1. Title:

Divergent, coexisting, *Pseudomonas aeruginosa* lineages in chronic cystic fibrosis lung infections

2. Authors:

David Williams[1], Benjamin Evans[1,2], Sam Haldenby[1], Martin J. Walshaw[3], Michael A. Brockhurst[4], Craig Winstanley[5], Steve Paterson[1]

3. Departments and institutions:

[1] Institute of Integrative Biology, Biosciences Building, University of Liverpool, Crown Street, Liverpool, L69 7ZB

[2] Current address: Department of Life Sciences, Anglia Ruskin University, East Road, Cambridge, CB1 1PT

[3] Liverpool Heart and Chest Hospital NHS Foundation Trust, Thomas Drive, Liverpool, L14 3PE

[4] Department of Biology, University of York, Wentworth Way, YO10 5DD

[5] Department of Clinical Infection, Microbiology and Immunology, Institute of Infection and Global Health, The Ronald Ross Building, University of Liverpool, 8 West Derby Street, Liverpool, L69 7BE

4. Corresponding author:

Prof. Steve Paterson

Institute of Integrative Biology,

Biosciences Building,

University of Liverpool,

Crown Street,

Liverpool L69 7ZB,
United Kingdom
Tel. +44 151 795 4521
Fax. +44 151 795 4408
email: S.Paterson@liverpool.ac.uk

5. Authors' contributions

MB, CW and SP conceived the study, obtained funding and contributed resources. DW, MB, CW and SP designed the study, interpreted the data and wrote the manuscript. DW performed evolutionary analyses. MJW acquired clinical samples. BE prepared DNA samples for sequencing. SH prepared sequence data and performed variant discovery. All authors read and approved the manuscript.

6. Sources of support

This work was supported by The Wellcome Trust (award 093306/Z/10). We acknowledge the technical assistance provided by staff at the Centre for Genomic Research, University of Liverpool, UK and are grateful to staff and patients at the Regional Adult Cystic Fibrosis Unit, Liverpool, UK.

7. Running head

Genetic diversity in chronic lung infections

8. Descriptor number: 9.16 Cystic Fibrosis: Basic Studies

9. Word counts. Body: 3354, abstract: 246

10. At a Glance Commentary:

Scientific Knowledge on the Subject: Chronic lung infections caused by *Pseudomonas aeruginosa* remain the major cause of morbidity and mortality among cystic fibrosis patients. Transmissible strains, such as the widespread (UK and North America) Liverpool Epidemic Strain, are especially problematic and require segregation of affected patients within CF units to prevent cross-infection. Surprisingly high levels of phenotypic diversity within individual patient sputum samples have been demonstrated, and genome sequencing of sequential isolates suggests that the pathogen accumulates mutations over time. However, little is known about the underlying genetic diversity within infecting populations, or the distribution of diversity between patients..

What This Study Adds to the Field: Using genomic analysis of sequential isolates, others have found evidence for rare replacements of one *P. aeruginosa* lineage by another. Here we show, using large-scale population genomic analyses, that the coexistence of distinct lineages of the *P. aeruginosa* Liverpool Epidemic Strain is typical, occurring in seven of nine chronically infected cystic fibrosis patients sampled. Genetic divergence between lineages *within* patients was greater than *between*, implying acquisition of diverse *P. aeruginosa* populations, and potentially acquisition of distinct lineages among the LES-infected cohort. Furthermore, evidence for ongoing homologous recombination within and between divergent lineages provides evidence for an alternative mode of potentially adaptive evolution by *P. aeruginosa* during chronic infection.

11. This article has an online data supplement, which is accessible from this issue's table of content online at www.atsjournals.org

Abstract

Rationale

Pseudomonas aeruginosa, the predominant cause of chronic airway infections of cystic fibrosis patients, exhibits extensive phenotypic diversity among isolates within and between sputum samples, but little is known about the underlying genetic diversity.

Objectives

To characterise the population genetic structure of transmissible *P. aeruginosa* Liverpool Epidemic Strain in chronic infections of nine cystic fibrosis patients, and infer evolutionary processes associated with adaptation to the cystic fibrosis lung.

Methods

We performed whole genome sequencing of *Pseudomonas aeruginosa* isolates and pooled populations and used comparative analyses of genome sequences, including phylogenetic reconstructions and resolution of population structure from genome-wide allele frequencies.

Measurements and Main Results

Genome sequences were obtained for 360 isolates from nine patients. Phylogenetic reconstruction of the ancestry of 40 individually sequenced isolates from one patient sputum sample revealed the coexistence of two genetically diverged, recombining lineages exchanging potentially adaptive mutations. Analysis of population samples for eight additional patients indicated coexisting lineages in six cases. Reconstruction of the ancestry of individually sequenced isolates from all patients indicated smaller genetic distances between than within patients in most cases.

Conclusions

Our population-level analysis demonstrates that coexistence of distinct lineages of *Pseudomonas aeruginosa* Liverpool Epidemic Strain within individuals is common. In

several cases, coexisting lineages may have been present in the infecting inoculum or assembled through multiple transmissions. Divergent lineages can share mutations *via* homologous recombination, potentially aiding adaptation to the airway during chronic infection. The genetic diversity of this transmissible strain within infections, revealed by high-resolution genomics, has implications for patient segregation and therapeutic strategies.

Total words in abstract: 246

Key words: bacteria, population genetics, genomics, homologous recombination

Introduction

Pseudomonas aeruginosa is the most common cause of airway infection in cystic fibrosis (CF) [1] and once established in the chronic stage is notoriously resistant to clearance by chemotherapy [2]. Chronic stage infections exhibit both adaptation and diversification. The genetic mechanisms underlying some chronic-stage adaptations, such as the switch to mucoid phenotype, have been well established for many years [3]. Other common adaptations include mutations in the gene encoding the key quorum sensing regulator LasR [4], loss of motility [5], auxotrophy [6], hypermutability [7], and increasing resistance to antibiotics [8]. It has been shown that evolutionary adaptation can occur rapidly in the airways of CF patients [9;10]. While most patients acquire their infecting *P. aeruginosa* from environmental sources with subsequent adaptation to the CF airway [11], there have been a number of transmissible strains identified [12]. Notable among these is the Liverpool Epidemic Strain (LES), which is the most abundant clone of *P. aeruginosa* isolated from CF patients in the UK [13;14], and has been reported in North America [15;16].

The genetic basis of diversification in chronic infections remains poorly understood. Whereas a number of studies have reported on the genetic adaptation of *P. aeruginosa* during CF lung infections by targeting specific genes [7] or whole genomes for sequencing [17;18], these studies have generally been optimised to capture genetic changes over time. At the expense of sampling depth within individuals, these studies sampled sequential isolates from individual CF patients or single isolates from many different CF patients. Consequently, it is essential that investigations of population-scale genetic diversity of *P. aeruginosa* be extended to consider diversity both within and between multiple chronically infected CF patients.

Evidence from phenotypic studies suggests widespread heterogeneity [19-22], and genome sequencing of paired *P. aeruginosa* isolates from three individuals revealed genetic diversity [23]. A recent study implicated spatial separation within the CF lung as causing diversification into distinct lineages in a single CF patient [24]. A report of a less common CF pathogen, *Burkholderia dolosa*, described unexpectedly high genetic diversity within patients using genome sequencing of pooled population samples [25]. To characterise the population structure of the *P. aeruginosa* LES populations in chronically infected CF patients, we assayed the genome sequence diversity among 40 isolates from a sputum sample for each of nine adults attending the same CF unit.

Methods

Acquisition of samples and isolation of *Pseudomonas aeruginosa*

Samples were collected from nine adult cystic fibrosis patients, each chronically infected with the *P. aeruginosa* LES, as described previously [19]. Briefly, a sputum sample was collected from each patient at a routine visit to the Regional Adult Cystic Fibrosis Unit in Liverpool, UK during January 2009. Sputum was treated with an equal volume of Sputasol (Oxoid), incubated at room temperature with shaking at 200 r.p.m. for 15 min, and then cultured on *Pseudomonas* selective agar under aerobic conditions with CN supplement (Oxoid) as described previously [19]. Forty LES colonies were selected to maximise colony morphology diversity and identified as described previously [19]. Details concerning age, sex and clinical status of patients CF01 and CF03-CF10 were given in our previous study [19]. Based on information available since this previous study, CF01, CF03, CF05 and CF07 are now known to have been LES positive since at least 1995; CF04, CF06, CF09 and CF10 since 1995 but

before 2004; CF08 since at least 2008. This study was approved by the local research ethics committee (REC reference 08/H1006/47).

Genomic DNA preparation and sequencing

Details of DNA extraction is outlined in the Online Data Supplement. Library preparation and whole-genome shotgun sequencing was performed by the Centre for Genomic Research at the University of Liverpool, UK using Illumina short read sequencing technology. Details of quality control of sequenced read data is outlined in the on-line supplement. The European Nucleotide Archive accession number for the study is PRJEB6642.

Variant calling and *de novo* genome assemblies

Reads were aligned to the *P. aeruginosa* LESB58 reference genome sequence (National Center for Biotechnology Information accession number: NC_011770) with the BWA-MEM aligner [26]. For individually sequenced isolates, paired-end reads were assembled *de novo* using SPAdes Genome Assembler version 3.0 [27]. Details of single nucleotide polymorphism, insertion and deletion discovery, prediction of genetic variant effects on protein sequences and *de novo* genome assembly are outlined in the on-line supplement. Genome assemblies were used to double-check homoplasies indicated by analysis of read alignments. The European Nucleotide Archive accession numbers for the assemblies of sequenced isolates are ERZ021677-716.

Phylogenetic reconstruction and hypothesis testing

Variable sites identified among the aligned sequencing reads for each genome sequence were combined into a multiple alignment for phylogenetic analysis. Genome phylogenies were reconstructed using the BIONJ algorithm [28] implemented in the APE version 3.1-2 [29] for the R statistical computing environment version 3.1.0. Statistical support for edges in phylogenies were split frequencies among a non-parametric bootstrap replicate sample of maximum-

likelihood phylogenies inferred using Garli version 2.01 [30] and a Bayesian sample of phylogenies inferred using MrBayes version 3.2.2 [31] and counted using methods of the DendroPy library for phylogenetic computing version 3.12.0 [32] in Python version 2.7.8. The HKY85 nucleotide substitution model was selected for use in Garli with JModelTest2 version 2.1.5 [33]. *P. aeruginosa* LESB58 genome sequence was used as an outgroup for rooting, the suitability of which was tested by phylogenetic reconstruction including sequences of distantly related LESlike 4, 5, 7, DK2 and PAO1 *P. aeruginosa* isolates. The scale-boot R package version 0.3-3 was used to perform the Approximately Unbiased (AU) test on a multiscale bootstrap for the alternative and maximum likelihood topologies against the sequence alignment [34] . Details of phylogenetic reconstructions, homoplasy identification and inference of homologous recombination are outlined in the Online Data Supplement.

Inference of divergent lineages from single nucleotide polymorphism frequencies

For each patient sputum sample, we tested for the coexistence of a pair of divergent lineages. A lineage is defined as a group of isolates with high genetic similarity. Divergent coexisting lineages in a sample are defined as each exhibiting more genetic differences from their most recent common ancestor (MRCA) than their MRCA has to the MRCA of all cohort patient isolates *i.e.*, of this epidemic, inferred from all samples. An alternative hypothesis is all isolates in a sample forming a single lineage where genetic divergence within the sample is less than to other samples. These hypotheses concern the deepest part of the phylogeny in each sample: the length of the root edge must be shorter than the lengths of the edges descending from the deepest bifurcation (the 'primary' edges) for the coexisting, divergent lineages hypothesis to be supported. The root and primary edges can be inferred from their corresponding root and primary peaks in the single nucleotide polymorphism frequency

distributions. Details of inference of divergent lineages from single nucleotide polymorphism frequencies are outlined in more detail in the Online Data Supplement. Monte Carlo simulation of nucleotide sequence evolution over the hypothesised phylogenies for corroboration was achieved using the PhyloSim R package version 2.1.1 [35].

Results

A single chronic CF infection sputum sample contains divergent, recombining *P. aeruginosa* lineages

Among 40 *P. aeruginosa* LES isolates from a CF patient sputum sample collected in 2009 (CF03) that were individually genome sequenced, we identified between 71 and 130 single nucleotide polymorphisms (SNPs) relative to the complete genome sequence of *P. aeruginosa* LESB58, isolated in 1988. Despite these 40 isolates being obtained from a single patient sample, reconstructions of their shared evolutionary history, by different methods, revealed two divergent lineages (clades; figure 1). One lineage had 13 members (CF03 lineage A), the other had 27 (CF03 lineage B).

CF03 lineage A was characterised by 55 shared SNPs while lineage B had 24 shared SNPs. There were 79 SNPs separating the coexisting CF03 lineages A and B but, by contrast, only 42 SNPs separate their most recent common ancestor (MRCA) from the LESB58 reference sequence and phylogenetic outgroup. The suitability of LESB58 as an outgroup is demonstrated by a phylogenetic reconstruction in which it is partitioned with distantly related isolates from the study sequences (figure 1, Online Data Supplement). CF03 lineage B genome sequences were more similar to LESB58 than to lineage A genome sequences, yet both lineages were from the same CF chronic infection sputum sample (figure 2). Mutations

exclusive to each lineage included those predicted to alter proteins associated with virulence factors (supplemental tables 1 and 2). For example, the 13 members of lineage A are predicted to have a truncated MexB multidrug efflux transporter (PLES_04241) while the 27 members of lineage B are predicted to have a truncated anti-sigma factor MucA (PLES_45801) involved in the regulation of alginate biosynthesis.

While mutually exclusive mutations define the two coexisting CF03 lineages, numerous mutations shared between a minority of each lineage (homoplasies) provide evidence of DNA transfer between cells via homologous recombination (figures 1 and 2). An alternative hypothesis excluding recombination between cells requires coincidental, parallel mutations to explain the phylogenetic distribution of these mutations among isolates, the probabilities of which are low ($< 1 \times 10^{-7}$) and listed in table 1. Two SNPs were common between a minority of isolates in each lineage indicating horizontal genetic transfer between the CF03 lineages. One of these is a non-synonymous (protein altering) replacement in *lysC* which codes for an aspartate kinase in LESB58 (locus ID: PLES_44121; *P. aeruginosa* PAO1 locus ID: PA0904; homoplasy 3 in figure 1 and table 1). Two homoplasies were detected within the larger clade indicating horizontal genetic transfer among members of CF03 lineage B. One of these intra-lineage transfers was a deletion predicted to cause the truncation of *mpl*, an ORF encoding a Mur family ligase in LESB58 (PLES_09561; PA4020; homoplasy 1 in figure 1 and table 1). The other intra-lineage transfer is a deletion of two codons within *glpT* which encodes a glycerol-3-phosphate transporter (PLES_56291; PA5235; homoplasy 4 in figure 1 and table 1).

Coexisting, divergent lineages in chronic CF infections are typical for the Liverpool Epidemic Strain.

To assess the prevalence of divergent *P. aeruginosa* LES lineages within chronic infections, we investigated a further eight CF patients. We used SNP frequencies, derived from sequencing an equimolar pool of genomic DNA from 40 isolates per patient, to estimate first, the genetic distance between the inferred MRCA of the epidemic and the MRCA of the patient sample (SNPs fixed in each sample minus SNPs fixed in all samples, root edge in a phylogeny) and second, the genetic distances between the lineages descending from the patient sample MRCA ('primary' edges to each lineage). Thus, if both primary lineage edges are longer than the root edge, divergent lineages are present. To validate the root and lineage edges in the phylogenies deduced from SNP frequencies within each pooled data sample, we, first, sequenced a pair of isolates from each patient to confirm that SNPs inferred as lineage-specific were always in linkage (orange and turquoise in figures 4 and 5) and, second, simulated sequence evolution along each phylogeny to ensure simulated SNP frequencies (red, blue and purple peaks in figures 4 and 5) agreed with observed SNP frequencies. Finally, we used the sequences of all 40 isolates from CF03 to validate these approaches. Both observed and simulated SNP frequency distributions for CF03 correspond to the phylogeny shape: two deeply divergent lineages descending from the sample MRCA (figure 3).

In addition to CF03, the presence of two divergent lineages was detected in six of the eight other patients (figure 4). Thus, chronic infections within these patients harbour a pair of *P. aeruginosa* lineages that differ from their MRCA by more SNPs than their MRCA differs from the inferred epidemic origin (epidemic MRCA inferred from the whole dataset). All members of one of the two lineages in sample CF09 shared an abundance of mutations including a 3 bp out-of-frame deletion in *mutS*, and non-synonymous SNPs in *mutM* and *uvrB*. Mutations at these loci have been shown to cause the “hyper-mutator” phenotype [36;37]

that in a previous report was identified in 36% of CF airway, but in no non-CF, *P. aeruginosa* infections [38]. Less structure was observed for the remaining two patients (figure 5), such that the most divergent lineages share more SNPs with each other than differentiates them from the epidemic MRCA. Thus, the edges arising from the first bifurcation at the patient sample MRCA were shorter than the root edge to the epidemic MRCA.

Comparisons of *P. aeruginosa* lineages among chronically infected CF patients suggests transmissions of diverse populations or multiple lineages.

We next investigated whether the differences between coexisting lineages can be explained exclusively by diversification within a patient following an initial infection. Alternative explanations include a single transmission of a diverse LES population from which lineages diverge, or multiple transmissions of different lineages to a patient causing superinfection. To elucidate patterns of diversification, we reconstructed the ancestry and evolutionary relationships among the 16 distinct *P. aeruginosa* LES lineages identified among the 9 CF patients. All of the individually sequenced isolates were included in the reconstruction using SNPs relative to the LESB58 genome: 40 isolates from patient CF03 and two each from the other eight patients (figure 6).

Within-patient diversification was clearly for patients CF06 and CF10. Thus, their isolates each grouped together to the exclusion of others, *i.e.* formed monophyletic clades with high bootstrap support. Furthermore, the relatively small diversity in their pooled samples was well represented in the sequenced isolate pairs (figure 5), indicating all isolates would group in their respective CF06 or CF10 clades (figure 6). Elsewhere in the phylogeny, isolates did not appear to group within patients. To confirm this, given that the phylogeny was poorly resolved

in the region where lineages diverged, we performed an explicit test of whether specific groupings were supported by the data using Shimodaira's AU test (see Online Data Supplement for details and discussion). Diversification exclusively within patients (monophyly) was rejected by the AU test for patients CF03 ($p = 0.0019$), CF05 ($p = 0.0006$) and CF07 ($p = 0.0087$) but not CF08 ($p = 0.1423$) or CF09 ($p = 0.3679$). CF03, CF05 and CF07 are therefore consistent with either superinfection or with transmission of genetically diverse innocula. The CF04 patient isolates are the only others that group together, but according to the population data represent only one of two lineages in CF04 (figure 4B). The phylogenetic placement of CF04 isolates implies a transmission of one lineage from CF01. The other CF04 lineage, containing 47 SNPs absent from CF01, supports either superinfection from another patient or transfer of a genetically diverse innoculum from CF01 and subsequent loss of a lineage from CF01.

Discussion

We observed high population genetic structure within *P. aeruginosa* infections such that in seven out of nine patients, divergent LES lineages were identified. These coexisting lineages were typically more closely related to lineages in other patients than to each other and include the broadest non-hypermutator *P. aeruginosa* genetic diversity within a single patient yet reported. In one such case, that we examined in more detail, genetic transfer between the divergent lineages by homologous recombination was evident. Genetic exchange between coexisting but divergent lineages can increase genetic variation and provides a mechanism by which adaptation to the lung environment may be accelerated.

Our study is the first to observe that multiple *P. aeruginosa* lineages coexisting within individual patients usually arise from genetic diversity acquired from other patients and that multiple coexisting lineages may be a general feature of CF chronic infection by the LES. This might reflect a unique trait of this strain or transmissible *P. aeruginosa* strains more generally, rather than a widespread characteristic of CF infections. Alternatively, such divergent lineages within *P. aeruginosa* infections may be relatively common, but would only be detectable using a systematic approach as employed here to characterise genetic diversity within infections. To date, no other studies have been designed to quantify the contemporary genetic diversity within each patient of a cohort. A recent report described coexisting lineages in a cystic fibrosis individual with evidence of spatial separation between the naso-pharynx and the lower lung correlating with genetic distance [24]. Another recent report described coexisting lineages in two patients dominated by hyper-mutators [39]. Other studies have generated data consistent with, but not conclusive of, coexisting, divergent lineages. Evidence for transmission of the abundant *P. aeruginosa* clone C lineage amongst siblings, with the possibility of subsequent coexistence of two clone C lineages, has been reported [40], while Chung *et al.* (2012) [23], found 54 SNPs and 38 indels differentiating a pair of non-hypermutator isolates from a single patient. This latter evolutionary distance is comparable to the divergences between co-existing lineages in our study. Our results are also consistent with a recent study of *Burkholderia dolosa*, a relatively rare CF airway pathogen, which included numerous isolates from single sputum samples and identified coexisting divergent lineages [25]. However, unlike the current report, comparative analyses between patient samples have not been performed and diversification has been suggested to have proceeded exclusively within each patient since initial infection, as opposed to acquisition of diverse inocula or divergent lineages.

Despite the extensive phenotypic diversity present amongst isolates [19], our genomic analysis indicates that the presence of discrete lineages, rather than a continuum of diversity within a sputum sample, is typical for LES infecting populations. Given the complexity and spatial heterogeneity of the CF airway, maintenance of the inter-lineage diversity may simply reflect the availability of sufficient niches to accommodate newly acquired invading *P. aeruginosa* populations. The apparent restriction to two distinct coexisting lineages in any one patient, may be a reflection of physiological compartmentalisation within the lungs, for example between the two lungs, or the lungs and paranasal sinuses [24] . An alternative explanation is a skewed community composition among lineages: one could be numerically dominant while many others are rare.

In the present study, we provide evidence for ongoing genetic exchange via homologous recombination among pathogenic *P. aeruginosa* populations during chronic infection. Evidence for recombination has been previously reported in sequence data from large *P. aeruginosa* isolate collections from environmental, animal and human sources [41;42],as well as in two studies of chronic CF airway infections [18;23]. When chronic infections contain significant genetic diversity as reported here, the potential for homologous recombination to generate novel genotypes is increased because of greater differences in genetic backgrounds within the patient. Epistatic effects such as a mutation being neutral to the CF airway in one *P. aeruginosa* genetic background, but adaptive in another [43;44], may be intensified by the greater genetic differences between coexisting lineages.

The cohort segregation policy adopted in Liverpool (UK) was designed to prevent transmission of the LES to patients free from *P. aeruginosa*, or infected with other strains [45], and as such the patients sampled in this study were not segregated from each other. While cohort segregation has been proven successful in halting the spread of the strain to new patients, our data indicate that it may not have prevented further transmission events amongst the LES-infected cohort. Although CF patients infected with LES are known to have a higher rate of mortality than those with other *P. aeruginosa* strains [46], at present, we do not know whether having multiple distinct lineages of LES is worse than having a single lineage. Nor can we be sure that these sub-lineages remain stable within patients. Our previous study, based on isolate phenotyping, suggested that populations were dynamic over a period of several months [19]. Further studies are needed to address the issue of lineage stability over time and to determine the clinical consequences of the coexistence of different lineages.

The accurate and rapid sequencing of bacterial genome sequences, made possible by the most recently available bench-top DNA sequencing platforms, provides numerous advantages over conventional methods for diagnosing hospital outbreaks and tracing transmission routes [47]. However, the potential for infections to be composed of multiple, divergent lineages has generally not been considered in diagnostics, where it is still typical for a single clone to be taken as representative of an infection. Our results demonstrate that this will, at least for chronic infections, vastly underestimate the diversity within infections. In particular, it is known that conventional antimicrobial susceptibility tests are not good predictors for response to therapy [48], which may be explicable in part by the high genetic diversity harboured within an infection, including at loci encoding antimicrobial resistance, and which would be missed by sampling only one or a few clones. Our use of whole genome sequencing and analysis of

366 the infection-specific population-level data from individual patients to diagnose infection with
367 multiple distinct lineages therefore holds promise for diagnostic clinical microbiology and
368 lineage-targeted therapies [48].

References

1. Driscoll J, Brody S, Kollef M. The Epidemiology, Pathogenesis and Treatment of *Pseudomonas aeruginosa* Infections. *Drugs* 2007;67:351-368.

2. Murray TS, Egan M, Kazmierczak BI. *Pseudomonas aeruginosa* chronic colonization in cystic fibrosis patients. *Curr Opin Pediatr* 2007;19:83-88.

3. Govan JR, Deretic V. Microbial pathogenesis in cystic fibrosis: mucoid *Pseudomonas aeruginosa* and *Burkholderia cepacia*. *Microbiol Rev* 1996;60:539-74.

4. Hoffman LR, Richardson AR, Houston LS, Kulasekara HD, Martens-Habbena W, Klausen M, Burns JL, Stahl DA, Hassett DJ, Fang FC, Miller SI. Nutrient availability as a mechanism for selection of antibiotic tolerant *Pseudomonas aeruginosa* within the CF airway. *PLoS Pathog* 2010;6:e1000712.

5. Goodman AL, Kulasekara B, Rietsch A, Boyd D, Smith RS, Lory S. A Signaling Network Reciprocally Regulates Genes Associated with Acute Infection and Chronic Persistence in *Pseudomonas aeruginosa*. *Dev Cell* 2004;7:745-754.

6. Thomas SR, Ray A, Hodson ME, Pitt TL. Increased sputum amino acid concentrations and auxotrophy of *Pseudomonas aeruginosa* in severe cystic fibrosis lung disease. *Thorax* 2000;55:795-797.

7. Ciofu O, Mandsberg LF, Bjarnsholt T, Wassermann T, Høiby N. Genetic adaptation of *P. 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* 2010;156:1108-1119.

8. Ashish A, Shaw M, Winstanley C, Ledson MJ, Walshaw MJ. Increasing resistance of the Liverpool Epidemic Strain (LES) of *Pseudomonas aeruginosa* (Psa) to antibiotics in cystic fibrosis (CF)--a cause for concern?. *J Cyst Fibros* 2012;11:173-9.

392 9. Wilder C, Allada G, Schuster M. Instantaneous within-patient diversity of *Pseudomonas*
393 *aeruginosa* quorum-sensing populations from cystic fibrosis lung infections. *Infect Immun*
394 2009;77:5631-5639.
395 10. Rau MH, Hansen SK, Johansen HK, Thomsen LE, Workman CT, Nielsen KF, Jelsbak L,
396 Høiby N, Yang L, Molin S. Early adaptive developments of *Pseudomonas aeruginosa* after
397 the transition from life in the environment to persistent colonization in the airways of human
398 cystic fibrosis hosts. *Environ Microbiol* 2010;12:1643-1658.
399 11. Römling U, Fiedler B, Boßhammer J, Grothues D, Greipel J, Von Der Hardt H, Tümmler
400 B. Epidemiology of chronic *Pseudomonas aeruginosa* infections in cystic fibrosis. *J Infect*
401 *Dis* 1994;170:1616-1621.
402 12. Fothergill JL, Walshaw MJ, Winstanley C. Transmissible strains of *Pseudomonas*
403 *aeruginosa* in cystic fibrosis lung infections. *Eur Respir J* 2012;40:227-238.
404 13. Winstanley C, Langille MG, Fothergill JL, Kukavica-Ibrulj I, Paradis-Bleau C,
405 Sanschagrin F, Thomson NR, Winsor GL, Quail MA, Lennard N, Bignell A, Clarke L, Seeger
406 K, Saunders D, Harris D, Parkhill J, Hancock RE, Brinkman FS, Levesque RC. Newly
407 introduced genomic prophage islands are critical determinants of *in vivo* competitiveness in
408 the Liverpool Epidemic Strain of *Pseudomonas aeruginosa*. *Genome Res* 2009;19:12-23.
409 14. Martin K, Baddal B, Mustafa N, Perry C, Underwood A, Constantidou C, Loman N,
410 Kenna DT, Turton JF. Clusters of genetically similar isolates of *Pseudomonas aeruginosa*
411 from multiple hospitals in the UK. *J Med Microbiol* 2013;62:988-1000.
412 15. Aaron SD, Vandemheen KL, Ramotar K, Giesbrecht-Lewis T, Tullis E, Freitag A,
413 Paterson N, Jackson M, Lougheed D, Dowson C, Kumar V, Ferris W, Chan F, Doucette S,
414 Fergusson D. Infection with transmissible strains of *Pseudomonas aeruginosa* and clinical
415 outcomes in adults with cystic fibrosis. *JAMA* 2010;304:2145-2153.

16. Jeukens J, Boyle B, Kukavica-Ibrulj I, Ouellet MM, Aaron SD, Charette SJ, Fothergill JL, Tucker NP, Winstanley C, Levesque RC. Comparative Genomics of Isolates of a *Pseudomonas aeruginosa* Epidemic Strain Associated with Chronic Lung Infections of Cystic Fibrosis Patients. *PLoS ONE* 2014;9:e87611.

17. 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. Genetic adaptation by *Pseudomonas aeruginosa* to the airways of cystic fibrosis patients. *Proc Natl Acad Sci U S A* 2006;103:8487-8492.

18. Marvig RL, Johansen HK, Molin S, Jelsbak L. Genome Analysis of a Transmissible Lineage of *Pseudomonas aeruginosa* Reveals Pathoadaptive Mutations and Distinct Evolutionary Paths of Hypermutators. *PLoS Genet* 2013;9:e1003741.

19. Mowat E, Paterson S, Fothergill JL, Wright EA, Ledson MJ, Walshaw MJ, Brockhurst MA, Winstanley C. *Pseudomonas aeruginosa* population diversity and turnover in cystic fibrosis chronic infections. *Am J Respir Crit Care Med* 2011;183:1674-1679.

20. Ashish A, Paterson S, Mowat E, Fothergill JL, Walshaw MJ, Winstanley C. Extensive diversification is a common feature of *Pseudomonas aeruginosa* populations during respiratory infections in cystic fibrosis. *J Cyst Fibros* 2013;12:790-793.

21. Workentine ML, Sibley CD, Glezerson B, Purighalla S, Norgaard-Gron JC, Parkins MD, Rabin HR, Surette MG. Phenotypic Heterogeneity of *Pseudomonas aeruginosa* Populations in a Cystic Fibrosis Patient. *PLoS ONE* 2013;8:e60225.

22. Mayer-Hamblett N, Ramsey BW, Kulasekara H, Wolter DJ, Houston L, Pope C, Kulasekara B, Armbruster C, Burns JL, Retsch-Bogart G, Rosenfeld M, Gibson RL, Miller SI, Khan U, Hoffman LR. *Pseudomonas aeruginosa* Phenotypes Associated with Eradication Failure in Children with Cystic Fibrosis. *Clin Infect Dis* 2014;59:624-631.

440 23. Chung JC, Becq J, Fraser L, Schulz-Trieglaff O, Bond NJ, Foweraker J, Bruce KD, Smith
441 GP, Welch M. Genomic variation among contemporary *Pseudomonas aeruginosa* isolates
442 from chronically infected cystic fibrosis patients. *J Bacteriol* 2012;194:4857-66.
443 24. Markussen T, Marvig RL, Gómez-Lozano M, Aanæs K, Burleigh AE, Høiby N, Johansen
444 HK, Molin S, Jelsbak L. Environmental Heterogeneity Drives Within-Host Diversification
445 and Evolution of *Pseudomonas aeruginosa*. *mBio* 2014;5:.
446 25. Lieberman TD, Flett KB, Yelin I, Martin TR, McAdam AJ, Priebe GP, Kishony R. Genetic
447 variation of a bacterial pathogen within individuals with cystic fibrosis provides a record of
448 selective pressures. *Nat Genet* 2013;46:82-87.
449 26. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform.
450 *Bioinformatics* 2009;25:1754-1760.
451 27. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM,
452 Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G,
453 Alekseyev MA, Pevzner PA. SPAdes: A new genome assembly algorithm and its applications
454 to single-cell sequencing. *J Comput Biol* 2012;19:455-477.
455 28. Gascuel O. BIONJ: an improved version of the NJ algorithm based on a simple model of
456 sequence data. *Mol Biol Evol* 1997;14:685-695.
457 29. Paradis E, Claude J, Strimmer K. APE: Analyses of phylogenetics and evolution in R
458 language. *Bioinformatics* 2004;20:289-290.
459 30. Zwickl DJ. Genetic algorithm approaches for the phylogenetic analysis of large biological
460 sequence datasets under the maximum likelihood criterion. The University of Texas at Austin,
461 School of Biological Sciences; 2006.
462 31. Ronquist F, Huelsenbeck J. MrBayes 3: Bayesian phylogenetic inference under mixed
463 models. *Bioinformatics* 2003;19:1572-1574.

32. Sukumaran J, Holder MT. DendroPy: a Python library for phylogenetic computing. *Bioinformatics* 2010;26:1569-1571.

33. Darriba D, Taboada GL, Doallo R, Posada D. jModelTest 2: more models, new heuristics and parallel computing. *Nat Methods* 2012;9:772-772.

34. Shimodaira H. Testing regions with nonsmooth boundaries *via* multiscale bootstrap. *J Statist Plann Inference* 2008;138:1227-1241.

35. Sipos B, Massingham T, Jordan G, Goldman N. PhyloSim - Monte Carlo simulation of sequence evolution in the R statistical computing environment. *BMC Bioinformatics* 2011;12:104.

36. Oliver A, Sánchez JM, Blázquez J. Characterization of the GO system of *Pseudomonas aeruginosa*. *FEMS Microbiol Lett* 2002;217:31 - 35.

37. Oliver A, Baquero F, Blázquez J. The mismatch repair system (*mutS*, *mutL* and *uvrD* genes) in *Pseudomonas aeruginosa*: molecular characterization of naturally occurring mutants. *Mol Microbiol* 2002;43:1641-1650.

38. Oliver A, Cantón R, Campo P, Baquero F, Blázquez J. High Frequency of Hypermutable *Pseudomonas aeruginosa* in Cystic Fibrosis Lung Infection. *Science* 2000;288:1251-1253.

39. Feliziani S, Marvig RL, Luján AM, Moyano AJ, Di Rienzo JA, Krogh Johansen H, Molin S, Smania AM. Coexistence and Within-Host Evolution of Diversified Lineages of Hypermutable *Pseudomonas aeruginosa* in Long-term Cystic Fibrosis Infections. *PLoS Genet* 2014;10:e1004651.

40. Cramer N, Wiehlmann L, Tümmler B. Clonal epidemiology of *Pseudomonas aeruginosa* in cystic fibrosis. *Int J Med Microbiol* 2010;300:526-533.

41. Wiehlmann L, Wagner G, Cramer N, Siebert B, Gudowius P, Morales G, Köhler T, van Delden C, Weinel C, Slickers P, Tümmler B. Population structure of *Pseudomonas aeruginosa*. *Proc Natl Acad Sci U S A* 2007;104:8101-8106.

42. Kidd TJ, Ritchie SR, Ramsay KA, Grimwood K, Bell SC, Rainey PB. *Pseudomonas aeruginosa* Exhibits Frequent Recombination, but Only a Limited Association between Genotype and Ecological Setting. *PLoS ONE* 2012;7:e44199.

43. Rakhimova E, Munder A, Wiehlmann L, Bredenbruch F, Tümmler B. Fitness of isogenic colony morphology variants of *Pseudomonas aeruginosa* in murine airway infection. *PLoS ONE* 2008;3:e1685.

44. Damkiær S, Yang L, Molin S, Jelsbak L. Evolutionary remodeling of global regulatory networks during long-term bacterial adaptation to human hosts. *Proc Natl Acad Sci U S A* 2013;110:7766-7771.

45. Ashish A, Shaw M, Winstanley C, Humphreys L, Walshaw MJ. Halting the spread of epidemic *Pseudomonas aeruginosa* in an adult cystic fibrosis centre: a prospective cohort study. *JRSM Short Reports* 2013;4:1.

46. Al-Aloul M, Crawley J, Winstanley C, Hart CA, Ledson MJ, Walshaw MJ. Increased morbidity associated with chronic infection by an epidemic *Pseudomonas aeruginosa* strain in CF patients. *Thorax* 2004;59:334-336.

47. Reuter S, Ellington MJ, Cartwright EJ, Köser CU, Török ME, Gouliouris T, Harris SR, Brown NM, Holden MT, Quail M, Parkhill J, Smith GP, Bentley SD, Peacock SJ. Rapid bacterial whole-genome sequencing to enhance diagnostic and public health microbiology. *JAMA Internal Medicine* 2013;173:1397-1404.

48. Smith AL, Fiel SB, Mayer-Hamblett N, Ramsey B, Burns JL. Susceptibility testing of *Pseudomonas aeruginosa* isolates and clinical response to parenteral antibiotic administration: Lack of association in cystic fibrosis. *Chest* 2003;123:1495-1502.

Figure Legends

Figure 1 - *P. aeruginosa* Liverpool Epidemic Strain population structure in cystic fibrosis sputum sample CF03 consists of two divergent, recombining lineages

Rooted Neighbor-Joining (BIONJ) phylogenetic reconstruction of 40 isolate genome sequences obtained from a single cystic fibrosis sputum sample (CF03, collected 2009) calculated from a distance matrix of single nucleotide polymorphism (SNP) counts. SNPs among whole genome sequence short reads mapped to the *P. aeruginosa* LESB58 reference genome sequence (collected 1988) which also serves an out-group for rooting. The three support values for each edge are the percent split frequency among a non-parametric bootstrap replicate sample of BIONJ and maximum-likelihood phylogenies and among a Bayesian sample of phylogenies. Only edges with at least 80% support by all three measures are labelled with the respective split frequencies. The isolate sequences sharing homoplasies are indicated with circles in the right-most columns which correspond to circled variants in figure 2; the column numbers relate to row numbers in table 1.

Figure 2 - Chromosome positions and predicted effects of mutations within and between CF03 lineages A and B.

Genome map of mutations in 40 isolate genome sequences (outer lanes) obtained from a single sputum sample (CF03, collected 2009). Lanes are ordered by phylogenetic relationships in the neighbour-joining phylogeny from figure 1, adapted and plotted at the lower left, and rooted using the LESB58 reference genome (collected 1988) as out-group. The outer 13 lanes correspond to CF03 lineage A and the 27 lanes inwards to CF03 lineage B. The next, wider, lane corresponds to the LESB58 reference genome sequence with prophage (light brown) and genomic island (dark green) regions indicated at their positions relative to the origin of replication indicated by the outer scale. The inner-most lane is a plot of percent guanine-cytosine (GC) in 5 Kbp regions calculated every 2.5 Kbp and filled red above and

below 50% GC. For each isolate genome sequence, the mutations identified from sequenced reads aligned to the reference sequence were classified as single nucleotide polymorphisms, small insertions or small deletions. Those that occurred in protein coding regions are further classified by the predicted effects on transcription to mRNA and translation to protein sequences. Mutation classes are plotted on the genome map in different colours indicated in the key. Regions in which no reads mapped to the reference chromosome for an isolate sample are indicated by gaps in the corresponding lane. Homoplasies are circled and correspond to those described in figure 1 and table 1.

Figure 3 - Simulation of sequence evolution is useful for corroboration of phylogenetic hypotheses deduced from SNP frequency distributions.

Left: histogram of observed SNP frequencies among individual genome sequences of 40 isolates in two divergent lineages; centre, Neighbor-Joining (BIONJ) phylogenetic reconstruction of the 40 isolates adapted from Figure 1; right, histogram of SNP frequencies among sequences simulated along the phylogeny in the center. The phylogeny is rooted using the inferred most recent common ancestor of all isolates in this study as an outgroup. Edge highlight colours correspond to histogram peak colours (solid) according to the SNPs represented: frequency of peak (position on *x*-axis) corresponds to edge length while area of peak (abundance) corresponds to the number of tips descendant from the edge. The simulation accurately reproduces the SNP frequency distribution (right) observed in the sequence data (left) and is thus useful to corroborate phylogenetic hypotheses deduced from SNP frequencies among isolates (see figure 4). The red and blue hatched peaks in the histograms represent diversity within, not between, each of the two deepest lineages (A and B) and are not relevant to a hypothesis of two divergent coexisting lineages.

Figure 4 - *P. aeruginosa* Liverpool Epidemic Strain populations in six of eight other CF patient sputum samples consist of two divergent lineages.

A further eight patients provided a sputum sample from each of which genomic DNA of forty isolates was sequenced in equimolar pools. SNP analysis revealed six samples consisted of two divergent lineages shown in A-F (CF01, CF04, CF05, CF07, CF08, and CF09). Left: histograms of SNP frequency distributions observed among pooled genome sequences of 40 isolates. Center: hypotheses of root edge and lineage edges plotted as phylogenies deduced from the observed SNP frequency distributions. Only the root and lineage edges are relevant to the hypothesis of two divergent lineages. Right: histogram of SNP frequencies among sequences simulated along the each phylogeny, recapitulating observed SNP peaks corresponding to the root and deepest edges (indicated by purple, red and blue). For each sample two isolates were sequenced separately, the observed SNP frequencies for which are indicated among the 40 in each pool with green, turquoise and orange (left). Edges within clades of lineages are not relevant to the hypotheses. The SNP mutations in the root edges are shared and derived in all descendants *i.e.*, fixed in the population, thus appear at the maximum frequency of 40: both the root edge and corresponding peak in the simulated SNP frequency histograms (right) are coloured purple. The red and blue peaks in the simulation histograms correspond to the lineage edges arising from the patient most recent common ancestor at the deepest bifurcation and should be at frequencies which sum to the total isolates (40). In all of these samples the deepest pair of lineages were considered divergent because they had more mutations since their most recent common ancestor (MRCA; red and blue peaks) than their MRCA had to that of all samples in the study (the outgroup; peak at frequency 40, purple on the right). Except for CF04 (B), the pairs of isolates were representatives of each divergent lineage so that the highest frequency peak in the distribution of observed SNPs exclusive to

each isolate (orange or turquoise) corresponds to one of the lineage peaks in the distribution of simulated SNPs (red or blue).

Figure 5 - *P. aeruginosa* Liverpool Epidemic Strain populations in two of eight other CF patient sputum samples consist of a single lineage.

The sampling approach, analysis and plotting is as described for figure 4. For these two samples (CF06 and CF10) the deepest pair of lineages were not considered divergent because both had fewer mutations since their most recent common ancestor (MRCA; red and blue edges and peaks) than their MRCA had to that of all samples in the study (purple peak edges and peaks): the purple peak was larger than the red and blue peaks.

Figure 6 - Complex patterns of *P. aeruginosa* Liverpool Epidemic Strain transmission among chronically infected CF patients.

Neighbor-joining (BIONJ) phylogenetic reconstruction of 40 *P. aeruginosa* isolate genome sequences from a patient CF03 sputum sample and two *P. aeruginosa* isolate genome sequences from eight other patient sputum samples, all collected in 2009. The distance matrix consisted of raw counts of shared single nucleotide polymorphisms. Mutations are relative to the *P. aeruginosa* LESB58 genome, collected in 1988, which also serves as an out-group for rooting. Support for each edge is as described in figure 1. The 13 sequences representing CF03 lineage A and the 27 sequences representing CF03 lineage B each form a monophyletic clade and are represented as a blue and red triangle. The edges to CF05 isolate 2 and CF09 isolate 2 are not to scale because they represent many more mutations than other edges. Patient sample isolates CF03, CF05, CF07-9 are paraphyletic consistent with some patient infections being from diverse inocula and/or acquisitions of multiple lineages.

Tables

Table 1 - Incidence of homoplasies likely to be involved in homologous recombination among sputum sample CF03 isolates with predicted effects on transcription of open reading frames and translation to polypeptides where applicable.

Homoplasy ID*	Incidence	Description	Annotation	Probability of independent mutations†	Position (bp)‡
1	Lineage B: isolates 13 and 24; isolates 14, 20 and 31	Deletion of one nucleotide in PLES_09561 (*mpl*) causing nonsense mutations and loss of stop codon extending the ORF	UDP-N-acetylmuramate:L-alanyl-gamma-D-glutamyl-meso-diaminopimelate ligase (Mur ligase family)	3.18×10^{-10}	1037925
2	Lineage A: isolate 25; lineage B: isolate 35; isolate 31.	SNP in non-protein coding region upstream of ORF PLES_37671 (*acnA*) on the forward strand and PLES_37661 (*ygdE*) on reverse strand	None. Between operons including ORFs coding for a putative RNA 2'-O-ribose methyltransferase and aconitate hydratase	1.84×10^{-8} §	4165056
3	Lineage A: isolate 4; lineage B: isolate 8	Non-synonymous SNP in PLES_44121 (*lysC*)	Aspartate kinase	1.84×10^{-8}	4847728
4	Lineage B: isolates 13 and 24; isolates 14 and 20.	Deletion of two codons in PLES_56291 (*glpT*)	sn-glycerol-3-phosphate transporter	2.87×10^{-10}	6229046

* homoplasy IDs correspond to phylogenetic distribution in figure 1 and are referred to in text.

† probability that a pair of homoplasies are caused by independent mutations occurring in different lineages at the same position in the genome (parallel evolution), as opposed to a

single ancestral mutation transferred between lineages with chromosomal integration by homologous recombination. Calculation considers only fourfold degenerate sites in protein coding regions of LESB58 reference chromosome [13] as a conservative estimate of total sites not under strong selection that would tolerate mutations. Insertions, deletions and SNPs are included even though the nearly neutral sites considered only concern SNPs, consequently the calculation is more conservative with respect to favouring a single mutation (see Methods).

‡ positions are according to the LESB58 reference chromosome.

§ the most recent common ancestor of isolates 31 and 35 was considered to have the homoplasy to simplify calculation. Consequently, the estimate is more conservative.

Distribution of homoplasic SNPs:

Isolate numbers:

1 2 3 4

42 SNPs to LESB58 (1988)

55 SNPs

100, 100, 100

100, 98, 100

100, 100, 100

88, 92, 100

100, 100, 100

93, 90, 100

CF03 lineage A 13 isolates

i12
i05
i10
i03
i16
i19
i15
i18
i09
i02
i04
i25
i23

24 SNPs

100, 100, 100

100, 99, 100

100, 97

100

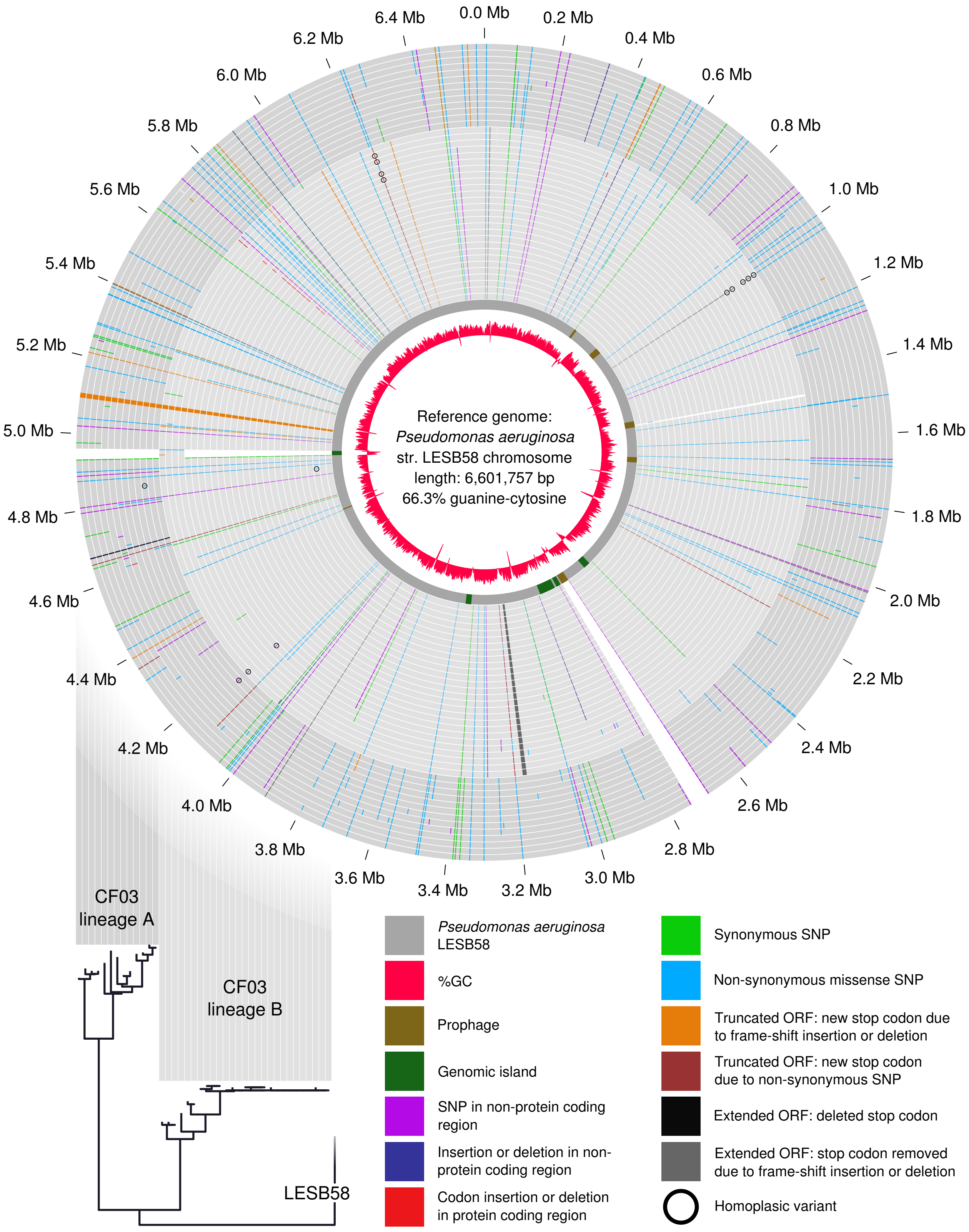
99, 98, 100

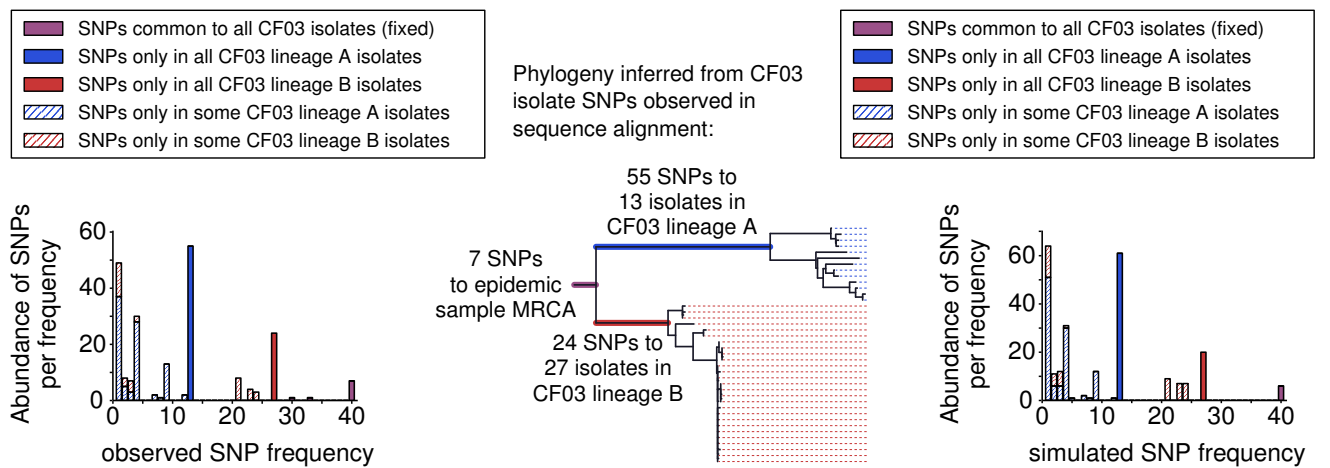
96, 93, 100

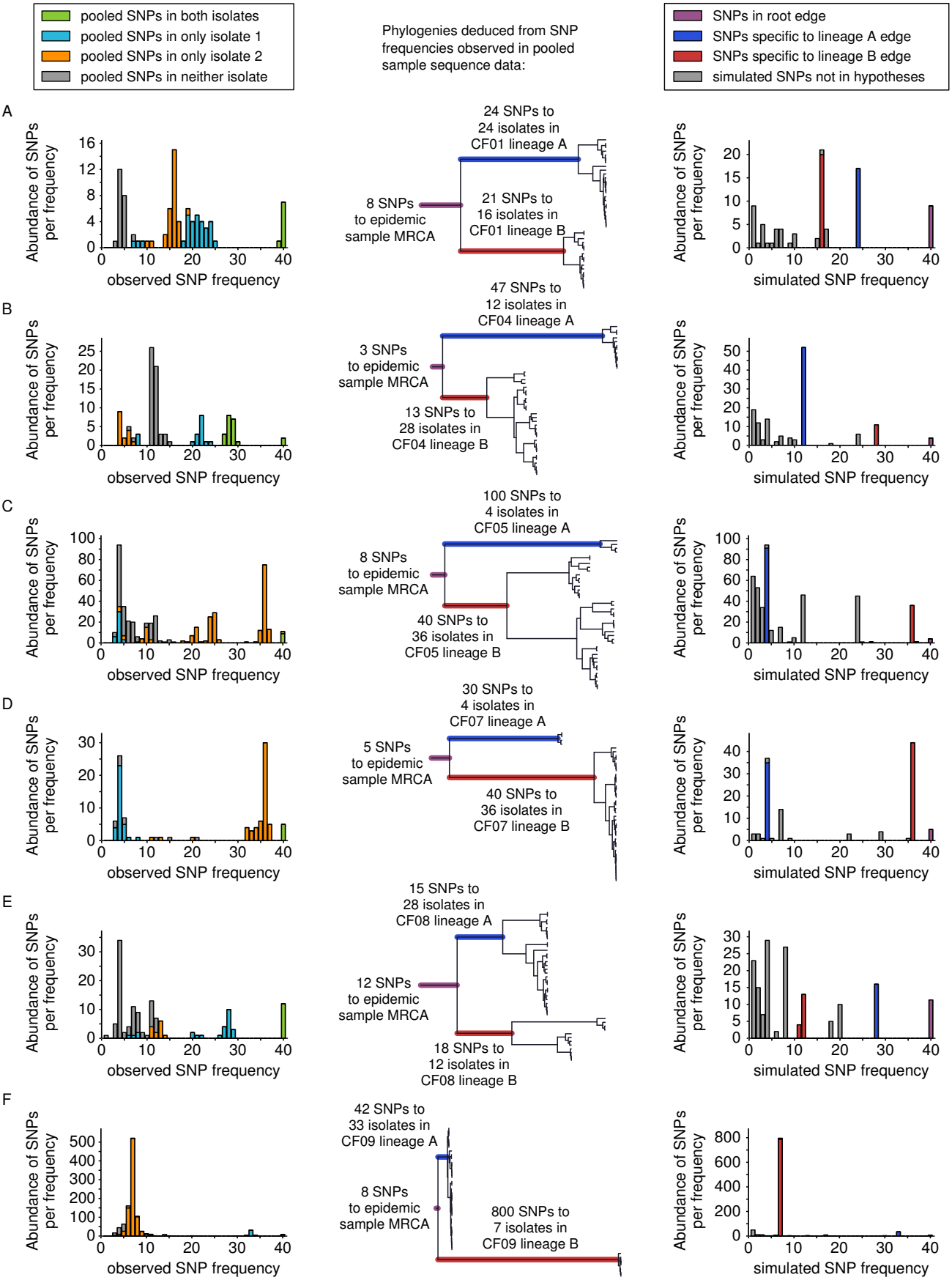
CF03 lineage B 27 isolates

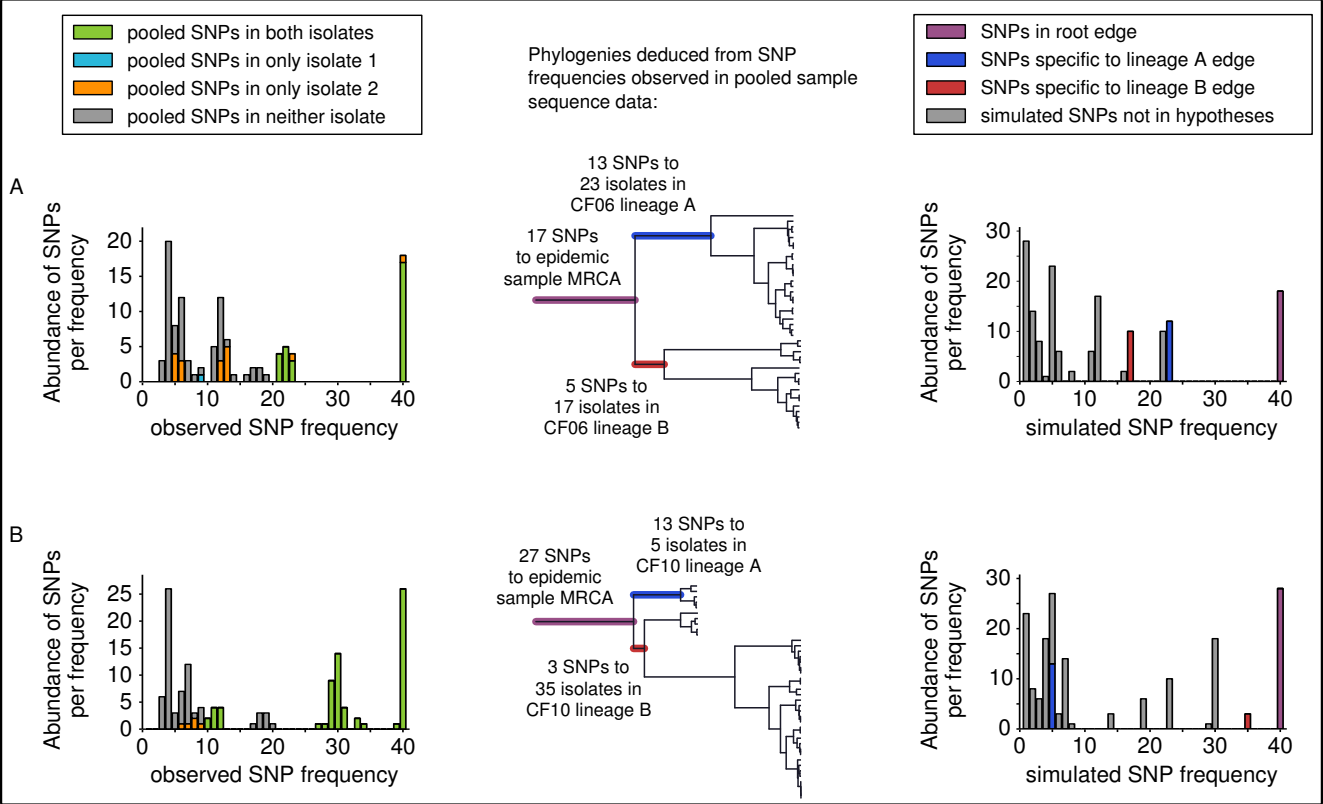
i31
i20
i14
i11
i24
i13
i35
i39
i28
i01
i40
i38
i37
i36
i30
i29
i17
i34
i33
i32
i27
i26
i22
i21
i08
i07
i06

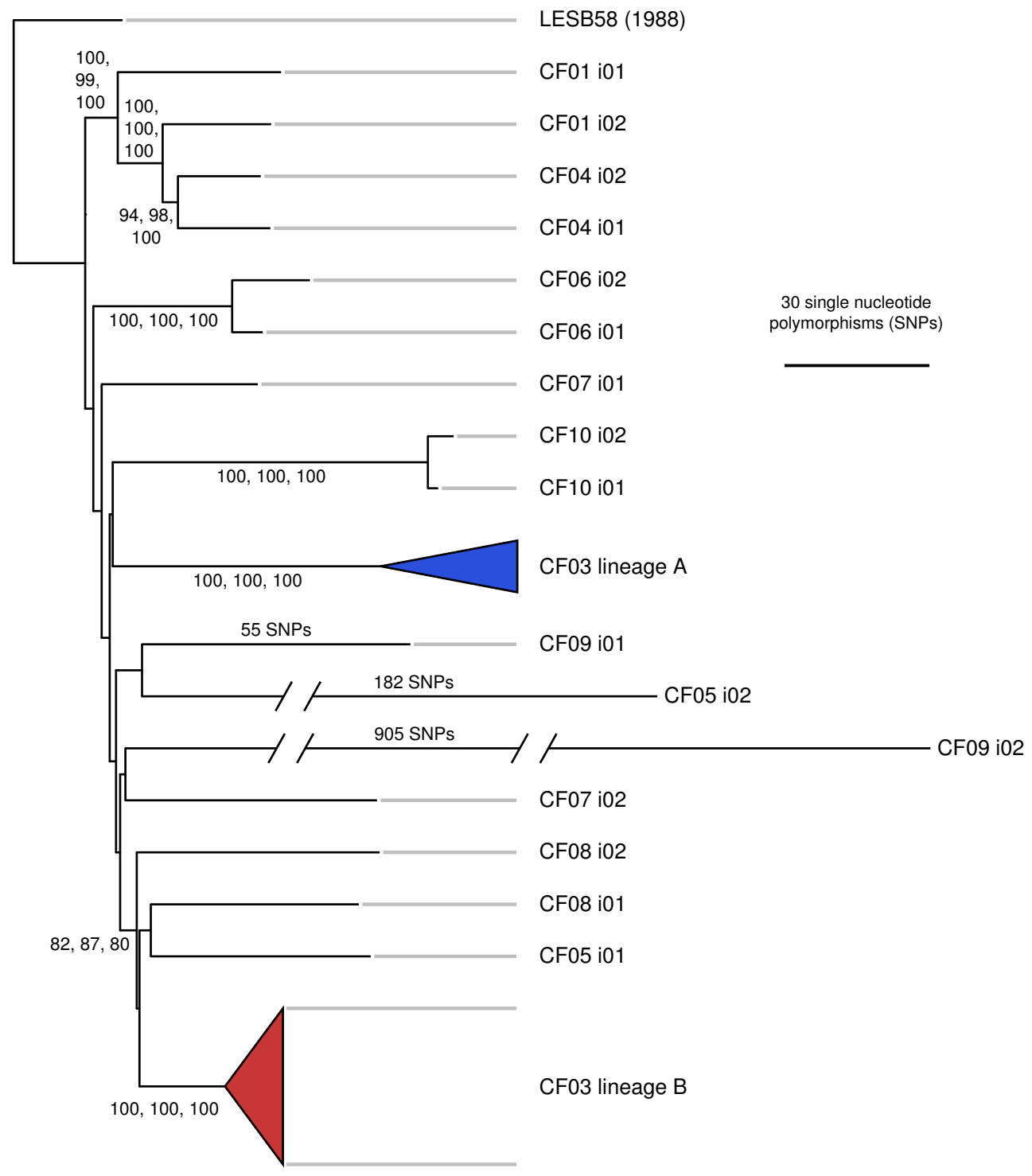
10 single nucleotide polymorphisms (SNPs)


0.0 Mb
0.2 Mb
0.4 Mb
0.6 Mb
0.8 Mb
1.0 Mb
1.2 Mb
1.4 Mb
1.6 Mb
1.8 Mb
2.0 Mb
2.2 Mb
2.4 Mb
2.6 Mb
2.8 Mb
3.0 Mb
3.2 Mb
3.4 Mb
3.6 Mb
3.8 Mb
4.0 Mb
4.2 Mb
4.4 Mb
4.6 Mb
4.8 Mb
5.0 Mb
5.2 Mb
5.4 Mb
5.6 Mb
5.8 Mb
6.0 Mb
6.2 Mb
6.4 Mb
Reference genome:
Pseudomonas aeruginosa
str. LESB58 chromosome
length: 6,601,757 bp
66.3% guanine-cytosine
CF03
lineage A
CF03
lineage B
LESB58
Pseudomonas aeruginosa
LESB58
%GC
Prophage
Genomic island
SNP in non-protein coding
region
Insertion or deletion in non-
protein coding region
Codon insertion or deletion
in protein coding region
Synonymous SNP
Non-synonymous missense SNP
Truncated ORF: new stop codon due
to frame-shift insertion or deletion
Truncated ORF: new stop codon
due to non-synonymous SNP
Extended ORF: deleted stop codon
Extended ORF: stop codon removed
due to frame-shift insertion or deletion
Homoplasic variant


SNPs common to all CF03 isolates (fixed)
SNPs only in all CF03 lineage A isolates
SNPs only in all CF03 lineage B isolates
SNPs only in some CF03 lineage A isolates
SNPs only in some CF03 lineage B isolates
Phylogeny inferred from CF03 isolate SNPs observed in sequence alignment:
55 SNPs to 13 isolates in CF03 lineage A
7 SNPs to epidemic sample MRCA
24 SNPs to 27 isolates in CF03 lineage B
Abundance of SNPs per frequency
observed SNP frequency
simulated SNP frequency
0
10
20
30
40
60


pooled SNPs in both isolates
pooled SNPs in only isolate 1
pooled SNPs in only isolate 2
pooled SNPs in neither isolate
Phylogenies deduced from SNP frequencies observed in pooled sample sequence data:
SNPs in root edge
SNPs specific to lineage A edge
SNPs specific to lineage B edge
simulated SNPs not in hypotheses
A
Abundance of SNPs per frequency
observed SNP frequency
24 SNPs to 24 isolates in CF01 lineage A
8 SNPs to epidemic sample MRCA
21 SNPs to 16 isolates in CF01 lineage B
simulated SNP frequency
B
47 SNPs to 12 isolates in CF04 lineage A
3 SNPs to epidemic sample MRCA
13 SNPs to 28 isolates in CF04 lineage B
C
100 SNPs to 4 isolates in CF05 lineage A
8 SNPs to epidemic sample MRCA
40 SNPs to 36 isolates in CF05 lineage B
D
30 SNPs to 4 isolates in CF07 lineage A
5 SNPs to epidemic sample MRCA
40 SNPs to 36 isolates in CF07 lineage B
E
15 SNPs to 28 isolates in CF08 lineage A
12 SNPs to epidemic sample MRCA
18 SNPs to 12 isolates in CF08 lineage B
F
42 SNPs to 33 isolates in CF09 lineage A
8 SNPs to epidemic sample MRCA
800 SNPs to 7 isolates in CF09 lineage B


pooled SNPs in both isolates
pooled SNPs in only isolate 1
pooled SNPs in only isolate 2
pooled SNPs in neither isolate
Phylogenies deduced from SNP frequencies observed in pooled sample sequence data:
SNPs in root edge
SNPs specific to lineage A edge
SNPs specific to lineage B edge
simulated SNPs not in hypotheses
A
Abundance of SNPs per frequency
observed SNP frequency
13 SNPs to 23 isolates in CF06 lineage A
17 SNPs to epidemic sample MRCA
5 SNPs to 17 isolates in CF06 lineage B
simulated SNP frequency
B
27 SNPs to epidemic sample MRCA
13 SNPs to 5 isolates in CF10 lineage A
3 SNPs to 35 isolates in CF10 lineage B


LESB58 (1988)
100, 99, 100
CF01 i01
100, 100, 100
CF01 i02
CF04 i02
94, 98, 100
CF04 i01
CF06 i02
100, 100, 100
CF06 i01
30 single nucleotide polymorphisms (SNPs)
CF07 i01
CF10 i02
100, 100, 100
CF10 i01
CF03 lineage A
100, 100, 100
55 SNPs
CF09 i01
182 SNPs
CF05 i02
905 SNPs
CF09 i02
CF07 i02
CF08 i02
CF08 i01
82, 87, 80
CF05 i01
CF03 lineage B
100, 100, 100

Online Data Supplement for:

Divergent, coexisting, *Pseudomonas aeruginosa* lineages in chronic cystic fibrosis lung infections.

David Williams, Benjamin Evans, Sam Haldenby, Martin J. Walshaw, Michael A. Brockhurst, Craig Winstanley, Steve Paterson

Methods

Genomic DNA preparation and sequencing

DNA was extracted from isolates grown in Luria Broth overnight at 37°C using the Wizard® Genomic DNA Purification Kit (Promega). The purity of the extracted bacterial DNA was assessed using the Nanodrop ND-1000 and the quantity of bacterial DNA was measured using the Qubit® 2.0 Fluorometer following the manufacturer's protocol (Life Technologies). For pools of 40 isolate genome samples, equimolar quantities were of genomic DNA from each isolate combined for sequencing together. An Illumina HiSeq 2000 was used to generate 100 bp paired reads from the ends of 500 bp fragments from all samples except isolates CF04 i02, CF05 i01 and i02, CF07 i01 and i02 and CF08 i01 for which 250 bp paired reads were generated from the ends of 500 bp fragments using an Illumina MiSeq. Sequenced read data in fastq files were trimmed for the presence of Illumina adapter sequences using Cutadapt version 1.2.1 [2]. The option -O 3 was used, so the 3' end of any reads which match the adaptersequence for 3 bp or more are trimmed. The reads are further trimmed using Sickle version 1.2 [3] with a minimum window quality score of 20. Reads shorter than 10 bp after trimming were removed. If only one of a read pair passed this filter, it was retained. Accession numbers for the read data generated from the pooled samples are ERS487747-57, for the isolates from sample CF03 are ERS487769-808 and for the pairs of isolates from the other samples are ERS508076-91.

Variant calling and *de novo* genome assemblies

The Genome Analysis Toolkit (GATK) [4] Indel Realigner module [5] was used to realign raw reads around indels. Subsequently, duplicate reads were identified and removed with Picard [6]. Single nucleotide polymorphism, insertion and deletion discovery was performed with GATK's Unified Genotyper module [5] with sample ploidy $n = 40$, and parameters set accordingly to permit detection of low frequency variants in each pooled sample (≤ 2.5 %). GATK was further utilised to filter sequence variants using standard conservative filtering parameters to provide high-quality variant calls [7]. False positive SNP variants at structural mismatches between sample genome sequence and the reference genome sequence *e.g.*, an absent prophage sequences, were mitigated by omitting variants at positions where aligned read coverage depth was less than 25% of the genome-wide median. Contiguous regions greater than 10 bp in the reference genome along which no sequenced reads were aligned were considered missing in the genome from which the reads were derived. For individually sequenced isolates, paired-end reads were assembled *de novo* using SPAdes Genome Assembler version 3.0 [8]. The default selection of *k*-mer lengths for multi-cell samples was used and BayesHammer read data error correction incorporated [9]. Read depth coverage used in assembly was a minimum of 131x and limited to 200x based on the 6 601 757 bp LES B58 reference genome.

Prediction of genetic variant effects

Single nucleotide polymorphisms (SNPs), insertions and deletions reported by GATK within annotated protein coding regions of the reference genome (LES B58) were classified on the basis of potential phenotypic effects. The effects of genetic variants were considered together if they occurred in the same open reading frame (ORF), in order of expected translation from the 5' end to the 3' end of the ORF sequence, and with respect to alterations to codons. If the predicted effect is an extension or truncation of an ORF because the transcription termination (stop) codon is altered, the effect is reported for the whole ORF else it is reported for the range of positions within the ORF that are affected. A position and codon change affected by a SNP relative to the reference genome, when reported in

isolation, may be different when considered in the context of an insertion or deletion towards the 5' end of the ORF sequence. In such cases, the insertion or deletion may alter codons towards the 3' end from those in the reference sequence for the reasons described below. These combinatorial effects are why accurate effect prediction requires genetic variants to be considered together if they occurred in the same open reading frame (ORF) and in the order of ORF translation.

For a deletion, after removal of the appropriate characters in the nucleotide sequence, codons are re-constructed by partitioning nucleotides into consecutive groups of three from the 5' end of the sequence. This causes codons completely spanned by the deletion to be removed and the codons in which the deletion starts and ends (which may be the same codon if the deleted region is ≤ 2 bp and ends on or before the last codon position) to be altered. If the length of the deletion is not a multiple of three, all of the downstream codons will change. If one of the new codons is a stop codon, the open reading frame is classified as a "truncated ORF: new stop codon due to frame-shift deletion". If none of the new codons is a stop codon and the original stop codon is no longer encoded, the open reading frame is classified as an "extended ORF: stop codon removed due to frame-shift deletion". Differences between the original codons and the reconstructed codons may cause a change in the protein sequence which was predicted by conceptual translation using the genetic code for Bacteria (NCBI translation table 11 [10] last accessed 19/07/2014). For an insertion, after adding the new characters at the appropriate position in the nucleotide sequence, codons are re-constructed by partitioning nucleotides into groups of three from the 5' end of the sequence. Similar to prediction of effects caused by deletions, loss of the original stop codon or acquisition of a new because of an insertion may cause classification as "truncated ORF: new stop codon due to frame-shift insertion" or "extended ORF: stop codon removed due to frame-shift insertion". For SNPs, conceptual translation of the codon in which the character change occurred predicts "non-synonymous missense SNP" if the amino acid changes to another and "synonymous SNP" if it does not. If a SNP changes a codon from an amino acid to a stop,

the ORF is classified as “truncated ORF: non-synonymous nonsense SNP”. If a SNP changes a codon from a stop to an amino acid, the ORF is classified as “extended ORF: non-synonymous missense SNP in stop codon”. If a stop codon is within a region of chromosome deemed absent because of a lack of aligned reads *i.e.*, a deletion that is too large to be called by the read alignment-based variant calling strategy described above, “extended ORF: deleted stop codon” is reported.

Phylogenetic reconstruction and hypothesis testing

The distance-matrix for each phylogenetic reconstruction was calculated as the pairwise sum of shared SNPs and was preferred over available models of nucleotide substitution because they did not account for the heterogeneity of substitutions across the phylogeny causing inaccurate edge lengths. Statistical support for each edge in phylogenies was obtained from split frequencies among a non-parametric bootstrap replicate sample of maximum-likelihood phylogenies inferred using Garli version 2.01 [11] and a Bayesian sample of phylogenies inferred using MrBayes version 3.2.2 [12]. and counted using methods implemented in the DendroPy library for phylogenetic computing version 3.12.0 [13] in Python version 2.7.8. The HKY85 nucleotide substitution model was selected for use in Garli with JModelTest2 version 2.1.5 [14] according to all available selection criteria. Other Garli options were set to the defaults. Flat prior distributions were set in MrBayes with nucleotide substitution models set to “mixed” running 10 000 000 generations of two independent runs at a sampling frequency of 500. After discarding the initial 1000 samples in each run, the estimated sample size for all parameters was greater than 1300 as calculated by Tracer 1.6 [15]. For testing phylogenetic hypotheses in which both lineages from a patient sample formed a monophyletic group, maximum likelihood phylogenies were inferred using Garli version 2.0 as described above except under a topological constraint for the desired patient-specific clade recording site-wise log-likelihoods. The scale-boot R package version 0.3-3 [16] was used to calculate third order p-values of the Approximately Unbiased (AU) test on a multiscale bootstrap for the alternative and maximum likelihood topologies against the aligned sequence data.

Homoplasy identification and inference of homologous recombination

Genetic variants among individually sequenced genomes were initially identified by analysis of reads aligned to LESB58 genome sequence alignment. Those variants that were homoplastic on the genome phylogeny were double-checked by identification of the same variant in *de novo* assembled sequence contigs (minimum n50 of *de novo* assemblies was 380 454 bp). A 1 000 bp region around the putative variant in the reference genome sequence, including the variant character itself, was used as a query for the BLAST-like Alignment Tool [17] to search the contigs derived from the respective sample. Putative homoplasic variants were only reported if they were found in a single alignment among the contigs with at least 99% sequence identity and the variant position was aligned to the same character. While some of these homoplasic mutations had fewer reads aligned to the LESB58 reference sequence than permitted by the global filter applied above (less than 25% of the genome-wide median) for some isolate read sets, they were corroborated by *de novo* read assemblies and had GATK 'Genotype Qualities' of at least 90% and 99% in most cases for the respective variant calls. Homologous recombination was inferred as a cause for a group of homoplasic variants where multiple independent acquisitions of the variant were preferred over loss of an ancestral variant in a maximum parsimony model (PARS) [18] implemented in Python 2.7. Acquisitions had 1.5 penalty weighting to favour losses of a variant in the maximum parsimony scoring scheme to minimise false positive detection of a recombination events due to failures to detect mutations (false negatives). It should be noted that an apparent loss of a mutation could be because of a failure to detect the sequence variation or because an ancestor of the isolate in question received a region of chromosome with the ancestral state by homologous recombination. Thus the frequency of reported homoplasies, omitting apparent losses, is likely to be a conservative estimate of recombination rates. More homoplasies were evident if considering variants detected by either *de novo* assemblies or aligned reads, or caused by rare losses of sequence variants. The probability of each observed homoplasy (table 1) being caused by two

independent mutations at the same site was calculated as:

$$(1 - ((\text{total nearly neutral sites} - 1) / \text{total nearly neutral sites})^{\text{total mutations between genomes}})^2$$

where nearly neutral sites were considered to be the 1,121,045 fourfold degenerate sites in protein coding regions of the LESB58 reference chromosome as a conservative estimate of chromosome positions that would tolerate SNPs. Mutations between genomes included indels and SNPs.

Inference of divergent lineages from single nucleotide polymorphism frequencies

The following logic forms the basis of these deductions and determines the relationship between the SNP frequency distribution and the deepest part of a hypothetical phylogeny. First, SNP mutations common to all 40 isolates form the root edge in a phylogeny (coloured purple in figures 3, 4 and 5) and will occur at a frequency of 40, therefore contributing to the highest-frequency peak in a SNP frequency histogram (the 'root' peak). The abundance of SNPs at this frequency will determine the length of the root edge in the population phylogeny and the area of the root peak, at the maximum frequency, in a SNP frequency histogram. Second, the deepest divergence in the phylogeny gives rise to the two primary edges, coloured red and blue in figure 3, corresponding to the primary lineages. Each primary lineage must consist of mutually exclusive SNPs contributing to peaks whose frequencies must sum to 40 (the 'primary' peaks). The primary lineages descending from the root lineage must sum to 40 because any one isolate must be in one or other of the primary lineages with 40 isolates per sample. The abundance of mutations specific to, and universal within, each primary lineage will determine the length of the edges leading to the clade corresponding to each lineage and the height of the pair of peaks in a frequency histogram whose frequency sums to 40. Where the area of the root peak in a SNP frequency histogram is smaller than the area of each of the primary peaks, two divergent lineages would be hypothesised (figure 4); the length of the root edge is smaller than the edge lengths descending from the deepest bifurcation. Where the area of the highest-frequency peak in a SNP

frequency histogram is larger than the area of each of the 'primary' peaks, a single lineage would be hypothesised (figure 5). Other peaks, all of which are at lower frequencies than the higher-frequency primary peak, are relevant to the internal structure of each primary lineage and were not considered in population structures described here. For all samples, SNPs were relative to the MRCA of all isolates in this sample, not the more distantly related LESB58 reference strain. Thus the divergence of lineages for each sample is in the context of the total *P. aeruginosa* diversity sampled across all nine CF patients. We tested the accuracy of peak identification by surveying the frequencies of SNPs from an isolate genome sequence within each population. The collection SNPs from a particular a genome must constitute all of the root peak, and because they are linked, all of one of the primary peaks and be excluded from all of the other primary peak.

Simulations of sequence evolution

Monte Carlo simulation of nucleotide sequence evolution over the hypothesised phylogenies was achieved using the PhyloSim R package version 2.1.1 [19]. The HKY85 nucleotide substitution model was used with base frequencies A = 0.18, C = 0.27, G = 0.33, T = 0.22 and a transition to transversion ratio of 1.35:1, as reported by JModelTest 2 for the CF03 nucleotide alignment. The genomes of the sampled organisms assembled to more than 6 Mbp and the total amount of sequence evolution separating the samples was sufficiently small for mutations at the same position to be very rare. Sequence simulations on millions of sites had unfeasable memory requirements so simulations were performed on shorter sequences, but mutations occurring at the same sites during a simulation would be unrealistic so had to be minimised. Simulations were therefore run on sequences 20 times longer than the observed number of variable sites (the sample multiple sequence alignment length) except for sample CF09 in which the observed data contained many more mutations and a sequence 100 times longer was required.

Discussion

The impact of natural selection on phylogenetic reconstruction

P. aeruginosa lineages may experience periods of strong purifying or diversifying selection during chronic infection of the CF airway and transmissions between patients. However, in the case of decades-old chronic infections where near-complete genome sequences are available, these evolutionary dynamics are not expected to affect the accuracy of reconstruction. As genetic divergence increases and positions in a sequence alignment become saturated with substitutions, modelling the substitution process to incorporate selection dynamics becomes complex to point that accuracy may suffer. In the current context, the evolutionary distances in question are small (less than 100 nucleotide changes between most sequences) and the amount of sequence data available is large: whole 6.5 Mb genomes. This means we do not have to model substitution processes or worry about the dynamics of selection with respect to reconstruction accuracy because with only a few hundred SNPs across a 6.5Mbp genome, multiple changes at the same site are expected to be very rare and therefore negligible to phylogenetic inferences. We need only count the number of nucleotide changes that have occurred across the genome. Including SNPs in regions of the genome which may have been subject to strong diversifying or purifying selection in phylogenetic reconstructions, which probably includes many SNPs predicted to cause changes in protein sequences, is therefore not expected affect reconstruction accuracy. Figure S2 shows a phylogenetic reconstruction with the same sampling as figure 6 but including only SNPs that are not predicted to cause changes to protein sequences. The phylogeny in figure S2 is congruent with that in figure 6.

Low bootstrap support in phylogenies and use of the Approximately Unbiased test

With regards to interpreting a phylogeny with branch support, it is important to note that a bipartition defined by an edge or branch in a phylogeny is a very specific hypothesis which can be falsified by a

difference of just one operational taxonomic unit (OTU) placed on the other side of the bipartition. When traversing a topology from one clade of interest (e.g. CF03 lineage A in figure 6) to another (e.g. CF03 lineage B) over bipartitions that include groups that are *not* of interest, it may be that no well-supported edges are encountered, but the reason for the poor edge support may not be related to the groups of interest. Poor edge support may be because other nearby clades that are not of interest are not confidently placed in one or more bipartitions encountered i.e., among the bootstrap replicate phylogenetic reconstructions their placements vary relative to other groups separated at that bipartition. Such clades are commonly referred to as 'rogue taxa' in the literature. Many short edges close to each other can exacerbate this phenomenon. As more isolate sequences are added to a tree, the density of branches and the number of possible topologies both increase at a rate much higher than the increase in information added to the sequence alignment. This has the effect of lowering support for many individual branches, especially short branches, and is a limitation of relying solely on bootstrap support of branches. For this reason we employed the bootstrap-based Approximately Unbiased (AU) test of Shimodaira to assess whether the alternative hypotheses of patient sample monophyly is supported by the data *i.e.*, single infections with subsequent diversification within the host, without further transmissions to or from the host.

Tables

Table S1. Mutations specific and exclusive to the 13 isolates in lineage A of sample CF03 that form a clade in the phylogeny in figure 1.

LESB58

Locus ID	Annotation	Mutation	Predicted effect
PLES_00731	putative permease	SNP*	Synonymous codon change
PLES_04141	putative transcriptional regulator	Out-of-frame insertion, multiple of 3	Non-synonymous codon change and codon gain
PLES_04241	Resistance-Nodulation-Cell Division (RND) multidrug efflux transporter MexB	Insertion, frame-shift	ORF truncated: new stop codon prior to original
PLES_04321	putative outer membrane ferric siderophore receptor	SNP	Synonymous codon change
PLES_05031	putative acyl-CoA dehydrogenase	SNP	Non-synonymous codon change
PLES_06611	30S ribosomal protein S7	SNP	Non-synonymous codon change
PLES_06901	50S ribosomal protein L17	SNP	Non-synonymous codon change
PLES_08611	beta-lactamase	SNP	Non-synonymous codon change
PLES_09191	transcriptional regulator NrdR	SNP	Non-synonymous codon change
PLES_09411	hypothetical protein	SNP	Non-synonymous codon change
PLES_09561	UDP-N-acetylmuramate:L-alanyl-gamma-D-glutamyl-meso-diaminopimelate ligase	SNP	Non-synonymous codon change
PLES_11511	queuine tRNA-ribosyltransferase	SNP	Non-synonymous codon change
PLES_11641	chaperone protein HscA	SNP	Non-synonymous codon change
PLES_15531	transposase A	SNP	Non-synonymous codon change
PLES_18021	putative transporter	SNP	Synonymous codon change
PLES_19001	DNA gyrase subunit A	SNP	Non-synonymous codon change
PLES_19001	DNA gyrase subunit A	SNP	Non-synonymous codon change
PLES_22161	putative transcriptional regulator	SNP	Non-synonymous codon change
PLES_27541	putative lysophospholipas	SNP	Synonymous codon change
PLES_27681	putative Resistance-	SNP	Synonymous codon change

	Nodulation-Cell Division		
	(RND) efflux transporter toluate 1,2-dioxygenase subunit		
PLES_27771	alpha putative AGCS	SNP	Non-synonymous codon change
	sodium/alanine/glycine		
PLES_30521	symporter	SNP	Synonymous codon change
PLES_30681	putative glycosyl transferase	SNP	Synonymous codon change
PLES_31471	hypothetical protein putative transcriptional	SNP	Non-synonymous codon change
PLES_32251	regulator methylcrotonyl-CoA	SNP	Non-synonymous codon change
PLES_33111	carboxylase subunit alpha	SNP	Non-synonymous codon change
PLES_33731	ribokinase membrane-bound lytic murein	SNP	Non-synonymous codon change ORF extended: new stop codon
PLES_35151	transglycosylase D	Insertion, frame-shift	after original
PLES_36151	ExsD transcriptional regulator protein	SNP	Non-synonymous codon change
PLES_36231	PcrR	SNP	Non-synonymous codon change
PLES_36561	serine-threonine kinase Stk1	SNP	Synonymous codon change ORF truncated: new stop codon
PLES_39841	transcriptional regulator LasR putative mechanosensitive ion	Deletion, frame-shift	prior to original
PLES_40051	channel family protein alkaline protease secretion	SNP	Non-synonymous codon change
PLES_40651	protein AprE	SNP	Synonymous codon change
PLES_46901	putative deacetylase	SNP	Non-synonymous codon change
PLES_47691	hypothetical protein putative mechanosensitive ion	SNP	Synonymous codon change
PLES_47731	channel family protein putative mechanosensitive ion	SNP	Non-synonymous codon change
PLES_47731	channel family protein	SNP	Synonymous codon change
PLES_47971	penicillin-binding protein 3	SNP	Non-synonymous codon change ORF truncated: new stop codon
PLES_49041	hypothetical protein	Insertion, frame-shift	prior to original

PLES_49861	motility regulator	SNP	Non-synonymous codon change
PLES_51171	glucose-6-phosphate isomerase	SNP	Non-synonymous codon change
PLES_52931	putative short-chain dehydrogenase	SNP	Synonymous codon change ORF truncated: new stop codon
PLES_53041	nicotinamidase	Deletion, frame-shift	prior to original
PLES_53861	transport protein MsbA	SNP	Non-synonymous codon change
PLES_56241	ABC transporter permease	SNP	Non-synonymous codon change ORF truncated: new stop codon
PLES_58431	glycosyltransferase WbpY	Deletion, frame-shift	prior to original
PLES_59111	hypothetical protein	SNP	Non-synonymous codon change ORF truncated: new stop codon
PLES_59241	hypothetical protein putative major facilitator	Deletion, frame-shift	prior to original
PLES_59441	superfamily transporter	SNP	Non-synonymous codon change

* Single nucleotide polymorphism

Table S2. Mutations specific and exclusive to the 27 isolates in lineage B of sample CF03 that form a clade in the phylogeny in figure 1.

LESB58

Locus ID	Annotation	Mutation	Predicted effect
	potassium transporter		ORF extended: new stop codon
PLES_00151	peripheral membrane protein	Deletion, frame-shift	after original
PLES_04651	hypothetical protein	SNP*	Non-synonymous codon change
PLES_08851	4-hydroxyphenylacetate 3-monooxygenase large chain	SNP	Non-synonymous codon change
PLES_11491	hypothetical protein	SNP	Non-synonymous codon change
PLES_14881	alginate-c5-mannuronan-epimerase AlgG	SNP	Non-synonymous codon change
PLES_16081	malate:quinone oxidoreductase	SNP	Non-synonymous codon change
PLES_16401	putative transcriptional regulator	SNP	Synonymous codon change

PLES_20021	protein PelG	SNP	Synonymous codon change
	putative D-alanyl-D-alanine	Out-of-frame deletion,	Non-synonymous codon change
PLES_20131	carboxypeptidase	multiple of 3	and codon loss ORF extended: new stop codon
PLES_28971	peptide synthase putative non-ribosomal peptide	Deletion, frame-shift	after original
PLES_29991	synthetase putative alcohol dehydrogenase	SNP	Non-synonymous codon change
PLES_30291	(Zn-dependent)	SNP	Non-synonymous codon change
		Out-of-frame deletion,	Non-synonymous codon change
PLES_30711	putative glycosyl transferase putative transcriptional	multiple of 3	and codon loss
PLES_33451	regulator membrane-bound lytic murein	SNP	Non-synonymous codon change ORF extended: new stop codon
PLES_35151	transglycosylase D putative sensor/response	Deletion, frame-shift	after original
PLES_37161	regulator hybrid	SNP	Non-synonymous codon change
PLES_41151	putative amino acid permease putative major facilitator	SNP	Non-synonymous codon change
PLES_42131	superfamily transporter	SNP	Synonymous codon change ORF truncated: new stop codon
PLES_45801	anti-sigma factor MucA phospho-N-acetylmuramoyl-	Deletion, frame-shift	prior to original
PLES_47941	pentapeptide- transferase	SNP	Non-synonymous codon change
PLES_47971	penicillin-binding protein 3 S-adenosyl-methyltransferase	SNP	Non-synonymous codon change ORF truncated: new stop codon
PLES_47991	MraW formate dehydrogenase-O,	Deletion, frame-shift	prior to original
PLES_51961	major subunit formate dehydrogenase-O,	SNP	Non-synonymous codon change
PLES_51961	major subunit	SNP	Non-synonymous codon change
PLES_52461	ABC transporter permease	SNP	Non-synonymous codon change
PLES_53471	hypothetical protein	SNP	Non-synonymous codon change
PLES_54671	glucosyltransferase MdoH	Deletion, frame-shift	ORF truncated: new stop codon

PLES_55931	two-component sensor EnvZ	SNP	prior to original Non-synonymous codon change ORF truncated: new stop codon
PLES_56861	putative choline transporter	Deletion, frame-shift	prior to original

* Single nucleotide polymorphism

Figures

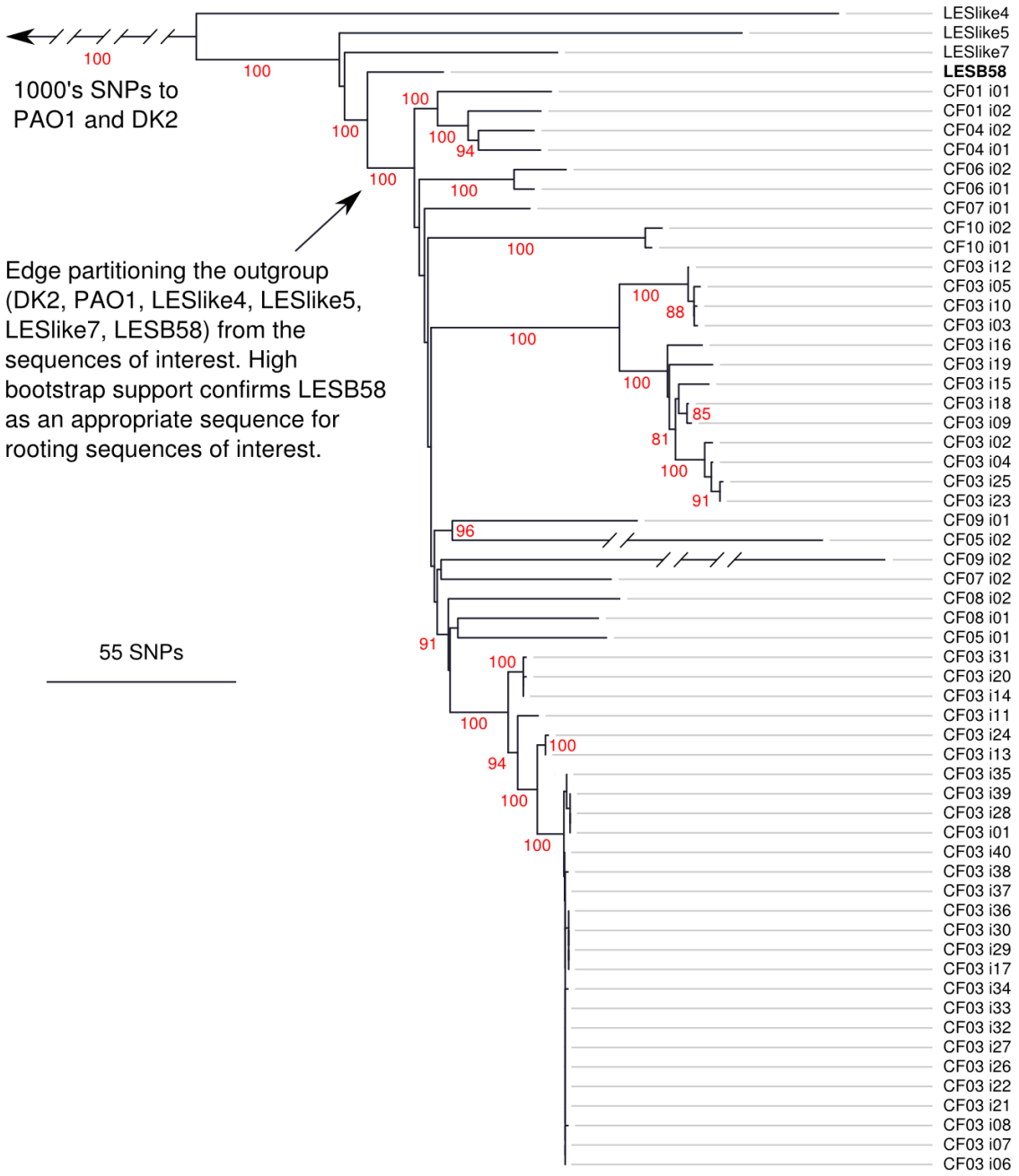


Figure S1: LESB58 is an appropriate outgroup for rooting the isolates sampled from the nine patients. Neighbor-joining (BIONJ) phylogenetic reconstruction using all SNPs of 40 *P. aeruginosa* isolate genome sequences from a patient CF03 sputum sample and two *P. aeruginosa* isolate genome sequences from eight other patient sputum samples, all collected in 2009 and, as potential outgroups, *P. aeruginosa* strains PAO1 (wound isolate, Australia, 1954 with a few decades *in vitro*), DK2 (CF airway

isolate, Denmark, 2007) and LESlike 4 (CF airway isolate, Canada 2005), LESlike 5 (CF airway isolate, Canada 2007), LESlike 7 (CF airway isolate, Canada 2006) and LESB58 (CF airway isolate, Liverpool 1988). The edge with 100/100 bootstrap support, indicated with the arrow and partitioning the 2009 sputum sample isolates from the potential outgroup samples confirms LESB58 as a suitable outgroup.

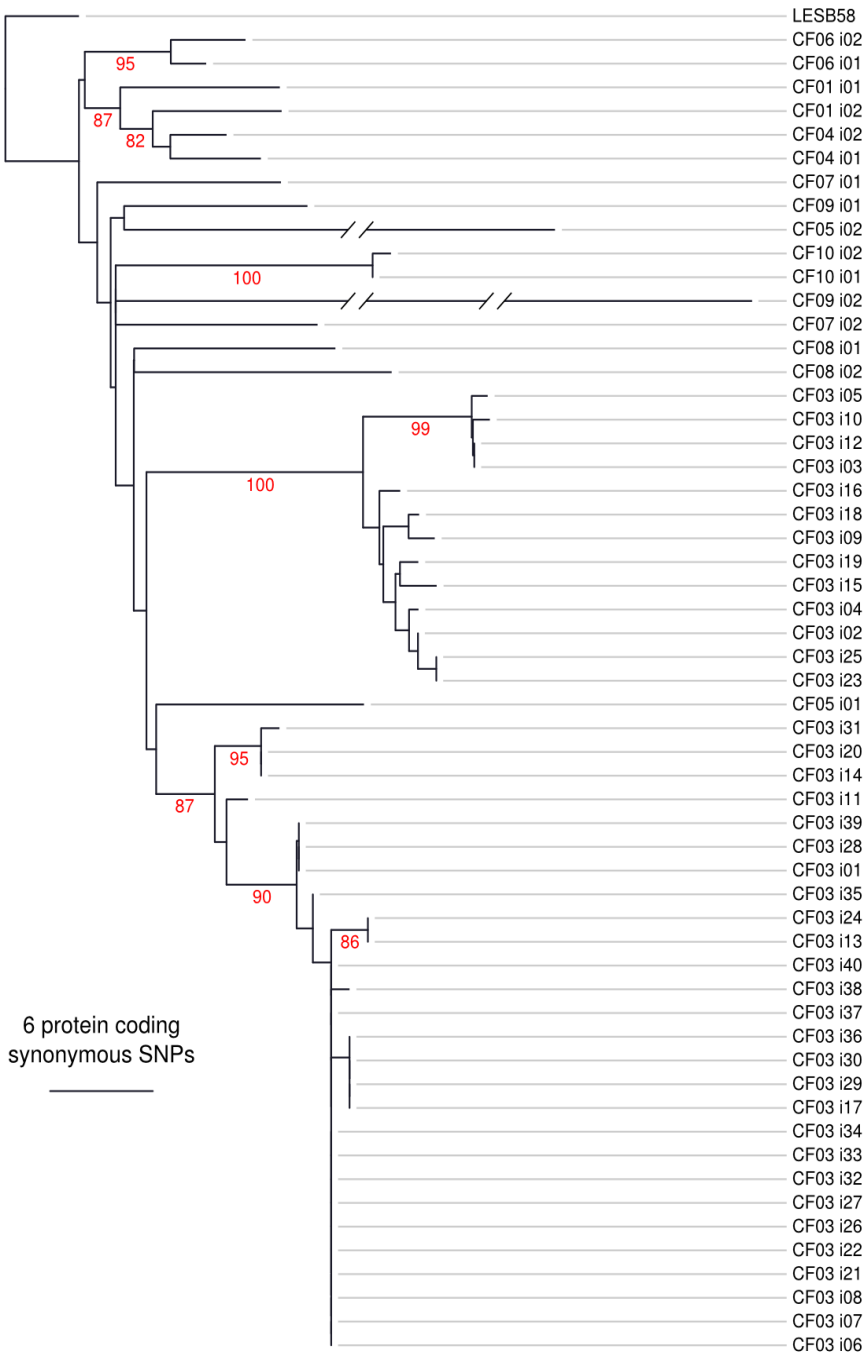


Figure S2: phylogenetic reconstruction from SNPs in protein coding synonymous positions is less resolved but congruent with phylogenetic reconstruction from SNPs in all positions. Neighbor-joining (BIONJ) phylogenetic reconstruction from synonymous SNPs of 40 *P. aeruginosa* isolate genome sequences from a patient CF03 sputum sample and two *P. aeruginosa* isolate genome

sequences from eight other patient sputum samples, all collected in 2009 with the genome sequence of LESB58 (CF airway isolate, Liverpool 1988) as an outgroup. Edges are labelled with bootstrap support where it is greater than 80%. No edges with bootstrap support greater than 80% are in disagreement with the phylogeny in figure 6 inferred from all SNPs.

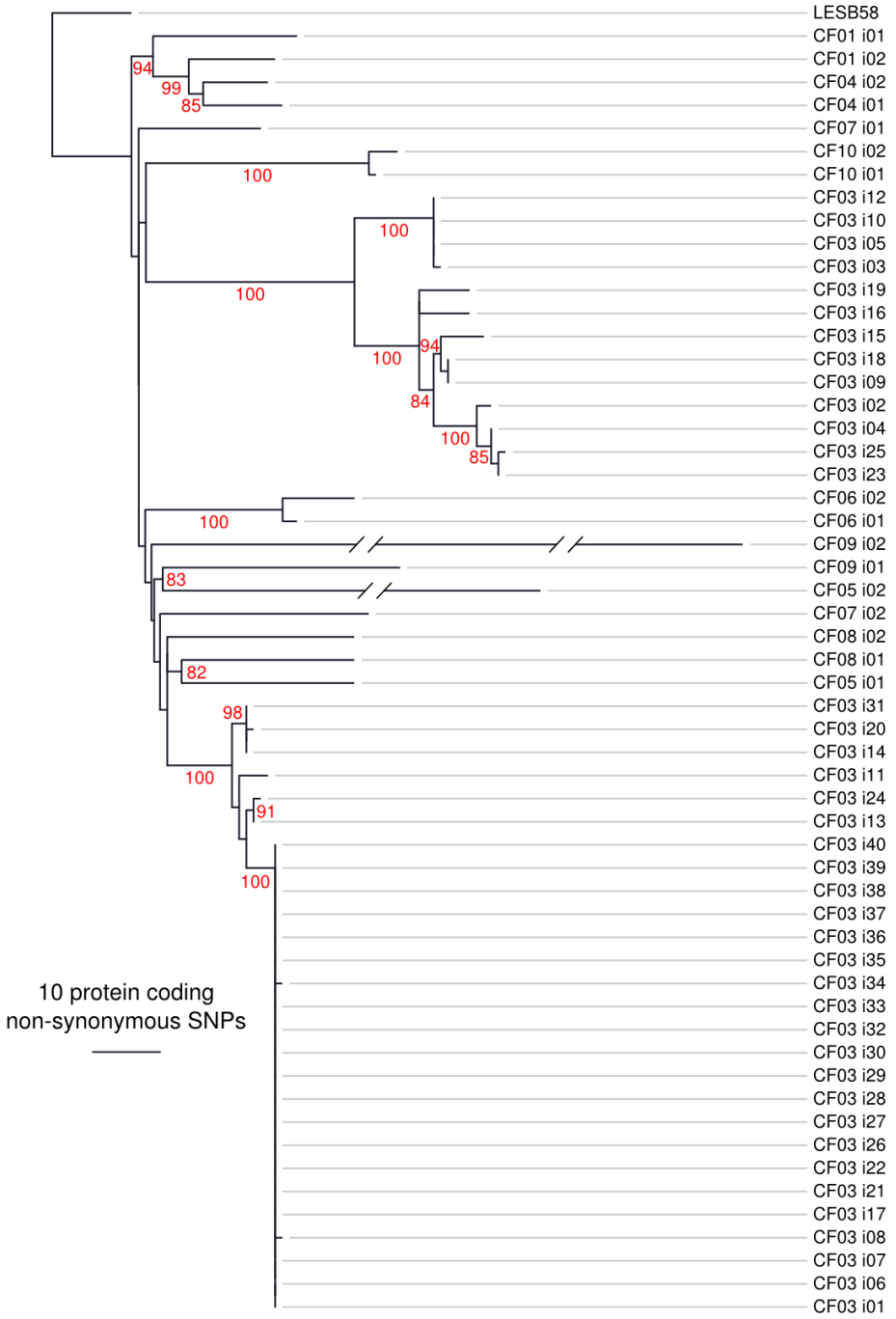


Figure S3: phylogenetic reconstruction from SNPs in protein coding non-synonymous positions is less resolved but congruent with phylogenetic reconstruction from SNPs in all positions.

Neighbor-joining (BIONJ) phylogenetic reconstruction from non-synonymous SNPs of the same sataset used in Figure S2. Edges are labelled with bootstrap support where it is greater than 80%. No

edges with bootstrap support greater than 80% are in disagreement with the phylogeny in figure 6 inferred from all SNPs.

References

1. Fothergill JL, Mowat E, Ledson MJ, Walshaw MJ, Winstanley C. Fluctuations in phenotypes and genotypes within populations of *Pseudomonas aeruginosa* in the cystic fibrosis lung during pulmonary exacerbations. *J Med Microbiol* 2010;59:472-481.

2. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetjournal* 2011;17:1.

3. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.29) [https://github.com/najoshi/sickle]

4. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;20:1297-1303.

5. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 2011;43:491-498.

6. Picard Sequence Alignment/Map file manipulation library [http://picard.sourceforge.net/]

7. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella KV, Altshuler D, Gabriel S, DePristo MA. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Curr Protoc Bioinform* 2002;43:11.10.1-11.10.33.

8. 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. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012;19:455-477.

9. Nikolenko S, Korobeynikov A, Alekseyev M. BayesHammer: Bayesian clustering for error correction in single-cell sequencing. *BMC Genomics* 2013;14:S7.

10. NCBI translation table 11 [http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi?mode=c#SG11]

11. Zwickl DJ. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. The University of Texas at Austin, School of Biological Sciences; 2006.

12. Ronquist F, Huelsenbeck J. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 2003;19:1572-1574.

13. Sukumaran J, Holder MT. DendroPy: a Python library for phylogenetic computing. *Bioinformatics* 2010;26:1569-1571.

14. Darriba D, Taboada GL, Doallo R, Posada D. jModelTest 2: more models, new heuristics and parallel computing. *Nat Methods* 2012;9:772-772.

15. Tracer v1.6 [http://tree.bio.ed.ac.uk/software/tracer/]

16. Shimodaira H. Testing regions with nonsmooth boundaries *via* multiscale bootstrap. *J Statist Plann Inference* 2008;138:1227-1241.

17. Kent WJ. BLAT—The BLAST-Like Alignment Tool. *Genome Res* 2002;12:656-664.

18. Mirkin BG, Fenner TI, Galperin MY, Koonin EV. Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol Biol* 2003;3:1-34.

19. Sipos B, Massingham T, Jordan G, Goldman N. PhyloSim - Monte Carlo simulation of sequence evolution in the R statistical computing environment. *BMC Bioinformatics* 2011;12:104.