome⁴⁰) and conditions (for example, obesity⁴¹, pregnancy⁴², metabolic⁴³ and mental health⁴⁴). This community lends itself well to these investigations; its composition can be approximated via faecal matter—readily available without intervention—and consists of a dense microbial community with little host DNA contamination. However, many other important human-associated communities have high levels of non-microbial DNA, including communities of the skin⁴⁵, tissue biopsies¹⁹ and oral microbiome samples⁴⁶. The respiratory microbiome is a low biomass community, often contaminated with DNA from endothelial cells or from the DNA associated with an acute immune response in diseases such as asthma and cystic fibrosis⁴⁷. Because of the nature of such samples, metagenomic sequencing must be paired with the in silico removal of most sequencing reads due to host contamination, meaning that only the most abundant members of the community can be assembled into MAGs. We demonstrate that culture-enriched metagenomics—in conjunction with traditional, culture-independent sequencing—can improve the resolution of these communities.

Within this Article, we show that the cystic fibrosis lung microbiota is culturable, and that culture enrichment increases OTU diversity (Fig. 2). This follows directly from ref. 16, in which, using T-RFLP and 454 sequencing, it was possible to identify a culturable majority within the cystic fibrosis lung microbiota¹⁶. Although culture enrichment consistently recapitulated >97% of the direct sequencing results, a few organisms were not cultured. Many, including members of the Spirochaetes, SR1, Tenericutes and Saccharibacteria, consist of organisms that are difficult to culture^{48–50}, but were also at low abundance in these samples. In contrast, the organisms recovered only by culture are common members of the cystic fibrosis lung microbiota²⁵, including *Rothia*, *Prevotella*, *Veillonella*, *Fusobacterium* and *Streptococcus* species. Selective media allow for the proliferation of low-abundance organ-

isms, which can be below the level of detection of standard sequencing approaches. It is not uncommon, for example, for the cystic fibrosis lung microbiota to reach a density of at least 10^8 c.f.u. ml⁻¹ (refs. 51–53). If amplicon sequencing produces 50,000 reads per sample, an organism identified by a single read would equate to $\leq 0.002\%$ relative abundance or 2×10^4 c.f.u. ml⁻¹. This is not to dismiss culture-independent approaches; culture enrichment benefits from being combined with direct sequencing in order to maintain the relative abundance ratios of the original community and to recover important uncultured organisms.

To most effectively combine culture-independent and -dependent approaches, we designed the PLCA, which determines the cultured metagenomic sequencing necessary to recapitulate the original microbial community, as determined by amplicon sequencing. We show that the PLCA recovers targeted OTUs as well as a substantial number of additional OTUs (Fig. 4). Importantly, PLCA-assisted culture-enriched metagenomics vastly improves the taxonomic and functional outputs of sequencing. The inability of traditional metagenomic sequencing to distinguish microorganism from host can result in the need for incredible sequencing depths of samples with high host contamination. Combining culture with culture-independent sequencing is one way of mitigating host contamination due to culture's ability to enrich for viable microorganisms.

The PLCA is not specific to the cystic fibrosis microbiome and can be used on any microbial community in which most of its membership is culturable. The culture conditions chosen will impact the ability to culture the sample and the performance of the PLCA. Ideally, a combination of selective, non-selective and enrichment media should be used. Furthermore, the PLCA is also not specific to any one 16S rRNA gene processing pipeline. Here, we have used a 16S rRNA gene sequencing pipeline, which has been validated previously against other available approaches⁵⁴; however, 16S rRNA gene sequencing processing is a moving target, with new (or improved) methodologies constantly being published and compared/validated against already existing methods. Because the PLCA is agnostic to how the data are processed, it can be applied to datasets processed with any method. As the field continues to progress, the PLCA will only benefit from the improvements in taxonomic resolution available from these technological improvements.

The combination of culture enrichment with direct sequencing enhances the observed taxonomic diversity and provides greater insight into human-associated microbial communities. We have shown that culture-enriched metagenomics provides deeper resolution of these communities. With these data, we can better predict mechanisms of antimicrobial resistance, virulence factors and, in general, gain a better understanding of each organism's gene repertoire. Furthermore, having these organisms in culture means that we can carry out *in vitro*, mechanistic studies to better understand these communities in the context of human health and disease. Culture-enriched metagenomics, as exemplified here for cystic fibrosis sputum, provides an approach for the study of microbiome samples that are composed mostly of human rather than microbial DNA. This method can be applied to any community where the majority of its members are culturable.

Methods

Sputum collection and culture enrichment. On receiving informed consent, sputum samples were collected from 4 December 2013 to 6 October 2014 from willing participants visiting the Calgary Adult Cystic Fibrosis Clinic (ethical approval was granted by the Calgary Health Region Ethics Board, REB-24123). Two samples were collected from each patient (with the noted exceptions): one at the onset of pulmonary exacerbation (as defined by ref. 55) and a second during a follow-up appointment (one week to four months) following the resolution of symptoms and antibiotic discontinuation. In one case, a patient was not able to produce a follow-up sputum sample; in another, a patient experienced two exacerbations before a follow-up appointment, so three samples were collected.

Samples were transported to an anaerobic environment within 1 min of expectoration and plated within 4 h of production. Samples were homogenized by passage through a 18 G needle and 1 ml syringe. Once homogeneous, 300 µl was set aside for direct sequencing. The remainder was used for culture enrichment. The 13 solid agar media used in this study included the following (Extended Data Fig. 1): actinomycetes isolation agar (AIA, BD), brain heart infusion agar (BHI, BD), cooked meat broth with 1.5% agar (Beef, Fluka), Columbia agar base with 5% sheep's blood (CBA, BD), GC powder (BD) with 5% haemoglobin and 0.5% IsoVitalX (CHOC, BD), Columbia CNA agar with 5% sheep's blood (CNA, BD), fastidious anaerobe agar (FAA, Acumedia), tryptic soy agar with 0.1 µg ml⁻¹ kanamycin, 7.5 µg ml⁻¹ vancomycin, 10 µg ml⁻¹ vitamin K, 0.05 ng ml⁻¹ haemin and 5% laked blood (KVLB), MacConkey agar (MAC, BD), mannitol salt agar (MSA, BD), McKay medium⁵⁶, phenylethyl alcohol agar with 5% sheep's blood (PEA, BD), and tryptic soy agar with 1.5% yeast extract (TSY, BD). To Beef, BHI and TSY, the following additional additives were included: 10 µg ml⁻¹ colistin sulfate, 0.5 mg ml⁻¹ L-cysteine, 1.0 ng ml⁻¹ vitamin K and 10 ng ml⁻¹ haemin. These media were chosen based on previous successful isolation of bacterial species from cystic fibrosis sputum in ref. 16, and include non-selective, selective and enrichment media types (as defined in ref. 57; Extended Data Fig. 1). Comparisons of this culturing method to that employed by a typical microbiology laboratory in Fig. 3 included the following media types: aerobic CBA, MAC, MSA and anaerobic CHOC.

Culture enrichment was performed by placing 100 µl of sputum diluted to 10^{-3} and 10^{-5} in BHI with 0.05% L-cysteine onto each of the above media. Two sets of plating were performed, one incubated aerobically (5% CO₂, 37 °C) and another anaerobically (5% CO₂, 5% H₂, 90% NO₂, 37 °C), resulting in 52 plates per sample.

After 3–5 days (aerobic) and 5–7 days (anaerobic) of growth, plates were imaged and growth acquired by adding 2 ml of BHI broth to each plate and lifting the colonies. A 1 ml sample of this broth was frozen directly for DNA extraction while the remaining 1 ml was frozen in skimmed milk (final concentration 10%) for any potential growth or re-isolation experiments. For the first few culture enrichment sample sets, plates with no visible growth were processed like any other plate (see below); however, we consistently were unable to obtain visible PCR products on a 2% agarose gel from those plates that did not have visible growth. Thus, any plate that resulted in no visible bacterial colonies was discarded and omitted from all downstream processing.

To demonstrate the reproducibility of the sputum sample collection and culture enrichment methods, we carried out an additional experiment where two sputum samples from each of three patients were collected in the clinic before and after physiotherapy (3 × 2 biological replicates). The consistency of these biological replicates indicated the similarity of sputum communities when collected in quick succession. These samples were then plated on six media in triplicate (6 × 3 technical replicates across six sputum samples, $n = 108$). The results demonstrate the consistency in replicate sputum samples and in culture enrichment (Extended Data Fig. 10).

DNA isolation and Illumina sequencing. Genomic DNA was isolated from culture-enriched plates and sputum as previously described⁵⁸, with the exception of using lifted colonies/homogenized sputum as input instead of Copan swabs, as performed in ref. 16. Dilutions resulting from the same culture conditions were combined into one genomic DNA isolation for a maximum of 26 culture-enriched samples per sputum sample. The variable 3 region of the 16S rRNA gene was amplified using universal primers as adapted from refs. 58,59. The PCR reaction consisted of 5 pmol of each primer, 1 ng template DNA, 200 µM deoxynucleoside triphosphates, 1.5 mM MgCl₂ and 1 U Taq polymerase. The PCR protocol used was as follows: 95 °C for 5 min, followed by 30 cycles of 95 °C for 30 s, 50 °C for 30 s and 72 °C for 30 s, with a final 72 °C for 7 min. The presence of a PCR product was verified by electrophoresis (2% agarose gel). PCR products were sequenced using the Illumina MiSeq platform using 2 × 250 paired-end reads.

DNA from select culture-enriched samples and the sputum sample were sonicated to 300 bp and library preparations were made using the NEBNext DNA Library Prep Master Mix Set for Illumina (New England Biolabs) and sequenced using the Illumina HiSeq platform with 2 × 250 paired-end reads.

All sequencing results are publicly available (BioProject ID PRJNA503799).

16S rRNA sequence processing and analysis. 16S rRNA paired-end reads were processed using sllp⁶⁴. Briefly, reads were trimmed of any remaining primers using cutadapt⁶⁰ and discarded using sickle based on a quality threshold of 30 (<https://github.com/najoshi/sickle>). Paired-end reads were assembled using PANDASEq⁶¹. OTUs were picked using AbundantOTU+ with a 97% clustering threshold⁶² and chimaeras removed using USEARCH⁶³ as implemented in QIIME⁶⁴. The Ribosomal Database Project classifier⁶⁵ was used to assign taxonomy against the 4 February 2011 release of the Greengenes database⁶⁶, and a phylogeny was created by pruning the Greengenes phylogeny to those taxa present in the dataset. OTU tables were created with QIIME⁶⁴. Any OTU that was not assigned a bacterial taxonomy or where there was only one instance across the full dataset (singleton) was culled. Any sample with <1,000 reads was discarded (Supplementary Table 5). The result of this culling process, in combination with only sequencing plates with visual growth, resulted in a total of 531 samples (20 sputum samples and 511

plates). The mean sequence depth across this dataset was 68,160 reads per sample (range 2,032–159,381), with a mean number of OTUs of 94.1 (range 10–311).

Taxonomic summaries over multiple samples were performed by calculating the maximum relative abundance across samples, and normalizing to 100%. Principal coordinate analysis (PCoA) plots were calculated using phyloseq⁶⁷ and ggplot2⁶⁸ in R after proportional normalization⁶⁹. An OTU was considered present in the direct or culture-enriched sequencing if it had a relative abundance of >0.01% (all exceptions noted). Phylogenies were decorated with GraPhlAn⁷⁰. Heatmaps were generated with pheatmap⁷¹. In the sequencing depth experiments shown in Extended Data Fig. 3, rarefaction was performed at varying depths using QIIME's alpha rarefaction function.

Recovery of isolates from frozen culture-enriched stocks. Improved isolation of *Stenotrophomonas maltophilia* from frozen skimmed milk stocks of select plates was performed using a selective medium as described previously⁷². Isolates were Sanger sequenced using the 8F (5'-AGAGTTTTCCTGGCTCAG3') and 926R (5'-CCGTCATTCCTTTTTRAGTTT-3') primers to the 16S rRNA gene, resulting in a 900 nt product. The identity of the isolates were confirmed by comparisons to the Human Oral Microbiome Database (HOMD) and to NCBI's 16S ribosomal RNA sequences (Bacteria and Archaea) Database.

The PLCA. Taking 16S rRNA gene sequencing results as input, the PLCA calculates, for each sample, the optimal subset of cultured plates that should be included in culture-enriched metagenomics in order to recapitulate the microbial community. The PLCA (Supplementary Fig. 3a,b) first identifies any OTU above the user-supplied relative abundance threshold that was only cultured on a single plate, and that plate is added to the 'plate set' for culture-enriched sequencing. Next, for all OTUs not already identified in the plate set, the plate with the most OTUs present above the threshold is added to the plate set. This continues until all OTUs are accounted for in the plate set and this list is output to the user. The PLCA incorporates a user-adjustable relative abundance threshold to determine which cultured OTUs the algorithm should include in the resulting plate set. In the adjusted PLCA, a second threshold determines the cutoff of OTU inclusion from the direct sequencing. Altering these thresholds results in varying plate and OTU recovery (including OTUs below the threshold, which are included as a consequence of being present on a plate that is part of the optimal plate set; Fig. 4a and Extended Data Fig. 7). The plate set for the adjusted PLCA is not always a direct subset of the de novo PLCA because, when only the organisms present in the sputum are considered, there is often a better combination of plates that minimizes the number of total plates needed for sequencing.

The PLCA is freely available from <http://github.com/fwhelan/PLCA>.

Metagenomic sequence processing and analysis. Resultant Illumina paired-end reads from the 20 sputum samples and their associated adjusted PLCA plate sets (and one additional de novo PLCA plate set for comparison) were processed first by using cutadapt to trim Illumina adapters and primers⁶⁰. Sickle, with a quality threshold of 30, was used to remove low-quality sequences (<https://github.com/najoshi/sickle>). The direct sequencing was decontaminated of host-associated reads using DeconSeq⁷³. Metagenomic assembly of the direct sequencing reads was conducted using Megahit⁷⁴. A co-assembly of the culture-enriched reads (as determined by the de novo PLCA and/or adjusted PLCA) was also conducted using Megahit⁷⁴. The results of all assemblies were separately binned using MaxBin-2.2.1⁷⁵; MaxBin was chosen based on its performance in the CAMI Challenge⁷⁶. Quality was assessed with checkM⁷⁷, and taxonomic assignments for each bin determined using KrakenUniq (formerly KrakenHLL)⁷⁸ and supplementary scripts (<https://github.com/shekas3/BinTaxaAssigner>). Assembly statistics, including mean/maximum contig length and N50 values, are provided in Supplementary Table 6. MAGs were defined as metagenomic bins containing ≥70% completeness and <10% contamination as previously described⁷⁹; non-MAG bins are any bin that does not meet the criteria of a MAG. Only contigs >1,000 bp were binned. Bowtie2 was used to map direct sequencing reads onto culture-enriched metagenomic bins. Even though extensive decontamination and quality control measures were performed, the direct sequencing suffered from considerable host DNA contamination, resulting in reads within these bins mapping with high stringency to the human genome. This resulted in bins from the direct sequencing that were much larger than the size of the closest reference genome (Supplementary Fig. 4). In contrast, culture-enriched bins with high 'contamination' indicated the inability of the binning algorithm to differentiate between closely related species. For example, b6 in de novo and b9 in adjusted had bin lengths almost double the closest reference and contained taxonomic signatures of two *Streptococcus* sp. (Supplementary Fig. 4).

To compare the taxonomic composition of 16S rRNA gene and metagenomic sequencing, we used the Kraken2 (version 2.0.7) classifier⁸⁰ to classify metagenomic reads from each culture-enriched sample against the 2011 Greengenes database⁶⁶.

To show that the PLCA was not specific to the cystic fibrosis lung microbiota, the algorithm was applied to a previously published culture enrichment study of the gut microbiota¹². The de novo and adjusted PLCA were applied, with default thresholds, to sample IBS3, and the taxonomic recovery from culture-enriched

metagenomic sequencing was predicted based on the 16S rRNA gene sequencing profiles of the resulting plate set.

Functional annotations of each metagenomic bin were performed with a variety of software. Virulence gene counts were determined by use of blastn (E value cutoff of 10⁻⁸)⁸¹ against PATRIC's virulence factor library⁸² and counting the number of hits to unique virulence genes per bin. COG functional category counts were determined using egg-nog-mapper with default parameters^{83,84}. Phage counts were determined by Phaster⁸⁵ and predictions of antibiotic resistance genes were conducted with CARD (2.0.2) in conjunction with the RGI (4.1.0)⁸⁶. Secondary metabolites were predicted with PRISM 3⁸⁷. The presence of CRISPRs were determined using MinCED⁸⁸. The direct sequencing was plagued with host DNA contamination, resulting in excessive bin lengths, and, possibly, abundant type I errors in the identified functionality of the community (Fig. 6, Supplementary Fig. 4 and Supplementary Table 7). All bar charts were made using R's ggplot2⁶⁸ and heatmaps were visualized using R's pheatmap⁷¹. The haplotype diversity of the open reading frames of each bin was calculated using Hansel and Gretel³⁷.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All sequencing results are publicly available (BioProject ID PRJNA503799). The PLCA algorithm is available from <https://github.com/fwhelan/PLCA>.

Code availability

All code developed by the authors is available under a GNU licence at <http://github.com/fwhelan/PLCA> and <https://github.com/shekas3/BinTaxaAssigner>.

Received: 5 November 2018; Accepted: 18 November 2019;
Published online: 20 January 2020

References

1. Van Leeuwenhoek, A. Microscopical observations about animals in the scurf of the teeth. *Philos. Trans. R Soc. Lond. B Biol. Sci.* **14**, 568–574 (1683).
2. Turnbaugh, P. J. et al. The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice. *Sci. Transl. Med.* **1**, 6ra14 (2009).
3. Huttenhower, C. et al. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
4. Spor, A., Koren, O. & Ley, R. Unravelling the effects of the environment and host genotype on the gut microbiome. *Nat. Rev. Microbiol.* **9**, 279–290 (2011).
5. Rothschild, D. et al. Environment dominates over host genetics in shaping human gut microbiota. *Nature* **555**, 210–215 (2018).
6. Olesen, S. W. & Alm, E. J. Dysbiosis is not an answer. *Nat. Microbiol.* **1**, 16228 (2016).
7. Shade, A. Diversity is the question, not the answer. *ISME J.* **11**, 1–6 (2017).
8. Finegold, S. M., Attebery, H. R. & Sutter, V. L. Effect of diet on human fecal flora: comparison of Japanese and American diets. *Am. J. Clin. Nutr.* **27**, 1456–1469 (1974).
9. Goodman, A. L. et al. Extensive personal human gut microbiota culture collections characterized and manipulated in gnotobiotic mice. *Proc. Natl Acad. Sci. USA* **108**, 6252–6257 (2011).
10. Lagier, J.-C. et al. Microbial culturomics: paradigm shift in the human gut microbiome study. *Clin. Microbiol. Infect.* **18**, 1185–1193 (2012).
11. Rettedal, E. A., Gumpert, H. & Sommer, M. O. A. Cultivation based multiplex phenotyping of human gut microbiota allows targeted recovery of previously uncultured bacteria. *Nat. Commun.* **5**, 4714 (2014).
12. Lau, J. T. et al. Capturing the diversity of the human gut microbiota through culture-enriched molecular profiling. *Genome Med.* **8**, 72 (2016).
13. Hilt, E. E. et al. Urine is not sterile: use of enhanced urine culture techniques to detect resident bacterial flora in the adult female bladder. *J. Clin. Microbiol.* **52**, 871–876 (2014).
14. Myles, I. A. et al. A method for culturing Gram-negative skin microbiota. *BMC Microbiol.* **16**, 60 (2016).
15. Thompson, H., Rybalka, A., Moazzez, R., Dewhirst, F. E. & Wade, W. G. In-vitro culture of previously uncultured oral bacterial phylotypes. *Appl. Environ. Microbiol.* **81**, 8307–8314 (2015).
16. Sibley, C. D. et al. Culture enriched molecular profiling of the cystic fibrosis airway microbiome. *PLoS ONE* **6**, e22702 (2011).
17. Oh, J. et al. Biogeography and individuality shape function in the human skin metagenome. *Nature* **514**, 59–64 (2014).
18. Wang, W.-L. et al. Application of metagenomics in the human gut microbiome. *World J. Gastroenterol.* **21**, 803–814 (2015).
19. Zhang, C. et al. Identification of low abundance microbiome in clinical samples using whole genome sequencing. *Genome Biol.* **16**, 265 (2015).
20. Lim, Y. W. et al. Metagenomics and metatranscriptomics: windows on CF-associated viral and microbial communities. *J. Cyst. Fibros.* **12**, 154–164 (2013).

21. Huang, Y. J. & LiPuma, J. J. The microbiome in cystic fibrosis. *Clin. Chest Med.* **37**, 59–67 (2015).
22. Zhao, J. et al. Decade-long bacterial community dynamics in cystic fibrosis airways. *Proc. Natl Acad. Sci. USA* **109**, 5809–5814 (2012).
23. Whelan, F. J. et al. Longitudinal sampling of the lung microbiota in individuals with cystic fibrosis. *PLoS ONE* **12**, e0172811 (2017).
24. Lagier, J.-C. et al. Culture of previously uncultured members of the human gut microbiota by culturomics. *Nat. Microbiol.* **1**, 16203 (2016).
25. Surette, M. G. The cystic fibrosis lung microbiome. *Ann. Am. Thorac. Soc.* **11**(Suppl. 1), S61–S65 (2014).
26. Field, T. R., Sibley, C. D., Parkins, M. D., Rabin, H. R. & Surette, M. G. The genus *Prevotella* in cystic fibrosis airways. *Anaerobe* **16**, 337–344 (2010).
27. van der Gast, C. J. et al. Partitioning core and satellite taxa from within cystic fibrosis lung bacterial communities. *ISME J.* **5**, 780–791 (2011).
28. Tunney, M. M. et al. Detection of anaerobic bacteria in high numbers in sputum from patients with cystic fibrosis. *Am. J. Respir. Crit. Care Med.* **177**, 995–1001 (2008).
29. Parkins, M. D. & Floto, R. A. Emerging bacterial pathogens and changing concepts of bacterial pathogenesis in cystic fibrosis. *J. Cyst. Fibros.* **14**, 293–304 (2015).
30. Pop, M. et al. Individual-specific changes in the human gut microbiota after challenge with enterotoxigenic *Escherichia coli* and subsequent ciprofloxacin treatment. *BMC Genomics* **17**, 440 (2016).
31. Coleman, F. T. et al. Hypersusceptibility of cystic fibrosis mice to chronic *Pseudomonas aeruginosa* oropharyngeal colonization and lung infection. *Proc. Natl Acad. Sci. USA* **100**, 1949–1954 (2003).
32. Jorth, P. et al. Regional isolation drives bacterial diversification within cystic fibrosis lungs. *Cell Host Microbe* **18**, 307–319 (2015).
33. Lieberman, T. D. et al. Genetic variation of a bacterial pathogen within individuals with cystic fibrosis provides a record of selective pressures. *Nat. Genet.* **46**, 82–87 (2014).
34. Pompilio, A. et al. *Stenotrophomonas maltophilia* phenotypic and genotypic diversity during a 10-year colonization in the lungs of a cystic fibrosis patient. *Front. Microbiol.* **7**, 1551 (2016).
35. Ferretti, P. et al. Mother-to-infant microbial transmission from different body sites shapes the developing infant gut microbiome. *Cell Host Microbe* **24**, 133–145 (2018).
36. Li, S. S. et al. Durable coexistence of donor and recipient strains after fecal microbiota transplantation. *Science* **352**, 586–589 (2016).
37. Nicholls, S. M. et al. Probabilistic recovery of cryptic haplotypes from metagenomic data. Preprint at <https://www.biorxiv.org/content/10.1101/117838v1> (2017).
38. Creevey, C. J., Doerks, T., Fitzpatrick, D. A., Raes, J. & Bork, P. Universally distributed single-copy genes indicate a constant rate of horizontal transfer. *PLoS ONE* **6**, e22099 (2011).
39. Frank, D. N. et al. Molecular–phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proc. Natl Acad. Sci. USA* **104**, 13780–13785 (2007).
40. Collins, S. M. A role for the gut microbiota in IBS. *Nat. Rev. Gastroenterol. Hepatol.* **11**, 497–505 (2014).
41. Ley, R. E., Turnbaugh, P. J., Klein, S. & Gordon, J. I. Microbial ecology: human gut microbes associated with obesity. *Nature* **444**, 1022–1023 (2006).
42. Gohir, W., Whelan, F. J., Surette, M. G., Moore, C. & Jonathan, D. Pregnancy-related changes in the maternal gut microbiota are dependent upon the mother's periconceptional diet. *Gut Microbes* **6**, 310–320 (2015).
43. Zeevi, D. et al. Personalized nutrition by prediction of glycemic responses. *Cell* **163**, 1079–1095 (2015).
44. Naseribafrouei, A. et al. Correlation between the human fecal microbiota and depression. *Neurogastroenterol. Motil.* **26**, 1155–1162 (2014).
45. Grice, E. A. & Segre, J. A. The skin microbiome. *Nat. Rev. Microbiol.* **9**, 244–253 (2011).
46. Wang, J. et al. Metagenomic sequencing reveals microbiota and its functional potential associated with periodontal disease. *Sci. Rep.* **3**, 1843 (2013).
47. Dickson, R. P., Erb-Downward, J. R., Martinez, F. J. & Huffnagle, G. B. The microbiome and the respiratory tract. *Annu. Rev. Physiol.* **78**, 481–504 (2015).
48. Chi, B., Chauhan, S. & Kuramitsu, H. Development of a system for expressing heterologous genes in the oral spirochete *treponema denticola* and its use in expression of the *treponema pallidum* flaA gene. *Infect. Immun.* **67**, 3653–3656 (1999).
49. Camanocha, A. & Dewhirst, F. E. Host-associated bacterial taxa from *Chlorobi*, *Chloroflexi*, GN02, Synergistetes, SR1, TM7 and WPS-2 phyla/candidate divisions. *J. Oral Microbiol.* **6**, 25468 (2014).
50. Marcy, Y. et al. Dissecting biological 'dark matter' with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. *Proc. Natl Acad. Sci. USA* **104**, 11889–11894 (2007).
51. Meyer, K. C., Sharma, A., Rosenthal, N. S., Peterson, K. & Brennan, L. Regional variability of lung inflammation in cystic fibrosis. *Am. J. Respir. Crit. Care Med.* **156**, 1536–1540 (1997).
52. Stressmann, F. A. et al. Does bacterial density in cystic fibrosis sputum increase prior to pulmonary exacerbation? *J. Cyst. Fibros.* **10**, 357–365 (2011).
53. Quigley, E. M. Gut bacteria in health and disease. *Gastroenterol. Hepatol.* **9**, 560–569 (2013).
54. Whelan, F. J. & Surette, M. G. A comprehensive evaluation of the sl1p pipeline for 16S rRNA gene sequencing analysis. *Microbiome* **5**, 100 (2017).
55. Fuchs, H. J. et al. Effect of aerosolized recombinant human DNase on exacerbations of respiratory symptoms and on pulmonary function in patients with cystic fibrosis. The Pulmozyme Study Group. *N. Engl. J. Med.* **331**, 637–642 (1994).
56. Sibley, C. D. et al. McKay agar enables routine quantification of the '*Streptococcus milleri*' group in cystic fibrosis patients. *J. Med. Microbiol.* **59**, 534–540 (2010).
57. Whelan, F. J., Rossi, L., Stearns, J. C. & Surette, M. G. Culture and molecular profiling of the respiratory tract microbiota. *Methods Mol. Biol.* **1894**, 49–61 (2018).
58. Whelan, F. J. et al. The loss of topography in the microbial communities of the upper respiratory tract in the elderly. *Ann. Am. Thorac. Soc.* **11**, 513–521 (2014).
59. Bartram, A. K., Lynch, M. D. J., Stearns, J. C., Moreno-Hagelsieb, G. & Neufeld, J. D. Generation of multimillion-sequence 16S rRNA gene libraries from complex microbial communities by assembling paired end illumina reads. *Appl. Environ. Microbiol.* **77**, 3846–3852 (2011).
60. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).
61. Masella, A. P., Bartram, A. K., Truszkowski, J. M., Brown, D. G. & Neufeld, J. D. PANDAseq: paired-end assembler for illumina sequences. *BMC Bioinformatics* **13**, 31 (2012).
62. Ye, Y. Identification and quantification of abundant species from pyrosequences of 16S rRNA by consensus alignment. In *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine* 153–157 (IEEE, 2011).
63. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).
64. Caporaso, J. G. et al. QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* **7**, 335–336 (2010).
65. Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* **73**, 5261–5267 (2007).
66. DeSantis, T. Z. et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* **72**, 5069–5072 (2006).
67. McMurdie, P. J. & Holmes, S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE* **8**, e61217 (2013).
68. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer, 2009).
69. McMurdie, P. J. & Holmes, S. Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput. Biol.* **10**, e1003531 (2014).
70. Asnicar, F., Weingart, G., Tickle, T. L., Huttenhower, C. & Segata, N. Compact graphical representation of phylogenetic data and metadata with GraPhlAn. *PeerJ* **3**, e1029 (2015).
71. Pheatmap: pretty heatmaps. R Package v.1.0.12 (CRAN, 2012).
72. Denton, M., Hall, M., Todd, N., Kerr, K. & Littlewood, J. Improved isolation of *Stenotrophomonas maltophilia* from the sputa of patients with cystic fibrosis using a selective medium. *Clin. Microbiol. Infect.* **6**, 395–396 (2000).
73. Schmieder, R. & Edwards, R. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PLoS ONE* **6**, e17288 (2011).
74. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
75. Wu, Y.-W., Tang, Y.-H., Tringe, S. G., Simmons, B. A. & Singer, S. W. MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. *Microbiome* **2**, 26 (2014).
76. Sczyrba, A. et al. Critical assessment of metagenome interpretation—a benchmark of metagenomics software. *Nat. Methods* **14**, 1063–1071 (2017).
77. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
78. Breitwieser, F. P. & Salzberg, S. L. KrakenHLL: confident and fast metagenomics classification using unique k-mer counts. *Genome Biol.* **19**, 198 (2018).
79. Lee, S. T. M. et al. Tracking microbial colonization in fecal microbiota transplantation experiments via genome-resolved metagenomics. *Microbiome* **5**, 50 (2017).
80. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46 (2014).

81. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
82. Mao, C. et al. Curation, integration and visualization of bacterial virulence factors in PATRIC. *Bioinformatics* **31**, 252–258 (2015).
83. Huerta-Cepas, J. et al. Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper. *Mol. Biol. Evol.* **34**, 2115–2122 (2017).
84. Huerta-Cepas, J. et al. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.* **44**, D286–D293 (2016).
85. Arndt, D. et al. PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res.* **44**, W16–W21 (2016).
86. Jia, B. et al. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res.* **45**, D566–D573 (2017).
87. Skinnider, M. A. et al. Genomes to natural products PRediction Informatics for Secondary Metabolomes (PRISM). *Nucleic Acids Res.* **43**, 9645–9662 (2015).
88. Bland, C. et al. CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* **8**, 209 (2007).

Acknowledgements

This research was funded in part by a Canadian Institutes of Health Research (CIHR) Doctoral Scholarship, a Cystic Fibrosis Canada (CFC) Studentship and a Marie Skłodowska-Curie Individual Fellowship (GA no. 793818) awarded to F.J.W., and grants from CIHR, CFC and a Tier 1 Canada Research Chair to M.G.S. We thank the patients and healthcare professionals at the Calgary Adult Cystic Fibrosis Clinic for their participation and assistance with this study. We acknowledge critical intellectual conversations with J.T. Lau and J.C. Szamosi. We wish to acknowledge that this research

was conducted on traditional territory shared by the Haudenosaunee confederacy and the Anishinaabe nations as well as the peoples of the Treaty 7 region in Southern Alberta.

Author contributions

F.J.W. is the primary author of this prepared manuscript. H.R.R. and M.D.P. collected patient information and enrolled willing participants for this study. B.W. collected, processed and cultured all sputum samples in addition to all biological and technical controls. F.J.W. and S.A.S. isolated DNA from culture/sputum material, and ran PCR reactions to amplify the 16S rRNA gene variable 3 region. S.A.S. performed the enrichment of *Stenotrophomonas*. F.J.W. prepared DNA for metagenomic sequencing. F.J.W. processed and analysed all 16S rRNA gene and metagenomic sequencing results. S.S. provided code for the taxonomic assignment of metagenomic bins. F.J.W., M.D.P. and M.G.S. conceptualized the experimental outline. F.J.W. conducted all data analyses and wrote the manuscript. All authors edited and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41564-019-0643-y>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41564-019-0643-y>.

Correspondence and requests for materials should be addressed to M.G.S.

Reprints and permissions information is available at www.nature.com/reprints.

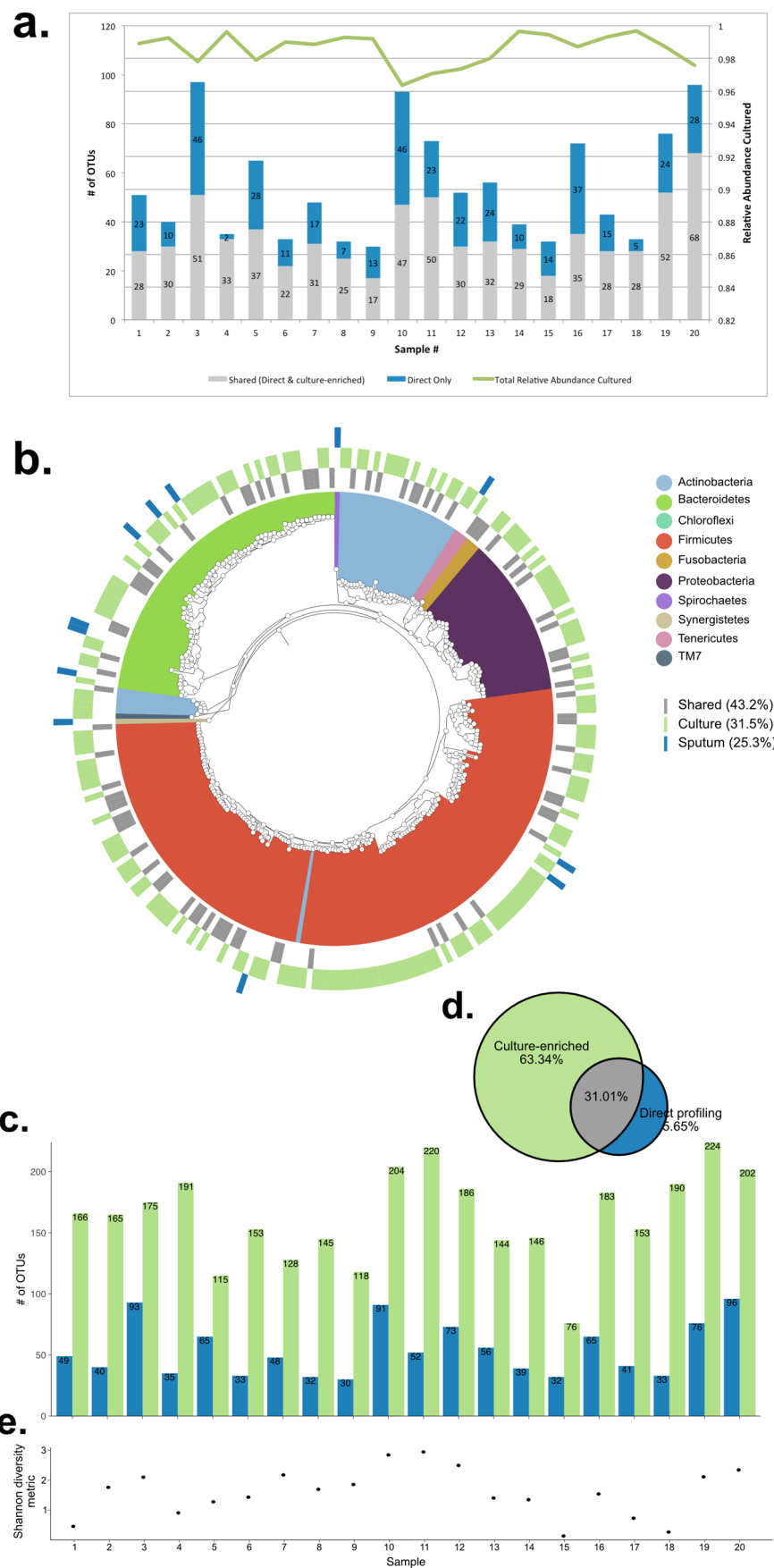
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

Extended Data Table 1: Media used for culture-enrichment.

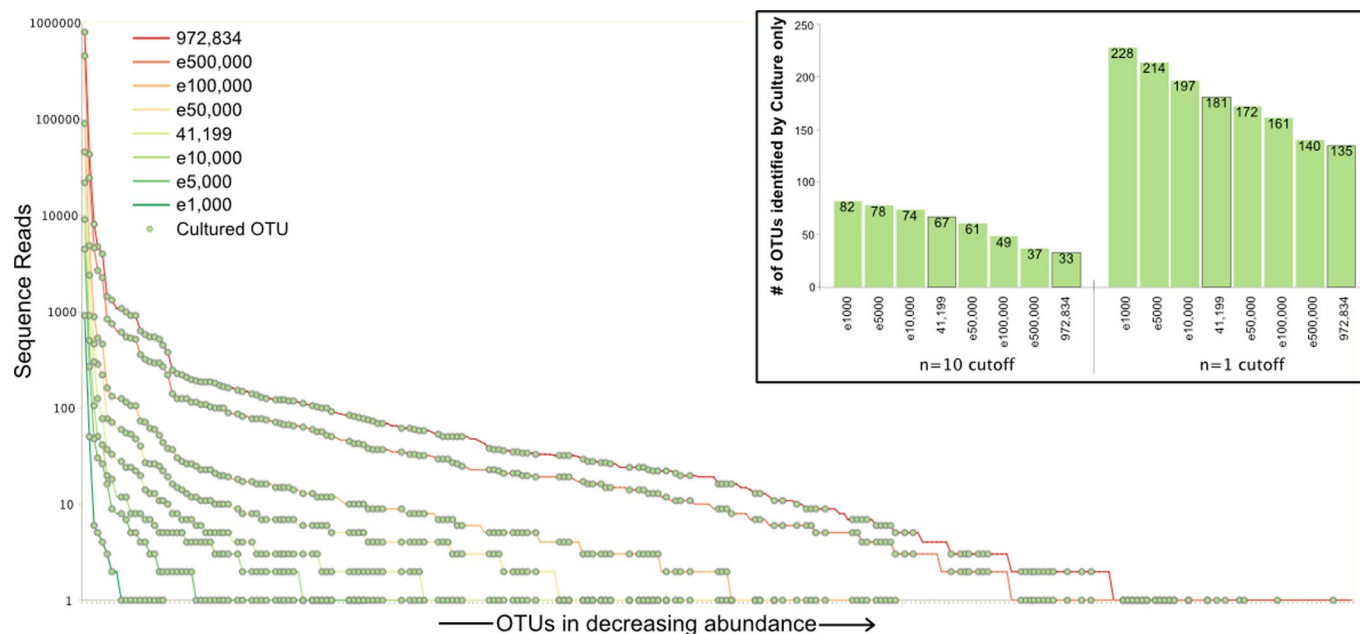
Image	Abbreviation	Media	Additional Supplements	Selective/ Non-Selective/ Enrichment
	AIA	Actinomycetes isolation agar	5 mL of glycerol	Enrichment
	Beef	Cooked meat broth with 1.5% agar	10µg/mL colistin sulfate, 0.5mg/mL L-Cysteine, 1.0ng/mL Vitamin K, 10ng/mL hemin	Non-Selective
	BHI	Brain heart infusion agar	10µg/mL colistin sulfate, 0.5mg/mL L-Cysteine, 1.0ng/mL Vitamin K, 10ng/mL hemin	Non-Selective
	CBA	Columbia agar	5% sheep's blood	Enrichment
	CHOC	GC powder	5% hemoglobin, 0.5% IsoVitaleX	Enrichment
	CNA	Columbia CNA agar	5% sheep's blood	Selective
	FAA	Fastidious anaerobe agar		Enrichment
	KVLB	Tryptic soy agar	0.1g/mL kanamycin, 7.5g/mL vancomycin, 10g/mL Vitamin K, 0.05ng/mL hemin, 5% laked blood	Selective
	MAC	MacConkey agar		Selective
	McKay	McKay agar [56]		Enrichment
	MSA	Mannitol salt agar		Selective
	PEA	Phenylethyl alcohol agar	5% sheep's blood	Selective
	TSY	Tryptic soy agar	1.5% yeast extract, 10µg/mL colistin sulfate, 0.5mg/mL L-Cysteine, 1.0ng/mL Vitamin K, 10ng/mL hemin	Non-Selective

Extended Data Fig. 1 | Media used for culture enrichment.

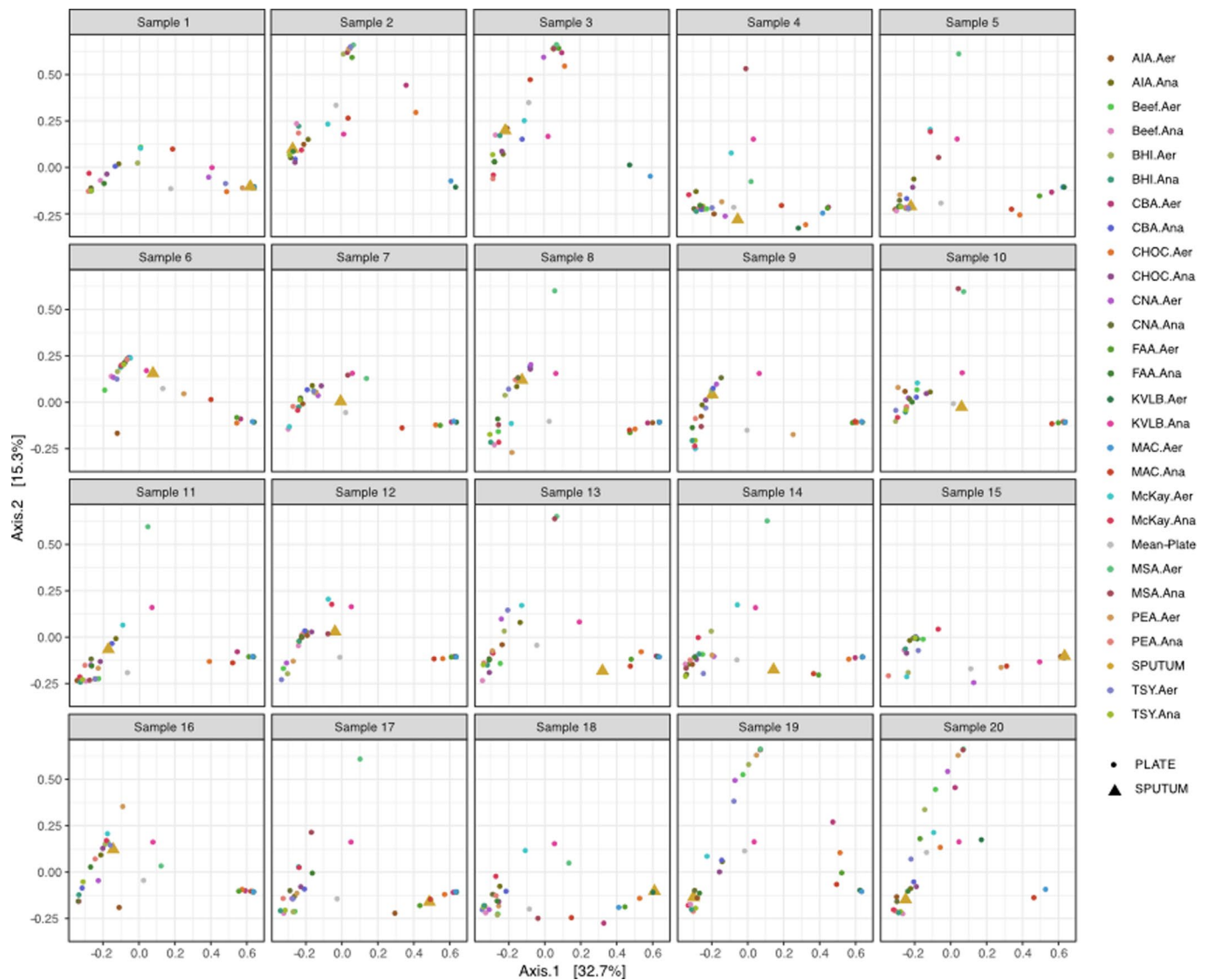


Extended Data Fig. 2 | See next page for caption.

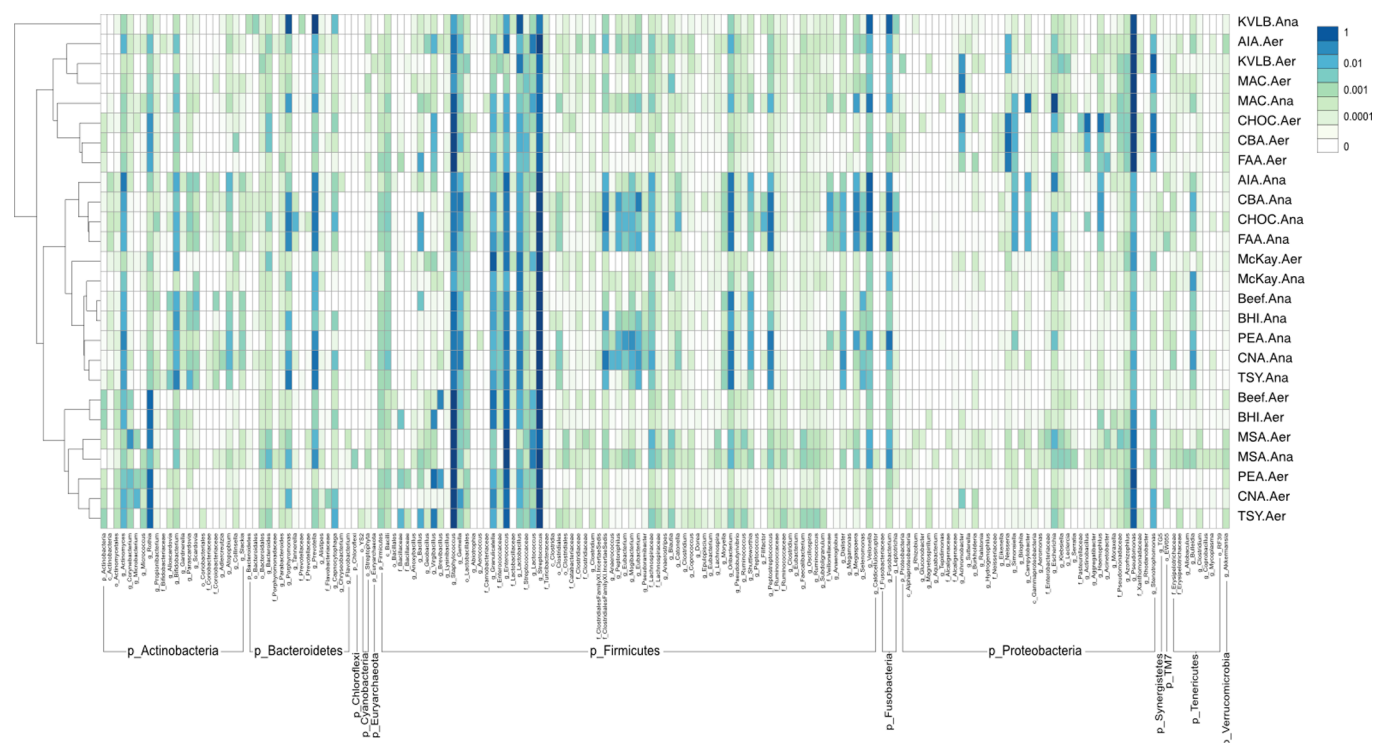
Extended Data Fig. 2 | Using more stringent abundance thresholds, the vast majority of the cystic fibrosis lung microbiota is still captured by culture-enriched methods. Culture-enriched 16S rRNA gene sequencing increases OTU recovery. Using a cutoff of $\geq 0.1\%$ relative abundance, 65.6% of the lung microbiota was identified only by culture across samples (**a**), with only 4.3% of OTUs not being cultured (**b**). **c**. In each sample, more OTUs were recovered via culture-enriched sequencing than by direct sequencing. **d**. Overall, an average of 65.5% of all OTUs were identified only via culture. **e**. The α -diversity of the original samples subject to culture enrichment.



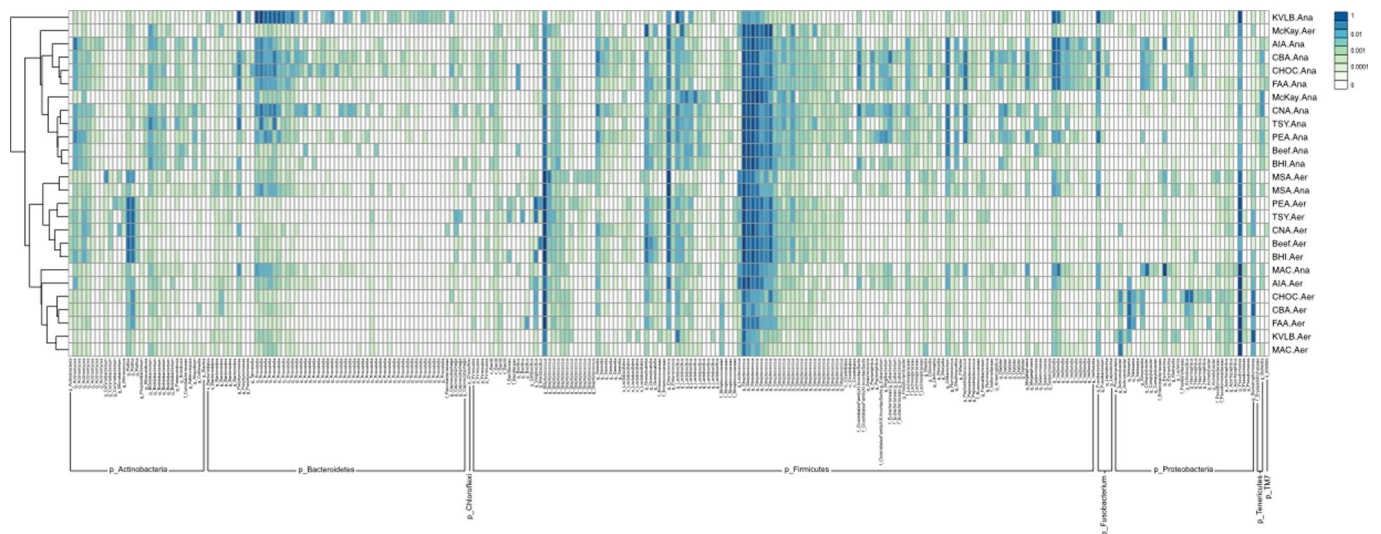
Extended Data Fig. 3 | Culture enriches for low abundance taxa. The first sample in our dataset was re-sequenced to a depth of 972,834 reads. A rank order curve, and associated quantification of the recovered OTUs (inset), illustrates the number of OTUs present in the original (41,199) and re-sequenced sample as well as at rarefied depths, indicating the increased recovery of cultured OTUs by direct sequencing as sequencing depth increases. The inset shows, at two different read cutoffs, that with increasing sequencing depth, the number of OTUs seen only in culture decreases.



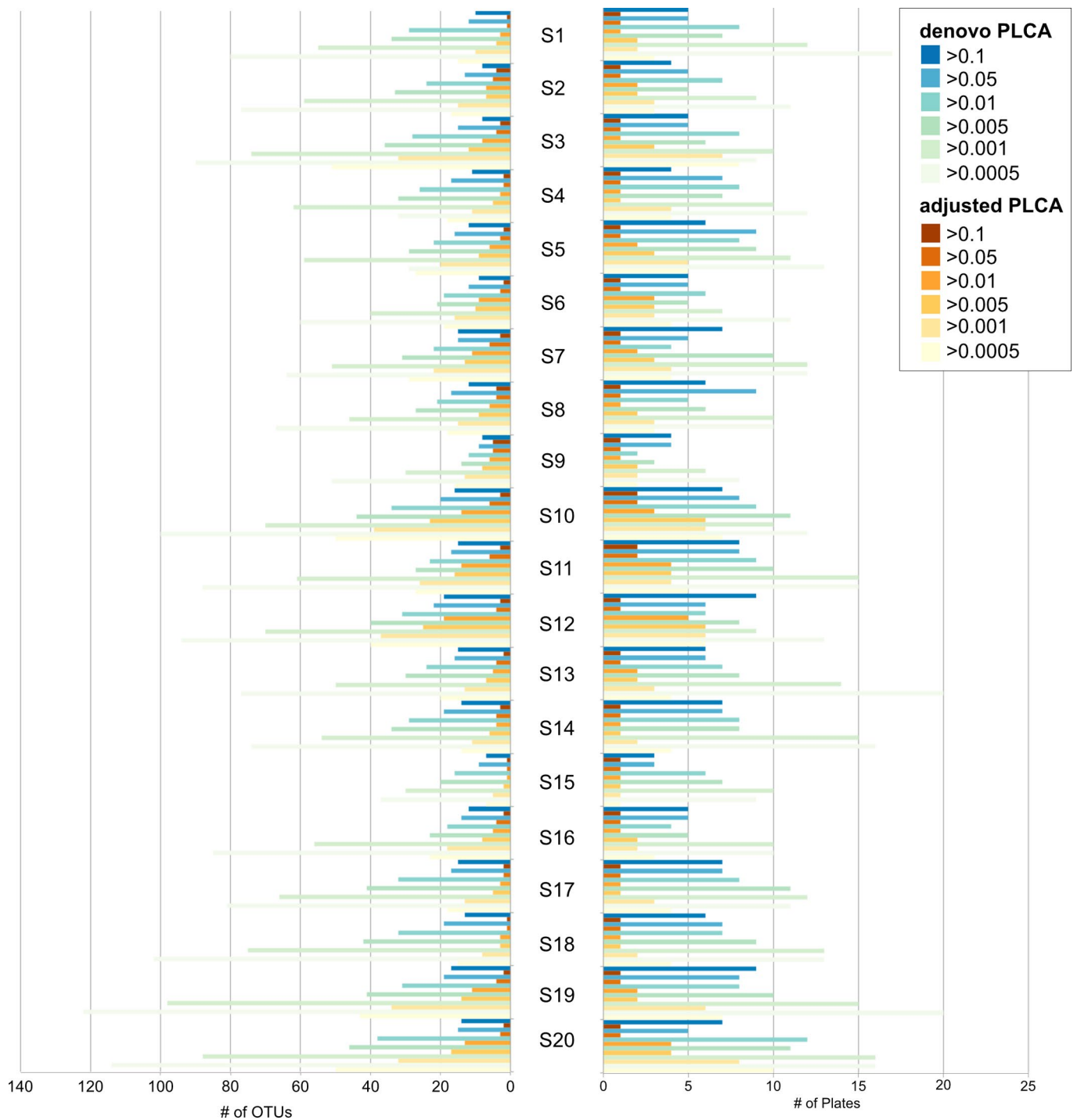
Extended Data Fig. 4 | No specific culture condition consistently recapitulates the originating sputum sample. Principle coordinate analyses based on the Bray Curtis β -diversity metric indicates that, across all 20 samples, no culture condition consistently recapitulates the microbial community of the originating sputum sample. For a list of media abbreviations see **Extended Data Table 1**; Aer = aerobic, Ana = anaerobic.



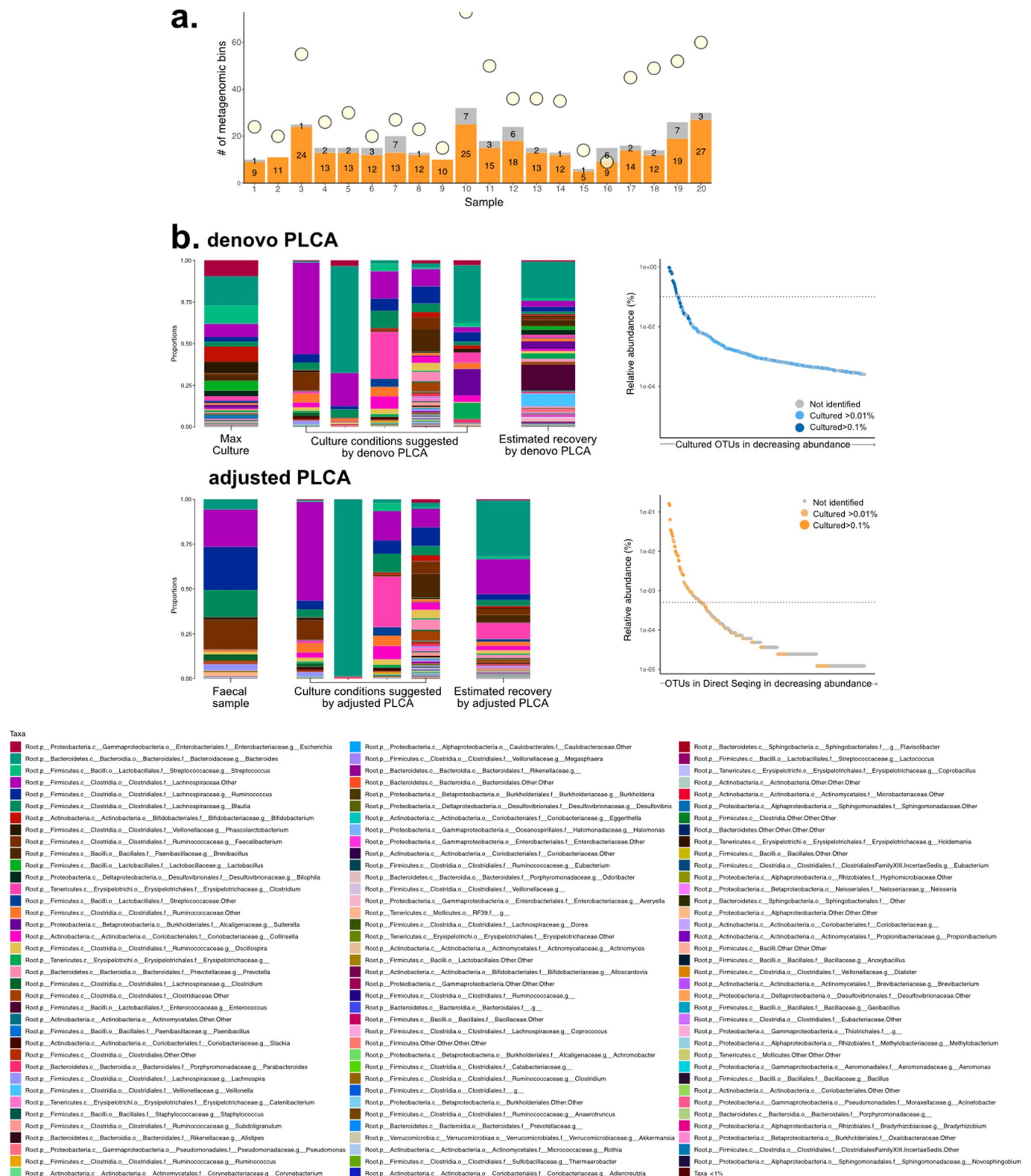
Extended Data Fig. 5 | The importance of selective and non-selective culture conditions in capturing the genus-level diversity of the cystic fibrosis microbiome. This Figure is a labelled version of Fig. 3c.



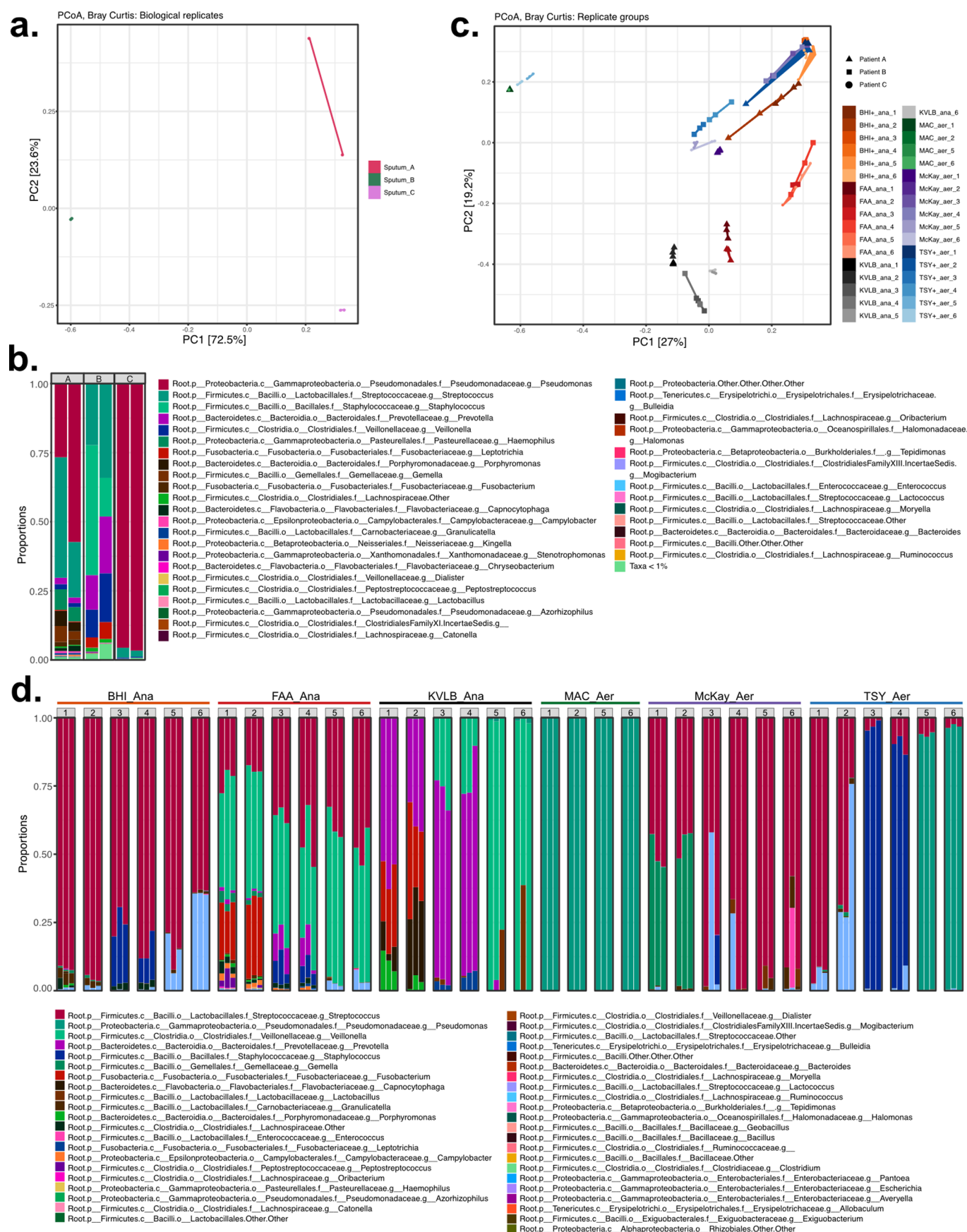
Extended Data Fig. 6 | The importance of selective and non-selective culture conditions in capturing the OTU-level diversity of the cystic fibrosis microbiome. A heatmap indicates the breadth of culture conditions necessary to culture this community at the OTU level.



Extended Data Fig. 7 | The optimal plate sets needed to sequence all samples (S) in this dataset, and the number of OTUs which would be obtained.



Extended Data Fig. 9 | The PLCA consistently recovers targeted OTUs and is not specific to the cystic fibrosis lung microbiota. **a.** When the adjusted PLCA was applied to the 20 samples in this cystic fibrosis dataset, it consistently recovered the targeted OTUs (orange), though some (gray) were not recovered as metagenomic bins due to inadequate sequencing depth, or the inability to separate species into separate bins. The overlaying numbers represent the number of OTUs in each category. Additional species obtained as a consequence of being present on a plate with a targeted OTU are shown as yellow dots. **b.** The PLCA is not specific to the cystic fibrosis lung microbiota or to a particular set of culture conditions; here, we apply the PLCA to previously published culture-enriched gut microbiota data (reference ¹²). Even through the culture conditions used by Lau *et al.* differ from those used in this study, the PLCA still predicts successful recovery of almost all species at abundances above the PLCA thresholds (dotted line).



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | Biological (a-b) and technical (c-d) replicates of sputum and culture-enriched 16S rRNA gene sequencing. PCoA plots of the Bray Curtis distance between sputum profiles show close clustering of 3 sputum biological replicates prior to plating, $n=6$ (**a**), and cultured replicates (technical replicates after plating of 3 sputum samples $\times$ 2 culture replicates of 6 representative media types, $n=108$, **c**). Polygons/lines connect the replicates in each plot; specifically between duplicate sputum sampling of 3 individuals (labelled A-C) in **a** and between triplicate platings of sputums in **c**. In **c**, the legend labels refer to the media type (for example BHI), environmental condition (Aer, aerobic or Ana, anaerobic), and replicate number (1-6). Taxonomic summaries of biological samples prior to plating (**b**) and 3 technical replicates each of 2 sputum samples from 3 patients (**d**) show the consistency of these techniques. Labelling of biological (A-C) and technical (1-6) replicates is consistent with parts **a** and **c**. In cases of no visible growth (that is MAC_Aer_3 and MAC_Aer_4) no samples were collected. In some cases, there was some difference in abundance of organisms between replicates (for example KVLB 5,6 and McKay 5,6) on which there were low bacterial counts and only at the lowest dilution plated (the cultured organisms representing $\sim 10^4$ CFU/ml and less than 1% in the original sputum sample). These will be subject to more variability in plating.

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- | | |
|-------------------------------------|--|
| n/a | Confirmed |
| ☒ | ☐ The exact sample size (<i>n</i>) for each experimental group/condition, given as a discrete number and unit of measurement |
| ☐ | ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☒ | ☐ The statistical test(s) used AND whether they are one- or two-sided
<i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i> |
| ☒ | ☐ A description of all covariates tested |
| ☒ | ☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| ☐ | ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| ☒ | ☐ For null hypothesis testing, the test statistic (e.g. <i>F</i> , <i>t</i> , <i>r</i>) with confidence intervals, effect sizes, degrees of freedom and <i>P</i> value noted
<i>Give P values as exact values whenever suitable.</i> |
| ☒ | ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| ☒ | ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| ☒ | ☐ Estimates of effect sizes (e.g. Cohen's <i>d</i> , Pearson's <i>r</i>), indicating how they were calculated |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection	16S rRNA gene sequencing data was collected via Illumina MiSeq v2 chemistry, and processed using sl1p v3.9 (default settings). Metagenomic shotgun sequencing data was collected via Illumina HiSeq 1500 rapid v2 chemistry, and processed using the following softwares: cutadapt v1.9.1, sickle v1.33, DeconSeq v0.4.3, Megahit v1.1.1-2-g02102e1, Maxbin v2.2.1, checkM v1.0.9, KrakenUniq v0.5, bowtie2 v2.2.6
Data analysis	Data analyses were performed using the following software (and packages): GraPhlAn v0.9.7, R v3.4.3 (ggplot2, pheatmap, phyloseq), QIIME v1.8.0 (rarefaction, taxa summaries), Hansel v0.0.81 & Gretel v0.0.81, minCED v0.2.0, eggno-mapper v1, CARD v2.0.2/RGI v4.1.0, Phaster (https://phaster.ca/), PRISM3. Additionally, computer code generated for this project is publicly available at the following Github repositories: https://github.com/shekas3/BinTaxaAssigner , https://github.com/fwhelan/PLCA .

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All sequencing results are publically available (BioProject IDL PRJNA503799).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	20 samples were used in this study: no sample size calculations were performed. The number of samples included were dependent on participant availability/consent in the study. This sample size was sufficient to answer our hypothesis of whether this human microbial community was culturable, and whether culture-enriched metagenomics could more readily identify low abundant organisms in these samples.
Data exclusions	Low abundant operational taxonomic units (OTUs) were excluded from this study. The rationale for this was to remove any possible contaminants, PCR errors, and/or any sequences which were erroneously assigned in silico.
Replication	All bioinformatic processing and analysis of this data was performed in order to make these results reproducible. Biological and technical replicates included in the manuscript show the reproducibility of sequencing sputum samples in short succession and of culturing these samples on various media types.
Randomization	Randomization was not relevant to this study. The only "groups" analyzed were in comparisons of cultured and not cultured samples, in which randomization does not apply.
Blinding	Blinding is not relevant to this study as all analyses were performed automatically with bioinformatic software i.e. there were no opportunities for human judgment to interfere with results.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics	All patients were adult (>18yrs) with cystic fibrosis. As within patient variability was the primary interest, no specific age, sex, or gender was used as a recruitment criteria. Clinical status (cftr genotype), FEV1, or other clinical indicators were also not used as requirement criteria.
Recruitment	Patients were approached at the onset of exacerbation as to whether they would like to be participants in this study. Willing patients were recruited that met the criteria for a pulmonary exacerbation during (using established criteria as defined by Fuchs et al.). Two samples were collected from each patient (with the noted exceptions): one at the onset of pulmonary exacerbation and a second during a follow up appointment (1week-4months) following the resolution of symptoms and antibiotic discontinuation.
Ethics oversight	Ethical approval granted by the Calgary Health Region Ethics Board, REB-24123.

Note that full information on the approval of the study protocol must also be provided in the manuscript.