AMERICAN SOCIETY FOR MICROBIOLOGY | mBio®

RESEARCH ARTICLE

Strain-Resolved Dynamics of the Lung Microbiome in Patients with Cystic Fibrosis

Marija Dmitrijeva,[a,b] Christian R. Kahlert,[c,d] Rounak Feigelman,[a,b]* Rebekka L. Kleiner,[e]* Oliver Nolte,[f] Werner C. Albrich,[d] Florent Baty,[e] Christian von Mering[a,b]

[a]Department of Molecular Life Sciences, University of Zurich, Zurich, Switzerland

[b]Swiss Institute of Bioinformatics, Zurich, Switzerland

[c]Infectious Diseases and Hospital Epidemiology, Children's Hospital of Eastern Switzerland, St. Gallen, Switzerland

[d]Infectious Diseases and Hospital Epidemiology, Cantonal Hospital St. Gallen, St. Gallen, Switzerland

[e]Pneumology and Sleep Medicine, Cantonal Hospital St. Gallen, St. Gallen, Switzerland

[f]Human Microbiology, Centre for Laboratory Medicine, St. Gallen, Switzerland

ABSTRACT In cystic fibrosis, dynamic and complex communities of microbial pathogens and commensals can colonize the lung. Cultured isolates from lung sputum reveal high inter- and intraindividual variability in pathogen strains, sequence variants, and phenotypes; disease progression likely depends on the precise combination of infecting lineages. Routine clinical protocols, however, provide a limited overview of the colonizer populations. Therefore, a more comprehensive and precise identification and characterization of infecting lineages could assist in making corresponding decisions on treatment. Here, we describe longitudinal tracking for four cystic fibrosis patients who exhibited extreme clinical phenotypes and, thus, were selected from a pilot cohort of 11 patients with repeated sampling for more than a year. Following metagenomics sequencing of lung sputum, we find that the taxonomic identity of individual colonizer lineages can be easily established. Crucially, even superficially clonal pathogens can be subdivided into multiple sublineages at the sequence level. By tracking individual allelic differences over time, an assembly-free clustering approach allows us to reconstruct multiple lineage-specific genomes with clear structural differences. Our study showcases a culture-independent shotgun metagenomics approach for longitudinal tracking of sublineage pathogen dynamics, opening up the possibility of using such methods to assist in monitoring disease progression through providing high-resolution routine characterization of the cystic fibrosis lung microbiome.

IMPORTANCE Cystic fibrosis patients frequently suffer from recurring respiratory infections caused by colonizing pathogenic and commensal bacteria. Although modern therapies can sometimes alleviate respiratory symptoms by ameliorating residual function of the protein responsible for the disorder, management of chronic respiratory infections remains an issue. Here, we propose a minimally invasive and culture-independent method to monitor microbial lung content in patients with cystic fibrosis at minimal additional effort on the patient's part. Through repeated sampling and metagenomics sequencing of our selected cystic fibrosis patients, we successfully classify infecting bacterial lineages and deconvolute multiple lineage variants of the same species within a given patient. This study explores the application of modern computational methods for deconvoluting lineages in the cystic fibrosis lung microbiome, an environment known to be inhabited by a heterogeneous pathogen population that complicates management of the disorder.

KEYWORDS cystic fibrosis, longitudinal study, lung sputum, metagenomics, strain typing

Citation Dmitrijeva M, Kahlert CR, Feigelman R, Kleiner RL, Nolte O, Albrich WC, Baty F, von Mering C. 2021. Strain-resolved dynamics of the lung microbiome in patients with cystic fibrosis. mBio 12:e02863-20. https://doi.org/10.1128/mBio.02863-20.

Editor David S. Guttman, University of Toronto

Address correspondence to Christian von Mering, mering@imls.uzh.ch.

* Present address: Rounak Feigelman, Paragon Genomics, Hayward, California, USA; Rebekka L. Kleiner, Department of Internal Medicine, Herisau Hospital, Herisau, Switzerland.

Received 9 October 2020
Accepted 30 January 2021
Published 9 March 2021

Cystic fibrosis (CF) is a monogenic, autosomal recessive, and life-shortening disease that predominantly affects the Caucasian population (1). The disease involves multiple organ systems but has its most severe consequences in the airways, where it leads to decreased mucociliary clearance followed by mucus plugging. Subsequently, the mucosal airways of CF patients are chronically inflamed and colonized by allochthonous microorganisms. The resulting respiratory symptoms include difficulty breathing, persistent cough, expectoration of sputum, and recurrent pulmonary infections. Respiratory failure accounts for more than half of CF patient deaths (2, 3). Nevertheless, improvements in CF management, such as antibiotic therapy and administration of mucolytic drugs, have increased the median life expectancy for patients, turning CF into a predominantly adult disorder (4). More recent therapies aim to directly ameliorate residual function of the protein encoded by the *CFTR* gene locus and have been shown to slow the rate of lung function decline in a subset of CF patients (5). Chronic respiratory infections, however, seem to persist even though respiratory symptoms improve (6). Therefore, improved characterization of persistent respiratory pathogens is needed to develop tailored therapies that control their composition and abundance.

Several common pathogens colonizing the lungs of CF patients are known. *Pseudomonas aeruginosa* is predominant in the adult CF population (2, 3). However, aggressive antimicrobial therapies aimed at reverting initial colonization by *P. aeruginosa* (7, 8) have recently led to a decline in its prevalence (2). Another key pathogen in CF is *Staphylococcus aureus*, which accounts for the majority of infections in young patients and has become increasingly more prevalent among all CF patients (2, 3). Other pathogens recognized in CF include members of the *Burkholderia cepacia* complex, mycobacteria such as *Mycobacterium avium* and *Mycobacterium abscessus*, *Stenotrophomonas maltophilia*, and members of the *Achromobacter* genus (2, 3, 9). Although the latter pathogens are present in a small fraction of CF patients, they are often multidrug resistant and, thus, challenging with regard to the treatment options in the clinic. Finally, anaerobic bacteria such as members of the *Prevotella* genus have also been identified in CF patient sputum using specialized culture techniques (10), but these typically are not assessed during routine clinical diagnosis, and their role as pathogens in CF patients has yet to be defined.

Culture-independent approaches are increasingly complementing and expanding on the findings of traditional microbiology approaches. For instance, studies using sequencing to characterize the lung microbiome have noted the presence of anaerobic bacteria not recognized as typical CF pathogens, such as *Prevotella* and *Veillonella*, in a sizable portion of the patients (11–13). In addition, culture-independent approaches uncovered a high level of variability across the lung microbiomes of CF patients (13–18). In late-stage patients, however, the microbiome generally tends to be lower in diversity and becomes dominated by one or a few of the commonly recognized CF pathogens (13, 14, 17). Several efforts have compared patient-matched samples from different clinical states but have not found significant reproducible changes between samples taken at baseline and at exacerbation, which suggests that the CF lung microbiome is resilient over time (11, 12, 14–16). Most culture-independent studies of the lung microbiome in CF, however, have been performed using 16S rRNA sequencing and, thus, provide only limited insights into the functions or strain identities of lung microbial communities.

Whole-genome sequencing (WGS) and metagenome sequencing improve on the limited taxonomic and functional resolution of 16S rRNA sequencing. Metagenomics allows us to survey bacterial, viral, and fungal populations at once, giving a more complete picture of microbial relative abundances in the CF lung microbiome (17, 19). Consequently, a larger portion of microbiome inhabitants can be classified at the species level (17), and prominent CF pathogens have been classified at the strain level (17, 18, 20). Moreover, multiple subpopulations of specific pathogens have been detected in CF through metagenome sequencing (17, 18). However, so far only single reference points per patient were typically sequenced, limiting haplotype deconvolution and preventing insights into the temporal dynamics of these lineages.

Here, we describe longitudinal sputum sampling in CF patients over the course of one and a half years, conducting metagenomics sequencing of spontaneously expectorated sputum at multiple time points. The aim of the study was to investigate the advantages of collecting longitudinal data of CF patients for monitoring and characterizing lineage successions *in situ*. We successfully classified most of the lung microbiome members at the species and genus level and confirmed the presence of pathogens identified during routine clinical diagnosis. Importantly, we show how longitudinal metagenomics data can be used to deconvolute distinct lineage variants of the same species within a given patient. We introduce an assembly-free approach that can delineate nearly complete, lineage-specific genomes even when their sequence divergence is fairly low. Our study introduces culture-independent methods that can be used in the future for monitoring pathogen lineages in CF.

RESULTS

Patient-specific lung microbiomes. We monitored four CF patients selected from a larger cohort over the course of 19 months (Fig. 1; see also Fig. S1, S2, and S3 in the supplemental material). A summary of patient information, clinical parameters, and prescribed medication is available in Data Set S1 at https://string-db.org/suppl/Dataset_S1_Strain-resolved_Microbiome_Dynamics_in_Cystic_Fibrosis.xlsx. During the course of our study, the patients produced sputum spontaneously, either during routine clinical check-ups or during exacerbations. In total, 25 samples were collected. We extracted total DNA from the collected sputum, enriched for nonmethylated DNA, and sequenced it using the Illumina HiSeq 4000 platform. Sequencing depth varied between 21 million reads and 179 million reads (Fig. 1F and Fig. S1F, S2F, and S3F). Human DNA constituted between 70% and 93% of all reads. We observed no significant associations between the total DNA concentration in the sample and the sequencing depth or the fraction of nonhuman reads. Nonhuman DNA predominantly originated from bacteria; viruses (including bacteriophages) accounted for, at most, 1.5% of reads, and fungi accounted for, at most, 0.15% of the reads (as profiled by MiCoP) (21). More than 95% percent of the bacteria at each time point could be identified to at least the genus level using the mOTUs software (22) (Fig. 1D and Fig. S1D, S2D, S3D, and Data Set S1 at the URL mentioned above, mOTUs), indicating that the lung microbiomes of these patients largely consisted of previously characterized bacterial clades.

Lung microbiome compositions showed marked differences between the four patients (Fig. S4). For instance, the lung microbiome profile of patient CFR06 contained between 60% and 93% of anaerobic bacteria, including the oral anaerobes *Prevotella*, *Parvimonas*, and *Fusobacterium*, and was the only patient devoid of any detectable *Pseudomonas* (Fig. S1D). Typical CF pathogens identified by the clinical microbiology laboratory, *A. xylosoxidans*, *H. influenzae*, and *S. aureus* (Fig. S1C), and their corresponding genera accounted for less than a fourth of the bacterial content (Fig. S1D). This patient was several years younger than the other subjects and displayed a milder form of CF, with the highest average forced expiratory volume (FEV), a measure of lung function (95% confidence interval [CI], 56.8% ± 6.4% versus 30.7% ± 0.7% in CFR07, 40.0% ± 3.1% in CFR09, 32.5% ± 3.3% in CFR11) (Fig. S1A). Around 7 months into the study, the patient had an exacerbation that was treated with a combination of piperacillin-tazobactam and intravenous tobramycin (Fig. S1B). Following this event, the lung microbiome composition of CFR06 experienced a major shift: *Prevotella buccae*, *S. aureus*, and *A. xylosoxidans/insuavis* all decreased in relative abundance. Concomitantly, *Prevotella oris*, *Fusobacterium nucleatum*, and *Gemella morbillorum* increased in relative abundance (Fig. S1D).

In contrast, the lungs of patients CFR07 and CFR09 were colonized predominantly by *P. aeruginosa* (Fig. S2C and D and S3C and D), with samples from these patients clustering together (Fig. S4). Patient CFR07 displayed a stable clinical phenotype, with no

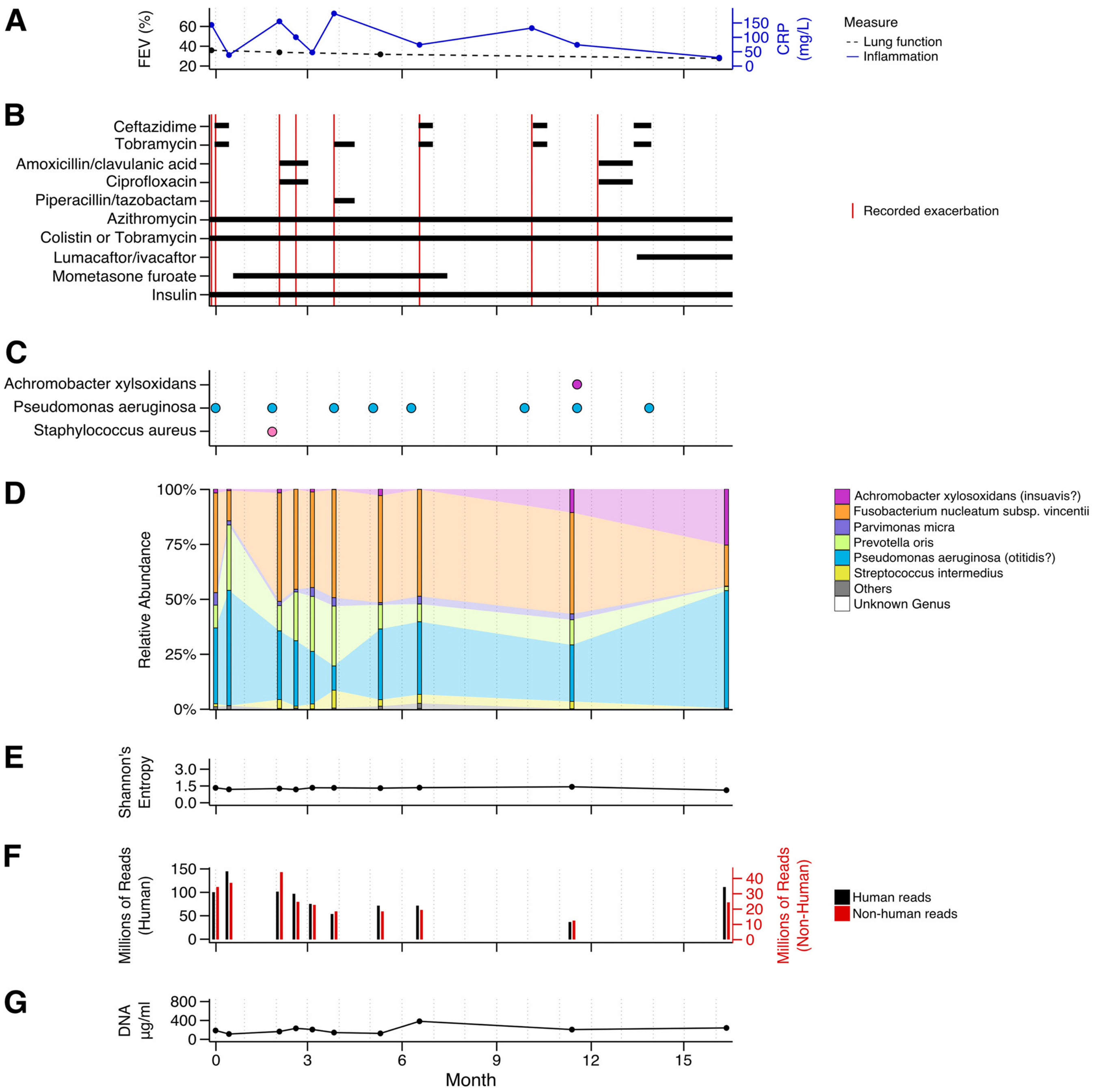

FIG 1 Study report of patient CFR11 displaying the dynamics of multiple parameters over time. (A) Percent forced expiratory volume (FEV) (black) and concentration of C-reactive protein (CRP) (blue), with actual measurements shown as dots. (B) Medication assigned to the patient during the course of the study and recorded exacerbation events (in red). (C) Bacteria identified in the clinical microbiology laboratory. (D) Relative abundance profiles generated by mOTUs, with actual measurements shown as bars. Selected species and their corresponding genera with more than 5% relative abundance at at least one time point are shown color-coded. Less abundant species are grouped into "Others" (gray). Taxa that could not be identified by mOTUs on the genus level are grouped into "Unknown Genus" (white). (E) Shannon's diversity index (entropy) calculated based on the relative abundance profiles generated by mOTUs, with actual measurements shown as dots. (F) Number of reads per sample. Human reads are indicated in black and plotted on the left axis. Reads that did not concordantly map to the human genome are depicted in red and plotted on the right axis. (G) Concentration of total DNA isolated from patient sputum, with actual measurements shown as dots.

exacerbations recorded during the course of the study, and retained FEV at 30% (Fig. S2A and B). *P. aeruginosa* accounted for more than 90% of all bacteria in this patient's lung microbiome, resulting in the lowest microbiome diversity of all examined patients (average Shannon's diversity index with 95% CI, 0.48 ± 0.09 versus 2.29 ± 0.44 in CFR06, 1.79 ± 0.93 in CFR09, and 1.29 ± 0.06 in CFR11) (Fig. S2E). The lung microbiome of patient CFR09 contained, in addition to *P. aeruginosa*, up to 17.5% and 9.2% of the genera *Prevotella* and *Veillonella*, respectively, and various low-abundance genera that comprised up to 35% of the lung microbiome (Fig. S3D).

Finally, patient CFR11 presented the most severe course of disease (Fig. 1) and died shortly after study completion. The patient experienced multiple exacerbations, accompanied by high levels of inflammation, with FEV gradually declining from 36% to 28% (Fig. 1A and B). The lung microbiome of CFR11 was dominated by *P. aeruginosa* (Fig. 1C and D) and the oral anaerobes *P. oris* and *F. nucleatum* (Fig. 1D). In addition, *Parvimonas micra*, *Streptococcus intermedius*, and *A. xylosoxidans/insuavis* were present in lower relative abundances, with the fraction of *Achromobacter* increasing at later time points (Fig. 1D), to the point of also being detected using standard clinical microbiology procedures (Fig. 1C). Thus, in three of the patients, the most abundant bacteria remained the same throughout the course of the study, and only CFR06 displayed a major sustained shift in lung microbiome composition (Fig. S4). From all identified bacteria, the most relevant from a clinical perspective were *A. xylosoxidans* (identified in two patients, CFR06 and CFR11) and *P. aeruginosa* (identified in three patients, CFR07, CFR09, and CFR11). Therefore, we set out to assess to what extent cultivation-independent sequencing would allow us to classify these pathogens in more detail.

Strain-level classification of *Achromobacter*. Seven of the sputum samples contained sufficient reads to provide a 2-fold median coverage of the *A. xylosoxidans* pangenome (one sample from CFR06 and six samples from CFR11), and the corresponding assembled contigs showed more than 80% expected genome completeness according to CheckM (23) (Data Set S1 at https://string-db.org/suppl/Dataset_S1_Strain-resolved_Microbiome_Dynamics_in_Cystic_Fibrosis.xlsx, assembly reports). From a selection of 22 fully sequenced *A. xylosoxidans* reference genomes, *A. xylosoxidans* FDAARGOS_147, a strain isolated from a patient at Children's National Hospital in Washington, DC, was the only genome with more than 50% gene family overlap with the *Achromobacter* contigs from the patients (Fig. S5A and B). Therefore, we decided to compare marker gene sequence identity to place our samples within the *Achromobacter* genus.

From 144 *Achromobacter* genomes in the NCBI genome database (November 2018) (24), *Achromobacter* genus trees were constructed using the sequences from the 10 single-copy genes used by mOTUs, sequences from the seven genes used in standard *Achromobacter* MLST (25), or pairwise genome average nucleotide identities (see Materials and Methods). The three genus trees were more consistent with each other (average normalized Robinson-Foulds distance, 0.62) than to a phylogenetic tree informed by a single marker gene, such as 16S rRNA (average normalized Robinson-Foulds distance, 0.90). All three trees revealed some discrepancies with the NCBI taxonomy (Fig. 2, NCBI columns). Out of the eight clades containing more than one genome, only *A. marplatensis* and *A. insolitus* were monophyletic in all three trees. Conversely, our analyses clustered together genomes assigned to different species with more than 90% bootstrap support. For example, one well-supported cluster contained genomes from *A. ruhlandii*, *A. denitrificans*, and *A. xylosoxidans* (Fig. 2A). As has been noted previously, taxonomic classification within the *Achromobacter* genus has inconsistencies (26). The Genome Taxonomy Database (GTDB) is a recent effort to redefine prokaryotic taxonomy and improve on such inconsistencies in species assignment (27). To determine whether this approach could be of help here, we downloaded the species assignments from GTDB (as of April 2019). Indeed, seven of the nine multigenome *Achromobacter* clades defined by GTDB were monophyletic in all three trees, and all nine clades were monophyletic in the mOTU tree (Fig. 2, GTDB columns).

On all three *Achromobacter* genus trees, we observed that our patient-derived *Achromobacter* genomes and *A. xylosoxidans* FDAARGOS_147 clustered with two *A. insuavis* genomes. The lineage in patient CFR11 was 99.99% identical to *A. insuavis* AXX-A, a strain observed at the Laboratory of Bacteriology at the Faculty of Medicine in Dijon, France (Fig. 2A, zoom-in). The lineage in patient CFR06 clustered

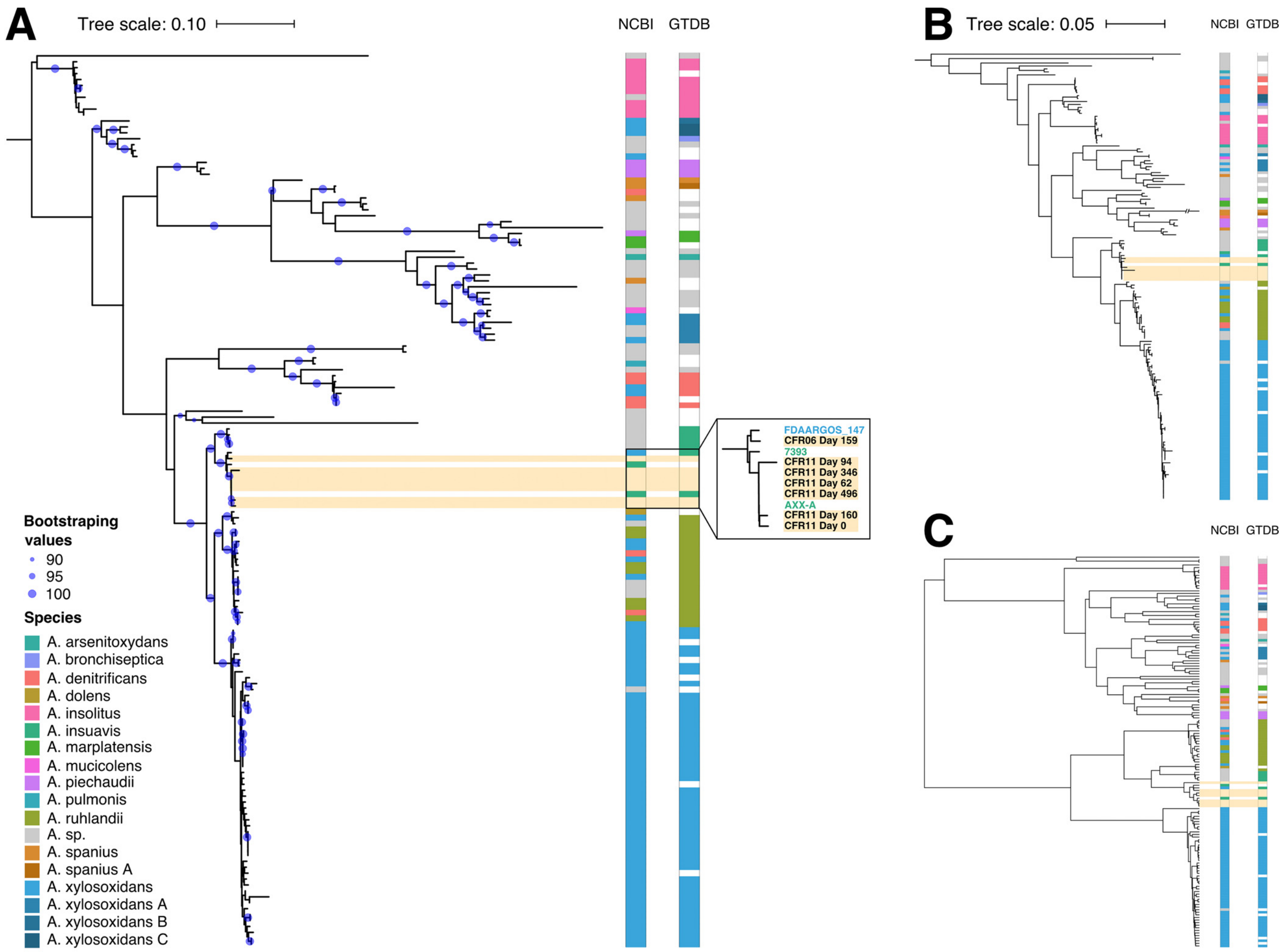

FIG 2 Strain-typing of patient-specific *Achromobacter* in the context of a sequence-based phylogeny of the genus. (A) Maximum-likelihood tree based on sequences of 10 single-copy genes used by mOTUs. Colors on the right of the tree depict species assignments according to NCBI and GTDB taxonomies and apply to panels B and C as well. Blue circles indicate branch confidence values (≥90) based on 100 bootstraps of the tree. (B) Maximum-likelihood tree based on sequences of seven housekeeping genes from the *Achromobacter* MLST database. (C) UPGMA clustering of pairwise average nucleotide identities of the comprising *Achromobacter* genomes.

with *A. xylosoxidans* FDAARGOS_147 and was 99.24% identical to it (Fig. 2A, zoom-in). Taken together, these results indicated that the clinical laboratory misidentified this pathogen, incorrectly reporting *A. xylosoxidans* instead of *A. insuavis*.

Strain-level classification of *P. aeruginosa*. Next, we asked how well strain identification performs in the case of *P. aeruginosa*, a species for which more comprehensive reference information is available. Patients CFR07, CFR09, and CFR11 all carried *Pseudomonas* in sufficient amounts to allow 99% estimated genome coverage (Data Set S1 at https://string-db.org/suppl/Dataset_S1_Strain-resolved_Microbiome_Dynamics_in_Cystic_Fibrosis.xlsx, assembly reports). To identify the lineages, we compared our samples to a representative set of 359 *P. aeruginosa* genomes based on gene family presence/absence profiles (28) (Fig. S5C) and marker gene sequence identity (Fig. S5D). Patients CFR09 and CFR11 harbored a lineage with single-copy gene sequences 100% identical to those of *P. aeruginosa* PAER4_119 (first sampled in Poland), although some differences in gene family content were present (Fig. S5C and D). The samples from CFR07, however, contained gene families distinct from those of the other patients and clustered with other reference genomes (Fig. S5C). Indeed, the tree based on marker genes revealed that CFR07 is infected by a new lineage that was, at most, 99.73% identical to the genomes in the representative set (Fig. S5D).

Our work with *Pseudomonas* indicated that more than one variant of the lineage was likely present in some of the patient samples, most notably in CFR11. Both PanPhlAn and CheckM presented corresponding warnings but could not further delineate these lineage variants.

Delineation of *P. aeruginosa* lineage variants without cultivation or genome assembly. To distinguish and track *P. aeruginosa* lineage variants in patient CFR11, we took advantage of the repeated samplings in our time series data. Under the assumption that the relative abundances of competing *P. aeruginosa* populations in a patient would vary over time, any population-specific sequence variants should similarly vary over time. This would allow reconstructing constituent genomes through clustering sequence variants by their shared temporal behavior. To test this approach, we mapped all apparent *Pseudomonas* reads from patient CFR11 to the closest reference genome (*P. aeruginosa* PAER4_119), which served as a scaffold (Fig. 3, steps 1 and 2). We then called single-nucleotide variants (SNVs) using metaSNV (29) and determined their allele frequencies at each time point (Fig. 3, steps 2 and 3). Finally, we determined clusters of SNVs displaying similar changes in allele frequencies over time (Fig. 3, step 4).

A total of 3,451 SNVs were called, of which 3,079 had allele frequencies detected at every time point, with at least one allele frequency not equal to one. Repeated t-distributed stochastic neighbor embedding (t-SNE) runs at slightly varying settings consistently yielded seven distinct clusters of SNVs in addition to a pool of lower-frequency SNVs that could not be reliably clustered (Fig. 4A). Of the seven clusters, each showed a clearly distinct pattern of allele frequencies over time (Fig. 4B). Cluster 3 appeared to be a linear sum of clusters 2 and 6 ($P < 2.2E-16$; comparison between the sum of SNV allele frequencies from the aforementioned clusters over 10 time points to that of the same SNVs but with the time points shuffled) and to inversely correlate with cluster 1 ($P < 2.2E-16$), not differing from it significantly in the extent of temporal variation ($P = 0.38$; comparison of standard deviations in individual SNV allele frequencies between clusters). Clusters 6 and 7 followed a somewhat shared pattern over time, with the exception of the first time point ($P < 2.2E-16$), and did not significantly differ in their extent of temporal variation ($P = 0.10$). Cluster 5 exhibited significantly less temporal variation than cluster 6 ($P = 3.4E-25$), and cluster 4 exhibited the least temporal variation.

To investigate the clusters more carefully, we plotted the spatial positioning of the cluster-specific SNVs in the reference genome (Fig. S7A). We found no association between the distance of SNVs from the same cluster on the chromosome and the similarity in their temporal profiles (Fig. S7C), indicating that the differences in allele frequency patterns were not simply due to recombination of selected genomic regions containing multiple SNVs. Neighboring SNVs from clusters 1, 2, 3, 6, and 7 were located within the expected range of distances, indicating homogeneous distribution, but neighboring SNVs from clusters 4 and 5 were closer to each other than expected ($P < 1E-04$), indicating the concentration of SNVs in selected genomic regions (Fig. S7B). Together with the fact that these clusters exhibited less SNV temporal profile cohesiveness than clusters 1, 2, 3, 6, and 7 (data not shown), we interpret such SNVs to reflect intragenomic polymorphisms in relation to the reference genome (e.g., at tandem-repeat regions) that were artifactually clustered together. Hence, clusters 4 and 5 were discarded. Considering the remaining clusters, the most parsimonious interpretation of the data appears to be the presence of three distinct *P. aeruginosa* lineage variants in the patient (reflected in clusters 1, 2, and 6/7, respectively). In this scenario, cluster 3 would consist of SNVs that are ancestrally shared between two of the variants and whose frequencies reflect the sum of their relative abundances.

Because t-SNE is a nondeterministic algorithm, we sought to validate our observations. Therefore, we used the called SNVs to perform principal-component analysis (PCA) combined with hierarchical clustering and to run DESMAN, a tool developed for grouping SNVs into haplotypes by assessing the variation of nucleotide base frequencies across samples and by using a Bayesian model to resolve possible sequencing errors and SNVs that are shared between more than one strain (30). The clusters generated based on PCA were largely consistent, the only deviation being a merging of

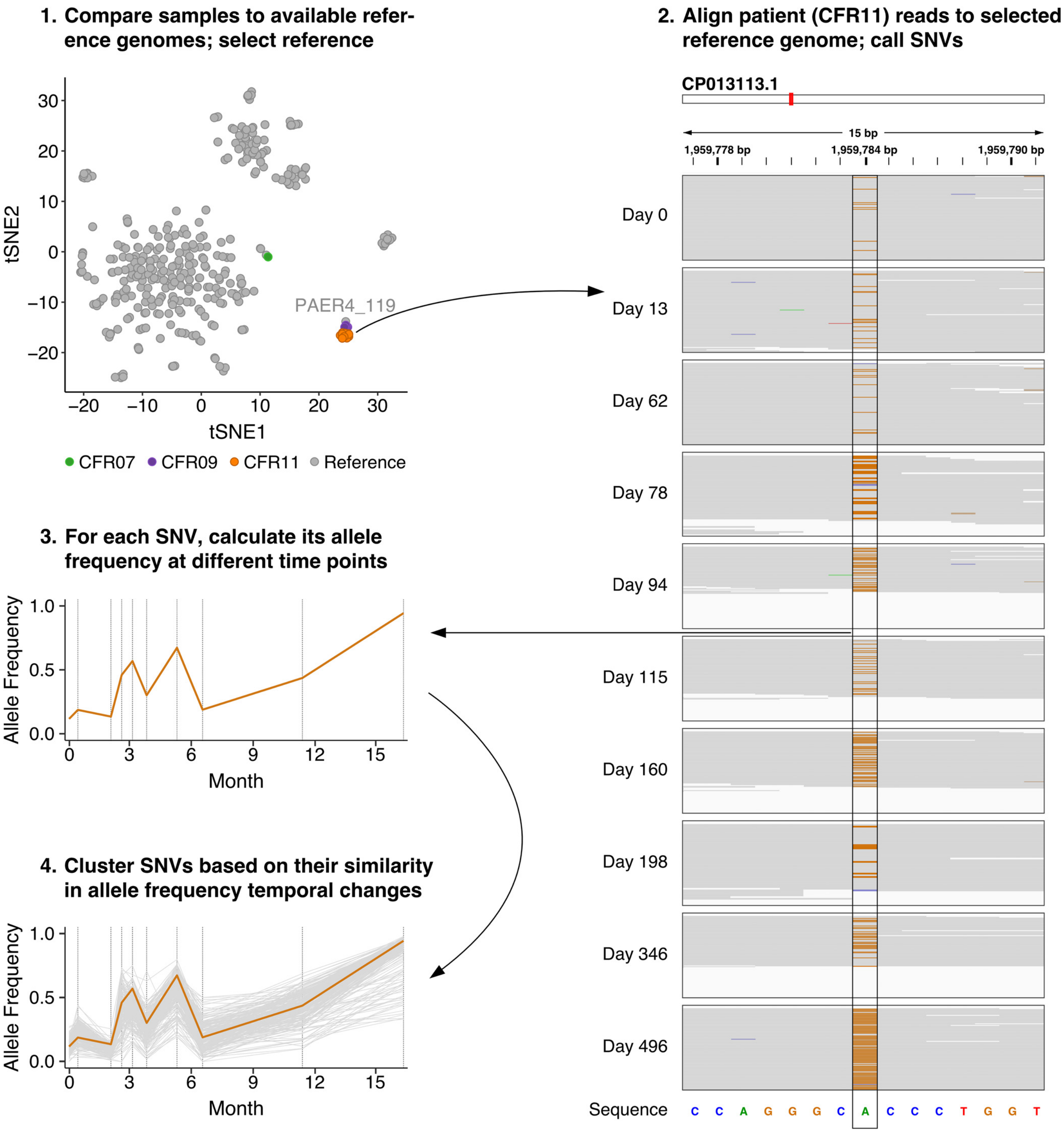

FIG 3 Identification of lineage variants through assessment of temporal changes in SNV allele frequencies in the metagenomics data of patient CFR11. (Step 1) Selection of a reference genome based on generated gene family presence/absence profiles. (Step 2) Read mapping of CFR11 samples to the reference genome (CP013113.1). A pile-up of a selected region containing an SNV (1,959,777 to 1,959,791 bp) is shown for every time point. The reference sequence is displayed on the bottom. Gray indicates read base pairs that are identical to the reference sequence. Orange indicates that a substitution to guanine has occurred. (Step 3) The change in allele frequency over time for the selected SNV. (Step 4) A group of SNVs that show a similar pattern of temporal changes in allele frequencies. The selected SNV is depicted in orange. The explicit steps performed and tools used in this approach can be found in a flow chart in Fig. S6.

t-SNE clusters 6 and 7 (Fig. S8A and B). The three haplotypes yielded by DESMAN coincided with clusters 1, 2, and 6 (Fig. S8C and D). Thus, we could validate the majority of SNVs that were clustered together in t-SNE (Fig. S8E). In addition, we could confirm via additional long-read sequencing that SNVs that were observed to cluster together by

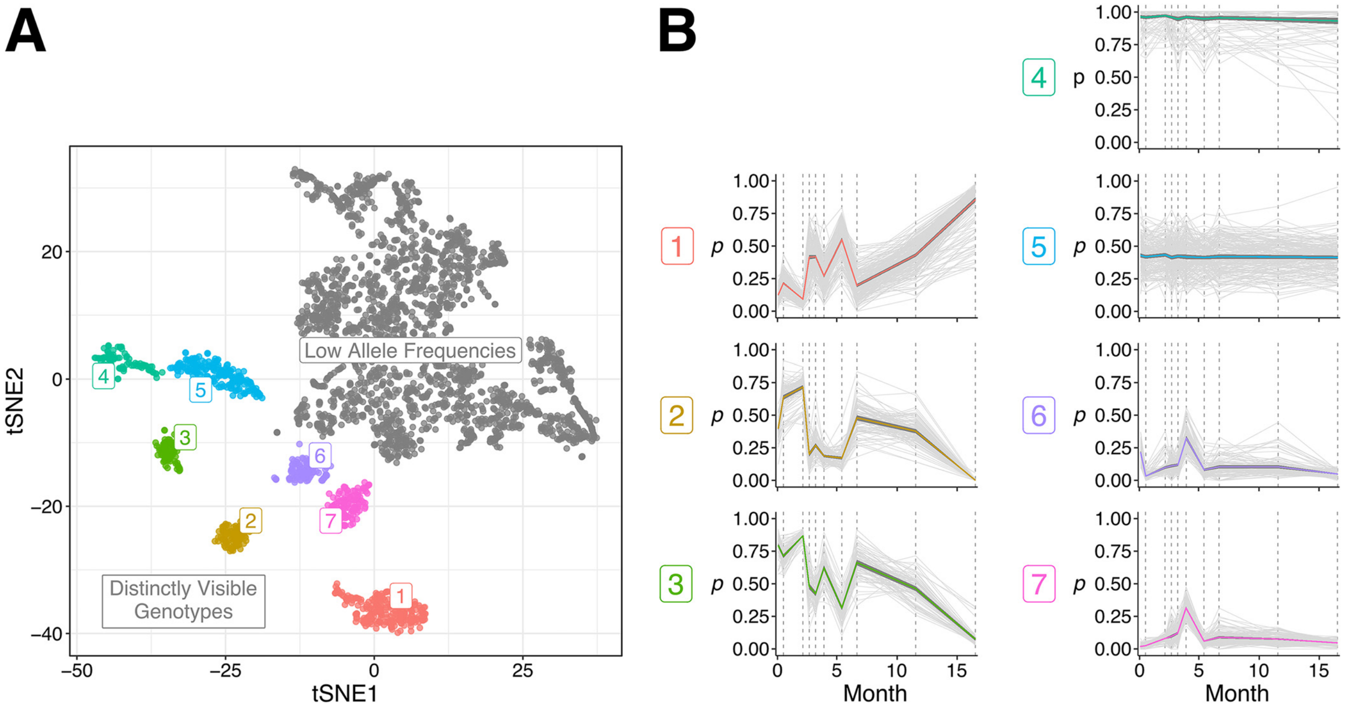

FIG 4 Clustering of SNVs detected in patient CFR11 based on their temporal changes in allele frequencies. (A) A t-SNE plot depicting the clustering pattern of 3,079 SNVs called in patient CFR11. Most SNVs occur at low allele frequencies (gray). The remaining SNVs form seven distinctly visible genotypes that are labeled and colored accordingly. (B) Changes in the allele frequencies (*p*) of SNVs belonging to each distinct genotype over time. The colored line indicates mean allele frequency of the genotype. Dark gray ribbons indicate the 95% confidence intervals.

all three methods indeed occurred on the same DNA molecule significantly more often than expected based on their individual allele frequencies alone ($P < 1E-04$) (Fig. S9). Taken together, our results suggest the coexistence of three lineage variants that, notably, would have been impossible to distinguish using the 16S rRNA gene alone (Fig. S10). Likewise, at the observed pairwise divergence of less than 0.01% between the variant genomes, traditional genome assembly approaches also would likely not be able to distinguish these (31).

Subsequently, we focused on SNVs that were assigned to the same cluster or haplotype by all three methods (Fig. S8E). Out of the 563 SNVs from all three lineage variants, 502 overlapped a gene in *P. aeruginosa* PAER4_119, with 459 genes containing at least one SNV (Data Set S1 at https://string-db.org/suppl/Dataset_S1_Strain-resolved_Microbiome_Dynamics_in_Cystic_Fibrosis.xlsx, diagnostic SNV genes). We next wondered whether genes known to be mutated in CF (32) would be preferentially mutated in our lineage variants. Only variant 2 had a borderline significant enrichment of mutations in these genes compared to the rest of the genome ($P = 0.03$, Fisher's exact test), while variants 1 and 3 had no enrichment compared to the rest of the genome (*P* values of 0.21 and 0.15, respectively, Fisher's exact test).

At least 100 SNVs separated each lineage variant from the other (Table 1). The rate of mutations in *P. aeruginosa* has been estimated to be, at most, 5.5 SNVs per year (33–35), unless a hypermutator phenotype develops (34, 36). To test for poten-

TABLE 1 Number of SNVs consistently clustered by three different approaches (t-SNE, PCA, and DESMAN)

SNV type	No. of SNVs
Lineage variant 1 specific	204
Lineage variant 2 specific	106
Lineage variant 3 specific	186
Shared between 2 and 3	67

tial hypermutator mutations, we mapped reads to the six DNA repair genes known to be affected in hypermutator strains: *mutS*, *mutL*, *uvrD*, *mutM*, *mutY*, and *mutT* (35–38). We detected only one mutation that might disrupt gene function, a frameshift deletion in the *mutS* gene. However, this mutation was observed only in reads corresponding to lineage variant 1 (not variant 2 or 3), and the mutation was positioned toward the 3′ end of the gene, close to the stop codon of the predicted open reading frame.

Detection of *P. aeruginosa* variant-specific structural genome differences. To determine whether longitudinal metagenomics sequencing provides sufficient evidence to detect large-scale genomic variation between lineage variants, we mapped reads to the reference *P. aeruginosa* PAER4_119 genome and calculated the average read coverage for windows of 1,000 bp along the genome. After normalizing for higher read coverage around the origin of replication (39), any large-scale genomic differences between lineage variants should become evident as time point-dependent deviations in read coverage.

Indeed, the mean coverage in 70 windows was at least two standard deviations from the overall mean coverage, combining into seven regions of the genome that spanned more than two windows each (Fig. 5, circos diagram). Of these, the regions located around 4.4 Mbp, 4.9 Mbp, and 5.5 Mbp coincided with predicted phage sequences; structural variations at phage insertions are to be expected. A fourth region around 2.7 Mbp showed inconsistent and overall low coverage in our sample data and was not considered further. This left three regions of interest. Manual inspection allowed us to pinpoint region borders more precisely at 1,639,504 bp to 1,649,747 bp and 3,024,303 bp to 3,088,958 bp for two of the regions. The third region was revealed to be composite, its borders at 3,228,483 bp to 3,235,055 bp and 3,241,118 bp to 3,243,808 bp. We further refer to these selected regions as regions 1, 2, and 3.

Each of the three regions showed significant correlations to the relative abundance profiles of one of our inferred lineage variants (Fig. 5, line plots). Region 1 highly correlated with the relative abundance profile of variant 1 (95% CI, $0.9731 < r < 0.9986$; $P = 6.3E-09$), indicating that it was present in that lineage but absent from variants 2 and 3. The genes in region 1 included multiple metabolism-related genes, but its overall functional significance was difficult to assess. Region 2 correlated with the relative abundance profile of variant 3 (95% CI, $0.8447 < r < 0.9913$; $P = 8.3E-06$). This region contained pyoverdine synthetases and genes from the type VI secretion system. We observed that two SNVs in the region mapped to variant 1, suggesting that it is present in all three lineage variants but amplified at least 2-fold in variant 3. Several dozen paired-end reads spanned from the end of the region back to its start, suggesting the additional copies in variant 3 are either arranged as tandem duplications or form an excised plasmid. Finally, the average coverage of region 3 correlated very well with the combined relative abundance profile of variants 2 and 3 (95% CI, $0.9713 < r < 0.9985$; $P = 8.2E-09$), indicating that it was absent from variant 1. This region contained multiple genes of the *psl* operon, which plays a role in biofilm generation in *P. aeruginosa*. Taken together, it becomes clear that the information contained in the relative read coverage over time, when correlated with the relative proportions of SNVs, allows precise and confident mapping of large-scale genomic structure variants to their respective lineage variants.

DISCUSSION

In this study, we sought insights into the temporal dynamics of the lung microbiome in CF by using noninvasive DNA sequencing of lung sputum. By repeated sampling of the lung microbiome over several months, we were able to distinguish persistent pathogens from other, more transient community members. The taxonomic identification of pathogens from sequencing data were generally in line with the clinical microbiology laboratory reports, although in the case of *Achromobacter* and *Pseudomonas*, the sequence-based identifications proved to be more precise.

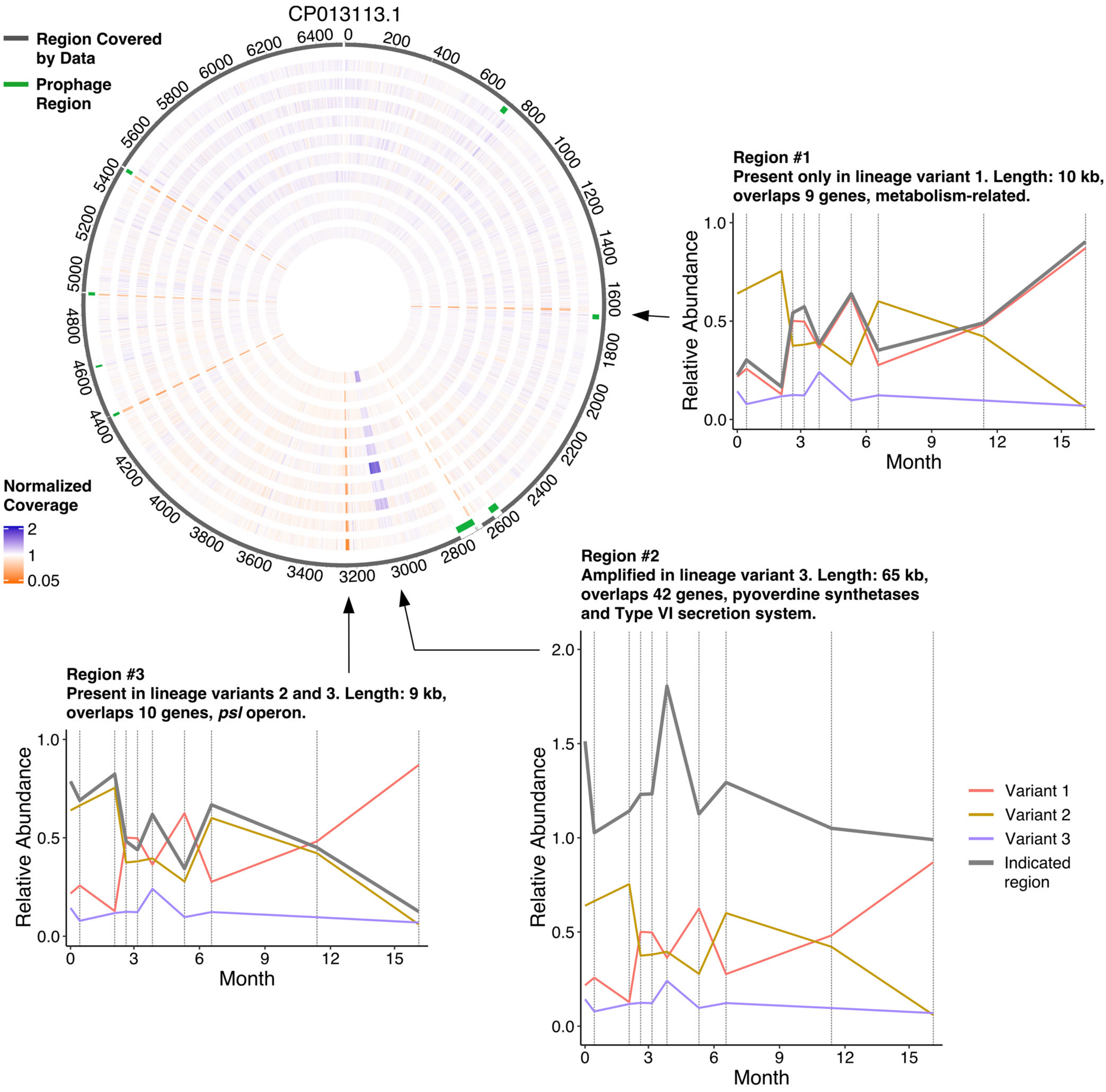

FIG 5 Assessment of temporal variation in the coverage of specific regions in the genome of *P. aeruginosa* in patient CFR11. The circos diagram provides an overview of the genome coverage profiles with chromosomal coordinates in kb. The dark gray outer circle depicts regions of the reference genome that have a coverage of at least 5% from the average coverage at at least one time point. The second outermost circle depicts detected phage regions in green. The remaining circles depict normalized coverage profiles for each of the 10 time points sampled for patient CFR11 (innermost, day 0; outermost, day 496). Orange regions indicate lower than average coverage, and purple-blue regions indicate higher than average coverage. Insets highlight three regions that display variant-specific coverage profiles. Each variant is depicted in a distinct color, and the average coverage of the selected region is depicted in gray.

Moreover, with *P. aeruginosa* in patient CFR11, at least three distinct lineage variants were observed. Importantly, distinguishing these variants would not have been possible without the longitudinal repeated samplings, showing that tracking patients over time provides valuable added information. To our knowledge, only two other shotgun metagenomics studies with multiple reference points per CF patient have been published (20, 40). Both studies perform strain typing for recognized pathogens but do not explore longitudinal genomic variation on a sublineage level.

Longitudinal data from patient CFR11 allowed us to delineate at least three lineage variants with distinct temporal dynamics for one of the best-covered pathogens, *P. aeruginosa*. Multiple studies have described genotypically and/or phenotypically distinct *P. aeruginosa* subpopulations (38, 41–48) and provided insights into their abundance fluctuations over time (43, 48) by sequencing cultured isolates from CF patients' lungs. Shotgun metagenomics sequencing approaches have indicated that *P. aeruginosa* is polymorphic in some patients (17, 18), as are some other CF pathogens (17, 18, 49); however, these studies were limited to a single time point for most patients, providing no insight into subpopulation dynamics. Outside CF, in a more controlled *in vitro* setting, a conceptually similar approach to ours has been used to study the molecular evolution of *E. coli* populations over 60,000 generations, leading to the recognition of coexisting clades (50). Here, we provide a proof of principle that this is also possible in a clinical setting, *in vivo*, without prior knowledge of which pathogen strains to expect in a patient.

The emergence of phenotypically and genotypically distinct subpopulations of *P. aeruginosa* in CF through lineage diversification has previously been shown to be driven by spatial heterogeneity (44, 51). Lung regions differ in oxygen and carbon dioxide concentrations (52), patterns of ventilation and deposition (53), and disease burden (54). General microbial community composition differs depending on lung region as well (55, 56). Nevertheless, other studies have found no clustering of *P. aeruginosa* isolates based on region of isolation (57) or have shown identical phenotypes and genotypes in upper and lower airways (58, 59). Sputum sequencing does not provide us with information on the spatial distribution of our lineage variants within the lung, but we have observed strong temporal changes in the relative variant abundances over the course of the study. These could be reflective of shifts in the lung compartments sampled in the sputum or be indicative of general shifts in the complete lung, an interesting question to explore in future research.

Lineage diversification within a patient makes infections in CF an unclear example of strain mixing, as in the case of fecal microbiota transplantation (60). Thus, methods that rely on all subpopulations being represented in a reference database would provide limited insights (61–63). Multiple tools, however, have been developed to reconstruct haplotypes based on genetic variation with or without a reference (30, 60, 64–68). Tools assessing variation in a set of marker genes (60, 65, 67), while allowing subpopulation identification when diagnostic SNVs happen to be present in these markers, preclude insights into subpopulation-specific mutations in other genes that could be of potential interest due to adaptation to the particular lung environment. MetaPalette does use the entire genome (66), but it is unclear whether its "k-mer painting" approach would be able to discern and reconstruct distinct sublineages that differ only by about 1 in 10,000 nucleotides. Of the remaining tools, to our knowledge only EVORhA (64) has been used in a clinical setting (69). This tool explicitly reconstructs haplotypes from reads mapped against a reference, but it does not use the information in longitudinally related samples, instead focusing on abundance differences within each single sample. Moreover, EVORhA has been criticized for artificially inflating the number of haplotypes detected (68–70), including by a study that also used PacBio sequencing for validation (70). Very recently, a promising new method for haplotype reconstruction was published, displaying better performance on synthetic benchmarks and strain mixtures than existing tools (71). This tool (mixtureS) likewise only works on samples individually, but it does employ an expectation maximization algorithm for the final step in strain identification. It has not yet been tested on highly similar lineages in a time course setting, however.

Our approach to longitudinal CF microbiome tracking using short-read metagenomics data still has a number of limitations. Linking SNVs from the whole genome predominantly based on allele frequencies can be obscured by recombination events and the presence of mobile genetic elements. Moreover, SNV linkage requires sufficient data in terms of the number of time points and in terms of sequence read coverage depth. Although we have also performed variant calling and SNV clustering on patients CFR07 and CFR09, the smaller number of time points prevented us from performing lineage deconvolution on *P. aeruginosa* in a manner similar to that for CFR11. In patient CFR06, *Achromobacter* was covered sufficiently for variant calling at only one time point. Finally, although we had *Achromobacter* data from five time points in patient CFR11, clustering of SNVs showed no apparent sublineages.

We could not perform subpopulation analysis of *A. insuavis* in a manner similar to that with *P. aeruginosa*, but we could detect an apparent case of clinical species misidentification in both patients CFR06 and CFR11. The observed pathogen lineage likely belongs to *A. insuavis*, not *A. xylosoxidans*. The misidentification of *Achromobacter* species by conventional clinical methods is not uncommon (26, 72, 73) due to the difficulty of distinguishing species based on 16S rRNA sequence alone (26, 74, 75) and lack of representative spectra in matrix-assisted laser desorption ionization–time-of-flight (MALDI-TOF) databases commonly used by clinical microbiology laboratories (73, 76). Genotyping of several CF patient cohorts using *Achromobacter*-specific marker sequences (26, 72) has revealed *A. insuavis* was the second-most prevalent species after *A. xylosoxidans* (73, 77–80) or at least accounted for a considerable fraction of *Achromobacter* infections (72). *A. insuavis* is also one of the few *Achromobacter* species capable of chronic infection (77, 79, 80), and our observations in patient CFR11 are in line with previous findings. Overall, our results from *P. aeruginosa* and *A. insuavis* show that clear and reliable pathogen identification at various taxonomic resolutions is possible without the need for cultivation based on community-wide sequencing data alone.

The limited availability of genetic data for characterization was partly due to an excess of human DNA; up to 93% of generated reads mapped to the human genome, which is not unexpected in studies of lung sputum (17, 18). To enrich for nonhuman material, we performed depletion of methylated DNA. The depletion worked to some extent based on data from paired samples, and we obtained more than 25% nonhuman reads in some samples, which is more than a 2-fold improvement on the numbers from previous studies (17, 18). However, it did not work equally well for each sample. A recent assessment of human DNA depletion methods in human saliva samples showed the limited effectiveness of currently available kits and introduced a new depletion method that decreased the fraction of human reads to 8.53% (81). This method has yet to be applied to sputum. Another recent study proposed a microfluidics-based method to enrich microbial DNA in samples from human airways (82). The implementation of methods enriching for nonhost material in oral and sputum samples looks promising, as this would lead to a decrease in sequencing costs and provide more sequencing material to study the less abundant bacteria.

In general, due to the lack of absolute abundance data, we also cannot be certain whether the observed change in the relative abundance of a specific bacterium could be in response to other bacteria growing and/or dying. In addition, as no explicit dead cell depletion has been performed, some changes in relative abundance could be influenced by the presence of DNA from dead cells, which have been known to accumulate in CF mucus (83). Absolute quantification has already provided novel insights into the gut microbiome (84), and, more recently, the application of quantitative PCR for the absolute quantification of bacteria CF lung microbiome has challenged the existence of a CF lung microbiome in early childhood (85). Combined with WGS, these quantitative approaches present a promising venue to increase interpretability in future studies.

In conclusion, we have demonstrated how metagenomics sequencing of time series data in CF patients can complement routine clinical diagnostics. Combined with recent advances in targeted depletion of human material in samples (81, 82), sequencing costs might sink soon to a point that would allow routine use of workflows such as ours in the clinic; a recent case report estimated a similar procedure would take less than 48 h (40). Noninvasive, whole-genome sequencing of sputum can provide better taxonomic resolution for pathogens than the current methods routinely used in the clinic. Unlike 16S rRNA sequencing, classification can be made on a sublineage level. In addition, by using data from multiple time points, multiple lineage variants of the same species can be tracked within a given patient, including the assignment of variant-specific SNVs and variant-specific large-scale genomic changes. Coupled to a growing database of previously observed strains (ideally including the results of past antibiotics resistance tests as well as clinical outcomes), precise computational lineage identification should enable continuous improvements in monitoring pulmonary infections in CF and assist in making decisions on disease management.

MATERIALS AND METHODS

Sputum sample collection. A cohort of 11 CF patients was monitored over the course of 2 years. All study participants provided informed consent. The study was approved by the Cantonal Ethics Committee, St. Gallen (EKSG 13/112). For the study, participants collected spontaneously produced sputum either at home on the same morning as their doctor consultation or directly at the hospital. All participants have been trained since childhood on how to provide sputum for clinical analysis and were particularly encouraged to brush their teeth and drink water prior to sputum collection. The sputum samples were collected at the Cantonal Hospital St. Gallen, weighed, and aliquoted into sterile tubes. Sputum samples from cohort patients who exhibited extreme clinical phenotypes during the course of the study were selected to undergo shotgun metagenomics sequencing.

Clinical microbiology pathogen identification. All samples were subjected to standard clinical microbiology procedures used for CF sputum in an ISO 15089 certified laboratory. Sputum samples were preprocessed with a liquefying agent (Copan SL-solution; RUWAG, Bettlach, Switzerland) before streaking on agar plates. Columbia, chocolate, MacConkey, and CNA agars (Becton, Dickinson, Allschwil, Switzerland) were streaked to support growth of the bacterial spectrum present in the upper airways. For the specific detection of CF-associated pathogens, selective chromogenic plates (bioMérieux, Geneva, Switzerland) were incubated: PAID agar for *P. aeruginosa*, SAID agar for *S. aureus*, and BCSA for *Burkholderia* species (*Achromobacter* species usually grow well on this agar as well). All plates were visually inspected after 16 to 24 h of incubation at 36°C with or without 5% (vol/vol) CO_2 per standard protocol (86), followed by a second inspection after another day of incubation. Colonies suggestive of CF-associated pathogens or showing indicative growth on selective media were subjected to MALDI-TOF analysis on a Bruker MALDI Biotyper (Bruker Daltonics, Bremen, Germany) using the standard direct smear protocol. Per manufacturer recommendations, species identification was considered reliable at a score above 2.000. In cases where no CF-associated pathogen was seen after both inspections, the culture was reported as respiratory tract flora.

DNA extraction, treatment, and sequencing. After dilution in Sputolysin (Calbiochem Corp., San Diego, CA, USA), total DNA was extracted using the High Pure PCR template preparation kit (Roche, Basel, Switzerland) per the manufacturer's instructions. DNA concentration was measured using an ACTgene UV99 spectrophotometer at a wavelength of 260 nm, and samples were stored at −20°C. As the starting material was not limiting and sufficient amounts of DNA were available, no extra amplification step was deemed necessary, and no extraction blanks for PCR/sequencing contamination control were processed.

After DNA isolation, samples were subjected to methylated DNA depletion using the NEBNext microbiome enrichment kit (New England Biolabs Inc., Ipswich, MA, USA) to enrich for microbial DNA. As a control, we included day 0 samples from all patients without performing depletion. Depletion of methylated DNA did not have a consistent effect on the total number of reads obtained (data not shown). Relative microbial DNA content increased in three out of four patients by up to 2.3-fold but did not exceed 27% (data not shown).

Next-generation sequencing libraries were prepared using the TruSeq DNA Nano library preparation kit (Illumina, Inc., CA, USA) per the manufacturer's instructions. The libraries were sequenced using the Illumina HiSeq 4000 platform (Illumina, Inc., CA) in paired-end mode (2 × 125 bp). Reads were quality checked with FastQC (87).

Removal of the host genome reads, contig assembly, and annotation. Reads were aligned to human genome build 38 (88) using BowTie2 (version 2.3.1) (89), reporting at most one alignment per read and writing read pairs that did not align concordantly to a separate file. Reads that did not align concordantly to the human genome were used for downstream analysis and assembly. We assembled reads into contigs using metaSPAdes (version 3.10.1) (90) with the metagenomic sample data flag. The contigs were then searched against the NCBI nucleotide database (as of 24 June 2017) using BLASTn (version 2.6.0) (91). During the search, an E value cutoff of 1E−15 was used, and the five closest matching sequences were retained. For taxonomic annotation, we only considered matching sequences that had a bit score within a 10% range of the maximum scoring match. Contigs were assigned to the most recent common ancestor of the considered matches. Assembly completeness and contamination were assessed using the lineage workflow in CheckM (23). Phages and viruses were largely excluded from this analysis due to their poor representation in databases and lack of a standardized taxonomy.

Taxonomic profiling and diversity estimation. Raw reads were trimmed and filtered based on quality using sickle (version 1.33) (92). Trimmed and filtered reads were profiled using mOTUs (version 2.0.1) (profile at molecular operational taxonomic unit [mOTU], genus, and family taxonomic level; output scaled read counts) (22) and MetaPhlAn (version 2.7.1) (profile at all taxonomic levels) (93). For

MetaPhlAn input, all trimmed and filtered reads were pooled in the same file. The two methods exhibited several disagreements in species delineation, but the generated taxonomic profiles (compared on a sample-by-sample basis) highly correlated at the genus level (see Data Set S1 at https://string-db.org/suppl/Dataset_S1_Strain-resolved_Microbiome_Dynamics_in_Cystic_Fibrosis.xlsx, mOTUs MetaPhlAn comparison).

The amount of viral and fungal content was estimated with MiCoP (repository cloned August 2020) (21). The run-bwa.py script was used first to map trimmed and filtered reads to the viral and fungal databases provided by the authors. Viral and fungal contents were then profiled using the compute_abundances.py script with default detection thresholds to call organisms as present. Results were output as raw counts.

To determine the aerobe and anaerobe content, detected species were mapped to oxygen tolerance data from BacDive (as of August 2019) (94). Unclassified species from a known genus were labeled as aerobe or anaerobe only when all species of this genus were labeled as aerobes or anaerobes. Otherwise, the label "unknown" was assigned.

Diversity was calculated based on relative abundances obtained from mOTUs using Shannon's diversity index.

Strain identification with PanPhlAn. For *A. xylosoxidans*, a total of 22 genomes and their annotations were downloaded from the Integrated Microbial Genomes and Microbiomes Database (as of May 2018) (95). These genomes were used to create a pangenome using PanPhlAn (version 1.2.3.6) (28). We used the pooled trimmed and filtered reads as input to the PanPhlAn software to generate gene family presence/absence profiles for both sample and reference genomes, setting the strain similarity percentage threshold to zero to show results from all reference genomes. To call gene family presence, default thresholds were used.

For *P. aeruginosa*, a total of 2,226 genomes were downloaded from the *Pseudomonas* Genome Database (as of July 2018) (96). Because of the large number of genomes, we could not use the complete set of genomes for PanPhlAn and had to generate a set of representative genomes. Pairwise genomic distances were calculated using the Mash (version 2.0) sketch and dist commands (97). We discarded outlier genomes with an average distance of more than 0.1 and with less than 90% estimated completeness according to BUSCO (version 3.0.2), using the *Gammaproteobacteria* OrthoDB v9 database and the Augustus *E. coli* gene prediction model (98, 99). Remaining genomes were clustered at a distance threshold of 0.005. From each cluster, we selected the genome with the smallest average distance to all other cluster members, yielding 359 representative genomes. The representative genomes were annotated using Prokka (version 1.12) (100) and used to generate the pangenome using PanPhlAn (version 1.2.3.6) (28) in the same manner as that for *A. xylosoxidans*.

Phylogenetic tree generation. A total of 145 genomes from the *Achromobacter* genus were downloaded from the NCBI Genome database (as of November 2018) (24) but one was discarded due to low estImated completeness. We searched these genomes using BLASTn (version 2.6.0) (91) with an E value cutoff of 1E−15 against the mOTUs database (version 2.0.1) and against the PubMLST database of the *Achromobacter* genus (as of July 2017). For our samples, we searched the contigs assigned to the *Achromobacter* genus against the same databases. We used the coordinates output by the search to extract the corresponding gene sequences from the genomes. If the extracted gene sequence was shorter than the sequences of this gene in the databases, we padded the gene sequence with "X." If a gene was absent from the genome, we introduced a string of X's that was the length of this gene.

We then produced two types of composite sequences. Based on the mOTUs database search, sequences were created by concatenating the single-copy genes *COG0012*, *COG0016*, *COG0018*, *COG0172*, *COG0215*, *COG0495*, *COG0525*, *COG0541*, *COG0533*, and *COG0552*. Based on the PubMLST database search, sequences were created by concatenating the housekeeping genes *eno*, *gltB*, *lepA*, *nrdA*, *nuoL*, *nusA*, and *rpoB*. Composite sequences that were more than half X's were omitted from further analysis. The remaining sequences were aligned using MUSCLE (version 3.8.1551) (101). Based on the alignments, maximum likelihood trees were constructed using RAxML (version 8.2.10) (102) under the GTRCAT model (random seed 1234). *Bordetella pertussis* (NC_002929.2) was used as an outgroup to root the trees. One hundred bootstraps (random seed 1234) were performed on the trees to estimate branch confidence.

For the third tree, the downloaded *Achromobacter* genomes and/or sample contigs were annotated using Prokka (version 1.12) (100). The obtained gene sequences were used as the input for the ANIcalculator (version 1.0) (103). Genes annotated as rRNA, tRNA, or tmRNA were excluded from the calculation. Based on the calculated pairwise average nucleotide identities, a distance matrix was created and genomes were clustered using the unweighted pair group method with arithmetic mean (UPGMA) function in the R package phangorn (104). The number of monophyletic clades was calculated using the check_monophyly function in the ETE Toolkit (version 3.0) (105).

For the 16S rRNA *Achromobacter* tree, rRNA sequences were predicted using barrnap (version 0.9) for the *Bacteria* kingdom, using the default E value of 1E−06 and rejecting all sequences that were less than 80% of the length threshold (106). In genomes with multiple predicted 16S rRNA sequences, we selected the sequence that had the highest average alignment score across all predicted singleton 16S rRNA sequences. Furthermore, we discarded the sequences from *Achromobacter* sp. strain KAs 3-5 and *Achromobacter* sp. strain BFMG1, as a quick search revealed these sequences were from a different family. A total of 127 sequences (including *B. pertussis*) were used to generate the alignment and tree using the same procedure as that for the *Achromobacter* mOTUs tree. The Robinson-Foulds metric was calculated using the compare function in the ETE toolkit (version 3.0) (105). Only genomes present in all trees were considered for the comparison.

The *P. aeruginosa* tree in Fig. S5D was generated using the same procedure as that for the *Achromobacter* mOTUs tree. No outgroup was used during tree generation, but midpoint rooting was used during tree visualization. All trees were visualized using iTOL (107).

Calling single-nucleotide variations. Filtered and trimmed sample reads were mapped to the *P. aeruginosa* PAER4_119 (CP013113.1) and *A. xylosoxidans* FDAARGOS_147 (CP014060.1; data not used further) genomes using the ngless framework provided by the developers of metaSNV (version 0.8.1) (108–111). The framework filtered out reads that did not map uniquely, mapped at an identity of less than 97%, or had less than a 45-bp match with the reference. SNVs were called using metaSNV (version 1.0.3) (29), under default thresholds.

Comparison of SNV temporal profiles from the clusters in Fig. 4. To investigate the relationship between the temporal profiles of the seven SNV clusters, an SNV from each considered cluster was drawn at random. Allele frequencies from all 10 time points were added or subtracted between drawn SNVs in accordance with the claims made, and the mean absolute error relative to the expected value (0 or 1) was calculated. A total of 10,000 draws were performed. To generate a random distribution, one of the drawn SNVs had the time points shuffled prior to performing arithmetic operations. The real and random mean absolute error distributions were compared to each other using a two-sided Kolmogorov-Smirnov test.

Temporal variation distributions were generated by using the standard deviation of allele frequencies from 10 time points for each SNV within the considered t-SNE cluster. The seven clusters were then pairwise compared using a two-sided Mann-Whitney U test. A Bonferroni correction was then applied to the obtained *P* values.

Haplotype detection with DESMAN. Prior to running DESMAN (version 2.1.1) (30), we filtered out SNVs that clustered together on the *P. aeruginosa* PAER4_119 chromosome (more than 8 per 1,000 bp), because these could have biased the relative abundance calculations. The Variant_Filter.py script was used to further select SNVs, resulting in 1,287 SNVs used by DESMAN to determine the relative abundance for three haplotypes. Ten runs, each consisting of 100 iterations, were performed using different random seeds (1 to 10).

Long-read sequencing and analysis. DNA isolated from day 94 and day 346 samples of patient CFR11 was additionally subjected to long-read sequencing. The sequencing libraries were prepared using the SMRTbell Express template preparation kit 2.0 (Pacific Biosciences of California, Inc.).

Prior to sequencing, size selection was performed on the DNA. Fifteen micrograms of genomic DNA (gDNA) was mechanically sheared to an average size distribution of 10 to 20 kb using a Megaruptor 3.0 device (Diagenode) and a Femto pulse gDNA analysis assay (Agilent). Ten micrograms of sheared gDNA was DNA damage repaired and end repaired using polishing enzymes. A ligation and a nuclease treatment reaction were performed to create the SMRT bell template per the manufacturer's instructions. A Blue Pippin device (Sage Science) was used to size select the SMRT bell template and enrich the big fragments that were longer than 8 kb. A ready-to-sequence SMRT bell-polymerase complex was created using the Sequel II binding kit 2.0 and Internal Control 1.0 (Pacific Biosciences of California, Inc.) per the manufacturer's instructions.

The Pacific Biosciences Sequel II instrument was programmed to sequence the library on 1 Sequel II SMRT Cell 8M (Pacific Biosciences of California, Inc.), taking one 30-h movie per cell, using the Sequel II sequencing kit 2.0 (Pacific Biosciences of California, Inc.). After the run, read quality was assessed using the "run QC" module in the PacBio SMRT Link software.

Reads with an average quality above 20 were mapped to the *P. aeruginosa* PAER4_119 (CP013113.1) genome using minimap2 (version 2.17-r941) (112) with "PacBio vs reference mapping" preset parameters. We selected reads that mapped to the reference with a sequence identity of at least 97% and spanned at least two variant-specific SNVs. Diagnostic reads with haplotypes not matching any of the three variants were labeled "incompatible." For the expected read count estimation, haplotypes for each diagnostic read were drawn randomly in proportion to the allele frequencies output by metaSNV for the corresponding sample. The procedure was repeated 10,000 times to generate a background distribution.

Analysis of *P. aeruginosa* genome coverage. Read mapping was performed as described in "Calling single-nucleotide variations," above. Average read coverage was calculated for windows of 1,000 bp. To account for higher coverage near origins of replication (39), we fit a quadratic polynomial function to every sample's coverage profile. Each window's coverage was then normalized to the value output by the function at the chromosome position in the middle of the window. The circos diagram in Fig. 5 was generated using the R package circlize (113).

Phage region detection. Locations of phage sequences on the *P. aeruginosa* PAER4_119 (CP013113.1) genome were determined through the use of multiple tools: PHASTER (114), Phage Web (115), Phigaro (116), and PhiSpy (117, 118). For both PHASTER and Phage Web, the web interface was used to search for the GenBank accession number (as of August 2020). We kept the default Phage Web settings for phage region identification: at least 80% sequence identity during BLAST, at least six coding sequences in a prophage region, and 80% of the elements in identified prophage regions to be used for integrity analysis. Phigaro (version 2.2.5) was run locally in basic mode. PhiSpy (version 4.1.17) was run locally three times using different training sets, genericAll, *Pseudomonas*, and 208964.452 (based on the *P. aeruginosa* PAO1 genome), with otherwise default parameters. Between all tools and runs, 17 potential regions were detected, although only five regions were detected by more than one tool. We adjusted phage region boundaries based on the overlap between multiple tools and by looking at alterations in read coverage. An additional sixth phage region was selected because it was detected in all PhiSpy runs and displayed similar coverage abnormalities.

Data availability. All samples, after removal of reads that mapped to the human genome, have been deposited in the European Nucleotide Archive (PRJEB32062). Data Set S1, which includes study information gathered for each patient (clinical information, prescribed medication, lung microbiome taxonomic profiles, and assembly reports) and a list of *P. aeruginosa* genes containing lineage variant-specific SNVs, can be downloaded at https://string-db.org/suppl/Dataset_S1_Strain-resolved_Microbiome_Dynamics_in_Cystic_Fibrosis.xlsx.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

FIG S1, PDF file, 0.8 MB.

FIG S2, PDF file, 0.3 MB.

FIG S3, PDF file, 0.4 MB.

FIG S4, PDF file, 0.3 MB.

FIG S5, PDF file, 1 MB.

FIG S6, PDF file, 0.3 MB.

FIG S7, PDF file, 1 MB.

FIG S8, PDF file, 1.2 MB.

FIG S9, PDF file, 0.5 MB.

FIG S10, PDF file, 0.5 MB.

ACKNOWLEDGMENTS

We thank João F. Matias Rodrigues and Janko Tackmann for helpful discussions on procedures and the manuscript. The Functional Genomics Centre Zurich is gratefully acknowledged for DNA library preparation, sequencing, and initial data processing, and we especially thank Andrea Patrignani for assistance with long-read sequencing.

M.D., R.F., and C.V.M. were supported by the Swiss National Science Foundation (grant 31003A_160095). This work was partly funded by an internal grant from the Cantonal Hospital St. Gallen. The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

We have no competing interests to declare.

REFERENCES

1. Elborn JS. 2016. Cystic fibrosis. Lancet 388:2519–2531. https://doi.org/10.1016/S0140-6736(16)00576-6.
2. Cystic Fibrosis Foundation. 2018. Cystic Fibrosis Foundation patient registry 2017 annual data report. https://www.cff.org/Research/Researcher-Resources/Patient-Registry/2017-Patient-Registry-Annual-Data-Report.pdf. Accessed 23 January 2020.
3. Zolin A, Orenti A, Naehrlich L, van Rens J, Fox A, Krasnyk M, Jung A, Mei-Zahav M, Cosgriff R, Storms V. 2019. ECFSPR annual report 2017. https://www.ecfs.eu/sites/default/files/general-content-images/working-groups/ecfs-patient-registry/ECFSPR_Report2017_v1.3.pdf. Accessed 23 January 2020.
4. Burgel P-R, Bellis G, Olesen HV, Viviani L, Zolin A, Blasi F, Elborn JS, ERS/ECFS Task Force on Provision of Care for Adults with Cystic Fibrosis in Europe. 2015. Future trends in cystic fibrosis demography in 34 European countries. Eur Respir J 46:133–141. https://doi.org/10.1183/09031936.00196314.
5. Burgener EB, Moss RB. 2018. Cystic fibrosis transmembrane conductance regulator modulators: precision medicine in cystic fibrosis. Curr Opin Pediatr 30:372–377. https://doi.org/10.1097/MOP.0000000000000627.
6. Hisert KB, Heltshe SL, Pope C, Jorth P, Wu X, Edwards RM, Radey M, Accurso FJ, Wolter DJ, Cooke G, Adam RJ, Carter S, Grogan B, Launspach JL, Donnelly SC, Gallagher CG, Bruce JE, Stoltz DA, Welsh MJ, Hoffman LR, McKone EF, Singh PK. 2017. Restoring cystic fibrosis transmembrane conductance regulator function reduces airway bacteria and inflammation in people with cystic fibrosis and chronic lung infections. Am J Respir Crit Care Med 195:1617–1628. https://doi.org/10.1164/rccm.201609-1954OC.
7. Hansen CR, Pressler T, Høiby N. 2008. Early aggressive eradication therapy for intermittent Pseudomonas aeruginosa airway colonization in cystic fibrosis patients: 15 years experience. J Cystic Fibrosis 7:523–530. https://doi.org/10.1016/j.jcf.2008.06.009.
8. Mogayzel PJ, Naureckas ET, Robinson KA, Brady C, Guill M, Lahiri T, Lubsch L, Matsui J, Oermann CM, Ratjen F, Rosenfeld M, Simon RH, Hazle L, Sabadosa K, Marshall BC, Cystic Fibrosis Foundation Pulmonary Clinical Practice Guidelines Committee. 2014. Cystic Fibrosis Foundation pulmonary guideline. Pharmacologic approaches to prevention and eradication of initial Pseudomonas aeruginosa infection. Ann Am Thorac Soc 11:1640–1650. https://doi.org/10.1513/AnnalsATS.201404-166OC.
9. Gilligan PH. 2014. Infections in patients with cystic fibrosis: diagnostic microbiology update. Clin Lab Med 34:197–217. https://doi.org/10.1016/j.cll.2014.02.001.
10. Tunney MM, Field TR, Moriarty TF, Patrick S, Doering G, Muhlebach MS, Wolfgang MC, Boucher R, Gilpin DF, McDowell A, Elborn JS. 2008. Detection of anaerobic bacteria in high numbers in sputum from patients with cystic fibrosis. Am J Respir Crit Care Med 177:995–1001. https://doi.org/10.1164/rccm.200708-1151OC.
11. Fodor AA, Klem ER, Gilpin DF, Elborn JS, Boucher RC, Tunney MM, Wolfgang MC. 2012. The adult cystic fibrosis airway microbiota is stable over time and infection type, and highly resilient to antibiotic treatment of exacerbations. PLoS One 7:e45001. https://doi.org/10.1371/journal.pone.0045001.
12. Carmody LA, Zhao J, Schloss PD, Petrosino JF, Murray S, Young VB, Li JZ, LiPuma JJ. 2013. Changes in cystic fibrosis airway microbiota at pulmonary exacerbation. Ann Am Thorac Soc 10:179–187. https://doi.org/10.1513/AnnalsATS.201211-107OC.
13. Coburn B, Wang PW, Diaz Caballero J, Clark ST, Brahma V, Donaldson S, Zhang Y, Surendra A, Gong Y, Elizabeth Tullis D, Yau YCW, Waters VJ, Hwang DM, Guttman DS. 2015. Lung microbiota across age and disease stage in cystic fibrosis. Sci Rep 5:10241. https://doi.org/10.1038/srep10241.
14. Zhao J, Schloss PD, Kalikin LM, Carmody LA, Foster BK, Petrosino JF, Cavalcoli JD, VanDevanter DR, Murray S, Li JZ, Young VB, LiPuma JJ. 2012. Decade-long bacterial community dynamics in cystic fibrosis airways. Proc Natl Acad Sci U S A 109:5809–5814. https://doi.org/10.1073/pnas.1120577109.

15. Price KE, Hampton TH, Gifford AH, Dolben EL, Hogan DA, Morrison HG, Sogin ML, O'Toole GA. 2013. Unique microbial communities persist in individual cystic fibrosis patients throughout a clinical exacerbation. Microbiome 1:27. https://doi.org/10.1186/2049-2618-1-27.
16. Carmody LA, Zhao J, Kalikin LM, LeBar W, Simon RH, Venkataraman A, Schmidt TM, Abdo Z, Schloss PD, LiPuma JJ. 2015. The daily dynamics of cystic fibrosis airway microbiota during clinical stability and at exacerbation. Microbiome 3:12. https://doi.org/10.1186/s40168-015-0074-9.
17. Losada PM, Chouvarine P, Dorda M, Hedtfeld S, Mielke S, Schulz A, et al. 2016. The cystic fibrosis lower airways microbial metagenome. ERJ Open Res 2:00096-15. https://doi.org/10.1183/23120541.00096-2015.
18. Feigelman R, Kahlert CR, Baty F, Rassouli F, Kleiner RL, Kohler P, Brutsche MH, von Mering C. 2017. Sputum DNA sequencing in cystic fibrosis: noninvasive access to the lung microbiome and to pathogen details. Microbiome 5:20. https://doi.org/10.1186/s40168-017-0234-1.
19. Hauser PM, Bernard T, Greub G, Jaton K, Pagni M, Hafen GM. 2014. Microbiota present in cystic fibrosis lungs as revealed by whole genome sequencing. PLoS One 9:e90934. https://doi.org/10.1371/journal.pone.0090934.
20. Bacci G, Taccetti G, Dolce D, Armanini F, Segata N, Di Cesare F, et al. 2020. Untargeted metagenomic investigation of the airway microbiome of cystic fibrosis patients with moderate-severe lung disease. Microorganisms 8:1003. https://doi.org/10.3390/microorganisms8071003.
21. LaPierre N, Mangul S, Alser M, Mandric I, Wu NC, Koslicki D, Eskin E. 2019. MiCoP: microbial community profiling method for detecting viral and fungal organisms in metagenomic samples. BMC Genomics 20:423. https://doi.org/10.1186/s12864-019-5699-9.
22. Milanese A, Mende DR, Paoli L, Salazar G, Ruscheweyh H-J, Cuenca M, Hingamp P, Alves R, Costea PI, Coelho LP, Schmidt TSB, Almeida A, Mitchell AL, Finn RD, Huerta-Cepas J, Bork P, Zeller G, Sunagawa S. 2019. Microbial abundance, activity and population genomic profiling with mOTUs2. Nat Commun 10:1014. https://doi.org/10.1038/s41467-019-08844-4.
23. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 25:1043–1055. https://doi.org/10.1101/gr.186072.114.
24. NCBI Resource Coordinators. 2018. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 46:D8–D13. https://doi.org/10.1093/nar/gkx1095.
25. Jolley KA, Bray JE, Maiden MCJ. 2018. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. Wellcome Open Res 3:124. https://doi.org/10.12688/wellcomeopenres.14826.1.
26. Spilker T, Vandamme P, LiPuma JJ. 2012. A multilocus sequence typing scheme implies population structure and reveals several putative novel Achromobacter species. J Clin Microbiol 50:3010–3015. https://doi.org/10.1128/JCM.00814-12.
27. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, Hugenholtz P. 2018. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat Biotechnol 36:996–1004. https://doi.org/10.1038/nbt.4229.
28. Scholz M, Ward DV, Pasolli E, Tolio T, Zolfo M, Asnicar F, Truong DT, Tett A, Morrow AL, Segata N. 2016. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. Nat Methods 13:435–438. https://doi.org/10.1038/nmeth.3802.
29. Costea PI, Munch R, Coelho LP, Paoli L, Sunagawa S, Bork P. 2017. metaSNV: a tool for metagenomic strain level analysis. PLoS One 12:e0182392. https://doi.org/10.1371/journal.pone.0182392.
30. Quince C, Delmont TO, Raguideau S, Alneberg J, Darling AE, Collins G, Eren AM. 2017. DESMAN: a new tool for de novo extraction of strains from metagenomes. Genome Biol 18:181. https://doi.org/10.1186/s13059-017-1309-9.
31. Ayling M, Clark MD, Leggett RM. 2019. New approaches for metagenome assembly with short reads. Brief Bioinform 21:584–594. https://doi.org/10.1093/bib/bbz020.
32. Marvig RL, Sommer LM, Molin S, Johansen HK. 2015. Convergent evolution and adaptation of Pseudomonas aeruginosa within patients with cystic fibrosis. Nat Genet 47:57–64. https://doi.org/10.1038/ng.3148.
33. Smith EE, Buckley DG, Wu Z, Saenphimmachak C, Hoffman LR, D'Argenio DA, Miller SI, Ramsey BW, Speert DP, Moskowitz SM, Burns JL, Kaul R, Olson MV. 2006. Genetic adaptation by Pseudomonas aeruginosa to the airways of cystic fibrosis patients. Proc Natl Acad Sci U S A 103:8487–8492. https://doi.org/10.1073/pnas.0602138103.
34. Cramer N, Klockgether J, Wrasman K, Schmidt M, Davenport CF, Tümmler B. 2011. Microevolution of the major common Pseudomonas aeruginosa clones C and PA14 in cystic fibrosis lungs. Environ Microbiol 13:1690–1704. https://doi.org/10.1111/j.1462-2920.2011.02483.x.
35. Marvig RL, Johansen HK, Molin S, Jelsbak L. 2013. Genome analysis of a transmissible lineage of Pseudomonas aeruginosa reveals pathoadaptive mutations and distinct evolutionary paths of hypermutators. PLoS Genet 9:e1003741. https://doi.org/10.1371/journal.pgen.1003741.
36. Feliziani S, Marvig RL, Luján AM, Moyano AJ, Di Rienzo JA, Krogh Johansen H, Molin S, Smania AM. 2014. Coexistence and within-host evolution of diversified lineages of hypermutable Pseudomonas aeruginosa in long-term cystic fibrosis infections. PLoS Genet 10:e1004651. https://doi.org/10.1371/journal.pgen.1004651.
37. Ciofu O, Mandsberg LF, Bjarnsholt T, Wassermann T, Høiby N. 2010. Genetic adaptation of Pseudomonas aeruginosa during chronic lung infection of patients with cystic fibrosis: strong and weak mutators with heterogeneous genetic backgrounds emerge in mucA and/or lasR mutants. Microbiology 156:1108–1119. https://doi.org/10.1099/mic.0.033993-0.
38. Williams D, Evans B, Haldenby S, Walshaw MJ, Brockhurst MA, Winstanley C, Paterson S. 2015. Divergent, coexisting Pseudomonas aeruginosa lineages in chronic cystic fibrosis lung infections. Am J Respir Crit Care Med 191:775–785. https://doi.org/10.1164/rccm.201409-1646OC.
39. Korem T, Zeevi D, Suez J, Weinberger A, Avnit-Sagi T, Pompan-Lotan M, Matot E, Jona G, Harmelin A, Cohen N, Sirota-Madi A, Thaiss CA, Pevsner-Fischer M, Sorek R, Xavier R, Elinav E, Segal E. 2015. Growth dynamics of gut microbiota in health and disease inferred from single metagenomic samples. Science 349:1101–1106. https://doi.org/10.1126/science.aac4812.
40. Güemes AGC, Lim YW, Quinn RA, Conrad DJ, Benler S, Maughan H, Edwards R, Brettin T, Cantú VA, Cuevas D, Hamidi R, Dorrestein P, Rohwer F. 2019. Cystic fibrosis rapid response: translating multi-omics data into clinically relevant information. mBio 10:e00431-19. https://doi.org/10.1128/mBio.00431-19.
41. Foweraker JE, Laughton CR, Brown DFJ, Bilton D. 2005. Phenotypic variability of Pseudomonas aeruginosa in sputa from patients with acute infective exacerbation of cystic fibrosis and its impact on the validity of antimicrobial susceptibility testing. J Antimicrob Chemother 55:921–927. https://doi.org/10.1093/jac/dki146.
42. Ashish A, Paterson S, Mowat E, Fothergill JL, Walshaw MJ, Winstanley C. 2013. Extensive diversification is a common feature of Pseudomonas aeruginosa populations during respiratory infections in cystic fibrosis. J Cyst Fibros 12:790–793. https://doi.org/10.1016/j.jcf.2013.04.003.
43. Diaz Caballero J, Clark ST, Coburn B, Zhang Y, Wang PW, Donaldson SL, Tullis DE, Yau YCW, Waters VJ, Hwang DM, Guttman DS. 2015. Selective sweeps and parallel pathoadaptation drive Pseudomonas aeruginosa evolution in the cystic fibrosis lung. mBio 6:e00981-15. https://doi.org/10.1128/mBio.00981-15.
44. Jorth P, Staudinger BJ, Wu X, Hisert KB, Hayden H, Garudathri J, Harding CL, Radey MC, Rezayat A, Bautista G, Berrington WR, Goddard AF, Zheng C, Angermeyer A, Brittnacher MJ, Kitzman J, Shendure J, Fligner CL, Mittler J, Aitken ML, Manoil C, Bruce JE, Yahr TL, Singh PK. 2015. Regional isolation drives bacterial diversification within cystic fibrosis lungs. Cell Host Microbe 18:307–319. https://doi.org/10.1016/j.chom.2015.07.006.
45. Marvig RL, Dolce D, Sommer LM, Petersen B, Ciofu O, Campana S, Molin S, Taccetti G, Johansen HK. 2015. Within-host microevolution of Pseudomonas aeruginosa in Italian cystic fibrosis patients. BMC Microbiol 15:218. https://doi.org/10.1186/s12866-015-0563-9.
46. Winstanley C, O'Brien S, Brockhurst MA. 2016. Pseudomonas aeruginosa evolutionary adaptation and diversification in cystic fibrosis chronic lung infections. Trends Microbiol 24:327–337. https://doi.org/10.1016/j.tim.2016.01.008.
47. Sherrard LJ, Tai AS, Wee BA, Ramsay KA, Kidd TJ, Ben Zakour NL, Whiley DM, Beatson SA, Bell SC. 2017. Within-host whole genome analysis of an antibiotic resistant Pseudomonas aeruginosa strain sub-type in cystic fibrosis. PLoS One 12:e0172179. https://doi.org/10.1371/journal.pone.0172179.
48. Williams D, Fothergill JL, Evans B, Caples J, Haldenby S, Walshaw MJ, Brockhurst MA, Winstanley C, Paterson S. 2018. Transmission and lineage displacement drive rapid population genomic flux in cystic fibrosis airway infections of a Pseudomonas aeruginosa epidemic strain. Microbial Genomics 4:e000167. https://doi.org/10.1099/mgen.0.000167.
49. Lieberman TD, Flett KB, Yelin I, Martin TR, McAdam AJ, Priebe GP, Kishony R. 2014. Genetic variation of a bacterial pathogen within individuals with

cystic fibrosis provides a record of selective pressures. Nat Genet 46:82–87. https://doi.org/10.1038/ng.2848.

50. Good BH, McDonald MJ, Barrick JE, Lenski RE, Desai MM. 2017. The dynamics of molecular evolution over 60,000 generations. Nature 551:45–50. https://doi.org/10.1038/nature24287.
51. Markussen T, Marvig RL, Gómez-Lozano M, Aanæs K, Burleigh AE, Høiby N, Johansen HK, Molin S, Jelsbak L. 2014. Environmental heterogeneity drives within-host diversification and evolution of Pseudomonas aeruginosa. mBio 5:e01592-14. https://doi.org/10.1128/mBio.01592-14.
52. Martin CJ, Marshall H, Cline F. 1953. Lobar alveolar gas concentrations: effect of reduced lung volumes. J Appl Physiol 6:209–212. https://doi.org/10.1152/jappl.1953.6.4.209.
53. Brown JS, Zeman KL, Bennett WD. 2001. Regional deposition of coarse particles and ventilation distribution in healthy subjects and patients with cystic fibrosis. J Aerosol Medicine 14:443–454. https://doi.org/10.1089/08942680152744659.
54. Gurney JW, Habbe TG, Hicklin J. 1997. Distribution of disease in cystic fibrosis: correlation with pulmonary function. Chest 112:357–362. https://doi.org/10.1378/chest.112.2.357.
55. Willner D, Haynes MR, Furlan M, Schmieder R, Lim YW, Rainey PB, Rohwer F, Conrad D. 2012. Spatial distribution of microbial communities in the cystic fibrosis lung. ISME J 6:471–474. https://doi.org/10.1038/ismej.2011.104.
56. Goddard AF, Staudinger BJ, Dowd SE, Joshi-Datar A, Wolcott RD, Aitken ML, Fligner CL, Singh PK. 2012. Direct sampling of cystic fibrosis lungs indicates that DNA-based analyses of upper-airway specimens can misrepresent lung microbiota. Proc Natl Acad Sci U S A 109:13769–13774. https://doi.org/10.1073/pnas.1107435109.
57. Sommer LM, Marvig RL, Luján A, Koza A, Pressler T, Molin S, Johansen HK. 2016. Is genotyping of single isolates sufficient for population structure analysis of Pseudomonas aeruginosa in cystic fibrosis airways? BMC Genomics 17:589. https://doi.org/10.1186/s12864-016-2873-1.
58. Mainz JG, Naehrlich L, Schien M, Käding M, Schiller I, Mayr S, Schneider G, Wiedemann B, Wiehlmann L, Cramer N, Pfister W, Kahl BC, Beck JF, Tümmler B. 2009. Concordant genotype of upper and lower airways P aeruginosa and S aureus isolates in cystic fibrosis. Thorax 64:535–540. https://doi.org/10.1136/thx.2008.104711.
59. Ciofu O, Johansen HK, Aanaes K, Wassermann T, Alhede M, von Buchwald C, Høiby N. 2013. P aeruginosa in the paranasal sinuses and transplanted lungs have similar adaptive mutations as isolates from chronically infected CF lungs. J Cyst Fibros 12:729–736. https://doi.org/10.1016/j.jcf.2013.02.004.
60. Smillie CS, Sauk J, Gevers D, Friedman J, Sung J, Youngster I, Hohmann EL, Staley C, Khoruts A, Sadowsky MJ, Allegretti JR, Smith MB, Xavier RJ, Alm EJ. 2018. Strain tracking reveals the determinants of bacterial engraftment in the human gut following fecal microbiota transplantation. Cell Host Microbe 23:229–240. https://doi.org/10.1016/j.chom.2018.01.003.
61. Albanese D, Donati C. 2017. Strain profiling and epidemiology of bacterial species from metagenomic sequencing. Nat Commun 8:2260. https://doi.org/10.1038/s41467-017-02209-5.
62. Ahn T-H, Chai J, Pan C. 2015. Sigma: strain-level inference of genomes from metagenomic analysis for biosurveillance. Bioinformatics 31:170–177. https://doi.org/10.1093/bioinformatics/btu641.
63. Hong C, Manimaran S, Shen Y, Perez-Rogers JF, Byrd AL, Castro-Nallar E, Crandall KA, Johnson WE. 2014. PathoScope 2.0: a complete computational framework for strain identification in environmental or clinical sequencing samples. Microbiome 2:33. https://doi.org/10.1186/2049-2618-2-33.
64. Pulido-Tamayo S, Sánchez-Rodríguez A, Swings T, Van den Bergh B, Dubey A, Steenackers H, Michiels J, Fostier J, Marchal K. 2015. Frequency-based haplotype reconstruction from deep sequencing data of bacterial populations. Nucleic Acids Res 43:e105. https://doi.org/10.1093/nar/gkv478.
65. Zolfo M, Tett A, Jousson O, Donati C, Segata N. 2017. MetaMLST: multi-locus strain-level bacterial typing from metagenomic samples. Nucleic Acids Res 45:e7. https://doi.org/10.1093/nar/gkw837.
66. Koslicki D, Falush D. 2016. MetaPalette: a k-mer painting approach for metagenomic taxonomic profiling and quantification of novel strain variation. mSystems 1:e00020-16. https://doi.org/10.1128/mSystems.00020-16.
67. Luo C, Knight R, Siljander H, Knip M, Xavier RJ, Gevers D. 2015. ConStrains identifies microbial strains in metagenomic datasets. Nat Biotechnol 33:1045–1052. https://doi.org/10.1038/nbt.3319.
68. Li X, Saadat S, Hu H, Li X. 2019. BHap: a novel approach for bacterial haplotype reconstruction. Bioinformatics 35:4624–4631. https://doi.org/10.1093/bioinformatics/btz280.
69. Gan M, Liu Q, Yang C, Gao Q, Luo T. 2016. Deep whole-genome sequencing to detect mixed infection of Mycobacterium tuberculosis. PLoS One 11:e0159029. https://doi.org/10.1371/journal.pone.0159029.
70. Nicholls SM, Aubrey W, Edwards A, de Grave K, Huws S, Schietgat L. 2019. Recovery of gene haplotypes from a metagenome. bioRxiv https://doi.org/https://doi.org/10.1101/223404.
71. Li X, Hu H, Li X. 2020. mixtureS: a novel tool for bacterial strain genome reconstruction from reads. Bioinformatics 2020:btaa728. https://doi.org/10.1093/bioinformatics/btaa728.
72. Spilker T, Vandamme P, LiPuma JJ. 2013. Identification and distribution of Achromobacter species in cystic fibrosis. J Cyst Fibros 12:298–301. https://doi.org/10.1016/j.jcf.2012.10.002.
73. Coward A, Kenna DTD, Perry C, Martin K, Doumith M, Turton JF. 2016. Use of nrdA gene sequence clustering to estimate the prevalence of different Achromobacter species among cystic fibrosis patients in the UK. J Cyst Fibros 15:479–485. https://doi.org/10.1016/j.jcf.2015.09.005.
74. Vandamme P, Moore ERB, Cnockaert M, De Brandt E, Svensson-Stadler L, Houf K, Spilker T, Lipuma JJ. 2013. Achromobacter animicus sp. nov., Achromobacter mucicolens sp. nov., Achromobacter pulmonis sp. nov. and Achromobacter spiritinus sp. nov., from human clinical samples. Syst Appl Microbiol 36:1–10. https://doi.org/10.1016/j.syapm.2012.10.003.
75. Vandamme P, Moore ERB, Cnockaert M, Peeters C, Svensson-Stadler L, Houf K, Spilker T, LiPuma JJ. 2013. Classification of Achromobacter genogroups 2, 5, 7 and 14 as Achromobacter insuavis sp. nov., Achromobacter aegrifaciens sp. nov., Achromobacter anxifer sp. nov. and Achromobacter dolens sp. nov., respectively. Syst Appl Microbiol 36:474–482. https://doi.org/10.1016/j.syapm.2013.06.005.
76. Dupont C, Michon A-L, Jumas-Bilak E, Nørskov-Lauritsen N, Chiron R, Marchandin H. 2015. Intrapatient diversity of Achromobacter spp. involved in chronic colonization of cystic fibrosis airways. Infect Genet Evol 32:214–223. https://doi.org/10.1016/j.meegid.2015.03.012.
77. Edwards BD, Greysson-Wong J, Somayaji R, Waddell B, Whelan FJ, Storey DG, Rabin HR, Surette MG, Parkins MD. 2017. Prevalence and outcomes of achromobacter species infections in adults with cystic fibrosis: a North American cohort study. J Clin Microbiol 55:2074–2085. https://doi.org/10.1128/JCM.02556-16.
78. Barrado L, Brañas P, Orellana MÁ, Martínez MT, García G, Otero JR, Chaves F. 2013. Molecular characterization of achromobacter isolates from cystic fibrosis and non-cystic fibrosis patients in Madrid, Spain. J Clin Microbiol 51:1927–1930. https://doi.org/10.1128/JCM.00494-13.
79. Amoureux L, Bador J, Bounoua Zouak F, Chapuis A, de Curraize C, Neuwirth C. 2016. Distribution of the species of Achromobacter in a French cystic fibrosis centre and multilocus sequence typing analysis reveal the predominance of A. xylosoxidans and clonal relationships between some clinical and environmental isolates. J Cyst Fibros 15:486–494. https://doi.org/10.1016/j.jcf.2015.12.009.
80. Gade SS, Nørskov-Lauritsen N, Ridderberg W. 2017. Prevalence and species distribution of Achromobacter sp. cultured from cystic fibrosis patients attending the Aarhus centre in Denmark. J Med Microbiol 66:686–689. https://doi.org/10.1099/jmm.0.000499.
81. Marotz CA, Sanders JG, Zuniga C, Zaramela LS, Knight R, Zengler K. 2018. Improving saliva shotgun metagenomics by chemical host DNA depletion. Microbiome 6:42. https://doi.org/10.1186/s40168-018-0426-3.
82. Shi X, Shao C, Luo C, Chu Y, Wang J, Meng Q, Yu J, Gao Z, Kang Y. 2019. Microfluidics-based enrichment and whole-genome amplification enable strain-level resolution for airway metagenomics. mSystems 4:e00198-19. https://doi.org/10.1128/mSystems.00198-19.
83. Surette MG. 2014. The cystic fibrosis lung microbiome. Annals ATS 11:S61–S65. https://doi.org/10.1513/AnnalsATS.201306-159MG.
84. Vandeputte D, Kathagen G, D'hoe K, Vieira-Silva S, Valles-Colomer M, Sabino J, Wang J, Tito RY, De Commer L, Darzi Y, Vermeire S, Falony G, Raes J. 2017. Quantitative microbiome profiling links gut community variation to microbial load. Nature 551:507–511. https://doi.org/10.1038/nature24460.
85. Jorth P, Ehsan Z, Rezayat A, Caldwell E, Pope C, Brewington JJ, Goss CH, Benscoter D, Clancy JP, Singh PK. 2019. Direct lung sampling indicates that established pathogens dominate early infections in children with cystic fibrosis. Cell Rep 27:1190–1204. https://doi.org/10.1016/j.celrep.2019.03.086.

86. Traub WH, Leonhard B. 1995. Antibiotic susceptibility tests with fastidious and nonfastidious bacterial reference strains: effects of aerobic versus hypercapnic incubation. Chemotherapy 41:18–33. https://doi.org/10.1159/000239320.
87. Andrews S, Krueger F, Segonds-Pichon A, Biggins L, Krueger C, Wingett S. 2010. FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Accessed 30 January 2020.
88. Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen H-C, Kitts PA, Murphy TD, Pruitt KD, Thibaud-Nissen F, Albracht D, Fulton RS, Kremitzki M, Magrini V, Markovic C, McGrath S, Steinberg KM, Auger K, Chow W, Collins J, Harden G, Hubbard T, Pelan S, Simpson JT, Threadgold G, Torrance J, Wood JM, Clarke L, Koren S, Boitano M, Peluso P, Li H, Chin C-S, Phillippy AM, Durbin R, Wilson RK, Flicek P, Eichler EE, Church DM. 2017. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Res 27:849–864. https://doi.org/10.1101/gr.213611.116.
89. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359. https://doi.org/10.1038/nmeth.1923.
90. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. 2017. metaSPAdes: a new versatile metagenomic assembler. Genome Res 27:824–834. https://doi.org/10.1101/gr.213959.116.
91. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421. https://doi.org/10.1186/1471-2105-10-421.
92. Joshi N, Fass J. 2011. Sickle: a sliding-window, adaptive, quality-based trimming tool for FastQ files (version 1.33). https://github.com/najoshi/sickle. Accessed 20 September 2019.
93. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. 2012. Metagenomic microbial community profiling using unique clade-specific marker genes. Nat Methods 9:811–814. https://doi.org/10.1038/nmeth.2066.
94. Reimer LC, Vetcininova A, Carbasse JS, Söhngen C, Gleim D, Ebeling C, Overmann J. 2019. BacDive in 2019: bacterial phenotypic data for high-throughput biodiversity analysis. Nucleic Acids Res 47:D631–D636. https://doi.org/10.1093/nar/gky879.
95. Chen I-MA, Markowitz VM, Chu K, Palaniappan K, Szeto E, Pillay M, Ratner A, Huang J, Andersen E, Huntemann M, Varghese N, Hadjithomas M, Tennessen K, Nielsen T, Ivanova NN, Kyrpides NC. 2017. IMG/M: integrated genome and metagenome comparative data analysis system. Nucleic Acids Res 45:D507–D516. https://doi.org/10.1093/nar/gkw929.
96. Winsor GL, Griffiths EJ, Lo R, Dhillon BK, Shay JA, Brinkman FSL. 2016. Enhanced annotations and features for comparing thousands of Pseudomonas genomes in the Pseudomonas genome database. Nucleic Acids Res 44:D646–D653. https://doi.org/10.1093/nar/gkv1227.
97. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM. 2016. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol 17:132. https://doi.org/10.1186/s13059-016-0997-x.
98. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31:3210–3212. https://doi.org/10.1093/bioinformatics/btv351.
99. Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM. 2018. BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol Biol Evol 35:543–548. https://doi.org/10.1093/molbev/msx319.
100. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069. https://doi.org/10.1093/bioinformatics/btu153.
101. Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797. https://doi.org/10.1093/nar/gkh340.
102. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313. https://doi.org/10.1093/bioinformatics/btu033.
103. Varghese NJ, Mukherjee S, Ivanova N, Konstantinidis KT, Mavrommatis K, Kyrpides NC, Pati A. 2015. Microbial species delineation using whole genome sequences. Nucleic Acids Res 43:6761–6771. https://doi.org/10.1093/nar/gkv657.
104. Schliep KP. 2011. phangorn: phylogenetic analysis in R. Bioinformatics 27:592–593. https://doi.org/10.1093/bioinformatics/btq706.
105. Huerta-Cepas J, Serra F, Bork P. 2016. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. Mol Biol Evol 33:1635–1638. https://doi.org/10.1093/molbev/msw046.
106. Seemann T. 2018. barrnap 0.9: rapid ribosomal RNA prediction. https://github.com/tseemann/barrnap. Accessed 29 September 2020.
107. Letunic I, Bork P. 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res 44:W242–W245. https://doi.org/10.1093/nar/gkw290.
108. Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv 13033997 [q-bio]. http://arxiv.org/abs/1303.3997.
109. Kultima JR, Sunagawa S, Li J, Chen W, Chen H, Mende DR, Arumugam M, Pan Q, Liu B, Qin J, Wang J, Bork P. 2012. MOCAT: a metagenomics assembly and gene prediction toolkit. PLoS One 7:e47656. https://doi.org/10.1371/journal.pone.0047656.
110. Kultima JR, Coelho LP, Forslund K, Huerta-Cepas J, Li SS, Driessen M, Voigt AY, Zeller G, Sunagawa S, Bork P. 2016. MOCAT2: a metagenomic assembly, annotation and profiling framework. Bioinformatics 32:2520–2523. https://doi.org/10.1093/bioinformatics/btw183.
111. Coelho LP, Alves R, Monteiro P, Huerta-Cepas J, Freitas AT, Bork P. 2019. NG-meta-profiler: fast processing of metagenomes using NGLess, a domain-specific language. Microbiome 7:84. https://doi.org/10.1186/s40168-019-0684-8.
112. Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34:3094–3100. https://doi.org/10.1093/bioinformatics/bty191.
113. Gu Z, Gu L, Eils R, Schlesner M, Brors B. 2014. circlize implements and enhances circular visualization in R. Bioinformatics 30:2811–2812. https://doi.org/10.1093/bioinformatics/btu393.
114. Arndt D, Grant JR, Marcu A, Sajed T, Pon A, Liang Y, Wishart DS. 2016. PHASTER: a better, faster version of the PHAST phage search tool. Nucleic Acids Res 44:W16–W21. https://doi.org/10.1093/nar/gkw387.
115. de Sousa AL, Maués D, Lobato A, Franco EF, Pinheiro K, Araújo F, Pantoja Y, da Costa da Silva AL, Morais J, Ramos RTJ. 2018. PhageWeb–web interface for rapid identification and characterization of prophages in bacterial genomes. Front Genet 9:644. https://doi.org/10.3389/fgene.2018.00644.
116. Starikova EV, Tikhonova PO, Prianichnikov NA, Rands CM, Zdobnov EM, Ilina EN, Govorun VM. 2020. Phigaro: high-throughput prophage sequence annotation. Bioinformatics 36:3882–3884. https://doi.org/10.1093/bioinformatics/btaa250.
117. Akhter S, Aziz RK, Edwards RA. 2012. PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies. Nucleic Acids Res 40:e126. https://doi.org/10.1093/nar/gks406.
118. McNair K, Decewicz P, Akhter S, Aziz RK, Daniel S, Edwards RA. 2019. PhiSpy. https://github.com/linsalrob/PhiSpy/. Accessed 18 August 2020.