Genetic variation of a bacterial pathogen within individuals with cystic fibrosis provides a record of selective pressures

Tami D Lieberman1, Kelly B Flett2, Idan Yelin3, Thomas R Martin4, Alexander J McAdam5, Gregory P Priebe2,6,7 & Roy Kishony1,3

Advances in sequencing technologies have enabled the identification of mutations acquired by bacterial pathogens during infection1–10. However, it remains unclear whether adaptive mutations fix in the population or lead to pathogen diversification within the patient11,12. Here we study the genotypic diversity of Burkholderia dolosa within individuals with cystic fibrosis by resequencing individual colonies and whole populations from single sputum samples. We find extensive intrasample diversity, suggesting that mutations rarely fix in a patient’s pathogen population—instead, diversifying lineages coexist for many years. Under strong selection, multiple adaptive mutations arise, but none of these sweep to fixation, generating lasting allele diversity that provides a recorded signature of past selection. Genes involved in outer-membrane components, iron scavenging and antibiotic resistance all showed this signature of within-patient selection. These results offer a general and rapid approach for identifying the selective pressures acting on a pathogen in individual patients based on single clinical samples.

Two opposing models of within-patient bacterial evolution have been proposed: a ‘dominant-lineage’ model, in which beneficial mutations drive superior lineages to dominate in the population, and a ‘diverse-community’ model, in which adaptive lineages rise to intermediate frequency and coexist with other lineages (Fig. 1)11–14. The diversity of within-patient pathogen populations has major implications for drug treatment and resistance7,15,16, for inferring transmission networks8,17,18 and for understanding evolutionary processes13,19. Here, to distinguish between these models and to understand the sources of genetic diversity, we compared the genomes of many bacterial cells of the same strain from the same clinical samples.

We focused on chronic infections with B. dolosa, a rare and deadly opportunistic pathogen that spread among 39 people with cystic fibrosis cared for at a single center in Boston starting in the 1990s (refs. 20,21). The airways of these patients were infected with very similar starting strains, and surviving patients have been colonized for years. A previous retrospective study of single-colony isolates identified specific B. dolosa genes that evolved under strong selective pressures during the outbreak8. Now, using sputum samples collected during clinical care, we characterize contemporary intraspecies diversity in five individuals from this outbreak who have been infected with B. dolosa since the early 2000s.

We used two genomic approaches, colony resequencing (patient 1) and deep population sequencing (patients 1–5), to identify single-nucleotide mutations and their frequencies in single sputum samples. In our colony resequencing approach, we isolated dozens of colonies from a clinical sample and analyzed their genomes individually by resequencing for years. A previous retrospective study of single-colony isolates identified specific B. dolosa genes that evolved under strong selective pressures during the outbreak8. Now, using sputum samples collected during clinical care, we characterize contemporary intraspecies diversity in five individuals from this outbreak who have been infected with B. dolosa since the early 2000s.

We used two genomic approaches, colony resequencing (patient 1) and deep population sequencing (patients 1–5), to identify single-nucleotide mutations and their frequencies in single sputum samples. In our colony resequencing approach, we isolated dozens of colonies from a clinical sample and analyzed their genomes individually by resequencing for years. A previous retrospective study of single-colony isolates identified specific B. dolosa genes that evolved under strong selective pressures during the outbreak8. Now, using sputum samples collected during clinical care, we characterize contemporary intraspecies diversity in five individuals from this outbreak who have been infected with B. dolosa since the early 2000s.

We used two genomic approaches, colony resequencing (patient 1) and deep population sequencing (patients 1–5), to identify single-nucleotide mutations and their frequencies in single sputum samples. In our colony resequencing approach, we isolated dozens of colonies from a clinical sample and analyzed their genomes individually by aligning reads to a B. dolosa reference genome, AU0158, a strain taken from a different patient in this outbreak. Because each colony originates from a single bacterium, this approach is equivalent to comparing different bacterial cells from the initial clinical sample. In the population sequencing approach22,23, we pooled hundreds of colonies from each clinical sample and sequenced the pool with deep coverage (~450×). We then aligned reads to AU0158 and identified fixed mutations, appearing in all reads, and polymorphisms, appearing in only a fraction of the reads. To remove false positive polymorphic sites caused by systematic sequencing or alignment errors24,25, we developed a set of thresholds and statistical tests that ensured accurate detection of polymorphisms (Supplementary Note). We calibrated this approach using an isogenic control for which we expected to detect no polymorphisms. For validation, we performed both methods on a single sample from patient 1, comparing diversity among 29 individual colonies to that detected with the population sequencing approach (Fig. 2). The population sequencing approach reliably detected polymorphisms where the minor allele frequency was greater than 3%, while decreasing the cost and labor required per sample.

We found that most of the mutations that arose during the course of infection did not fix, with sites remaining polymorphic within

1Department of Systems Biology, Harvard Medical School, Boston, Massachusetts, USA. 2Department of Medicine, Division of Infectious Diseases, Boston Children’s Hospital and Harvard Medical School, Boston, Massachusetts, USA. 3Faculty of Biology, Technion-Israel Institute of Technology, Haifa, Israel. 4Department of Medicine, Division of Respiratory Diseases, Boston Children’s Hospital and Harvard Medical School, Boston, Massachusetts, USA. 5Department of Laboratory Medicine, Boston Children’s Hospital and Harvard Medical School, Boston, Massachusetts, USA. 6Department of Anesthesiology, Perioperative and Pain Medicine, Division of Critical Care Medicine, Boston Children’s Hospital and Harvard Medical School, Boston, Massachusetts, USA. 7Department of Medicine, Division of Infectious Diseases, Brigham and Women’s Hospital and Harvard Medical School, Boston, Massachusetts, USA. Correspondence should be addressed to A.J.M. (alexander.mcadam@childrens.harvard.edu), G.P.P. (gregory.priebe@childrens.harvard.edu) or R.K. (roy_kishony@hms.harvard.edu).

Received 18 March; accepted 13 November; published online 8 December 2013; doi:10.1038/ng.2848
the patient. The colony resequencing approach performed for patient 1 identified 188 mutations occurring in some but not all isolates and only 10 mutations shared by all isolates. This dominance of polymorphisms, also seen in population sequencing of the same sample, strongly supports the diverse-community model (Fig. 3a,b). Similarly, for the four other patients, population sequencing on single samples identified a preponderance of polymorphisms compared to fixed mutations (≥73% of mutations; Fig. 3b). We found these excesses in polymorphisms despite the bias to overestimate mutations fixed during infection; some fixed mutations in a sputum sample might be polymorphic within the patient’s airways or might have become fixed before patient colonization (Supplementary Fig. 1).

The observed genomic diversity is a reflection of multiple coexisting lineages. Investigating the community structure of *B. dolosa* within patient 1, we found a deeply branched phylogeny with six lineages separated by at least five lineage-specific mutations (Fig. 3a). On average, pairs of isolates from this sample differed by 26 mutations, and, of all 406 possible isolate pairs, only 1 was identical. Thus, even within a single sputum sample, the population is so diverse that full identity of isolates is extremely rare.

In one patient (patient 5), the *B. dolosa* community had many more mutations than in other patients’ populations (*P* < 0.05, Grubbs’ test for outliers). This excess of mutations was due solely to increased transitions and not transversions, suggesting hypermutation (*P* < 0.01; Supplementary Fig. 2a). A search of the 199 mutated genes unique to the population of patient 5 identified a single mutation involved in DNA repair—a nonsynonymous mutation at a conserved position in *mutL*, defects of which are known to cause excess transitions (Supplementary Fig. 2b). The excess mutations were enriched in synonymous mutations relative to the mutations identified in the other patients, further supporting the presence of hypermutation (*P* < 0.001; Supplementary Fig. 2c). Although hypermutation is a common phenotype in many pathogens, hypothesized to accelerate the evolution of antibiotic resistance, it has not previously been described in members of the *Burkholderia cepacia* complex.

For how long have these diverging lineages coexisted? The time to the last common ancestor (LCA) of each nonhypermutating population from each patient can be estimated using the number of mutations accumulated since the LCA and the molecular clock previously

**Figure 1** Alternative models of within-patient evolution. (a) In the dominant-lineage model of within-host evolution, lineages with beneficial mutations sweep to fixation (green lines), eliminating their less fit ancestors or other temporarily arising genotypes (dashed lines). In this model, most observed mutations will be fixed, and polymorphic mutations will be rare, representing only recent mutational events (magenta lines). (b) In the diverse-community model, lineages coexist and compete for long stretches of time. In this model, most sampled mutations will be polymorphic.

**Figure 2** Two methods for studying genomic intraspecies diversity. (a) To study within-patient evolution, we cultured sputum samples from patients with cystic fibrosis on selective medium. In the colony resequencing approach (solid arrows; performed for one patient), we isolated multiple individual colonies from the same sample, independently called variants for each isolate via alignment of reads and compared variants between the isolates. In the deep population sequencing approach (dashed arrow; performed for five patients), we pooled hundreds of colonies from the same plate and analyzed the pool’s genomic DNA. We identified positions on the genome where some reads, originating from different colonies on the plate, disagreed with an inferred ancestral genome (Online Methods). (b) Allele frequency estimates in population sequencing (y axis) versus colony resequencing (x axis) for the same sputum sample from patient 1 for each mutated position. Mutations are classified as either fixed or polymorphic. Some mutations found in the colony-based approach were below the threshold in frequency (3%) or confidence in the pool-based approach (dashed line). Slight jitter was added to the x and y coordinates for each point to improve visibility (up to 2% change). As an example, the insets at top and right display a summary of the raw data at the indicated genomic position. Population sequencing (right) at this position shows 70% of aligned reads supporting a T (orange) and 30% of aligned reads supporting a G (black), consistent with the corresponding number of colonies in the individual isolates (T, 22; G, 7). Reads from each isolate (top) are mostly of identical calls (all T or all G). Green indicates a single read in one isolate supporting an A, likely a sequencing error. For further comparison of the two methods, see Supplementary Figure 7.
measured for this outbreak (2.1 SNPs/year; ref. 8). Given the phylogeny of isolates from patient 1, we calculated the distribution of the number of mutations since the LCA, \( d_{LCA} \), across the population (Fig. 4a). The mean value of \( d_{LCA} \) across isolates, \(<d_{LCA}>\), was 19.6 single-nucleotide mutations per genome (95% confidence interval, CI = 18.3–20.8), suggesting that the LCA existed 9.3 years ago (CI = 8.7–9.9). This calculation places the LCA of the isolates from this sample slightly earlier than the first B. dolosa culture from this patient (7.6 years before sample collection), suggesting that the B. dolosa population in patient 1 has been diverging since or perhaps before initial colonization. Although the population sequencing approach could not provide a distribution of \( d_{LCA} \), owing to a lack of information regarding linkage between mutations, we could still calculate \(<d_{LCA}>\) from the sum of the polymorphic mutation frequencies (see the Supplementary Note for derivation). Using this approach, the estimated time to the LCA for the population from patient 1 was 7.9 years. This value is slightly lower than that calculated with the clonal resequencing approach, likely owing to mutations left undetected by our conservative polymorphism caller (see the Supplementary Note for a discussion of error). For patients 2 and 4, the time to the LCA calculated by this population sequencing approach was several years less than the time since the first positive culture, suggesting fixation events sometime during these patients’ histories (Supplementary Table 1). For all the patients, we estimated that diverging lineages coexisted in each for at least 5 years (Fig. 4b).

To explore the drivers of this long-coexisting diversity, we examined the identities of the evolving genes. Interestingly, we found that, within each sample, several B. dolosa genes carried 2–4 coexisting polymorphisms (Supplementary Table 2). This clustering represents a significant departure from a neutral model, given the number of mutations and the distribution of gene lengths (\( P < 0.005 \) for patients 1–4; Fig. 5a and Online Methods). A similar analysis on the operon level further identified several operons enriched for polymorphisms (Supplementary Fig. 3 and Supplementary Table 3). An enrichment of nonsynonymous mutations in these multidiverse genes and operons suggests that they are drivers of adaptive change in vivo (normalized ratio of nonsynonymous to synonymous mutations, \( \text{dN/dS} = 7.0, \text{CI} = 2.3–34.9 \); Fig. 5b). Polymorphisms are thus concentrated in genes undergoing adaptive evolution.

To understand why polymorphisms cluster in some genes, we asked whether coexisting mutations in the same gene appeared in different lineages or were linked in a double mutant. Examining the genomes of single isolates, we found no isolates with doubly mutated genes (Supplementary Fig. 4). Similarly, with population sequencing, in 10 of 11 cases where polymorphic positions were close enough on the genome to be covered by the same short sequencing reads, we did not find reads that contained both variants (Fig. 5c and Supplementary Fig. 5). In some of these cases, the ancestral genotype was completely purged from the population (Fig. 5d). Thus, diversification is driven by multiple adaptive mutations in the same genes evolving in parallel within individual patients.
These findings provide a new signature of past selective pressures detectable in a single clinical sample: the coexistence of multiple polymorphisms within the same gene in a clinical sample. Sixteen B. dolosa genes displayed this multidiverse signature, including genes with homologs involved in outer-membrane synthesis, antibiotic resistance, iron scavenging, oxygen sensing, amino acid synthesis, lactate use and stress response. Additionally, some genes with less characterized biological roles displayed a multidiverse signature, including two transcriptional regulators with unknown targets in B. dolosa, an uncharacterized glucosamylase and two genes that encode hypothetical proteins (Supplementary Table 2). A similar signature for selection was seen in three operons, two involved in lipopolysaccharide transport and one containing a two-component regulatory system with unknown targets (Supplementary Table 3). Selection on many of these elements can be rationalized by the relevance of their annotated functions to conditions to which the bacteria are exposed during the course of the infection. Yet, further investigation will be required to understand the potential roles of some of these genes in antibiotic resistance, fitness and other aspects of pathogenesis.

We found that many of the selective forces acting on the pathogen were the same across patients (Fig. 5e). Often, genes showing a multidiverse signature for selection in one patient also carried mutations in other patients (P < 0.002, hypergeometric test). A prominent example was gyrA, a well-studied target of quinolones, which was mutated in all patient populations. Further support for commonality in mutational trajectories across patients emerged from a significant overlap between this list of 16 multidiverse genes and 17 genes previously found to be under parallel evolution across a larger group of patients, only 1 of whom (patient 2) was included in both studies (P < 0.001, hypergeometric test). Thus, the study of a single clinical sample can provide generalizable lists of the selective pressures felt within the human body.

Yet, some multidiverse signatures were patient specific. A penicillin-binding protein (BDAG_01166, homologous to PBP7) was affected by three nonsynonymous mutations in patient 1 but was not mutated in other patients. Such patient-specific parallel evolution might reflect patient-specific selective pressure or perhaps a fitness benefit dependent on previously acquired mutations. However, these hypotheses are hard to test because the genomic target for a selective force might include more than one gene. For example, populations from four of the five patients had a mutation in a homolog of the histidine kinase gene fixL (BDAG_01161; known to be under strong selection in these infections8), whereas the population from the fifth patient had a mutation in the corresponding response regulator gene.

To investigate the stability of these multidiverse signatures for selection, we collected a second sputum sample 14 d after initial sample collection from patient 2. Three of the four genes with the multidiverse signature at day 0 showed the same pattern at day 14. The absence of the signature in the fourth gene at the later time point does not reflect a relaxation in selection for mutant alleles but, rather, incomplete detection of genes under selection; this gene also had abundant nonsynonymous mutations at day 14, concentrated at a single nucleotide position (Supplementary Fig. 6). These results suggest that the multidiverse signature for selection is relatively stable and that multiple sample collections per patient can increase the sensitivity of detection.
Our results reject the dominant-lineage model of infection yet demonstrate that diversifying bacteria adapt under the pressure of natural selection. These observations are consistent with clonal interference: in large asexual populations, multiple beneficial mutations emerge and compete, impeding the ability of the corresponding lineages to reach fixation\textsuperscript{33–35}. In addition to large population size (10\textsuperscript{8} cells/ml of sputum), the branched structure of the airways may further hinder the capacity of any adaptive lineage to dominate and fix, and the immune system or niche-specific adaptations might directly promote diversity. Diversified by any of these means, lineages may then continue to evolve in parallel against common selective forces.

As \textit{B. dolosa} adapts to the airways of people with cystic fibrosis, mutations lead to diversification rather than fixation and replacement. Although it is possible that adaptive mutations lead to fixation more frequently in other infections, there is evidence that, at least in long-term colonization, diversity might be common\textsuperscript{14,36–38}. This long-term coexistence of diverse lineages records the genomic history of selection on the pathogen in its host. The ability to rapidly read off within-patient evolutionary history from the genotypic diversity within a single clinical sample may greatly accelerate the ability to survey the selective pressures acting on bacterial pathogens \textit{in vivo}—shifting from an epidemic-level investigation to a single-patient paradigm.

URLs. \textit{Burkholderia dolosa} sequencing project, http://www.broadinstitute.org/annotation/genome/burkholderia_dolosa.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Sequencing reads for all 29 isolates and 6 populations have been deposited in the NCBI Sequence Read Archive (SRA) under accession SRP030656.

ACKNOWLEDGMENTS
We are grateful to J.-B. Michel and members of the Kishony laboratory for insightful discussions and support, to the team at the Partners HealthCare Center for Personalized Genetic Medicine (PCPGM) for Illumina sequencing, to L. Williams and A. Palmer for discussions and technical assistance, and to Y. Gerardin, J. Meyer, L. Stone and R. Ward for their comments on the manuscript. T.D.L. and G.P.P. were supported in part by grants from the Cystic Fibrosis Foundation at L. Williams and A. Palmer for discussions and technical assistance, and to Y. Gerardin, J. Meyer, L. Stone and R. Ward for their comments on the manuscript.

AUTHOR CONTRIBUTIONS
T.D.L., G.P.P. and R.K. designed the study. A.J.M. and T.R.M. collected clinical samples. K.B.F., T.R.M., A.J.M. and G.P.P. conducted chart review and provided medical information. T.D.L. performed experiments. T.D.L., I.Y. and R.K. wrote the sequence analysis scripts. T.D.L. and R.K. analyzed the data. T.D.L., A.J.M., G.P.P. and R.K. interpreted the results and wrote the manuscript.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Mwangi, M.M. et al. Tracking the \textit{in vivo} evolution of multidrug resistance in \textit{Staphylococcus aureus} by whole-genome sequencing. \textit{Proc. Natl. Acad. Sci. USA} \textbf{104}, 9451–9456 (2007).
34. Gerrish, P.J. & Lenski, R.E. The fate of competing beneficial mutations in an asexual population. *Genetica* **102–103**, 127–144 (1998).

35. Hegreness, M., Shoresh, N., Hartl, D. & Kishony, R. An equivalence principle for the incorporation of favorable mutations in asexual populations. *Science* **311**, 1615–1617 (2006).

36. Mowat, E. et al. *Pseudomonas aeruginosa* population diversity and turnover in cystic fibrosis chronic infections. *Am. J. Respir. Crit. Care Med.* **183**, 1674–1679 (2011).

37. Schloissnig, S. et al. Genomic variation landscape of the human gut microbiome. *Nature* **493**, 45–50 (2013).

38. Ashish, A. et al. Extensive diversification is a common feature of *Pseudomonas aeruginosa* populations during respiratory infections in cystic fibrosis. *J. Cyst. Fibros.* doi:10.1016/j.jcf.2013.04.003 (1 May 2013).

39. Menard, A., de Los Santos, P.E., Graindorge, A. & Cournoyer, B. Architecture of *Burkholderia cepacia* complex σ70 gene family; evidence of alternative primary and clade-specific factors, and genomic instability. *BMC Genomics* **8**, 308 (2007).
Online Methods

Study cohort and sample collection. An epidemic clone of B. dolosa infected and colonized 39 individuals with cystic fibrosis in the Boston area over a 20-year period. We studied B. dolosa inpatient diversity in five surviving individuals still infected with B. dolosa. All subjects were male, had homozygous ∆F508 alterations, had not received lung transplants, were between 21 and 35 years of age, and had been colonized for between 7 and 10 years at the time of sample collection (Supplementary Table 1). Longitudinal microbial isolates from patient 2 were also included in a previous retrospective study (patient J in ref. 8).

For patient 1, both the colony resequencing and deep population sequencing approaches were performed on a single sputum sample (P1). For patient 2, population deep sequencing was performed on each of two sputum samples (P2 and P2T), collected 14 d apart. Between collections, patient 2 was treated for a pulmonary exacerbation, including a change in antibiotic regimen, but his condition did not improve, and B. dolosa density did not decrease. For patients 3–5, population sequencing was performed on a single sputum sample from each patient (P3–P5).

Expectorated sputum samples were collected at Boston Children’s Hospital after written informed consent was obtained under protocols approved by the institutional review boards at Boston Children’s Hospital and Harvard Medical School. Samples were liquefied with dithiothreitol and stored at −80 °C in 20% glycerol. B. dolosa was cultured from frozen samples. For population sequencing, a plate with 5,000 to 30,000 small colonies was chosen from a serial dilution.

See the Supplementary Note for more details on sample preparation.

Illumina sequencing. Genomic DNA was extracted using the MoBio UltraClean Microbial DNA Isolation kit according to the manufacturer’s instructions. Genomic libraries were constructed and barcoded using the Illumina-compatible Epicentre Nextra DNA Sample Prep kit following the manufacturer’s instructions (PCR amplification in the Nextra preparation does not introduce false positive polymorphisms; Supplementary Note). Genomic libraries were sequenced on the Illumina HiSeq 2000 platform by the Partners HealthCare Center for Personalized Genetic Medicine. Individual colonies were sequenced using single-end 50-bp reads, and pooled samples were sequenced using paired-end 50-bp reads. Reads were aligned to the B. dolosa draft genome AU0158 (GenBank accession AAYK00000000; see URLS), belonging to an isolate recovered from patient 0 of the outbreak. AU0158 consists of 233 contigs on 3 scaffolds (B. dolosa has 3 chromosomes). Standard approaches were used for read filtering and alignment (Supplementary Note).

Mutation identification, colony resequencing. An outgroup of three outbreak strains (A-0-0, G-9-8 and N-12-6d-8; previously sequenced in ref. 8) was included in the analysis to identify mutations fixed among the 29 isolates from patient 1. We considered genomic positions at which at least one pair of isolates was discordant on the called base and both members of the pair had FQ scores less than −40. We considered genomic positions at which at least one pair of isolates was discordant on the called base and both members of the pair had FQ scores less than −40. The CI for <dLCA> was calculated as the sum of the mutation frequencies at each polymorphic position called within that population, normalized by the size of the callable genome (Supplementary Note). For the pool-based approach, we defined the callable genome as the set of positions that met the chosen thresholds for coverage, average base quality, average mapping quality and average tial distance for each strand, irrespective of nucleotide call. See Supplementary Figure 6b and the Supplementary Note for a discussion of sources of error in estimating <dLCA> and time since the LCA.

Detection of parallel evolution within patients. We defined genes with a multidivisive signature of selection as genes for which within the same sputum sample there were multiple polymorphisms and multiple polymorphisms per 2,000 bp (to account for the fact that long genes are more likely to be mutated multiple times by chance). To determine whether the number of genes showing this signature represented a significant departure from what would be expected in a neutral model, we performed for each sputum sample 1,000 simulations in which we randomly shuffled the polymorphisms found across the callable genome and calculated how many genes showed a signature of selection (Fig. 5a).

This analysis was repeated at the operon and pathway levels, using the free version of FgenesB to identify operons and subsystem annotations provided by SEED as pathways (Supplementary Fig. 3). As in the gene analysis, we considered operons and pathways to have a signature of selection if they had both multiple polymorphisms and multiple polymorphisms per 2,000 nucleotides within the same patient.

dN/dS. Mutations were classified as nonsynonymous (N) or synonymous (S) according to annotations provided in the GenBank file. For ORFs in the draft genome without a provided reading frame, we used BLAST and RefSeq to identify the most likely reading frame in the neighborhood of the found mutations. For each dN/dS calculation, we used the particular spectrum of mutations observed to calculate the expected N/S ratio (for example, across the B. dolosa genome, A>C mutations are 10.6 times more likely to cause a nonsynonymous mutation than G>A mutations). The observed value of the N/S ratio was divided by this expectation to give dN/dS. CIs and P values were calculated according to binomial sampling. The dN/dS value reported (Fig. 5b) groups together the mutations found in genes and operons under selection: the same calculation for only genes gave a dN/dS value of 5.9 (95% CI = 1.9–29.6).

Parallel evolution across patients. We used the hypergeometric distribution to assess the significance of overlap between gene sets. Of the 225 B. dolosa genes mutated in patients 1–4, only 16 showed the multidivisive signature of selection within patients and only 29 genes were mutated in multiple patients (fixed or polymorphic), yet 7 genes were in common between these lists (P = 0.0015). Similarly, 13 of these 225 genes were also found on a list of 17 genes evolved in parallel across patients in a previous study. These 13 genes were enriched in the 16 genes under selection in this study (5-gene overlap; P = 0.0009). When this analysis was repeated without mutations from patient 2 (this patient was also included in the retrospective study), 11 of the 189 mutated genes were found in the previous study, and 13 genes showed a multidivisive signature of selection. The overlap between these lists of 11 and 13 genes was still significant (4 genes; P = 0.0035).

40. Guss, A.M. et al. Phylogenetic and metabolic diversity of bacteria associated with cystic fibrosis. ISME J. 5, 20–29 (2011).
41. Li, H. et al. The sequence alignment/map format and SAMtools. Bioinformatics 25, 2078–2079 (2009).
42. Felsenstein, J. PHYLIP-Phylogeny Inference Package (Version 3.2). Cladistics 5, 164–166 (1989).
43. Aziz, R.K. et al. The RAST Server: rapid annotations using subsystems technology. BMC Genomics 9, 75 (2008).