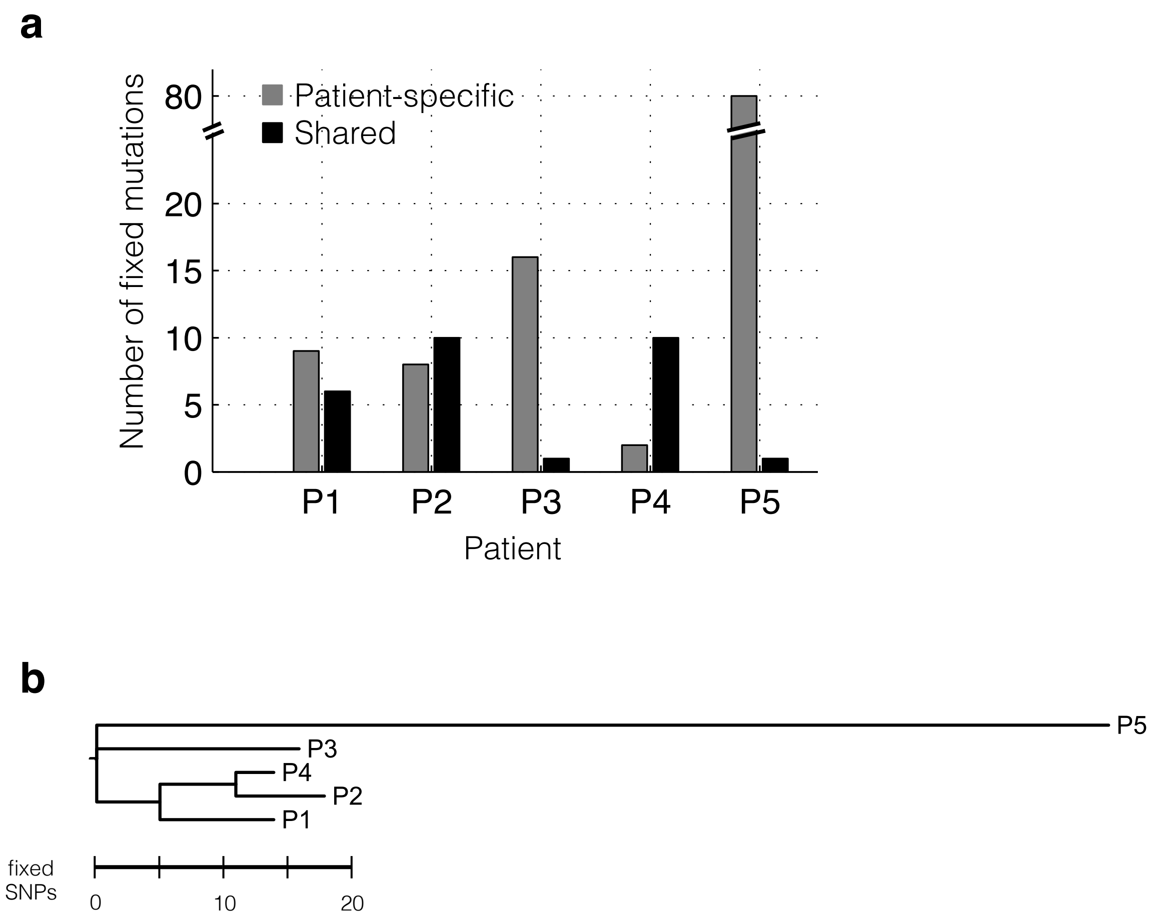
Supplementary information for “Genetic variation of a bacterial pathogen within a single clinical sample provides a record of its adaptive evolution in the patient”

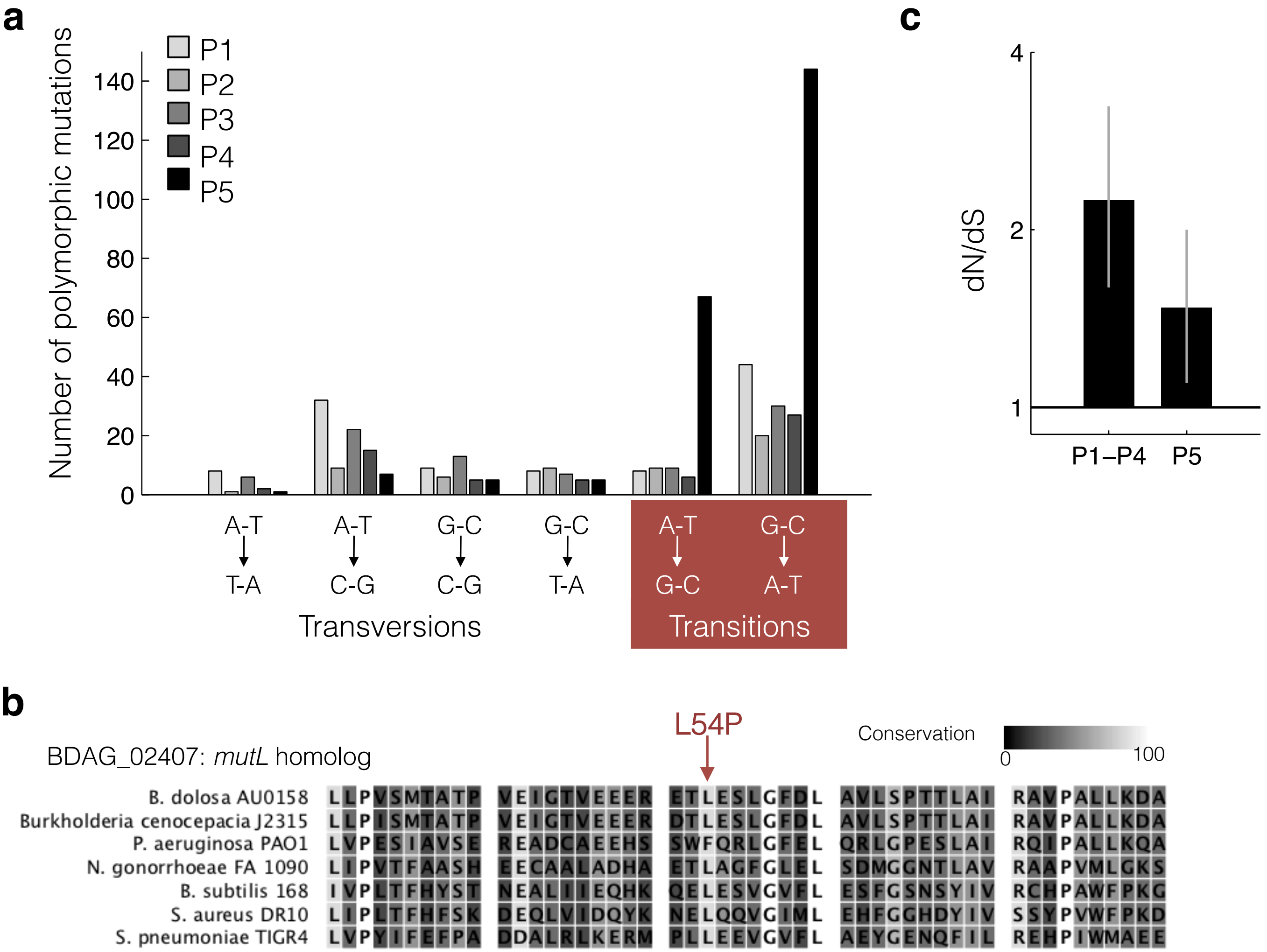
Tami D. Lieberman[1], Kelly B. Flett[2], Idan Yelin[3], Thomas R. Martin[4], Alexander J McAdam[5], Gregory P Priebe[2,6,7], & Roy Kishony[1,3]

1. Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA.
2. Division of Infectious Diseases, Department of Medicine, Boston Children’s Hospital; and Harvard Medical School, Boston, MA 02115, USA.
3. Faculty of Biology, Technion-Israel Institute of Technology, Haifa, Israel.
4. Division of Respiratory Diseases, Department of Medicine, Boston Children’s Hospital and Harvard Medical School, Boston, MA 02115, USA
5. Department of Laboratory Medicine, Boston Children’s Hospital and Harvard Medical School, Boston, MA 02115, USA.
6. Division of Critical Care Medicine, Department of Anesthesiology, Perioperative and Pain Medicine; Boston Children’s Hospital; and Harvard Medical School, Boston, MA 02115, USA.
7. Division of Infectious Diseases, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA.

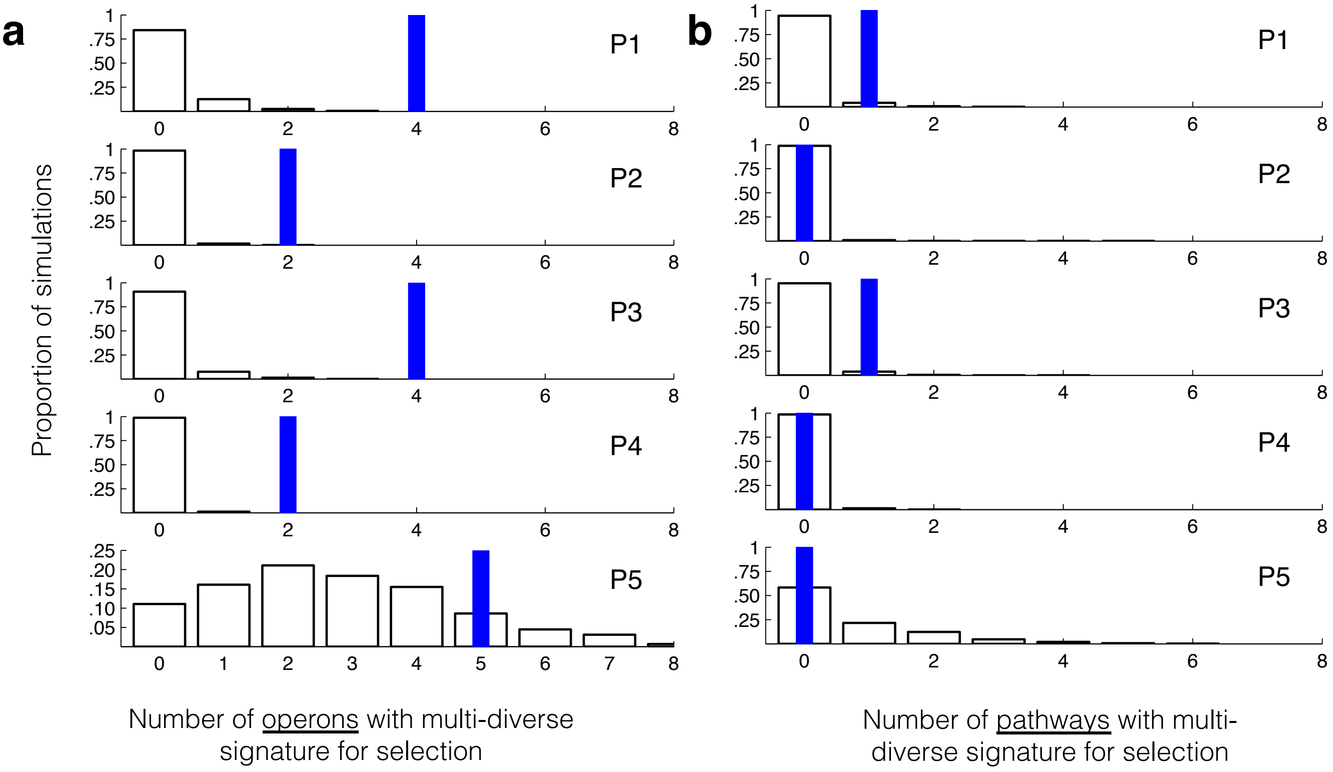
This file contains Supplementary Figures 1-8, Supplementary Tables 1-4, Supplementary Note, and Supplementary References (19 pages). There are two Excel-formatted Supplementary Tables describing the mutations found.



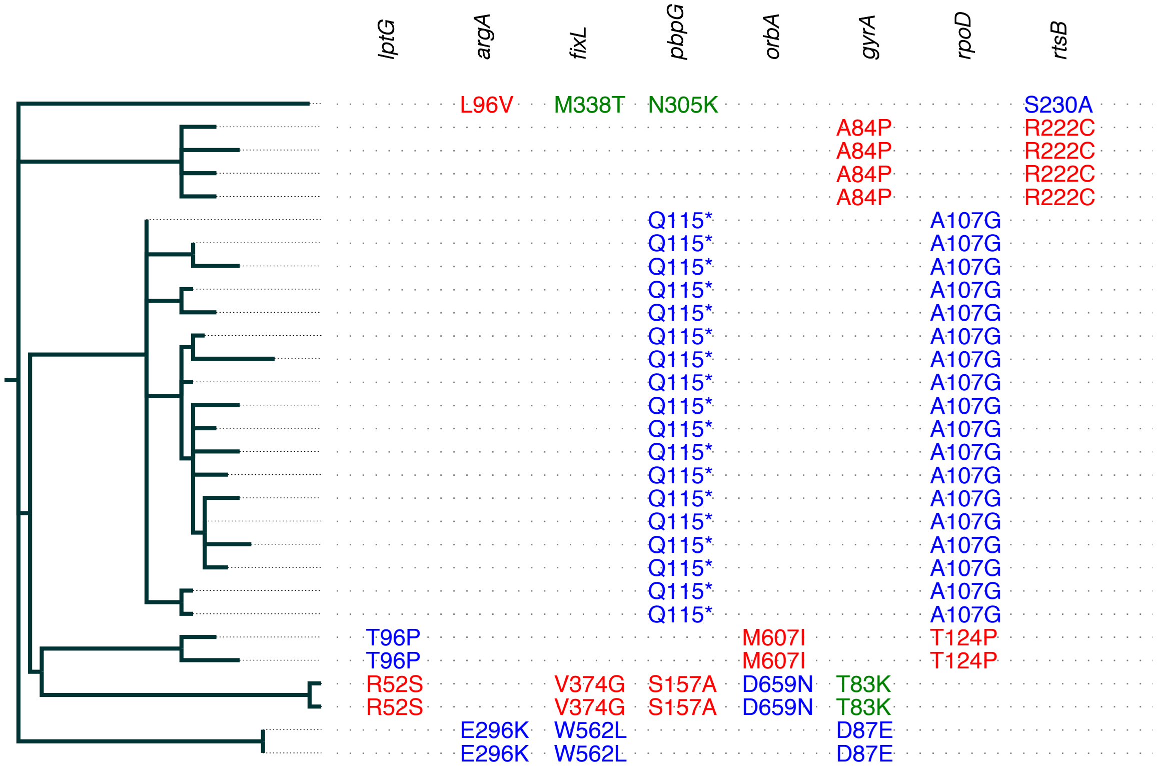
Supplementary Figure 1: Some fixed mutations may have arisen and fixed prior to patient colonization. (**a**) Number of patient-specific and shared fixed mutations per patient. Shared fixed mutations are found to be fixed in some, but not all sputum samples, while patient-specific mutations are fixed in only one patient's sample. (**b**) The fixed mutations for each patient define a patient LCA, and we generate a maximum parsimony phylogeny among patient LCAs. The presence of interior branches in this phylogeny illustrates that some shared fixed mutations likely arose in an LCA of multiple patients. This tree was generated using the dnapars package in Phylip[1] and visualized in Figtree.



Supplementary Figure 2: Excess mutations in Patient 5 are due to hypermutation. (**a**) Polymorphic mutations found within each patient's population were classified into 6 categories, without regard for strand (e.g. A->C is equivalent to T->G). P5 has an excess of both types of transition mutations (P<.001, Grubb's test for outliers) but not transversion (P<.001, Grubb's test for outliers) mutations, consistent with the known spectrum of mutations caused by *mutL* defects. (**b**) We scanned the annotations of the genes with point mutations in P5 for "DNA"' and manually inspected the results for roles in DNA repair. Only BDAG_02407 met this criteria. An NCBI BLAST search revealed this as a homolog of *mutL,* an essential component of mismatch repair. Other *mutL* sequences from NCBI gene were aligned and conservation was calculated using CLC Sequence Viewer 6. (**c)** P5's population has an increased relative rate of synonymous mutations. dN/dS was assessed as listed in the Methods for the intragenic mutations found in P1-P4 (non-hypermutators, n=278) and the intragenic mutations found in P5 (n=242). As described in the **Online Methods**, our approach for dNdS accounts for the decreased expected N/S in the hypermutators. Gray bars indicate 95% confidence interval. P < .001, one-side binomial z-test comparing the observed to the N/S expected under a dN/dS of 2.5 (P1-P4) and P5's spectrum of mutations.

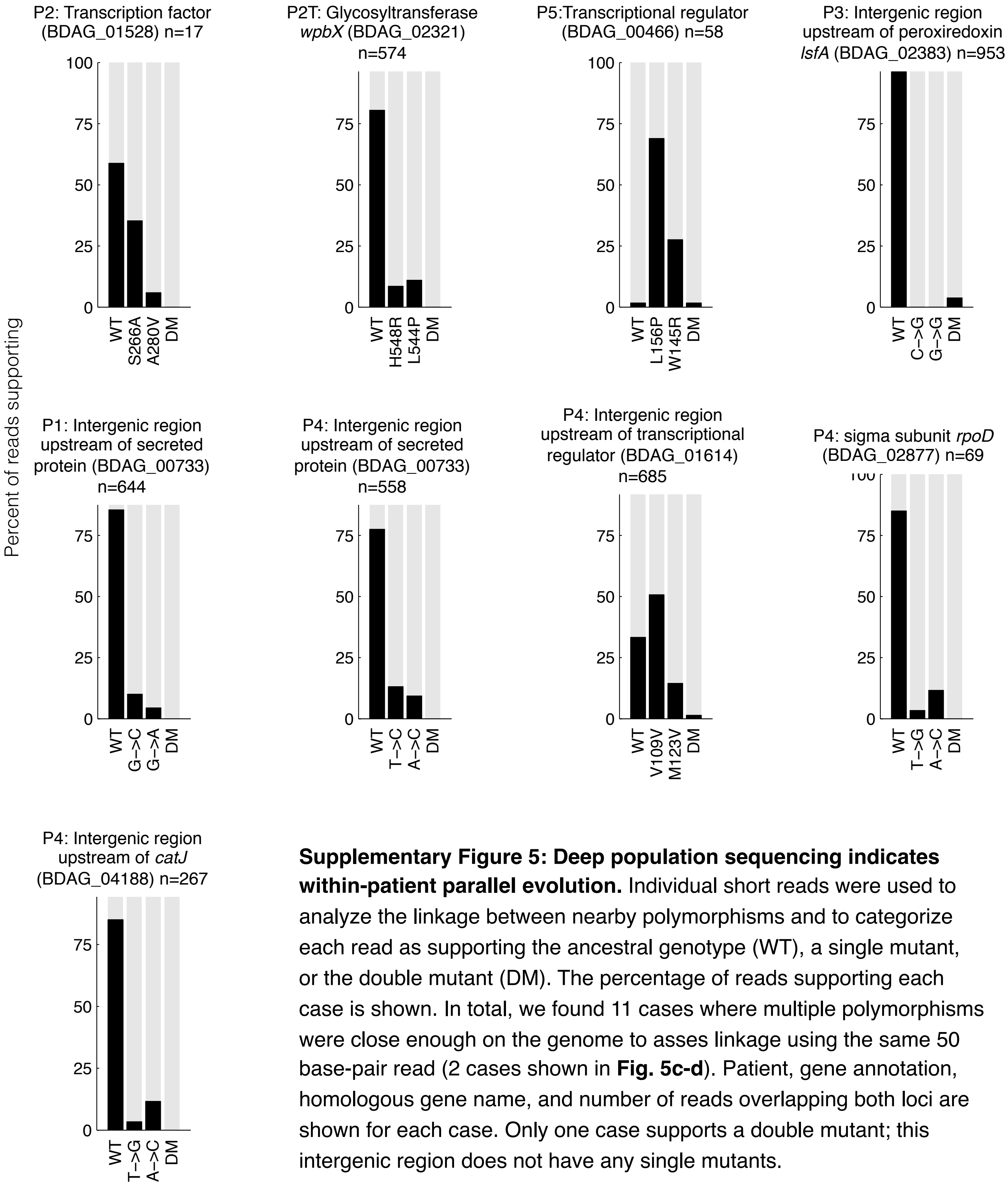


Supplementary Figure 3: Search for operon and pathways undergoing parallel evolution within patients. We expanded our search for selective pressures to the operon and pathway levels, applying the same requirements of multiple mutations and more than one mutation per 2000 bp. (**a**) Multiple operons have this multi-diverse signature for selection in each patient (blue bars), while under neutrality we expect none in Patients 1-4 (P1-4). In P1-P4, the observed number of operons with multiple mutations is greater than the number obtained in > 995/1000 simulations (white bars show the results of 1000 simulations). For most of the 9 operons, the mutations that triggered their identification were all concentrated in a single, previously identified gene. The operons not previously implicated are shown in **Supplementary Table 3**. (**b**) We did not find enrichment at the pathway level, suggesting that parallel evolution primarily acts at or below the gene level and supporting the idea that binning over larger regions of the genome can dilute the signal for selection. For each patient, we observe multiple mutations in the same pathway in over 18% of simulations (white bars). Relaxation of the requirement of mutations per bp in pathway brings up more multiply mutated genes in the simulations and not many more pathways. The two pathways multiply mutated per 2000 bp contain genes already found at the gene and operon level.

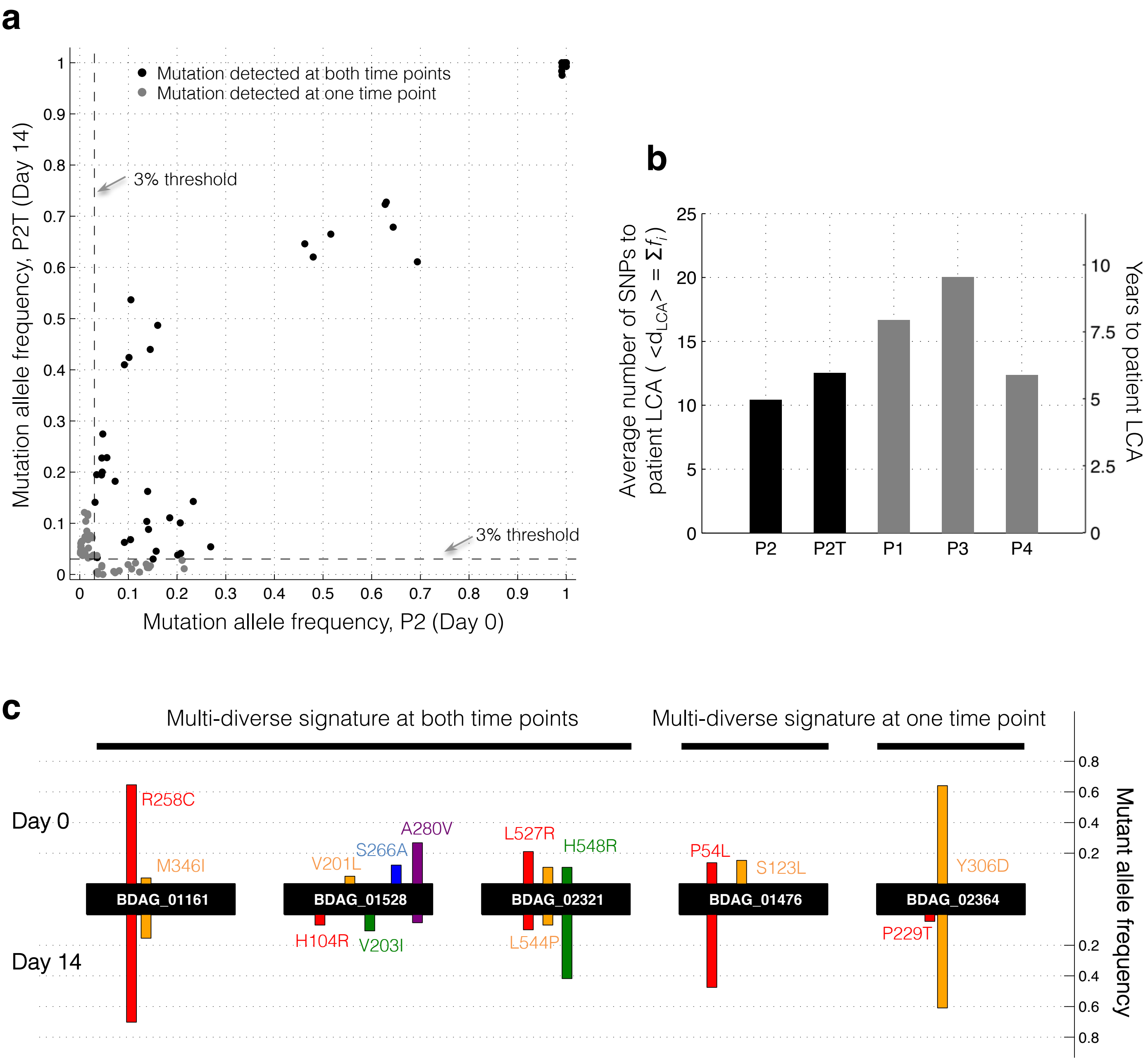


Patient 1 isolate phylogeny

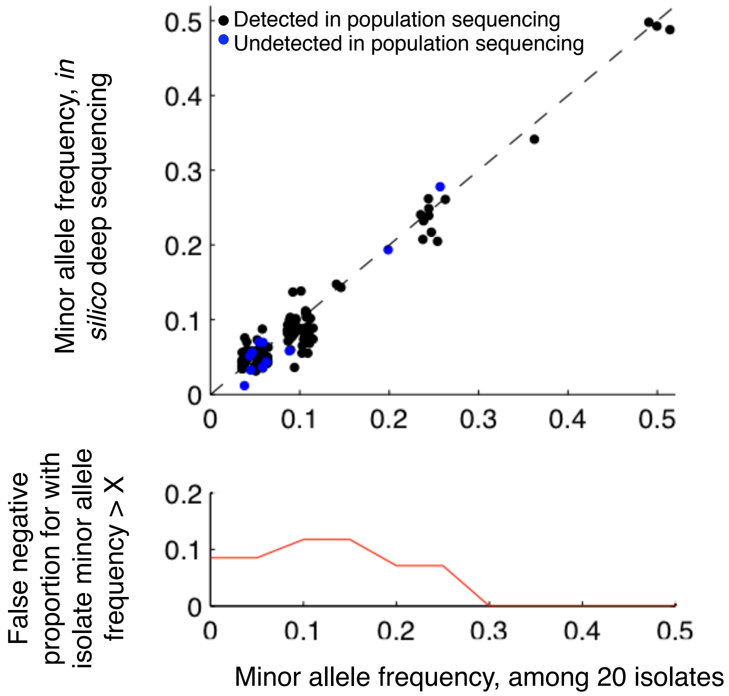
Supplementary Figure 4: Colony re-sequencing from Patient 1 indicates within-patient parallel evolution. Seven genes within Patient 1 showed a multi-diverse signature for selection in the colony re-sequencing approach. The gene names are listed at top (see **Supplementary Tables 2 and 3**) and the phylogeny of 29 colony isolates from Patient 1 is shown at left (same as **Fig. 3a**). For each isolate, any mutations found in that gene are indicated on the corresponding horizontal line and column. No isolate has multiple mutations in the same gene.



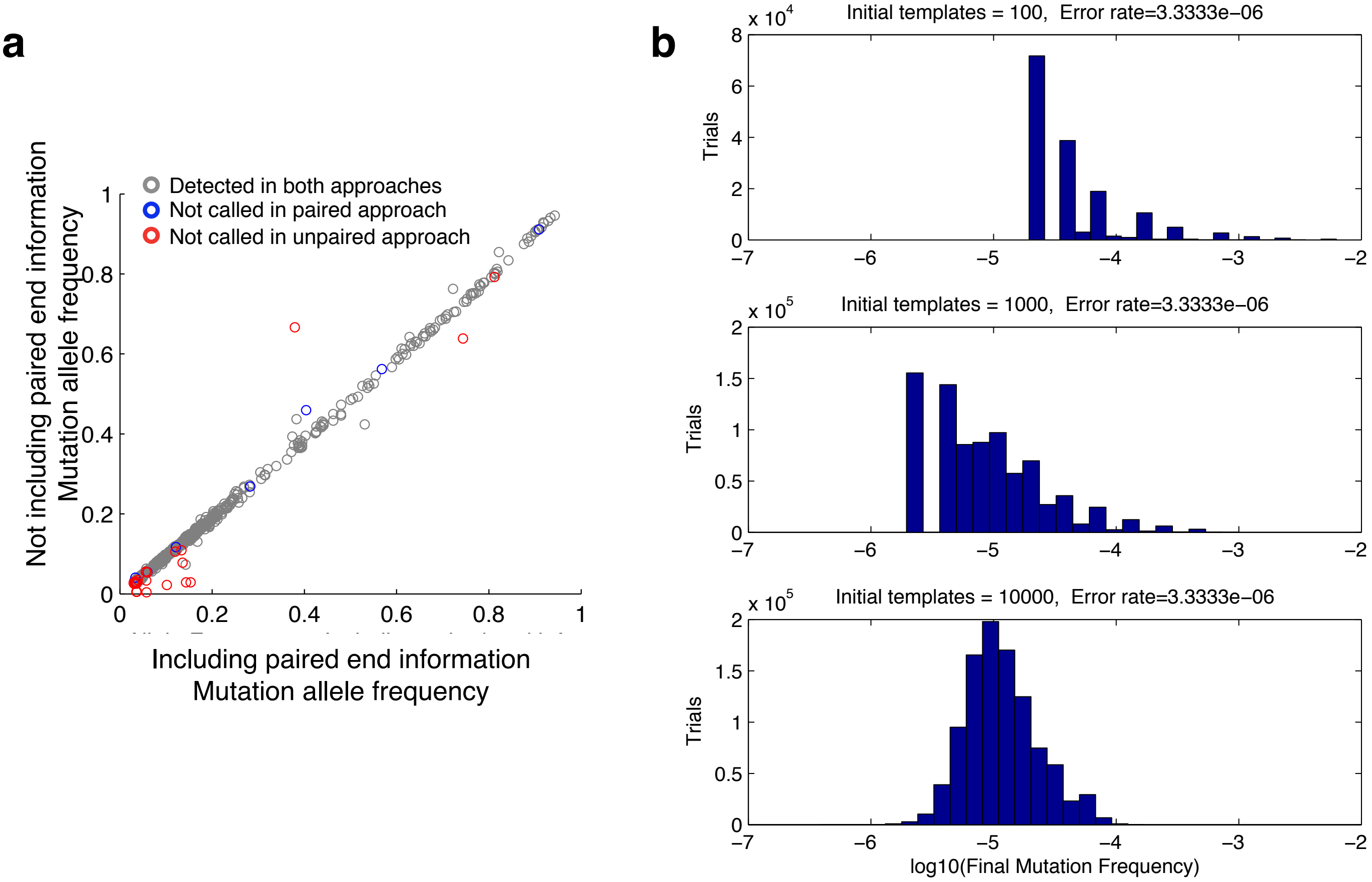
Supplementary Figure 5: Deep population sequencing indicates within-patient parallel evolution. Individual short reads were used to analyze the linkage between nearby polymorphisms and to categorize each read as supporting the ancestral genotype (WT), a single mutant, or the double mutant (DM). The percentage of reads supporting each case is shown. In total, we found 11 cases where multiple polymorphisms were close enough on the genome to asses linkage using the same 50 base-pair read (2 cases shown in **Fig. 5c-d**). Patient, gene annotation, homologous gene name, and number of reads overlapping both loci are shown for each case. Only one case supports a double mutant; this intergenic region does not have any single mutants.



Supplementary Figure 6: Comparison between two samples from same patient taken 14 days apart. (**a**) Scatter plot of mutant allele frequencies between samples shows agreement at very diverse positions with more variation in lower frequency mutations. (**b**) Both samples give similar estimates for time to patient LCA. See **Supplementary Note** for discussion of error. (**c**) Three genes show a multi-diverse signature for selection in both samples, and each sample has only one gene displaying this signature that is not present in the other sample. Moreover, these two sample-specific genes show abundant mutations at both time points. See **Online Methods** for a description of the the two samples taken from this patient.



Supplementary Figure 7: Deep population sequencing has low false positive rate and higher false negative rate. (a) We compared our approaches by mixing 20 isolates *in silico,* taking the same number reads from each of 20 isolates. We performed population deep sequencing analysis on this mixture of single-end reads, using the same isogenic control without paired-end information. All positives in the deep sequencing that were also found in the single colony isolates (no false positives). False negative positions (blue circles) fail the strand-bias filter or other quality filters (not shown). Jitter is added on the X-axis to improve visibility. (**b)** The proportion of positions called in the isolates not called in the *in silicio* deep sequencing (false negatives). 91% of positions called in the isolates were also called in deep sequencing (black circles).



Supplementary Figure 8: Paired-end information and Nextera amplification do not significantly affect polymorphism detection. (a) To investigate the effect of paired-end information on our results, we re-ran all population deep sequencing analyses, treating each read from a pair independently. For each genomic position, the mutation frequency from each approach is shown. Gray circles indicate positions called by both methods, blue circles indicate positions that did not pass all filters in the paired approach and red circles, and red circles indicate positions that did not pass all filters in the unpaired approach. 94% of positions called in the paired approach are also called in the unpaired approach and 98% of reads called in the paired approach are called in the paired approach. Bowtie2 sometimes fails to align unpaired reads well near short indels, and these positions have low coverage in the unpaired approach. Consistent with the fact that the paired approach discards pairs of reads when only one of the reads has poor sequencing quality (Illumina provided Phred scores), most of the discrepancy is attributable to noise around thresholds. The list of genes under selection within patients is identical using both approaches. **(b)** To understand the maximum possible frequency of errors introduced early during the 9-cycle PCR step of Nextera preparation, we performed simulations of PCR (described in **Supplementary Note**). Each simulation represents a unique single-nucleotide genomic position, and we ran 1 million simulations for each of 3 values of initial templates. We plot the histograms of final mutation frequency after 9 cycles for all genomic positions on a log scale (simulations with no mutations are not shown).

Supplementary Table 1: Patient information at time of sample collection.

Patient	Years since acquisition based on clinical data	Sample Name	Analysis performed	FEV1 (% Predicted)	Approximate *B. dolosa* density in sputum sample (CFU/mL)*	Previous antibiotics (within 30 days)	Antibiotics at sample collection
1	7	P1	Colony re-sequencing (29 isolates) AND Deep population sequencing	57	4 x 10^{7}	None	None
2	8	P2	Deep population sequencing	57	6 x 10^{7}	Aztreonam (inhaled)	Aztreonam (inhaled)
		P2T (14 days after P2)	Deep population sequencing	58	4 x 10^{8}	Aztreonam (inhaled) Ceftazidime Minocycline Ciprofloxacin	Ceftazidime Minocycline Ciprofloxacin
3	9	P3	Deep population sequencing	61	4 x 10^{7}	Aztreonam (inhaled) Levofloxacin Minocycline Trimethoprim/Sulfamethoxazole Meropenem Ceftazidime	Aztreonam (inhaled) Levofloxacin Minocycline Chloramphenicol Meropenem
4	8	P4	Deep population sequencing	32	4 x 10^{8}	Levofloxacin Trimethoprim/Sulfamethoxazole Azithromycin	Azithromycin
5	9	P5	Deep population sequencing	57	5 x 10^{8}	Levofloxacin Minocycline Ceftazidime (inhaled) Azithromycin	Levofloxacin Minocycline Ceftazidime (inhaled) Azithromycin

*Colony forming units (CFU) per mL of frozen sample was calculated by serial dilution and plating.

Supplementary Table 2: Genes with multi-diverse fingerprints for selection in one or more patients.

Predicted biological role	Gene number	Reference genome annotation	Annotated homolog [organism]*	Notes	CELLO[2] predicted localization	Patients mutated in [Mutations]**
Antibiotic resistance	*BDAG_ 01166*	D-alanyl-D-alanine carboxypeptidase	***pbpG*** [*Rubrivivax gelatinosus* IL144]	PBP7; Peptidoglycan biosynthesis; Beta-lactam resistance[3]; general stress[4].	Periplasmic	P1 [Q115*, S157A, N305K]
Antibiotic resistance	*BDAG_ 02180*	DNA gyrase subunit A	***gyrA*** [*Burkholderia gladioli* BSR3]	Fluroquinolone resistance; mutations in drug binding sites.	Cytoplasmic	P1 [A84P, T83K, T83M], P2 [*T83K*] P3 [*D87Y*], P4 [D87Y, T83K], P5 [*T83M*]
Outer membrane synthesis	*BDAG_ 00856*	hypothetical protein	***lptG*** *[Ralstonia sp.* PBA]	Permease YjgP/YjgQ family; LPS transport to the outer membrane[5].	Cytoplasmic Membrane	P1 [R52S, T96P], P2 [R268C] P3 [G323R], P4 [E63G, F306V]
Outer membrane synthesis	*BDAG_ 02321*	Glycosyltransferase	***wbpX*** [*Pseudomonas aeruginosa* PA7]	LPS synthesis[6].	Cytoplasmic	P2 [H548R, L527R]
Outer membrane synthesis	*BDAG_ 02311*	4-hydroxybenzoate polyprenyltransferase	***noeC*** [*Azorhizobium caulinodans* ORS 71]	D-arabinosylation; homologs *M. tuberculosis* and *P. aeruginosa* involved in cell-wall synthesis and pili glycosylation, respectively; adjacent to O-antigen biosynthesis genes here and in other organisms[7].	Cytoplasmic Membrane	P1 [W162*], P3 [A202A, S472R]
Iron scavenging	*BDAG_ 03997*	Outer membrane receptor protein	***huvA*** [*Vibrio anguillarum*]	Hemin transport system; Iron-regulated outer membrane heme receptor[8]; close homolog to several TonB-dependent heme receptors.	Outer membrane	P4 [V1M, L135R]
Iron scavenging	*BDAG_ 01606*	Outer membrane receptor protein	***orbA*** *[Burkholderia multivorans CF2]*	Siderophore receptor required for ferric ornibactin uptake[9]; mutations focused conserved barrel structure.	Outer membrane	P1 [M607I, D659N], P2 [*G547R*]
Lactate utilization	*BDAG_ 02124*	hypothetical protein	***lutC*** *[Bacillus licheniformis DSM 13]*	Iron-sulfur containing protein involved in lactate utilization[10]; also implicated in biofilm formation[11].	Cytoplasmic	P3 [A12E, E50A, G213G]
Oxygen-related gene regulation	*BDAG_ 01161*	PAS	***fixL*** [*Burkholderia rhizoxinica* HKI 454]	Two component system histidine kinase containing PAS domain, oxygen sensing[12]; also homologous to *P. aeruginosa bfiS*, biofilm development[13]; most mutations in or near heme binding pocket.	Cytoplasmic Membrane	P1 [M338T, V374G, W562L], P2 [R258C, M346I], P3 [*D445H*], P4 [M443R]
Unknown gene regulation	*BDAG_ 01528*	sigma54 specific transcriptional regulator, Fis family	YP_83538 [*Burkholderia cenocepacia* HI2424]	Transcriptional regulator containing PAS domain. Half of mutations focused in and near heme pocket, remaining near putative sigma54-interaction domain.	Cytoplasmic	P2 [V201L, S266A, A280V], P3 [L100R], P5 [A81V, H104R, E115A, A239T]
Unknown gene regulation	*BDAG_ 02877*	DNA-directed RNA polymerase sigma subunit	***rpoD*** [*Burkholderia cenocepacia* AU1054]	Homologous to primary sigma factor *rpoD*; mutated during experimental evolution of *B. cenocepacia*[14]*;* comparative analysis suggests *B. cepacia*[15] complex species have multiple recently evolved alternative primary sigma factors.	Cytoplasmic	P1 [A107G, T124P], P4 [V109V, M123V], P5 [*A87A, A95T*][+]
Stringent Response	*BDAG_ 02219*	Guanosine polyphosphate pyrophosphohydrolase	***spoT*** [*Burkholderia ambifaria* AMMD]	(p)ppGpp metabolism; Role in virulence in *B. psuedomallei*[16]*.*	Cytoplasmic	P3 [S412L, I650L], P5 [R257H]
Arginine biosynthesis	*BDAG_ 01143*	Acetylglutamate kinase	***argA*** [*Cupriavidus necator* N-1]	Arginine biosynthesis.	Cytoplasmic	P1 [L96V, E296K], P5 [*R420H*]
Unknown	*BDAG_ 00993*	Glucoamylase	*ZP_02893540* [*Burkholderia ambifaria* IOP40-10]	Glycoside hydrolase. Trehalose synthesis.	Cytoplasmic	P4 [R77L, W276*, W299L, E403K]
Unknown	*BDAG_ 01476*	hypothetical protein	***rodZ*** [*Edwardsiella ictaluri* 93-146]	Homology to *E. coli rodZ* (component of bacterial cytoskeleton[17]) is weak and is an region that covers 1/3 of the gene.	Periplasmic	P2 [P54L, S123L]
Unknown	*BDAG_ 00061*	Hypothetical protein	YP_367612.1 [*Burkholderia sp.* 383]	Homologs in other *Burkholderia* but not many other genuses; no domains with known function.	Periplasmic	P3 [G95W, Q150*]

*Bolded gene names indicate that a reverse-BLAST suggests that the most homologous *B. dolosa* gene to the annotated gene is the queried gene (Methods). When database searches did not offer a candidate annotated homolog, the best BLAST hit is shown.

**Mutations are fixed in the patient's population if italicized, polymorphisms otherwise.

[+]PATRIC[18] suggests that this gene is likely misannotated, as homology to other *Burkholderia rpoD* genes starts at amino acid 92. There is a mutation prior to this, in Patient 5, which is likely noncoding but treated here as synonymous to remain systematic. Amino acids numbers here are relative to the genebank entry for BDAG_02877.

Supplementary Table 3: Operons with multi-diverse fingerprints for selection in one or more patients.

Predicted biological role	Chromosome	Operon start*	Operon end*	Genes in operon mutated in at least one patient	Annotated homolog [organism]**	Patients mutated in [Mutations]***	Notes
Outer membrane synthesis	NZ_CH482380.1	729604	733200	*BDAG_00576*	***lptB*** [*Burkholderia sp.* KJ006]	P1 [D68A] P5 [*E243Q*]	LPS transport to the outer membrane[5].
				BDAG_00577	***lptA*** [*Burkholderia multivorans* ATCC 17616]	P1 [V70A]	
Outer membrane synthesis	NZ_CH482380.1	1052795	1055573	*BDAG_00856*	***lptG*** [*Ralstonia sp.* PBA]	P1 [R52S, T96P], P2 [R268C] P3 [G323R], P4 [E63G, F306V]	LPS transport to the outer membrane[5].
				BDAG_00857	***lptF*** [*Cupriavidus necator* N-1]	P3 [P289L], P4 [A154E]	
Unknown; implicated in various virulence-related pathways	NZ_CH482381.1	1240739	1242656	*BDAG_03702*	***rstB*** [*Serratia proteamaculans* 568]	P1 [R222C], P3 [R273P]	Two component histidine kinase and response regulator; *rtsAB* has been implicated as targets of the divalent cation sensing PhoQP system[19], though their targets and triggers are unknown: homologs of *rtsAB* are implicated in iron transport[20], biofilm-formation[21], degradation of virulence-related sigma factor RpoS[22], and acid shock and curli production[23].
				BDAG_03703	***rstA*** [*Burkholderia multivorans* ATCC 17616]	P3 [K108M], P4 [L176F]	

Operons that were detected exclusively on the basis of mutations focused in a single gene are not listed here and can be found in **Supplementary Table 2.**

*Determined by FgenesB
**Bolded gene names indicate that a reverse-BLAST suggests that the most homologous *B. dolosa* gene to the annotated gene is the queried gene (Methods). When database searches did not offer a candidate annotated homolog, the best BLAST hit is shown.
***Mutations are fixed in the patient's population if italicized, polymorphisms otherwise.

Supplementary Table 4: Coverage statistics

Sample	Type of reads	Percent of filtered reads aligned	Percent of reference genome callable*	Average coverage
Isogenic control	50bp paired end	94.8%	93.2%	708
P1	50bp paired end	96.3%	91.3%	582
P2	50bp paired end	95.3%	89.5%	489
P2T	50bp paired end	96.1%	90.7%	509
P3	50bp paired end	95.9%	89.8%	420
P4	50bp paired end	95.4%	89.8%	401
P5	50bp paired end	96.1%	89.1%	323
P1-01	50bp single end	96.5%	94.8%	36
P1-02	50bp single end	96.7%	94.7%	37
P1-03	50bp single end	96.7%	94.4%	34
P1-04	50bp single end	96.4%	94.7%	23
P1-05	50bp single end	96.5%	94.8%	37
P1-06	50bp single end	96.7%	93.3%	43
P1-07	50bp single end	96.9%	94.7%	51
P1-08	50bp single end	96.5%	94.8%	26
P1-09	50bp single end	96.9%	94.7%	40
P1-10	50bp single end	96.8%	94.4%	40
P1-11	50bp single end	96.8%	94.8%	60
P1-12	50bp single end	96.7%	94.8%	28
P1-13	50bp single end	96.6%	94.8%	36
P1-14	50bp single end	96.6%	94.4%	32
P1-15	50bp single end	96.6%	94.8%	29
P1-16	50bp single end	96.7%	94.7%	40
P1-17	50bp single end	96.8%	94.8%	48
P1-18	50bp single end	96.7%	94.7%	39
P1-19	50bp single end	96.6%	94.7%	25
P1-20	50bp single end	96.7%	94.8%	26
P1-21	50bp single end	96.5%	94.7%	34
P1-22	50bp single end	96.6%	94.7%	36
P1-23	50bp single end	96.8%	94.7%	48
P1-24	50bp single end	96.7%	93.4%	38
P1-25	50bp single end	97.1%	94.7%	39
P1-26	50bp single end	96.4%	94.6%	32
P1-27	50bp single end	96.9%	94.8%	37
P1-28	50bp single end	96.7%	94.8%	40
P1-29	50bp single end	96.4%	94.7%	25

*For isolates, callable positions are those with a consensus quality score (provided by samtools) score below -40. For population sequencing, callable positions are those that met coverage, base quality, mapping quality, and tail distance thresholds and for which the isogenic control had a major allele frequency of at least 98.5%

Supplementary Note

Sample collection

Expectorated sputum samples were collected at Boston Children's Hospital after written informed consent was obtained under protocols approved the Institutional Review Boards at both Boston Children's Hospital and Harvard Medical School. Samples were immediately placed on ice and then liquefied with dithiothreitol. 10-15 mL phosphate buffered saline containing 1mM of dithiothreitol was added to each sample. Each sample was incubated on ice for 1 hour, vortexing every 20 minutes. 50% glycerol was added to each sample to a final concentration of 20%, and samples were then frozen at -80°C.

Sample prep, colony re-sequencing approach

Using a sterile 1-microliter plastic loop, ice was scraped from the top of the frozen homogenized sputum from Patient 1 and streaked onto the *Burkholderia cepacia* complex specific media, OFPBL (oxidation-fermentation basal medium supplemented with polymyxin B, bacitracin, lactose, and agar, BD diagnostic, USA). Individual colonies were picked into individual culture tubes containing 10mL of LB. Cultures were incubated with shaking at 37° C for 20 hours. Aliquots from these cultures were used to make a frozen library in 15% glycerol in a microtiter plate in triplicate and frozen at -80° C for further use. Additionally, 1.8mL of this overnight culture was used for genomic DNA extraction.

Sample prep, deep population sequencing

Frozen sputum samples were thawed on ice. A 10-fold serial dilution was performed in PBS, and 0.8 mL of each dilution was plated on OFPBL using a disposable plastic spreader. After allowing the plates to dry, plates were incubated at 37°C for 48 hours. Variation in colony size was observed within each sample. This variation may cause differences between allele frequencies measured and *in vivo* allele frequencies. Our results are not very sensitive to such differences; the signature for selection reported does not depend on allele frequency (so long as mutations are above 3%) and all lineages, both those underrepresented and overrepresented, have been accumulating mutations since their LCA according to the molecular clock.

For each sample, a dilution plate was selected for harvesting which had between 5,000 and 30,000 colonies, such that the number of colonies was maximized while limiting competition with nearby colonies. From this plate, 2 mL of PBS was added, cells were scraped with a plastic loop, and transferred to a microcentrifuge tube. A 0.5 mL aliquot of each sample was frozen in 15% glycerol at -80°C for future use, and the remainder was pelleted and used for DNA extraction. For the isogenic control, the same procedure was used with the exception that the serial dilution was made starting from a single colony taken from a plate from Patient 1's sputum sample.

Initial data processing workflow for colony re-sequencing

We used custom MATLAB scripts to pipe together cutadapt[24] (remove adapter read-through), sickle[25] (trim low quality bases from reads), bowtie2[26] (align reads), and SAMtools[27] (call potential variants) and run them in parallel for many isolates on the Orchestra shared research cluster at Harvard Medical School. The following options were used:

```
>   cutadapt -a CTGTCTCTTATACACATCTCTGA reads_1.fastq > trimmed_reads_1.fastq
>   sickle se -f trimmed_reads_1.fastq -o filtered_reads_1.fastq  -s singles.fastq -q 20 -l 25
>   bowtie2 -X 2000 --no-mixed --very-sensitive --n-ceil 0,0.01 --un-conc unaligned.fastq -x
    refgenome_bowtie2  -U filteredreads.fastq -S aligned.sam
>   samtools view -bS -o aligned.bam aligned.sam
>   samtools sort aligned.bam aligned.sorted
>   samtools mpileup -q30 -S -ugf refgenome.fasta aligned.sorted.bam > sample
>   bcftools view -g sample > sample.vcf
>   bcftools view -vS sample.vcf > variant.vcf
```

Mutations were called using custom MATLAB scripts and the vcf files.

Initial data processing workflow for deep population sequencing

We used custom MATLAB scripts to pipe together cutadapt, sickle, bowtie2, and SAMtools and run them in parallel for many isolates on the Orchestra shared research cluster at Harvard Medical School. The following options were used:

```
> cutadapt -a CTGTCTCTTATACACATCTCTGA reads_1.fastq > trimmed_reads_1.fastq
> cutadapt -a CTGTCTCTTATACACATCTCTGA reads_2.fastq > trimmed_reads_2.fastq
> sickle pe -f trimmed _reads_1.fastq -r trimmed _reads_2.fastq -o filtered_reads_1.fastq -p filtered_reads_2.fastq -s singles.fastq -q 20 -l 50
> bowtie2 -X 2000 --no-mixed --very-sensitive --n-ceil 0,0.01 --un-conc unaligned.fastq -x refgenome_bowtie2 -1 filteredreads_1.fastq -2 filteredreads_2.fastq -S aligned.sam
> samtools view -bS -o aligned.bam aligned.sam
> samtools sort aligned.bam aligned.sorted
> samtools mpileup -q30 -s -O -B -d3000 –f refgenome.fasta aligned.sorted.bam > sample.pileup
> samtools mpileup -q30 -S -ugf refgenome.fasta aligned.sorted.bam > sample
> bcftools view -g sample > sample.vcf
> bcftools view -vS sample.vcf > variant.vcf
```

The strict removal of all trimmed reads in sickle (-l 50 option) was used to make comparing tail distances across reads comparable. Custom MATLAB scripts were used to call diverse positions using the sample.pileup file. Fixed positions were called using custom MATLAB scripts and the vcf files.

Polymorphic mutation calling (deep population sequencing)

Using our isogenic control as an approximate negative control, multiple isolates from Patient 1 as an approximate positive control, and an interactive MATLAB environment that enabled investigation of the raw data, we developed a set of filters to identify polymorphic positions with minor allele frequency above 3%. We set the thresholds for these filters conservatively, minimizing false positives. An *in silico* mixing of reads from the isolates calls no positions not detected in the isolates and calls 78% of positions found when treating isolates separately (**Supplementary Figure 7)**. Similarly, 85.6% of positions called polymorphic above 3% frequency in Patient 1 using the pooled approach were also detected in the isolates (compared to 91.5% expected due to binomial sampling). No position is detected at greater than 13.5% frequency in population sequencing that is not detected in the isolates. All fixed and polymorphic mutations found and their frequencies are listed **Supplementary Table 6.**

We considered a position to be polymorphic if it met the following quality thresholds in the given sample:

- **Minor allele frequency:** More than 3% of reads support a particular minor allele
- **Overall coverage:** At least 15 reads align in both the forward and reverse direction, and the total number of reads aligning is below the 99th percentile of covered positions in that sample.
- **Minor allele coverage:** At least 3 reads per major and minor allele aligning in both the forward and reverse direction (4 thresholds: forward minor, forward major, reverse minor, reverse major)
- **Base quality:** Average base quality (provided by sequencer) of greater than 19 for both the major and minor allele calls on both the forward and reverse strand
- **Mapping quality:** Average mapping quality (provided by aligner) of greater than 34 for reads supporting both the major and minor allele on both the forward and reverse strand
- **Tail distance:** Average tail distance of between 6 and 44 for reads supporting both the major and minor allele on both the forward and reverse strand
- **Indels:** Fewer than 20% of the reads aligning to that position support an indel at any position along that read
- **Strand bias:** A P-value of > 10^{-5} supporting a null hypothesis that the minor allele frequency is the same for reads aligning to both the forward and reverse strand (Fisher's exact test).
- **Tail distance bias:** P-values of > 10^{-5} supporting null hypotheses that the tail distances come from the same distribution for both the minor and major allele, for both the forward and reverse strand (t-test)

- **Isogenic control:** More than 98.5% of reads aligning to this genomic position in the isogenic control support a major allele

These filters remove false positive polymorphisms, which are not caused from randomly distributed sequencing error, but rather from systematic errors at particular genomic positions. For example, false positive polymorphisms occur from misalignment near small insertions and deletions, from recent duplications represented once in the reference genome, and from neighboring sequences that increase the probability of sequencing error. Most systematic errors have hallmarks in individual reads data. Errors induced by the nearby DNA sequence, for example, will produce polymorphism on reads aligning in either the 5'->3' or 3'->5' direction, but not both. We do not consider regions with very high coverage, which may reflect recent duplications or older duplications that are only listed once in the draft genome. We use the paired end information only to discard discordant read pairs and find that we would get similar results even without this information (see **Supplementary Fig. 8a**).

We find that false positive polymorphisms tend to be repeatable, showing nearly identical frequency and nucleotide identity across samples. Some genomic positions show repeatable polymorphism across samples and our isogenic control despite not having a hallmark for false-positive in the individual read data; such positions are removed by the requirement that the position be isogenic in the isogenic control (10-15 genomic positions per sample).

Estimation of false-positives from PCR amplification

While the Nextera kit involves PCR amplification, the large amount of template used and the limited number of cycles should prevent errors from PCR amplification to reach near the 3% minor allele frequency threshold.

Simulation: To understand the maximum possible frequency of errors introduced early during the 9-cycle PCR step of Nextera preperation, we performed simulations of PCR. We conservatively assume that all PCR errors create the same mutated nucleotide, we assume perfect amplification, and we model a single genomic position (nucleotide) at a time. We assume a uniform error rate across the genome of $3.3 *10^{-7}$, the error rate provided by Epicentre. During each cycle of the simulation, new mutated molecules are introduced according to a Poisson process with mean (numberOfMolecules * polymeraseErrorRate), the number of molecules doubles, and the number of previously mutated molecules doubles. Another parameter is the number of initial molecules covering each nucleotide. Larger numbers of initial molecules result in more buffering of PCR errors. Given that the final concentration of DNA from the PCR reaction is 300ng and the genome size is 6.4 Mb, even coverage predicts that there are $8.7x10^{4}$ copies of each nucleotide in the initial pool. Because genomic positions are certainly not represented evenly in this initial pool, we ran sets of simulation simulations using 10000, 1000, and 100 initial molecules (regions with lower abundances will not have enough coverage in the final library to meet the coverage threshold). For each number of initial molecules, we simulate 10^{6} nucleotide positions, a number larger than the number genomic positions with low representation in the initial pool. We find the maximum final frequency of mutation in 10^{6} simulations with 100 initial molecules is .1%, much below our detection threshold of 3%. Simulations with more starting molecules have even lower error. See **Supplementary Fig. 9b.**

Empirical analysis of isogenic control: We have also empirically estimated sample-prep error based on the isogenic control data. Positions introduced by PCR error will not be filtered out by the quality filters described in the Methods (all filters except for the filter which specifically depends on isogenic control). Thus, if PCR introduces false polymorphisms, we should see them in the isogenic control. 10 genomic positions in the isogenic control pass all quality filters. Some of these positions appear because of recent duplications (coverage relative to nearby regions), while others may be PCR error. Of these 10, 9 show consistent polymorphisms across all samples and the control (same nucleotides and approximate frequency), indicating that when PCR error emerges, it can easily be filtered out by the isogenic control because of reproducibility. Thus, we estimate 1 false positive polymorphic genomic position per sample introduced by PCR error per sample.

Genetic distance to LCA calculation (deep population sequencing)

Here, we show that the average number of mutations per cell in a population is equivalent to the sum of the mutation frequencies in that population. The number of mutations in a given cell can be represented as $\sum_0^P m_i$, where P is the number of positions on the genome, and $m_i = 1$ if a cell as a mutation at position i and $m_i = 0$ otherwise. Thus:

$$\langle \text{Number of muations per cell} \rangle >= \langle \sum_{i=1}^{P} m_i \rangle$$

Now, we consider a population of C cells. The presence of a mutation at position i in a cell j is now indicated by $m_{ij} = 1$.

$$\langle \sum_{i=1}^{P} m_i \rangle = \frac{\sum_{j=1}^{C} \sum_{i=1}^{P} m_{ij}}{C} = \frac{\sum_{i=1}^{P} \sum_{j=1}^{C} m_{ij}}{C} = \sum_{i=1}^{P} \frac{\sum_{j=1}^{C} m_{ij}}{C} = \sum_{i=1}^{P} f_i$$

Where f_i is the mutation frequency at position i on the genome.

Error sources in estimation of time to LCA

Potential sources of error in estimating the time to LCA include Poisson error in the number of mutations accumulated in each lineage since the LCA, underestimation due to limited sensitivity in detecting mutations, overestimation due to false positives, and underestimation due to incomplete sampling of the diversity in the lung. Additionally, possible errors in converting $<d_{LCA}>$ to years include potential overestimation of the molecular clock resulting in underestimation of time to LCA (if historical sampling in the clinic is biased towards colonies with faster growth rates) and deviations from clock-like evolution (e.g. extra doublings following antibiotic treatment). The confidence intervals of time to LCA presented for the colony re-sequencing approach are calculated according to a Poisson distribution. We assume that the number of mutations in each cell is drawn from a Poisson distribution with λ equal to the mean across cells. Poisson measurement error is minimal for the population sequencing, as we are sampling hundreds of lineages simultaneously. The differences in estimated time to LCA between the two samples taken from Patient 2 are larger than would be expected given only Poisson error, suggesting other factors contribute to our ability to date the LCA (**Supplementary Fig. 6b**).

Gene annotation

Genes were functionality annotated using a suite of online bioinformatics tools. Open reading frames were compared about RefSeq using NCBI's BLASTp and close homologs were scanned for gene names. For genes for which this did not indicate an obvious homolog, the Burkholderia Genome Database[28] and Microbial Genome Database for Comparative Analysis[29] (MBGD) were used to probe for candidate ortholog gene names and look for synteny (As *B. dolosa* is not in MBGD, homologs from other members of the *B. cepacia* complex were used to query MBGD).

When these searches produced candidate orthologs, reverse BLAST was used to test the orthology[30]; an ORF with that putative ortholog name from *E. coli* or other well-described non-*Burkholderia* genome was located in the NCBI gene database and used as a query for Blastp against the *B. dolosa* genome. If the original query came up as the best match and had an E value of < .001, this ortholog name was listed in **Figure 5e**. Otherwise, the locus tag was listed. Biological relevance was assigned using literature search and various databases, including WikiGenes[31], STRING[32], UniProt[33], and PATRIC[18]. Subcellular localization was predicted using CELLO[2]. Summary and gene-specific references can be found in **Supplementary Table 2**. Visualization for **Fig. 5e** was performed using Cytoscape[34].

Supplementary References

1 Felsenstein, J. PHYLIP-Phylogeny Inference Package (Version 3.2). *Cladistics* **5**, 164-166 (1989).

2 Yu, C. S., Chen, Y. C., Lu, C. H. & Hwang, J. K. Prediction of protein subcellular localization. *Proteins* **64**, 643-651, doi:10.1002/prot.21018 (2006).

3 Gomez, M. J. & Neyfakh, A. A. Genes involved in intrinsic antibiotic resistance of Acinetobacter baylyi. *Antimicrobial agents and chemotherapy* **50**, 3562-3567, doi:10.1128/AAC.00579-06 (2006).

4 Kenyon, W. J. *et al.* Sigma(s)-Dependent carbon-starvation induction of pbpG (PBP 7) is required for the starvation-stress response in Salmonella enterica serovar Typhimurium. *Microbiology* **153**, 2148-2158, doi:10.1099/mic.0.2007/005199-0 (2007).

5 Ruiz, N., Kahne, D. & Silhavy, T. J. Transport of lipopolysaccharide across the cell envelope: the long road of discovery. *Nature Reviews Microbiology* **7**, 677-683 (2009).

6 Rocchetta, H. L., Burrows, L. L., Pacan, J. C. & Lam, J. S. Three rhamnosyltransferases responsible for assembly of the A-band D-rhamnan polysaccharide in Pseudomonas aeruginosa: a fourth transferase, WbpL, is required for the initiation of both A-band and B-band lipopolysaccharide synthesis. *Molecular microbiology* **28**, 1103-1119 (1998).

7 Harvey, H., Kus, J. V., Tessier, L., Kelly, J. & Burrows, L. L. Pseudomonas aeruginosa D-arabinofuranose biosynthetic pathway and its role in type IV pilus assembly. *The Journal of biological chemistry* **286**, 28128-28137, doi:10.1074/jbc.M111.255794 (2011).

8 Mourino, S., Rodriguez-Ares, I., Osorio, C. R. & Lemos, M. L. Genetic variability of the heme uptake system among different strains of the fish pathogen Vibrio anguillarum: identification of a new heme receptor. *Applied and environmental microbiology* **71**, 8434-8441, doi:10.1128/AEM.71.12.8434-8441.2005 (2005).

9 Sokol, P. A., Darling, P., Lewenza, S., Corbett, C. R. & Kooi, C. D. Identification of a siderophore receptor required for ferric ornibactin uptake in Burkholderia cepacia. *Infection and immunity* **68**, 6554-6560 (2000).

10 Smaldone, G. T., Antelmann, H., Gaballa, A. & Helmann, J. D. The FsrA sRNA and FbpB protein mediate the iron-dependent induction of the Bacillus subtilis lutABC iron-sulfur-containing oxidases. *Journal of bacteriology* **194**, 2586-2593, doi:10.1128/JB.05567-11 (2012).

11 Snitkin, E. S. *et al.* Tracking a Hospital Outbreak of Carbapenem-Resistant Klebsiella pneumoniae with Whole-Genome Sequencing. *Science Translational Medicine* **4**, 148ra116-148ra116 (2012).

12 Crosson, S., McGrath, P. T., Stephens, C., McAdams, H. H. & Shapiro, L. Conserved modular design of an oxygen sensory/signaling network with species-specific output. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 8018-8023, doi:10.1073/pnas.0503022102 (2005).

13 Petrova, O. E. & Sauer, K. A novel signaling network essential for regulating Pseudomonas aeruginosa biofilm development. *PLoS pathogens* **5**, e1000668, doi:10.1371/journal.ppat.1000668 (2009).

14 Coutinho, C. P., de Carvalho, C. C., Madeira, A., Pinto-de-Oliveira, A. & Sá-Correia, I. Burkholderia cenocepacia phenotypic clonal variation during a 3.5-year colonization in the lungs of a cystic fibrosis patient. *Infection and immunity* **79**, 2950-2960 (2011).

15 Menard, A., de Los Santos, P. E., Graindorge, A. & Cournoyer, B. Architecture of Burkholderia cepacia complex sigma70 gene family: evidence of alternative primary and clade-specific factors, and genomic instability. *BMC genomics* **8**, 308, doi:10.1186/1471-2164-8-308 (2007).

16 Muller, C. M. *et al.* Role of RelA and SpoT in Burkholderia pseudomallei virulence and immunity. *Infection and immunity* **80**, 3247-3255, doi:10.1128/IAI.00178-12 (2012).

17 Alyahya, S. A. *et al.* RodZ, a component of the bacterial core morphogenic apparatus. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 1239-1244, doi:10.1073/pnas.0810794106 (2009).

18 Gillespie, J. J. *et al.* PATRIC: the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species. *Infection and immunity* **79**, 4286-4298, doi:10.1128/IAI.00207-11 (2011).

19 Eguchi, Y. *et al.* Signal transduction cascade between EvgA/EvgS and PhoP/PhoQ two-component systems of Escherichia coli. *Journal of bacteriology* **186**, 3006-3014 (2004).

20 Jeon, J. *et al.* RstA-promoted expression of the ferrous iron transporter FeoB under iron-replete conditions enhances Fur activity in Salmonella enterica. *Journal of bacteriology* **190**, 7326-7334 (2008).

21 Bilecen, K. & Yildiz, F. H. Identification of a calcium-controlled negative regulatory system affecting Vibrio cholerae biofilm formation. *Environmental microbiology* **11**, 2015-2029 (2009).

22 Cabeza, M. L., Aguirre, A., Soncini, F. C. & Véscovi, E. G. Induction of RpoS degradation by the two-component system regulator RstA in Salmonella enterica. *Journal of bacteriology* **189**, 7335-7342 (2007).

23 Ogasawara, H. *et al.* Genomic SELEX search for target promoters under the control of the PhoQP-RstBA signal relay cascade. *Journal of bacteriology* **189**, 4791-4799 (2007).

24 Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal* **17**, pp. 10-12 (2011).

25 Sickle (https://github.com/ucdavis-bioinformatics/sickle).

26 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature methods* **9**, 357-359 (2012).

27 Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).

28 Winsor, G. L. *et al.* The Burkholderia Genome Database: facilitating flexible queries and comparative analyses. *Bioinformatics* **24**, 2803-2804, doi:10.1093/bioinformatics/btn524 (2008).

29 Uchiyama, I., Higuchi, T. & Kawai, M. MBGD update 2010: toward a comprehensive resource for exploring microbial genome diversity. *Nucleic acids research* **38**, D361-365, doi:10.1093/nar/gkp948 (2010).

30 Wolf, Y. I. & Koonin, E. V. A tight link between orthologs and bidirectional best hits in bacterial and archaeal genomes. *Genome biology and evolution* **4**, 1286-1294, doi:10.1093/gbe/evs100 (2012).

31 Hoffmann, R. A wiki for the life sciences where authorship matters. *Nature genetics* **40**, 1047-1051, doi:10.1038/ng.f.217 (2008).

32 Szklarczyk, D. *et al.* The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic acids research* **39**, D561-568, doi:10.1093/nar/gkq973 (2011).

33 UniProt, C. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic acids research* **40**, D71-75, doi:10.1093/nar/gkr981 (2012).

34 Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* **13**, 2498-2504, doi:10.1101/gr.1239303 (2003).