AMERICAN SOCIETY FOR MICROBIOLOGY | mBio®

RESEARCH ARTICLE
Host-Microbe Biology

Gene Loss and Acquisition in Lineages of *Pseudomonas aeruginosa* Evolving in Cystic Fibrosis Patient Airways

Migle Gabrielaite,[a] Helle K. Johansen,[b,c] Søren Molin,[d] Finn C. Nielsen,[a] Rasmus L. Marvig[a]

[a]Center for Genomic Medicine, Rigshospitalet, Copenhagen, Denmark
[b]Department of Clinical Microbiology, Rigshospitalet, Copenhagen, Denmark
[c]Department of Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark
[d]The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Lyngby, Denmark

ABSTRACT Genome analyses have documented that there are differences in gene repertoire between evolutionary distant lineages of the same bacterial species; however, less is known about microevolutionary dynamics of gene loss and acquisition within bacterial lineages as they evolve over years. Here, we analyzed the genomes of 45 *Pseudomonas aeruginosa* lineages evolving in the lungs of cystic fibrosis (CF) patients to identify genes that are lost or acquired during the first years of infection. On average, lineage genome content changed by 88 genes (range, 0 to 473). Genes were more often lost than acquired, and prophage genes were more variable than bacterial genes. We identified convergent loss or acquisition of the same genes across lineages, suggesting selection for loss and acquisition of certain genes in the host environment. We found that a notable proportion of such genes are associated with virulence; a trait previously shown to be important for adaptation. Furthermore, we also compared the genomes across lineages to show that the within-lineage variable genes (i.e., genes that had been lost or acquired during the infection) often belonged to genomic content not shared across all lineages. In sum, our analysis adds to the knowledge on the pace and drivers of gene loss and acquisition in bacteria evolving over years in a human host environment and provides a basis to further understand how gene loss and acquisition play roles in lineage differentiation and host adaptation.

IMPORTANCE Bacterial airway infections, predominantly caused by *P. aeruginosa*, are a major cause of mortality and morbidity of CF patients. While short insertions and deletions as well as point mutations occurring during infection are well studied, there is a lack of understanding of how gene loss and acquisition play roles in bacterial adaptation to the human airways. Here, we investigated *P. aeruginosa* within-host evolution with regard to gene loss and acquisition. We show that during long-term infection *P. aeruginosa* genomes tend to lose genes, in particular, genes related to virulence. This adaptive strategy allows reduction of the genome size and evasion of the host's immune response. This knowledge is crucial to understand the basic mutational steps that, on the timescale of years, diversify lineages and adds to the identification of bacterial genetic determinants that have implications for CF disease.

KEYWORDS *Pseudomonas aeruginosa*, computational biology, evolution, genomics, host-pathogen interactions

Gene acquisition and gene loss are prominent in bacterial evolution and are also crucial during adaptation to new environments (1, 2). In contrast to point mutations, small insertions and deletions (microindels), inversions, and translocations that gradually alter existing genomic content, the acquisition or loss of entire genes rapidly confer large changes to the genomic content which alter bacterial phenotypes such as

Citation Gabrielaite M, Johansen HK, Molin S, Nielsen FC, Marvig RL. 2020. Gene loss and acquisition in lineages of *Pseudomonas aeruginosa* evolving in cystic fibrosis patient airways. mBio 11:e02359-20. https://doi.org/10.1128/mBio.02359-20.

Editor Joanna B. Goldberg, Emory University School of Medicine

Address correspondence to Migle Gabrielaite, migle.gabrielaite@regionh.dk, or Rasmus L. Marvig, rasmus.lykke.marvig@regionh.dk.

Received 20 August 2020
Accepted 22 September 2020
Published 27 October 2020

virulence, antibiotic resistance, and metabolic capability (3, 4). Thus, genome-wide analysis of the gene presence or absence is necessary to better understand bacterial evolution and adaptation (5).

While genome comparison of evolutionarily distant lineages of the same bacterial species gives insight into gene flux over the macroevolutionary scale, there is less knowledge of the pace at which and mechanisms by which genes are lost and acquired at the scale of microevolution, i.e., from studies of evolution of individual bacterial lineages (6, 7). Additionally, we have only a limited understanding of how lineage gene loss and acquisition are driven by selective versus genetic drift (1, 2).

Evolutionary studies on individual bacterial lineages are dependent on the ability to obtain multiple samples of the same lineage, which can be difficult in natural, *in vivo* environments that constantly change (8, 9), so studies are more easily performed *in vitro* (10–14). However, *Pseudomonas aeruginosa* infections in cystic fibrosis (CF) patients represent an infectious disease scenario in which the genomic evolution of individual bacterial lineages can be followed over the years and thus give an opportunity to research bacterial evolution and adaptation *in vivo* in the human host (15, 16). There is already a large pool of knowledge on the role of point mutations and microindels in evolution and adaptation of *P. aeruginosa* in CF patients, whereas gene loss and acquisition have been less extensively investigated (17–19). A better understanding of the genetic changes responsible for *P. aeruginosa* pathogenicity in CF patients is crucial to improve CF treatment strategies (20–22).

To better understand the role of gene loss and acquisition in within-host evolution and adaptation, we used genomic data from 474 longitudinally collected isolates of *P. aeruginosa* from children and young CF patients to investigate gene loss and acquisition in lineages of *P. aeruginosa* as they evolve from the initial invasion of CF airways and onward as they adapt to the human host. In total, 34 patients and 45 different clonal lineages were analyzed, and we aimed to identify gene loss or acquisition events in each of the different lineages to detect patterns across lineages ultimately leading to a better understanding of the genetic basis of bacterial adaptation in the human host.

RESULTS

***De novo* genome assembly and gene annotation.** We previously generated short-read sequencing data for the genomes of 474 isolates of *P. aeruginosa* sampled from the airways of 34 young CF patients to follow the genomic evolution of bacterial lineages within the host airways over the initial 0 to 9 years of infection (18). While the previous analysis aligned sequence reads to a *P. aeruginosa* reference genome to identify single nucleotide polymorphisms (SNPs) and small insertions and deletions (indels), we here used the same sequencing reads for *de novo* assembly of genomes to identify genes that are either lost or acquired during the course of infection.

We successfully *de novo* assembled the genomes of 446 isolates into 500 scaffolds or fewer (median, 172 scaffolds). The sizes of the assembled genomes ranged from 6,032,338 to 7,593,423 nucleotides (nt), and they contained 5,462 to 7,111 genes. The 446 assembled genomes represented 51 clone types as defined previously by Marvig et al. (2015) (18) (see Fig. S1 in the supplemental material). We grouped the isolates into 45 lineages; i.e., isolates of the same clone type and from the same patient were grouped together to allow identification of within-host accumulated gene differences (Fig. 1). In total, the 45 lineages encompassed 423 isolates distributed among 34 patients as 9 patients were infected with two ($n = 7$) or more ($n = 2$) clone types where multiple isolates were available (Fig. S1). The remaining 23 isolates with successful genome assembly were excluded from the analysis as there were no other clonal genomes available for the respective patients ($n = 22$) or the patient was infected multiple times with the same clone type and no other clonal genomes were available for that lineage ($n = 1$); i.e., at least two genomes were required for intralineage genome comparison.

Pan-genomes and identification of gene presence-absence. We analyzed 423 genomes in a two-step process to identify genes that showed variation within or

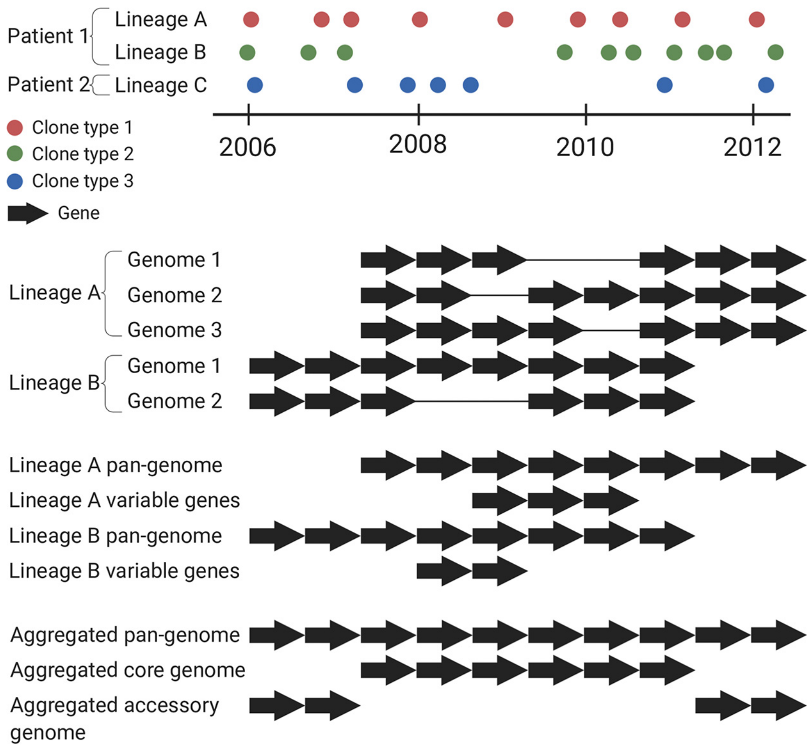


FIG 1 Schematic visualization of how bacterial lineages, lineage pan-genomes, within-lineage variable genes, and aggregated pan-genomes, core genomes, and accessory genomes were defined in this study.

between lineages. First, we compared the genomes of isolates of the same lineage to determine the full set of nonredundant genes found within the lineage, i.e., the lineage pan-genome. The lineage pan-genome consisted of (i) genes present in all isolates of the respective lineage (lineage core genome) and (ii) genes present in only some of the lineage isolates (lineage variable genes), i.e., genes that had been lost or acquired during the infection, referred to here as variable genes (Fig. 1). The lineage pan-genomes consisted of 5,607 to 7,008 genes longer than 150 bp, of which 0 to 473 were variable genes (median, 44 variable genes). A weak positive correlation (Pearson's correlation coefficient 0.15, P value $= 2.5 \times 10^{-3}$) was identified between the assembly quality (number of scaffolds) and the number of absent genes (Fig. S2A) which did not explain the observed variability in gene content. Furthermore, by aligning the raw sequencing reads to the pan-genomes of the corresponding lineages, we determined that only 52 of 13,246 genes (0.4%) were incorrectly identified as absent by GenAPI because of a lack of assemblies. These genes were treated as present in all further analyses.

Second, we compared the lineage pan-genomes to determine the full set of 14,462 nonredundant genes found across all lineages, i.e., the aggregated pan-genome (Fig. 1; see also Fig. 2). The aggregated pan-genome consisted of 4,887 genes shared across all lineage pan-genomes (aggregated core genome) and an aggregated accessory genome of 9,575 genes (genes present in only one or some lineage pan-genomes) (Fig. 1; see also Fig. 2). About half (4,932) of the aggregated accessory genes were unique for single lineages, and, overall, the lineage pan-genomes contained 0 to 540 (median, 78) of such lineage-specific genes (see Table S1 in the supplemental material). Furthermore, we found that all 335 genes reported to be essential genes in PAO1 and UCBPP-PA14 (23) were in the aggregated core genome; 29 of these genes were not present in one or more *P. aeruginosa* isolate genomes (Table S2).

Aggregated accessory genes were 15-fold more often variable within lineages than genes in the aggregated core genome (Table S1). While several factors might drive the

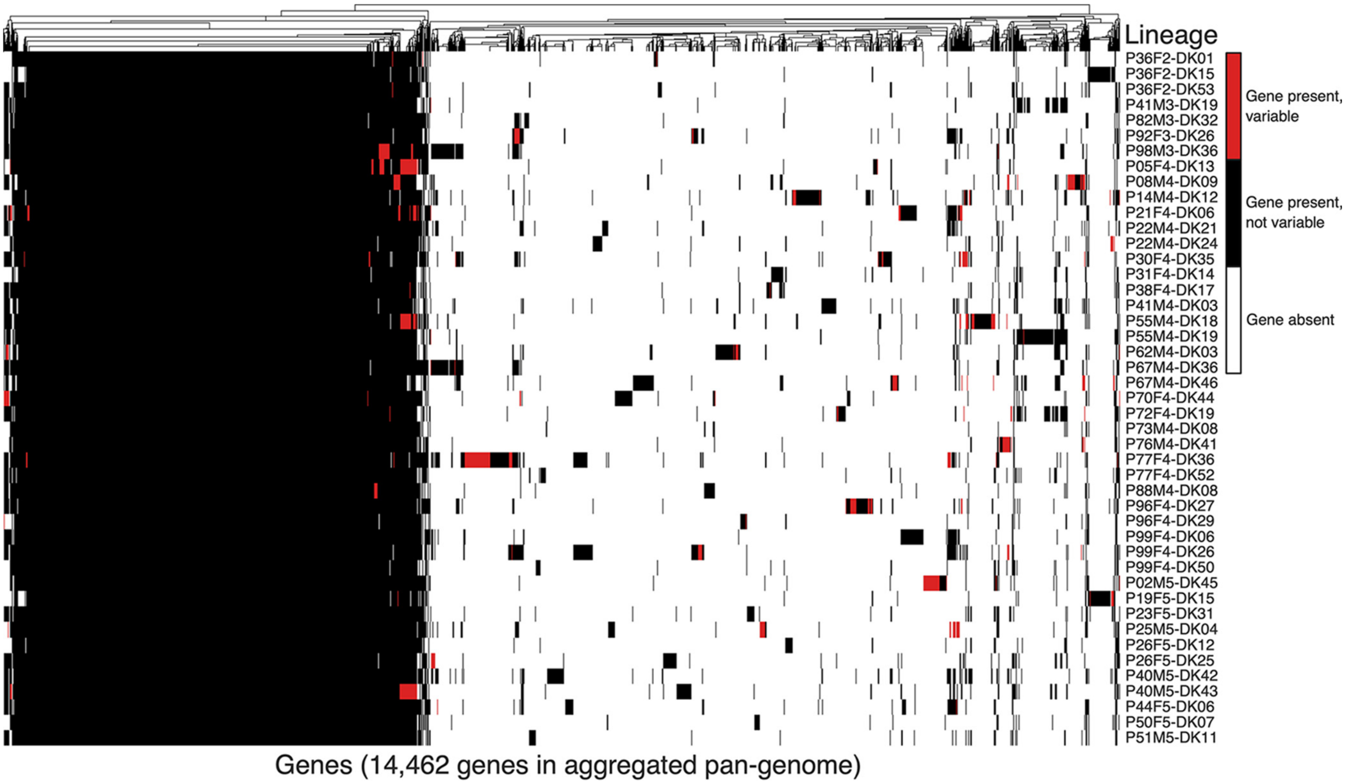


FIG 2 Presence or absence of 14,462 aggregated pan-genome genes in 45 lineages evolving in cystic fibrosis patients. Blue denotes that gene is present in all isolates of the lineage. Red denotes that the gene shows variable presence within the lineage. White denotes that the gene is not present in any of the isolates in the lineage.

higher turnover of aggregated accessory genes, one explanatory factor could be that the aggregated accessory genome has a larger amount of mobile genetic elements, such as prophage origin sequences. Therefore, we used the ACLAME database to identify and annotate phage and prophage sequences (longer than 150 bp) in the core and accessory genomes of the aggregated pan-genome, respectively. The accessory genome contained 116-fold more prophage genes, and these genes were highly variable over the course of infection; 58% of the prophage sequences in the accessory genome of the aggregated pan-genome were variable within lineages.

Changes in gene content in lineages over the course of infection. Next, we asked if the variable genes were either lost from or acquired in bacterial lineages. For this, we defined a gene as lost when it was present in the first isolate but absent in one or more of the later isolates and defined a gene as acquired when it was absent in the first isolate but present in one or more of the later isolates. Note that this definition of gene loss/acquisition might not be accurate as the first isolates might not represent the most recent common ancestor for the lineage. We found that the variable genes were more often lost. Of 3,955 variable genes, 3,411 were present in the first isolates and absent in the later ones, and the opposite was true for only 544 genes. Accordingly, we concluded that gene loss occurs at least 6 times more often than gene acquisition (Table S3).

Prophage sequences and plasmids are known to be mobile elements in bacterial genomes. Prophage genes were found in all 45 lineages by using the ACLAME database. Prophage genes were among the variable genes in 22 of the lineages, and the prophage genes were lost in 70% of cases (Table S3); i.e., they were present in the early isolates and absent in the later ones. In contrast, plasmid genes were not identified to be lost or acquired in any lineage (the PlasmidFinder database was used to define plasmid genes). In total, three lineages (P41M3-DK19, P92F3-DK26, and P72F4-DK19) carried a plasmid belonging to the replicon IncQ2_1.

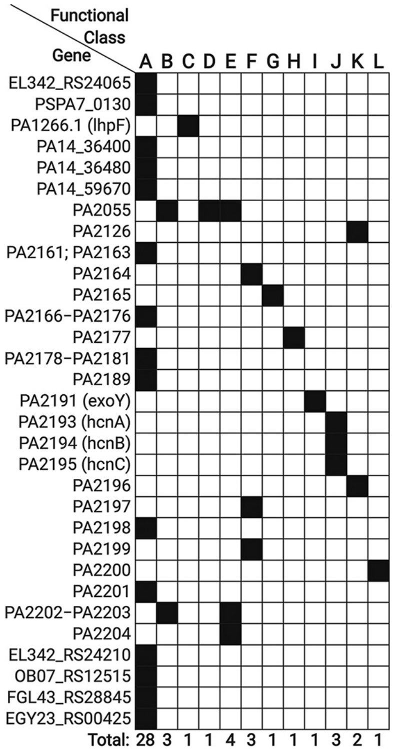
A total of 257 genes in the aggregated pan-genome were related to virulence as defined in the VFDB database. Of these, 17 genes were variable in at least one lineage, and in 8 of 17 cases, these genes were variable in more than one lineage. A two-sided Fisher's exact test showed that virulence genes were in general less often lost/acquired than other genes (P value $= 4.45 \times 10^{-9}$). Furthermore, by using the Resfinder database, seven genes were defined to be related to antibiotic resistance in the aggregated pan-genome. None of these genes were lost/acquired in any lineages, while each isolate had 5 to 7 antibiotic resistance genes (Table S1). Of 52 pathoadaptive genes reported by Marvig et al. (2015) (18), 9 were lost/acquired in lineages. No significant difference in loss/acquisition between pathoadaptive and nonpathoadaptive genes was found by performing a two-sided Fisher's exact test (P value $= 0.861$). Finally, we found that genes were 25-fold more often lost or acquired in a group than individually (see examples in Fig. S2B and Table S3); i.e., the loss/acquisition of 3,806 of 3,981 variable genes correlated with the loss/acquisition of other genes, while only 175 genes were lost/acquired alone.

Convergent evolution: same genes are variable across lineages. While variable genes made a small fraction of the aggregated pan-genome, and the majority of variable genes were lost/acquired in only one lineage, some genes were observed to be variable in multiple lineages.

We defined genes as highly variable if they were identified as variable in ≥4 lineages (among the top 2% of all variable genes; Table S4). To ensure that the high level of variation was not due to technical artifacts of analysis, all highly variable genes were manually checked as follows: (i) *P. aeruginosa* origin genes were mapped to a PAO1 reference genome, and the coverage of the gene alignments was manually assessed; (ii) for other genes, the aggregated pan-genome was subjected to BLAST analysis by using the BLAST+ suite (24) against isolate genomes and then the alignments were manually assessed. Of the 54 genes initially identified as variable in ≥4 lineages, 2 were removed after the manual check (a detailed explanation is available in Materials and Methods). Of the 52 manually confirmed variable genes, 47 genes were variable in 4 lineages, 4 genes were variable in 5 lineages, and 1 gene was variable in 10 lineages (Fig. 3).

We annotated the highly variable genes according to PseudoCAP functional classes (25) (if present in PAO1/UCBBP-PA14 reference genomes) or by a BLAST search against the National Center for Biotechnology (NCBI) nucleotide collection (nr/nt) database. Most of the highly variable genes (34 of 52) were genes with hypothetical function or genes of non-*Pseudomonas* origin. The second-largest group of highly variable genes encoded membrane proteins (4 genes). Since genes encoding membrane proteins make up around 10% of the *P. aeruginosa* genome, we tested using a Fisher exact test if genes encoding membrane proteins are more variable than expected in accounting for their abundance, and we concluded that such was not the case (P value $= 1.00$). Other highly variable genes encoded proteins involved in amino acid and nucleotide biosynthesis, antibiotic resistance and susceptibility, transport, secreted factors, transcriptional regulation, and metabolism as defined in the PseudoCAP database (Fig. 3).

Convergent evolution of locus with *hcnABC* and *exoY* genes. We found that a group of 34 genes was lost/acquired in four lineages (P21F4-DK06, P05F4-DK13, P55M4-DK18, and P40M5-DK43). The 34 genes were orthologs of genes PA2161 to PA2181 and genes PA2189 to PA2204 in the PAO1 reference genome. We noted that three of the lineages (P05F4-DK13, P55M4-DK18, and P40M5-DK43) did not have genes PA2182 to PA2188 (genes flanked by PA2161 to PA2181 and genes PA2189 to PA2204 in the PAO1 reference genome) in their lineage pan-genomes and that genes PA2182 to PA2188 were variable and congregated with genes PA2161 to PA2181 and genes PA2189 to PA2204 in the fourth lineage (P21F4-DK06), so we concluded that the 34 genes were likely lost/acquired together rather than in separate two events (Fig. 4 shows the genetic region of the group of 34 variable genes). Further, we aligned reads from each of the lineages to the PAO1 reference genome to show that parallel



Functional Class / Gene	A	B	C	D	E	F	G	H	I	J	K	L
EL342_RS24065	■											
PSPA7_0130	■											
PA1266.1 (lhpF)			■									
PA14_36400	■											
PA14_36480	■											
PA14_59670	■											
PA2055		■		■	■							
PA2126											■	
PA2161; PA2163	■											
PA2164						■						
PA2165							■					
PA2166−PA2176	■											
PA2177								■				
PA2178−PA2181	■											
PA2189	■											
PA2191 (exoY)									■			
PA2193 (hcnA)										■		
PA2194 (hcnB)										■		
PA2195 (hcnC)										■		
PA2196											■	
PA2197						■						
PA2198	■											
PA2199						■						
PA2200												■
PA2201	■											
PA2202−PA2203		■			■							
PA2204					■							
EL342_RS24210	■											
OB07_RS12515	■											
FGL43_RS28845	■											
EGY23_RS00425	■											
Total:	28	3	1	1	4	3	1	1	1	3	2	1

FIG 3 List of the most variable *Pseudomonas* origin genes and their function according to the PseudoCAP annotation. The genes code for proteins in the following categories: A—hypothetical, unclassified, unknown; B—membrane proteins; C—amino acid biosynthesis and metabolism related; D—antibiotic resistance and susceptibility; E—transport of small molecules; F—putative enzymes; G—energy metabolism; H—two-component regulatory systems; I—secreted factors (toxins, enzymes, alginate); J—central intermediary metabolism; K—transcriptional regulators; L—nucleotide biosynthesis and metabolism.

deletion/insertion of the 34 genes was the result of larger yet different deletions/insertions in the individual lineages (i.e., the 34 genes represented the shared overlapping of four different deletions/insertions in the same genomic region; see Fig. S2C and Fig. 4).

In 3 of 4 lineages, genes were present in early isolates and absent in late isolates. For lineage P40M5-DK43, the 34 genes were present in the later of only two isolates that were sampled less than a year apart, so it is likely that two isolates represent different sublineages where one sublineage did not lose the genes while another one did. Also, genes PA2161 to PA2204 were present in all 45 lineage pan-genomes, suggesting that genes PA2161 to PA2204 were present in the ancestor of lineage P40M5-DK43 and thus were lost during the course of infection.

A total of 17 of the 34 genes were annotated as "Hypothetical, unknown or unclassified" by PseudoCAP; other genes were annotated as coding for "Putative enzymes" (4 genes), "Transport of small molecules," "Membrane proteins," "Central

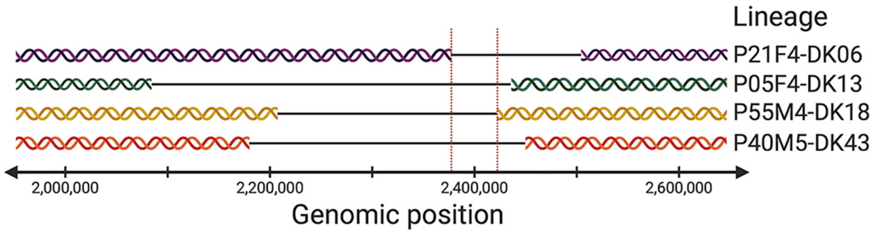


FIG 4 Genetic regions of four lineages where the same group of 34 genes is lost. Red vertical lines show the overlapping genetic regions lost in all four lineages.

intermediary metabolism" (3 genes), or "Energy metabolism," "Two-component regulatory systems," "Secreted factors," "Transcriptional regulators," and "Nucleotide biosynthesis and metabolism" (1 gene) (Fig. 3). Some of these proteins had more than one function assigned by PseudoCAP. A literature search indicated that four of the genes (*hcnABC* and *exoY* encoding hydrogen cyanide synthase and type III secreted protein, respectively) are known to play a role in the virulence and pathogenesis of *P. aeruginosa* (25, 26).

Convergent evolution in prophage-related genes and genomic islands. A total of 6 of the 52 highly variable genes were identified as prophage origin genes originating from different *P. aeruginosa* prophages, similarly to the genes from phi1 and phi2 *Pseudomonas* phages. The variable groups of 56 to 82 genes which included the 6 most variable prophage genes might represent yet-undescribed genomic islands (GIs) as they are adjacent to tRNA encoding genes, contain both prophage origin and *P. aeruginosa* origin sequences, and are longer than 10,000 bp.

As the mobility of genomic islands could explain the high variability of these gene regions, we predicted GIs with IslandViewer4. All six prophage origin genes were predicted to be part of GIs, e.g., exemplified by a 78-gene deletion in a genomic island encoding virulence factors in lineage P67M4-DK46 (Fig. 5). While IslandViewer4 predicted on average 40 GIs (range, 15 to 59) per lineage, we note that, as the analyzed genomes were not complete (i.e., in scaffolds), the GI prediction should be interpreted carefully; e.g., GIs were often predicted at the ends of scaffolds.

On average, 90% (range, 87% to 92%) of the genes in the predicted GIs code for hypothetical proteins. Therefore, it is difficult to define the function of most of the genes present in GIs. However, possible drivers of the loss of predicted GIs were identified; we identified homologs of genes coding for Clp protease (7 lineages) or of the *prtR* gene (5 lineages) as parts of predicted GIs which are known to be related to bacterial virulence and pathogenicity (27, 28). In addition, other probable virulence factors were identified in multiple lineages (Table S5).

Overall population structure: SNP and gene distances. We wanted to understand our results determined for lineage genomes in the context of the overall population structure of *P. aeruginosa*. Accordingly, we determined the genetic relationships of the 446 isolates based on either SNPs in the core genome (Pactyper [see Text S1 in the supplemental material]) or gene presence-absence (GenAPI). Both the SNP-based and the gene-based phylogenies clustered the isolates according to clone type and patient (Fig. 6). Also, both phylogenies showed that the lineages clustered into one of two groups overall with either reference strain PAO1 or UCBPP-PA14, respectively. Furthermore, we confirmed that we obtained the same population structure (i.e., the same clustering according to clone type, patient, and reference strain) when we reconstructed the phylogeny with the native *de novo* assemblies as input and also when we masked recombined regions (4,570 of 224,614 SNPs [2%] were within the masked regions; Fig. S3).

The core genome pairwise SNP distance between lineages was on average 31,909 (22 to 67,325) SNPs, while the gene content difference was on average 1,142 (13 to 2,250) genes (one random isolate was chosen to represent each lineage to avoid overrepresentation of some lineages and underrepresentation of others) (Fig. 7). Moreover, the average diversities between lineages corresponded to 19,853 (22 to 24,957) SNPs and 1,043 (28 to 2,250) gene differences in the PAO1 group, while those within the UCBPP-PA14 group consisted of 35,191 (44 to 67,325) SNPs and 1,217 (13 to 1,775) gene differences. Performing Wilcoxon's rank sum test on the distributions in the two groups, a significant difference between the groups was identified with a *P* value of $<2.2{\cdot}10^{-16}$ for pairwise SNP distance and a *P* value of 2.57×10^{-13} for pairwise gene difference distance, showing that the variability of SNPs and genes is lower within the PAO1 group than within the UCBPP-PA14 group.

Finally, we found that the ratios of core genome SNPs per difference in gene content were on average 3.5 (0.01 to 115.00; median, 0.66) and 30.2 (7.22 to 88.00; median,

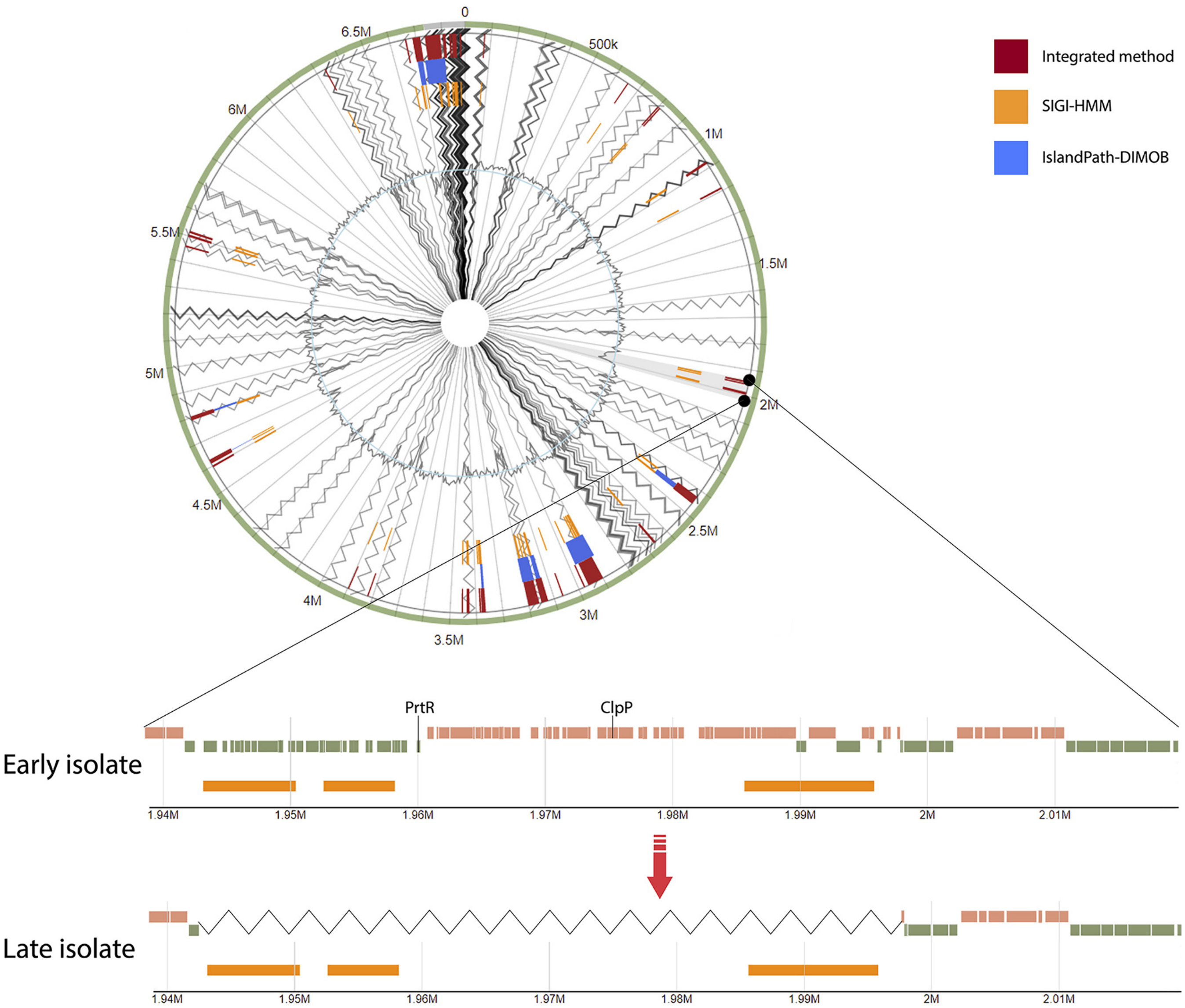


FIG 5 Genomic island predictions for early and late isolates of lineage P67M4-DK46. The zoomed region shows gene loss (zigzag line) in the late isolate. Possible virulence factors in the zoomed region are marked in the early isolate. Orange blocks in early and late isolates indicate predicted genomic islands.

28.95) within and between the clone types, respectively. Using a Wilcoxon's rank sum test, we concluded that the difference between two groups is statistically significant, with a *P* value of $<2.2 \times 10^{-16}$.

Reanalysis of other study data to compare sizes of pan-genomes and core genomes. We analyzed publicly available genome sequencing data for a collection of 1,139 isolates that were previously included in a pan-genome and core genome analysis by Freschi et al. (2019) (6), to compare the pan-genome and core genome sizes between different collections of isolates. Using the same method as that used for our own isolate collection (GenAPI [29]), we found that the 1,139 isolates previously analyzed by Freschi et al. (2019) (6) shared a core genome of 2,360 genes within a pan-genome of 38,017 genes. Using the method of Freschi et al. (2019) (SaturnV), the core genome and pan-genome were shown to consist of 619 and 43,703 genes, respectively. Defining the core genome as consisting of all genes present in at least 99% of the samples, the core genome consisted of 4,870 and 3,879 genes for GenAPI and SaturnV, respectively.

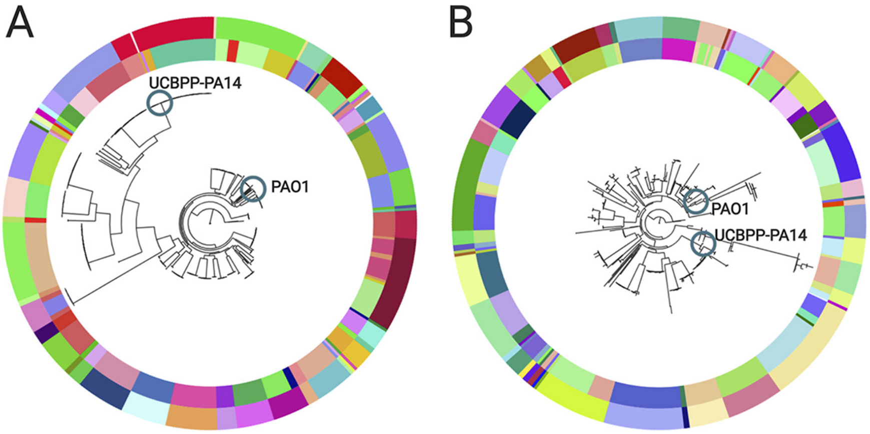


FIG 6 Phylogenetic trees of 446 *P. aeruginosa* samples (A) based on core genome SNPs and (B) based on gene presence-absence. The color of the outer circle of the trees denotes clone type, and the inner circle denotes the patient. Blue circles denote the position of reference genomes. The phylogenetic trees can be accessed on the Microreact webserver at https://microreact.org/project/KYbEXuFS0 (phylogenetic tree based on core genome SNP distances) and https://microreact.org/project/BkZdRqP-E (phylogenetic tree based on gene differences).

DISCUSSION

By analysis of genome sequences from 45 longitudinally sampled *P. aeruginosa* lineages from CF patients, we determined the microevolutionary dynamics of gene loss and acquisition in lineages of bacteria evolving in a human host environment. While similar analyses of within-host bacterial evolution investigated within-host gene loss and acquisition, our collection enables comparative analysis across multiple genotypically different strains (45 lineages distributed on 34 clone types) of the same species. Here, we not only identified events of gene loss or acquisition in the individual lineages

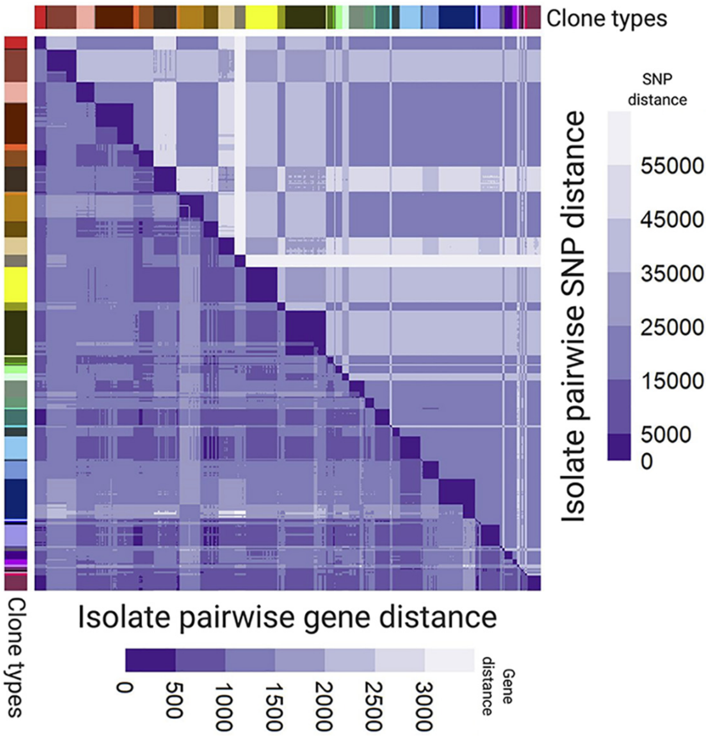


FIG 7 Pairwise SNP distances (top triangle) and gene distances (bottom triangle) between *P. aeruginosa* isolates with clone type annotation on the left side and on top.

but also analyzed this in the context of the gene variation across lineages, i.e., in the context of the species pan-genome and core genome.

Pan-genomes and core genomes. The aggregated pan-genome across lineages had 14,743 genes in total, and 4,887 genes were present in all 45 individual-lineage pan-genomes (i.e., those defined here as the aggregated core genome across lineages). Our findings are similar to reported from a study by Hilker et al. (2015) (30) that found the genomes of 21 *P. aeruginosa* strains to share a core of 4,748 genes of a pan-genome comprising 13,527 genes. In contrast, our calculated aggregated pan-genome and core genome sizes are significantly different from recent findings by Freschi et al. (2019) (6), where the pan-genomes and core genomes consisted of 54,272 and 665 genes, respectively. The difference may be explained in part by the fact that, while our isolates were collected only from CF patients in Denmark, the isolate collection used by Freschi et al. was more diverse. Also, by reanalysis of the data set from Freschi et al. with both our (GenAPI) and the original (SaturnV) methods, we show that the difference in the pan-genome and core genome sizes can in part be ascribed to differences in bioinformatics analyses as the pan-genome and core genome sizes reported by Freschi et al. converged toward findings from this and other studies (30–33) when we reanalyzed the data with GenAPI. We previously compared tools used for gene presence-absence identification (GenAPI, SaturnV, Roary, panX, EDGAR, Pandelos and BPGA) (29), and we suggest that the relatively small size of the core genome reported by Freschi et al. (665 genes) may in part be a consequence of false-negative calls of gene absence due to incomplete genome assemblies (34). Future studies based on long-read sequencing may overcome the issue of incomplete assemblies.

Within-host gene loss and acquisition. We found that genes were six times more often lost than acquired in lineages during within-host evolution. It remains unknown if the lack of gene acquisition is a consequence of limitations due to the availability of donor DNA, mechanisms of DNA uptake, or selection (either selection against the acquisition of genes or lack of selection for the acquisition of genes). Nonetheless, we found our results to be in line with previous hypotheses proposing that genomes are selectively reduced during the course of infection (34–36). Note that we defined a gene as lost if it was present in the first isolate but absent in one or more of the later isolates and vice versa. This definition of gene loss/acquisition might not be accurate as the first isolates may not represent the most recent common ancestor for the lineage, and as such, our analysis may be confounded by stochastic sampling of multiple coexisting sublineages. More sampling is required to resolve population heterogeneity (37–40).

Most of the genes that were lost or acquired within the host were part of the genome that was not shared across lineages (i.e., the aggregated accessory genome). The relative low turnover in the aggregated core genome of 4,887 genes shared by all lineages suggests that, while these genes are not essential *per se*, they may be generally important for survival under the conditions met by *P. aeruginosa* in the human host environment. Accordingly, they are maintained in the genomes. In contrast, 29 of the essential genes defined by Liberati et al. (2006) (23) were absent in some clinical isolates. This discrepancy could be explained by different conditions with respect to the human airways and laboratory media which were used in the study by Liberati et al. The lack of overlap of essential genes in different experiments reported previously by Poulsen et al. (2019) (41) corroborates the belief that *in vitro* experiments do not fully reflect the processes observed *in vivo*. Furthermore, some bacterial clones could compensate for a lost essential iron acquisition gene or antimicrobial resistance gene by cheating, i.e., by exploiting the molecules produced by other cells (42, 43).

In contrast, the prophage genes were the genes that were most often lost or acquired within the host, and prophages were putatively the major source of new genetic material. A total of 268 of 462 (58%) of the prophage genes in the aggregated pan-genome were variable within hosts, and despite taking up only 3% of the aggregated pan-genome, prophage genes constituted 9.4% of all within-host variable genes. This confirms the idea that prophage-facilitated gene flux is abundant and supports the conclusions from other studies indicating that prophages play an important role in *P. aeruginosa* CF infections (44, 45). The lack of plasmids and, therefore, their variability could be associated with the high fitness cost of carrying a plasmid in *P. aeruginosa* (46) and with the overall low number of identified plasmids in *P. aeruginosa* genomes (of 5,370 *P. aeruginosa* genomes in the NCBI database, only 70 have plasmids) (47). Finally, the absence of *Pseudomonas* plasmid annotations in the PlasmidFinder database could have led to a low number of identified plasmids among our isolates.

Convergent evolution and adaptive loss of virulence. The sampling from multiple lineages (and across multiple patients) allowed us to detect genes that were lost or acquired independently in parallel evolving lineages. While most genes were variable in only a single lineage, we found 52 genes to be variable within ≥4 lineages, which constitutes the top 2% most variable genes. The observed parallel loss or acquisition of the same genes across lineages may be driven by selection for loss and acquisition of certain genes in the host environment. It was previously hypothesized that virulence factors are selected against in CF infections, and in agreement with this, we found that 34 of the 52 highly variable genes were lost as part of a genomic region encoding the virulence factors hydrogen cyanide synthase (*hcnABC*) and type III secreted protein ExoY (*exoY*). It was also shown previously by Wee et al. (2018) (26) that selective pressures associated with loss of *hcnA*, *hcnB*, *hcnC*, and *exoY* genes exist, and Wee et al. also observed deletions of various sizes around the respective genes. Furthermore, the selective pressure associated with loss of *hcnABC* locus virulence genes was recently shown to possibly be related to the increased antibiotic resistance in multidrug-resistant strains (36).

We noticed that, while virulence genes *hcnABC* and *exoY* were among the most variable genes, in general, the virulence genes were less often variable within lineages. This may be counterintuitive if loss of virulence is beneficial for bacteria in chronic infections; nonetheless, we recognize that virulence factors may be downregulated rather than deleted as suggested previously by Rau et al. (2010) (44).

Genes were 25 times more often observed to be lost or acquired as groups of genes than as single-gene losses or acquisitions. This observation is in line with previous studies (35, 36) and illustrates how the presence of specific genetic elements enables and defines mobilization of genes: 6 of the 52 highly variable genes were prophage genes, and prophage regions often act as mobile elements. Accordingly, these six prophage genes were part of groups of 56 to 82 genes that were deleted together and constituted genomic islands.

All gene groups that were lost with the six highly variable prophage genes contained the gene orthologs coding for Clp protease as well as the *prtR* gene. PrtR is required for type III secretion system (28), and Clp protease induces virulence by regulating flagellar gene expression and ultimately increasing bacterial adhesion (27). Accordingly, the frequent loss of *prtR* and Clp protease genes adds to our observation that virulence factors are lost during infection. This loss of virulence may be positively selected in the host environments as the virulence factors activate the host immune response; hence, loss of virulence helps the bacteria to hide from the immune defense. Two of eight lineages with loss of Clp and *ptrR* genes also lost *hcnABC* loci and *exoY* genes, and as such, we observed no evidence that losses of the different virulence factors were mutually exclusive or concurrent (see Table S6 in the supplemental material).

Population structure. We described the population structure of our *P. aeruginosa* population of 446 isolates using both SNPs and gene absence/presence information, and in both ways, we identified two major phylogenetic clusters, one with PAO1 and one with UCBPP-PA14, in agreement with previous studies by Hilker et al. (2015) (30) and Stewart et al. (2014) (45). Furthermore, we showed that the levels of SNP and gene differences are significantly lower among PAO1-like isolates than among UCBPP-PA14-like isolates. Finally, we determined that there were significantly fewer SNPs per gene loss/acquisition in isolates belonging to the same clone type than in isolates from different clone types. We have previously shown for this data set that recombination of homologous DNA does not play a major role in microevolution within the CF host (18), and this in line with conclusions previously reported by Winstanley et al. (48). In contrast, it is likely that recombination of homologous DNA plays a relatively larger role over macroevolutionary scales in differentiating clone types and that such recombination plays a role in generating the larger amounts of differences in the number of SNPs per gene that we found in our comparisons of genomes across clone types.

Our study had several limitations. First, it is known that bacterial populations are highly heterogeneous across CF patient airway (49, 50); therefore, while we used the first isolate as a representative of the most recent common ancestor, that approach might not always have been valid. Furthermore, sequencing was mostly performed on single isolates from a sputum sample (218 of 312 cases), which additionally might have reduced the representation of true heterogeneity of bacterial lineages in the patient airway. To address these shortcomings, multiple isolates from each sputum sample should be sequenced. However, since we observed the same genetic variation tendencies across lineages, we believe that these limitations do not weaken the findings of this study. Short-read sequencing data were used in this study, which resulted in incomplete *de novo* assemblies increasing the uncertainty in gene loss and acquisition analysis. Nonetheless, we partly addressed this by using GenAPI for gene presence-absence identification as it performs better on the fragmented genome assemblies than other tools (29).

In summary, we used a genome-wide and hypothesis-free gene presence-absence analysis approach to identify the main patterns of *in vivo* bacterial microevolution. Our analysis adds to the knowledge of how prevalent loss or acquisition of genes is within bacteria evolving in the human host environment and provides a basis to further understand how gene loss and acquisition play a role in host adaptation.

MATERIALS AND METHODS

Bacterial isolates, determination of clone types, and lineage definition. This study used genomic data from a previously reported collection of 474 clinical isolates of *P. aeruginosa* that were sampled from 34 patients with CF attending the Copenhagen Cystic Fibrosis Center at the University Hospital, Rigshospitalet, Denmark (18). Genomes were sequenced as follows: genomic DNA was prepared from *P. aeruginosa* isolates on a QIAcube system using a DNeasy blood and tissue kit (Qiagen) and sequenced on an Illumina HiSeq 2000 platform, generating 100-bp paired-end reads and using a multiplexed protocol to obtain an average of 7,139,922 reads (range, 3,111,062 to 13,085,190) for each of the genomic libraries. On average, isolates had estimated genomic coverage of 107× (55× to 195×).

The clone type of each of the isolates was previously reported by Marvig et al. (2015) from a study that determined the clone types by an *ad hoc* analysis (18), and we furthermore confirmed the clone types by the use of Pactyper (https://github.com/MigleSur/Pactyper), which is a tool developed as part of this study for stable and discriminatory clone typing of bacterial genomes. Pactyper was run with default settings, which defined isolates to be of different clone types if they differed by more than 5,000 SNPs in a core genome of 4,760 genes (i.e., all genes shared by 446 of 474 genomes that were successfully assembled *de novo* [see below]). Isolates of the same clone type and from the same patient were defined as being part of the same lineage.

Bacterial genome assembly. Sequence reads from each isolate were error corrected and assembled *de novo* by SPAdes version 3.10.1 (51) using k-mer sizes from 21 to 127. Assembled contigs were joined to scaffolds per SPAdes default parameters. *De novo* assemblies of sequence reads from 28 of the 474 isolates (6%) were unsuccessful (>500 scaffolds in the final assembly); thus, those isolates were excluded from the analysis.

Genome annotation and identification of gene loss and acquisition within lineage pan-genomes. Genomes assembled *de novo* were annotated using Prokka version 1.11 (52) and a custom annotation database for *P. aeruginosa* species based on PAO1 (RefSeq assembly accession no. GCF_000006765.1) and UCBPP-PA14 (RefSeq assembly accession no. GCF_000014625.1) reference genomes. GenAPI was run with default settings to determine lineage pan-genomes as well as the presence/absence of genes in individual genomes (29). Note that GenAPI default settings include the specification that genes shorter than 150 nucleotides are excluded from the analysis. Genes identified as absent by GenAPI were confirmed as absent by aligning the raw sequencing reads to the pan-genome with bwa v0.7.15 (53). A total of 52 of 13,246 genes which were identified by GenAPI as absent had ≥50% of the gene covered with ≥10× coverage (mosdepth [54]) and therefore were defined as present in all succeeding analyses.

Aggregated pan-genome and visualization. An aggregated pan-genome was determined by gene clustering with GenAPI, which uses CD-HIT-EST version 4.6.1 software (55) and has the requirement for alignments to cover at least 80% of the query gene length and to have a minimum of 90% identity in the alignment. Every gene in the aggregated pan-genome was then aligned back to the individual lineage pan-genomes to determine if the gene was (i) nonpresent in the lineage pan-genome, (ii) present and variable within the lineage, or (iii) present and nonvariable within the lineage. A heat map for the aggregated pan-genome was made with all 45 lineages using R version 3.3.3 (56) and pheatmap library version 1.0.8 (57).

Identification of the most variable genes. For a gene to be considered highly variable, it had to be lost or acquired (variable) in at least 4 lineages (among the top 2% of all variable genes). All genes which were identified as highly variable were manually inspected as follows to confirm the results: read sequences from isolates of interest were aligned to the reference PAO1 (RefSeq assembly accession no. GCF_000006765.1) genome using bowtie2 version 2.3.2 (58) with the default parameters for paired-end sequencing. The sequence alignments at genomic positions of interest were visualized with IGV version 2.4.9 (59) and then manually assessed. Genes of non-*Pseudomonas* origin were manually inspected by evaluating their alignments to the pan-genome genes (from the GenAPI analysis).

In total, two genes (PA1352 and PA3457) were concluded to be falsely called as variable because the alignments to the reference PAO1 genome did not support the prediction of the genes being absent. These false calls were in genome regions which are complex and difficult to assemble *de novo*, i.e., calls of gene presence or absence were found to vary with the success of the assembly of the specific genome region rather than as a result of genuine gene presence or absence.

Reanalysis of *P. aeruginosa* genomes previously analyzed in another *P. aeruginosa* study. Analysis of the data set from a study previously reported by Freschi et al. (2019) (6) included 1,139 of 1,311 genomes as 172 genomes were not publicly available on the day of access (2 March 2019). All available samples were analyzed with SaturnV (https://github.com/ejfresch/saturnV) using the default settings and the "lazy" option and with GenAPI (https://github.com/MigleSur/GenAPI) (29) using the default settings.

Resistance, virulence, pathoadaptive and prophage origin gene identification. Resistance, plasmid, and virulence genes were identified by comparing the aggregated pan-genomes of 45 lineages with the corresponding databases by using ABRicate version 0.8 (60). The gene from the corresponding database was considered present if its identity was at least 98% and the alignment made up a minimum of 25% of the gene length.

The PlasmidFinder (61) database (263 sequences; retrieved 21 March 2018) was used for plasmid gene identification, the VFDB database (2,597 genes; retrieved 21 March 2018) (62) was used for virulence gene identification, the Resfinder database (2,280 genes; retrieved 21 March 2018) (63) was used for resistance gene identification, and the ACLAME database (54,945 genes; retrieved 7 June 2018) (64) was used for prophage origin sequence identification. For pathoadaptive gene identification, a list of 52 pathoadaptive genes reported previously by Marvig et al. (2015) (18) was compared to the aggregated pan-genome of the 45 lineages. All isolate assemblies were inspected for the presence of genes in the essential gene list reported previously by Liberati et al. (2006) by using ABRicate version 0.8 (60).

Fisher's exact test was performed to identify whether the numbers of genes from the corresponding database were significantly different between the within-host variable and nonvariable genes.

Genomic island identification. Genomic islands were predicted using the IslandViewer4 (65) webserver with PAO1 (RefSeq assembly accession no. GCF_000006765.1) as the reference genome. Genomic island prediction was performed for the annotated scaffold sequences. IslandViewer4 integrated tools—IslandPath-DIMOB and SIGI-HMM—were used for prediction of genomic islands.

Pairwise gene and SNP distance estimation between *P. aeruginosa* isolates. The gene distance between genomes was defined as the number of genes not present in one of the genomes as determined by GenAPI (i.e., genes present in one genome and absent in the other and vice versa). Pairwise SNP distance was determined using PacTyper (https://github.com/MigleSur/Pactyper), which uses sequence reads to call and compare SNPs across the core genome. The default thresholds of Pactyper require that sequence reads cover at least 80% of the core genome with not less than 10-fold coverage to ensure exclusion of genomes with poor sequencing coverage. The core genome was defined in this study with GenAPI analysis by including all genes shared by the 446 successfully sequenced *P. aeruginosa* genomes. The core genome contained 4,760 genes (4,705,617 nucleotides).

Phylogenetic tree generation. A SNP-based phylogenetic tree was generated with RAxML version 8.2.11 (66) (with the GTRCAT settings for nucleotide sequence analysis and "12345" as a random number seed) by alignments of the previously defined core genome, and PAO1 (RefSeq assembly accession no. GCF_000006765.1) and UCBPP-PA14 (RefSeq assembly accession no. GCF_000014625.1) were included as reference genomes. A SNP-based phylogenetic tree was also generated with Parsnp (67) both with and without the use of a PhiPack (67) recombination filter. A gene presence-absence-based phylogenetic tree was generated with RAxML version 8.2.11 (66) (with the BINCAT settings for nucleotide sequence analysis and "12345" as a random number seed) by using gene presence-absence information from GenAPI analysis, and PAO1 (RefSeq assembly accession no. GCF_000006765.1) and UCBPP-PA14 (RefSeq assembly accession no. GCF_000014625.1) were included as reference genomes. The Microreact webservice was used to visualize the phylogenetic trees (68).

Data availability. The sequences analyzed in this work are deposited in the Sequence Read Archive (SRA) under accession no. ERP004853.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

TEXT S1, DOCX file, 0.02 MB.

FIG S1, PDF file, 1.2 MB.

FIG S2, PDF file, 0.7 MB.
FIG S3, PDF file, 1.3 MB.
TABLE S1, DOCX file, 0.02 MB.
TABLE S2, DOCX file, 0.01 MB.
TABLE S3, DOCX file, 0.02 MB.
TABLE S4, DOCX file, 0.01 MB.
TABLE S5, DOCX file, 0.01 MB.
TABLE S6, DOCX file, 0.01 MB.

ACKNOWLEDGMENTS

This work was supported by the Danish Cystic Fibrosis Association (Cystisk Fibrose Foreningen) and the Danish National Research Foundation (grant number 126). H.K.J. was supported by the Novo Nordisk Foundation as a clinical research stipend (NNF12OC1015920), by Rigshospitalets Rammebevilling 2015-17 (R88-A3537), by Lundbeckfonden (R167-2013-15229), by Novo Nordisk Fonden (NNF15OC0017444), by RegionH Rammebevilling (R144-A5287), and by Independent Research Fund Denmark/Medical and Health Sciences (FTP-4183-00051).

We furthermore thank Anders Krogh for valuable discussions throughout the study. Also, Fig. 1, Fig. 3, Fig. 4, Fig. 7, and Fig. S1 were created with BioRender.

R.L.M., H.K.J., and S.M. conceived the study. R.L.M. and F.C.N. supervised the study. M.G., R.L.M., and H.K.J. defined the methodology. M.G. and R.L.M. designed the bioinformatics workflows for the analysis. M.G. conducted the analysis. M.G. and R.L.M. analyzed and interpreted the data. M.G. prepared the manuscript draft and visualizations. R.L.M., H.K.J., F.C.N., and S.M. reviewed and edited the draft.

We declare no conflicts of interest.

REFERENCES

1. Brockhurst MA, Harrison E, Hall JP, Richards T, McNally A, MacLean C. 2019. The ecology and evolution of pangenomes. Curr Biol 29:R1094–R1103. https://doi.org/10.1016/j.cub.2019.08.012.
2. Mira A, Martín-Cuadrado AB, D'Auria G, Rodríguez-Valera F. 2010. The bacterial pan-genome: a new paradigm in microbiology. Int Microbiol 13:45–57. https://doi.org/10.2436/20.1501.01.110.
3. Richardson EJ, Bacigalupe R, Harrison EM, Weinert LA, Lycett S, Vrieling M, Robb K, Hoskisson PA, Holden MT, Feil EJ, Paterson GK, Tong SY, Shittu A, van Wamel W, Aanensen DM, Parkhill J, Peacock SJ, Corander J, Holmes M, Fitzgerald JR. 2018. Gene exchange drives the ecological success of a multi-host bacterial pathogen. Nat Ecol Evol 2:1468–1478. https://doi.org/10.1038/s41559-018-0617-0.
4. Yu Z, Ding Y, Yin J, Yu D, Zhang J, Zhang M, Ding M, Zhong W, Qiu J, Li J. 2018. Dissemination of genetic acquisition/loss provides a variety of quorum sensing regulatory properties in Pseudoalteromonas. Int J Mol Sci 19:3636. https://doi.org/10.3390/ijms19113636.
5. Hall JP, Brockhurst MA, Harrison E. 2017. Sampling the mobile gene pool: innovation via horizontal gene transfer in bacteria. Philos Trans R Soc Lond B Biol Sci 372:20160424. https://doi.org/10.1098/rstb.2016.0424.
6. Freschi L, Vincent AT, Jeukens J, Emond-Rheault J-G, Kukavica-Ibrulj I, Dupont M-J, Charette SJ, Boyle B, Levesque RC. 2019. The Pseudomonas aeruginosa pan-genome provides new insights on its population structure, horizontal gene transfer and pathogenicity. Genome Biol Evol 11:109–120. https://doi.org/10.1093/gbe/evy259.
7. Nowell RW, Green S, Laue BE, Sharp PM. 2014. The extent of genome flux and its role in the differentiation of bacterial lineages. Genome Biol Evol 6:1514–1529. https://doi.org/10.1093/gbe/evu123.
8. Denef VJ, Banfield JF. 2012. In situ evolutionary rate measurements show ecological success of recently emerged bacterial hybrids. Science 336:462–466. https://doi.org/10.1126/science.1218389.
9. Lieberman TD, Michel JB, Aingaran M, Potter-Bynoe G, Roux D, Davis MR, Skurnik D, Leiby N, Lipuma JJ, Goldberg JB, McAdam AJ, Priebe GP, Kishony R. 2011. Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes. Nat Genet 43:1275–1280. https://doi.org/10.1038/ng.997.
10. Barrick JE, Yu DS, Yoon SH, Jeong H, Oh TK, Schneider D, Lenski RE, Kim JF. 2009. Genome evolution and adaptation in a long-term experiment with Escherichia coli. Nature 461:1243–1247. https://doi.org/10.1038/nature08480.
11. Palmer KL, Daniel A, Hardy C, Silverman J, Gilmore MS. 2011. Genetic basis for daptomycin resistance in enterococci. Antimicrob Agents Chemother 55:3345–3356. https://doi.org/10.1128/AAC.00207-11.
12. Linkevicius M, Sandegren L, Andersson DI. 2013. Mechanisms and fitness costs of tigecycline resistance in Escherichia coli. J Antimicrob Chemother 68:2809–2819. https://doi.org/10.1093/jac/dkt263.
13. Wong A, Rodrigue N, Kassen R. 2012. Genomics of adaptation during experimental evolution of the opportunistic pathogen Pseudomonas aeruginosa. PLoS Genet 8:e1002928. https://doi.org/10.1371/journal.pgen.1002928.
14. Ensminger AW, Yassin Y, Miron A, Isberg RR. 2012. Experimental evolution of Legionella pneumophila in mouse macrophages leads to strains with altered determinants of environmental survival. PLoS Pathog 8:e1002731. https://doi.org/10.1371/journal.ppat.1002731.
15. Yang L, Jelsbak L, Marvig RL, Damkiær S, Workman CT, Rau MH, Hansen SK, Folkesson A, Johansen HK, Ciofu O, Høiby N, Sommer MOA, Molin S. 2011. Evolutionary dynamics of bacteria in a human host environment. Proc Natl Acad Sci U S A 108:7481–7486. https://doi.org/10.1073/pnas.1018249108.
16. da Silva Filho LVRF, de Aguiar Ferreira F, Caldeira Reis FJ, de Britto MCA, Levy CE, Clark O, Ribeiro JD. 2013. Pseudomonas aeruginosa infection in patients with cystic fibrosis: scientific evidence regarding clinical impact, diagnosis, and treatment. J Bras Pneumol 39:495–512. https://doi.org/10.1590/S1806-37132013000400015.
17. Klockgether J, Cramer N, Fischer S, Wiehlmann L, Tümmler B. 2018. Long-term microevolution of Pseudomonas aeruginosa differs between mildly and severely affected cystic fibrosis lungs. Am J Respir Cell Mol Biol 59:246–256. https://doi.org/10.1165/rcmb.2017-0356OC.
18. 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.
19. La Rosa R, Johansen HK, Molin S. 2018. Convergent metabolic specialization through distinct evolutionary paths in Pseudomonas aeruginosa. mBio 9:e00269-18. https://doi.org/10.1128/mBio.00269-18.
20. Sheppard SK, Guttman DS, Fitzgerald JR. 2018. Population genomics of

bacterial host adaptation. Nat Rev Genet 19:549–565. https://doi.org/10.1038/s41576-018-0032-z.

21. Burgener EB, Sweere JM, Bach MS, Secor PR, Haddock N, Jennings LK, Marvig RL, Krogh Johansen H, Rossi E, Cao X, Tian L, Nedelec L, Molin S, Bollyky PL, Milla CE. 2019. Filamentous bacteriophages are associated with chronic Pseudomonas lung infections and antibiotic resistance in cystic fibrosis. Sci Transl Med 11:eaau9748. https://doi.org/10.1126/scitranslmed.aau9748.
22. Sweere JM, Van Belleghem JD, Ishak H, Bach MS, Popescu M, Sunkari V, Kaber G, Manasherob R, Suh GA, Cao X, de Vries CR, Lam DN, Marshall PL, Birukova M, Katznelson E, Lazzareschi DV, Balaji S, Keswani SG, Hawn TR, Secor PR, Bollyky PL. 2019. Bacteriophage trigger antiviral immunity and prevent clearance of bacterial infection. Science 363:eaat9691. https://doi.org/10.1126/science.aat9691.
23. Liberati NT, Urbach JM, Miyata S, Lee DG, Drenkard E, Wu G, Villanueva J, Wei T, Ausubel FM. 2006. An ordered, nonredundant library of Pseudomonas aeruginosa strain PA14 transposon insertion mutants. Proc Natl Acad Sci U S A 103:2833–2838. https://doi.org/10.1073/pnas.0511100103.
24. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos JS, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421. https://doi.org/10.1186/1471-2105-10-421.
25. Winsor GL, Lo R, Ho Sui SJ, Ung KS, Huang S, Cheng D, Ching WKH, Hancock RE, Brinkman FS. 2005. Pseudomonas aeruginosa Genome Database and PseudoCAP: facilitating community-based, continually updated, genome annotation. Nucleic Acids Res 33(Database issue): D338–D343. https://doi.org/10.1093/nar/gki047.
26. Wee BA, Tai AS, Sherrard LJ, Ben Zakour NL, Hanks KR, Kidd TJ, Ramsay KA, Lamont I, Whiley DM, Bell SC, Beatson SA. 2018. Whole genome sequencing reveals the emergence of a Pseudomonas aeruginosa shared strain sub-lineage among patients treated within a single cystic fibrosis centre. BMC Genomics 19:644. https://doi.org/10.1186/s12864-018-5018-x.
27. Ingmer H, Brøndsted L. 2009. Proteases in bacterial pathogenesis. Res Microbiol 160:704–710. https://doi.org/10.1016/j.resmic.2009.08.017.
28. Sun Z, Shi J, Liu C, Jin Y, Li K, Chen R, Jin S, Wu W. 2014. PrtR homeostasis contributes to Pseudomonas aeruginosa pathogenesis and resistance against ciprofloxacin. Infect Immun 82:1638–1647. https://doi.org/10.1128/IAI.01388-13.
29. Gabrielaite M, Marvig RL. 2020. GenAPI: a tool for gene absence-presence identification in fragmented bacterial genome sequences. BMC Bioinformatics 21:320. https://doi.org/10.1186/s12859-020-03657-5.
30. Hilker R, Munder A, Klockgether J, Losada PM, Chouvarine P, Cramer N, Davenport CF, Dethlefsen S, Fischer S, Peng H, Schönfelder T, Türk O, Wiehlmann L, Wölbeling F, Gulbins E, Goesmann A, Tümmler B. 2015. Interclonal gradient of virulence in the Pseudomonas aeruginosa pangenome from disease and environment. Environ Microbiol 17:29–46. https://doi.org/10.1111/1462-2920.12606.
31. Mathee K, Narasimhan G, Valdes C, Qiu X, Matewish JM, Koehrsen M, Rokas A, Yandava CN, Engels R, Zeng E, Olavarietta R, Doud M, Smith RS, Montgomery P, White JR, Godfrey PA, Kodira C, Birren B, Galagan JE, Lory S. 2008. Dynamics of Pseudomonas aeruginosa genome evolution. Proc Natl Acad Sci U S A 105:3100–3105. https://doi.org/10.1073/pnas.0711982105.
32. Valot B, Guyeux C, Rolland JY, Mazouzi K, Bertrand X, Hocquet D. 2015. What it takes to be a Pseudomonas aeruginosa? The core genome of the opportunistic pathogen updated. PLoS One 10:e0126468. https://doi.org/10.1371/journal.pone.0126468.
33. Klockgether J, Cramer N, Wiehlmann L, Davenport CF, Tümmler B. 2011. Pseudomonas aeruginosa genomic structure and diversity. Front Microbiol 2:150. https://doi.org/10.3389/fmicb.2011.00150.
34. Rau MH, Marvig RL, Ehrlich GD, Molin S, Jelsbak L. 2012. Deletion and acquisition of genomic content during early stage adaptation of Pseudomonas aeruginosa to a human host environment. Environ Microbiol 14:2200–2211. https://doi.org/10.1111/j.1462-2920.2012.02795.x.
35. Hocquet D, Petitjean M, Rohmer L, Valot B, Kulasekara HD, Bedel E, Bertrand X, Plésiat P, Köhler T, Pantel A, Jacobs MA, Hoffman LR, Miller SI. 2016. Pyomelanin-producing Pseudomonas aeruginosa selected during chronic infections have a large chromosomal deletion which confers resistance to pyocins. Environ Microbiol 18:3482–3493. https://doi.org/10.1111/1462-2920.13336.
36. Hwang W, Yoon SS. 2019. Virulence characteristics and an action mode of antibiotic resistance in multidrug-resistant Pseudomonas aeruginosa. Sci Rep 9:487. https://doi.org/10.1038/s41598-018-37422-9.
37. Colque CA, Albarracín Orio AG, Feliziani S, Marvig RL, Tobares AR, Johansen HK, Molin S, Smania AM. 2020. Hypermutator Pseudomonas aeruginosa exploits multiple genetic pathways to develop multidrug resistance during long-term infections in the airways of cystic fibrosis patients. Antimicrob Agents Chemother 18:e02142-19. https://doi.org/10.1128/AAC.02142-19.
38. 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.
39. 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.
40. 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.
41. Poulsen BE, Yang R, Clatworthy AE, White T, Osmulski SJ, Li L, Penaranda C, Lander ES, Shoresh N, Hung DT. 2019. Defining the core essential genome of Pseudomonas aeruginosa. Proc Natl Acad Sci U S A 116:10072–10080. https://doi.org/10.1073/pnas.1900570116.
42. Butaitė E, Baumgartner M, Wyder S, Kümmerli R. 2017. Siderophore cheating and cheating resistance shape competition for iron in soil and freshwater Pseudomonas communities. Nat Commun 8:414. https://doi.org/10.1038/s41467-017-00509-4.
43. Ghoul M, Griffin AS, West SA. 2014. Toward an evolutionary definition of cheating. Evolution 68:318–331. https://doi.org/10.1111/evo.12266.
44. Rau MH, Hansen SK, Johansen HK, Thomsen LE, Workman CT, Nielsen KF, Jelsbak L, Høiby N, Yang L, Molin S. 2010. Early adaptive developments of Pseudomonas aeruginosa after the transition from life in the environment to persistent colonization in the airways of human cystic fibrosis hosts. Environ Microbiol 12:1643–1658. https://doi.org/10.1111/j.1462-2920.2010.02211.x.
45. Stewart L, Ford A, Sangal V, Jeukens J, Boyle B, Kukavica-Ibrulj I, Caim S, Crossman L, Hoskisson PA, Levesque R, Tucker NP. 2014. Draft genomes of 12 host-adapted and environmental isolates of Pseudomonas aeruginosa and their positions in the core genome phylogeny. Pathog Dis 71:20–25. https://doi.org/10.1111/2049-632X.12107.
46. Kottara A, Hall JP, Harrison E, Brockhurst MA. 2018. Variable plasmid fitness effects and mobile genetic element dynamics across Pseudomonas species. FEMS Microbiol Ecol 94:fix172. https://doi.org/10.1093/femsec/fix172.
47. National Center for Biotechnology Information (NCBI). 2018. Pseudomonas aeruginosa. https://www.ncbi.nlm.nih.gov/genome/browse/#!/prokaryotes/187/.
48. 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.
49. Lieberman TD, Wilson D, Misra R, Xiong LL, Moodley P, Cohen T, Kishony R. 2016. Genomic diversity in autopsy samples reveals within-host dissemination of HIV-associated Mycobacterium tuberculosis. Nat Med 22:1470–1474. https://doi.org/10.1038/nm.4205.
50. 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.
51. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455–477. https://doi.org/10.1089/cmb.2012.0021.
52. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069. https://doi.org/10.1093/bioinformatics/btu153.
53. Li H. 2013. Aligning sequence reads, clone sequences and assembly. arXiv:1303.3997 [q-bio.GN].
54. Pedersen BS, Quinlan AR. 2018. Mosdepth: quick coverage calculation for genomes and exomes. Bioinformatics 34:867–868. https://doi.org/10.1093/bioinformatics/btx699.
55. Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT. Bioinformatics 28:3150–3152. https://doi.org/10.1093/bioinformatics/bts565.

56. R Core Team. 2018. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.
57. Kolde R. 2018. pheatmap: Pretty Heatmaps, R package version 1.0.10. https://CRAN.R-project.org/package=pheatmap.
58. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359. https://doi.org/10.1038/nmeth.1923.
59. Thorvaldsdóttir H, Robinson JT, Mesirov JP. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 14:178–192. https://doi.org/10.1093/bib/bbs017.
60. Seemann T. 2018. ABRicate. https://github.com/tseemann/abricate.
61. Carattoli A, Zankari E, Garciá-Fernández A, Larsen MV, Lund O, Villa L, Aarestrup FM, Hasman H. 2014. In silico detection and typing of plasmids using plasmidfinder and plasmid multilocus sequence typing. Antimicrob Agents Chemother 58:3895–3903. https://doi.org/10.1128/AAC.02412-14.
62. Chen L, Zheng D, Liu B, Yang J, Jin Q. 2016. VFDB 2016: hierarchical and refined dataset for big data analysis—10 years on. Nucleic Acids Res 44:D694–D697. https://doi.org/10.1093/nar/gkv1239.
63. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, Aarestrup FM, Larsen MV. 2012. Identification of acquired antimicrobial resistance genes. J Antimicrob Chemother 67:2640–2644. https://doi.org/10.1093/jac/dks261.
64. Leplae R, Lima-Mendez G, Toussaint A. 2010. ACLAME: a CLAssification of mobile genetic elements, update 2010. Nucleic Acids Res 38:D57–D61. https://doi.org/10.1093/nar/gkp938.
65. Bertelli C, Laird MR, Williams KP, Lau BY, Hoad G, Winsor GL, Brinkman FS, Simon Fraser University Research Computing Group. 2017. IslandViewer 4: expanded prediction of genomic islands for larger-scale datasets. Nucleic Acids Res 45:W30–W35. https://doi.org/10.1093/nar/gkx343.
66. 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.
67. Treangen TJ, Ondov BD, Koren S, Phillippy AM. 2014. The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. Genome Biol 15:524. https://doi.org/10.1186/s13059-014-0524-x.
68. Argimón S, Abudahab K, Goater RJ, Fedosejev A, Bhai J, Glasner C, Feil EJ, Holden MTG, Yeats CA, Grundmann H, Spratt BG, Aanensen DM. 2016. Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. Microb Genom 2:e000093. https://doi.org/10.1099/mgen.0.000093.