PLOS PATHOGENS

RESEARCH ARTICLE

A genome-wide association analysis reveals a potential role for recombination in the evolution of antimicrobial resistance in *Burkholderia multivorans*

Julio Diaz Caballero[1], Shawn T. Clark[2,3], Pauline W. Wang[4], Sylva L. Donaldson[4], Bryan Coburn[5], D. Elizabeth Tullis[6], Yvonne C. W. Yau[3,7], Valerie J. Waters[8], David M. Hwang[2,3,9], David S. Guttman[1,4]*

1 Department of Cell and Systems Biology, University of Toronto, Toronto, Ontario, Canada, **2** Latner Thoracic Surgery Laboratories, University Health Network, University of Toronto, Toronto, Ontario, Canada, **3** Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Ontario, Canada, **4** Centre for the Analysis of Genome Evolution and Function, University of Toronto, Toronto, Ontario, Canada, **5** Division of Infectious Diseases, Department of Medicine, University Health Network, University of Toronto, Toronto, Ontario, Canada, **6** Adult Cystic Fibrosis Clinic, St. Michael's Hospital, Toronto, Ontario, Canada, **7** Department of Pediatric Laboratory Medicine, Division of Microbiology, The Hospital for Sick Children, Toronto, Ontario, Canada, **8** Department of Pediatrics, Division of Infectious Diseases, The Hospital for Sick Children, University of Toronto, Toronto, Ontario, Canada, **9** Department of Pathology, University Health Network, Toronto, Ontario, Canada

* david.guttman@utoronto.ca

Citation: Diaz Caballero J, Clark ST, Wang PW, Donaldson SL, Coburn B, Tullis DE, et al. (2018) A genome-wide association analysis reveals a potential role for recombination in the evolution of antimicrobial resistance in *Burkholderia multivorans*. PLoS Pathog 14(12): e1007453. https://doi.org/10.1371/journal.ppat.1007453

Editor: David Weiss, Emory University School of Medicine, UNITED STATES

Received: June 11, 2018

Accepted: November 2, 2018

Published: December 7, 2018

Data Availability Statement: NCBI BioProject ID, BioSample IDs, and Genbank Accession IDs for genomes of study isolates are available in S3 Table.

Funding: This research was funded by an Emerging Team Grant awarded to DSG from the Canadian Institutes of Health Research (CIHR) and Cystic Fibrosis Canada (CMF108027, www.cihr-irsc.gc.ca/e/193.html). JDC was supported by an Ontario Trillium Scholarship (www.sgs.utoronto.ca/currentstudents/Pages/International-Student-

Abstract

Cystic fibrosis (CF) lung infections caused by members of the *Burkholderia cepacia* complex, such as *Burkholderia multivorans*, are associated with high rates of mortality and morbidity. We performed a population genomics study of 111 *B. multivorans* sputum isolates from one CF patient through three stages of infection including an early incident isolate, deep sampling of a one-year period of chronic infection occurring weeks before a lung transplant, and deep sampling of a post-transplant infection. We reconstructed the evolutionary history of the population and used a lineage-controlled genome-wide association study (GWAS) approach to identify genetic variants associated with antibiotic resistance. We found the incident isolate was basally related to the rest of the strains and more susceptible to antibiotics from three classes (β-lactams, aminoglycosides, quinolones). The chronic infection isolates diversified into multiple, distinct genetic lineages and showed reduced antimicrobial susceptibility to the same antibiotics. The post-transplant reinfection isolates derived from the same source as the incident isolate and were genetically distinct from the chronic isolates. They also had a level of susceptibility in between that of the incident and chronic isolates. We identified numerous examples of potential parallel pathoadaptation, in which multiple mutations were found in the same locus or even codon. The set of parallel pathoadaptive loci was enriched for functions associated with virulence and resistance. Our GWAS analysis identified statistical associations between a polymorphism in the *ampD* locus with resistance to β-lactams, and polymorphisms in an *araC* transcriptional regulator and an outer membrane porin with resistance to both aminoglycosides and quinolones. Additionally, these three loci were independently mutated four, three and two times,

Awards.aspx). STC was supported by an Ontario Graduate Scholarship (osap.gov.on.ca/OSAPPortal/en/A-ZListofAid/PRDR015090.html). BC was supported by a CIHR Fellowship (www.cihr-irsc.gc.ca/e/193.html). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

respectively, providing further support for parallel pathoadaptation. Finally, we identified a minimum of 14 recombination events, and observed that loci carrying putative parallel pathoadaptations and polymorphisms statistically associated with β-lactam resistance were over-represented in these recombinogenic regions.

Author summary

Cystic fibrosis (CF) is the most common lethal genetic disorder affecting individuals of European descent. Most CF patients die at a young age due to chronic lung infections. Among the organisms involved in these infections are bacteria from the *Burkholderia cepacia* complex (BCC), which are strongly associated with poor clinical prognosis. This study examines how the most prevalent BCC species among CF patients, *B. multivorans*, evolves within a single CF patient by studying the first *B. multivorans* isolate recovered from the patient, one hundred isolates recovered over a one year period during the chronic infection phase, and an additional ten isolates recovered after the reinfection of the transplanted lungs. We found that *B. multivorans* diversify phenotypically and genetically within the CF lung over the course of the infection, and evolves into a complex population during the chronic infection phase. We found that isolates collected from the post-transplant reinfection were more closely related to descendants of the original isolate rather than those recovered in the chronic infection. We identify genetic variants statistically associated with resistance to the antibiotics, and showed that some of these variants were found in regions that show patterns of recombination (genetic exchange) between strains. We also found that genes which were mutated multiple times during overall infection were more likely to be found in regions showing signals consistent with recombination. The presence of multiple independent mutations in a gene is a very strong signal that the gene helps bacteria adapt to their environment. Overall, this study provides insight into how pathogens adapt to the host during long-term infections, specific genes associated with antibiotic resistance, and the origin of new and recurrent infections.

Introduction

The *Burkholderia cepacia* complex (BCC) describes a highly diverse group of at least 20 closely related species within the genus *Burkholderia* that can cause serious opportunistic infections in humans [1, 2]. Individuals with the fatal genetic disease cystic fibrosis (CF) are particularly susceptible to chronic BCC infections, which are commonly associated with rapid decline in lung function, high rates of mortality and poor post-transplant outcome [3, 4]. Of the BCC species, *Burkholderia multivorans* and *Burkholderia cenocepacia* account for 85–97% of all BCC found in CF patients [5]; however, *B. multivorans* infections have surpassed *B. cenocepacia* in prevalence over the past decade [6]. Many BCC that are CF-associated are intrinsically virulent and antibiotic resistant, and strict infection control practices are required since these bacteria can be transmitted between patients [7–10]. Despite a wealth of knowledge describing the molecular basis of these pathogenic properties and their evolution in strains of the well-studied *B. cenocepacia*, little is known about the factors that govern these attributes in *B. multivorans* [9].

Dissecting the molecular basis of complex adaptive traits in bacterial pathogens, such as antimicrobial resistance, can be difficult since a single phenotype may be influenced by a large number of loci that interact with each other as well as their environment. Resistance in the

BCC is associated with alterations to outer membrane permeability, the expression of multi-drug efflux pumps and β-lactamases, and diversification of antimicrobial targets [11]. Consequently, methods that focus on identifying polymorphisms in single genes with large effects may miss the majority of loci that modulate phenotypes in more subtle ways. The development of genome-wide association studies (GWAS) has expanded our ability to identify loci of small effect size that have been associated with numerous diseases and other related phenotypes of interest in humans [12, 13]. In contrast, the application of GWAS to analyze bacterial behaviors has been slower to gain traction for a number of inter-related reasons: 1) clonal reproduction of microbes leads to confounding associations due to common ancestry, often referred to as population structure; 2) recombination in bacteria, which is more analogous to gene conversion than eukaryotic recombination, occurs at variable rates among different species and is not linked to reproduction; 3) the unpredictable nature of recombination results in the erratic breakdown of linkage disequilibrium between selected sites and distal neutral sites; and 4) selection can be extremely strong, resulting in the relatively rapid fixation of not only a selected allele, but entire genomes due to the linkage disequilibrium [14, 15].

Despite the challenges inherent in bacterial GWAS, several approaches have recently been proposed. These methods include using cluster membership [16–18], phylogenetic history [15, 19, 20], or lineage effects [21] to differentiate mutations leading to a phenotypic outcome from mutations related to the genetic background of the bacterial population. While these methods hold tremendous promise for identifying genetic variation underlying bacterial phenotypes of interest, they generally focus on cross sectional sampling of diverse isolates and populations. Their power has not been established for the fine-scale analysis of individual bacterial populations evolving over short time scales, with strong positive selection and restricted recombination [14, 22]. The application of fine-scale evolutionary analysis to bacterial populations is especially important in the context of clinically significant pathogen infections, where evolution is associated with adaptation to the host environment and antimicrobial treatment [23].

In this study, we take a fine-scale approach to microbial GWAS to examine the genetic basis of antimicrobial resistance within a *B. multivorans* population that had been sampled longitudinally from a single patient over a ten-year period. We characterized the genomic diversity in this population and assessed associations between all genetic variants and multiple antibiotic resistance phenotypes. We used a clustering-based approach to control for population structure and linkage disequilibrium and identified single nucleotide polymorphisms (SNPs) that were associated with resistance to β-lactams, aminoglycosides, and quinolones. In addition, we found that both multi-mutated loci (those that are potential targets of parallel pathoadaptation) and β-lactam resistance-associated variants were overrepresented in recombinogenic regions of the *B. multivorans* genome.

Results

We used a series of *B. multivorans* isolates that were cultured from respiratory specimens obtained from one adult male with CF (patient CF170) being treated at the CF Clinic of St. Michael's Hospital, Toronto, Canada. In a ten-year period, patient CF170 acquired an incident (i.e. initial) lung *B. multivorans* infection, developed a chronic *B. multivorans* lung infection, received a double lung transplant, and finally experienced a *B. multivorans* re-colonization of the allograft three years post-transplant. Isolates from each of these three phases of his *B. multivorans* infection are represented in this study (Fig 1). We defined these isolates as 1) the single isolate recovered from the patient's first culture-positive sputum specimen–the 'incident infection' isolate; 2) 100 isolates collected six to seven years post-incident infection from ten sputum specimens (ten isolates per specimen) over approximately a one-

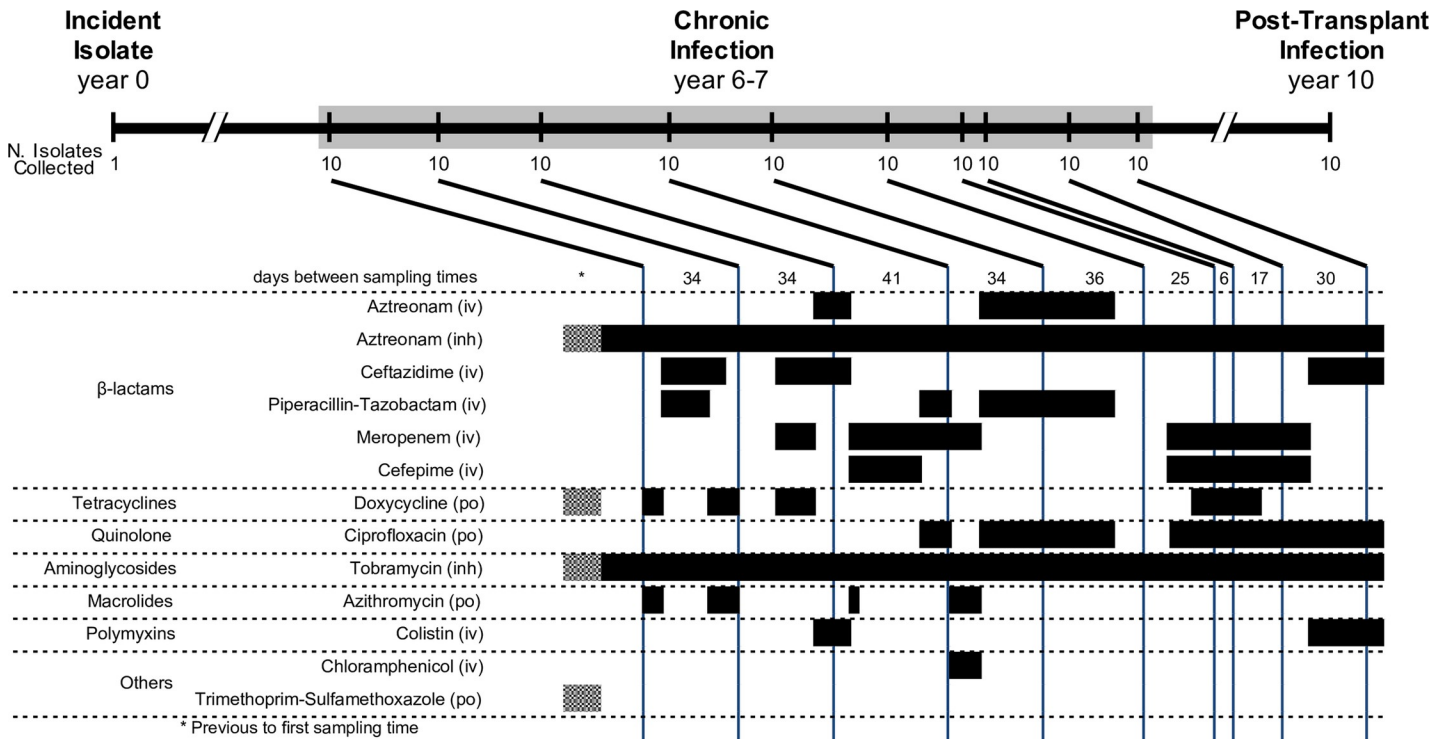

Fig 1. Time course of *B. multivorans* infection in study patient CF170. A total of 111 *B. multivorans* isolates from twelve collection times were used in this study (1 isolate from the initial infection, 10 isolates from each of 10 sputum samples collected during chronic infection, and 10 isolates from a sputum sample obtained during a post-transplant infection). Antibiotic treatment history during the chronic infection period is shown in the lower panel. Black bars indicate antibiotic administration, while hashed bars indicate intermittent exposure in that time block (only relevant prior to the start of chronic sampling). The method of antibiotic administration is shown as intravenous (iv), inhaled (inh), or oral (po).

https://doi.org/10.1371/journal.ppat.1007453.g001

year period–the 'chronic infection' isolates; and 3) ten isolates collected from a single expectorated sputum sample ten years after the incident infection, and three years after the patient underwent a double lung transplant–the 'post-transplant' isolates. Patient CF170 was being treated with alternating cycles of antibiotic therapy while chronically infected, with 13 antibiotics being administered at different intervals and durations over the course of the chronic infection sampling period (Fig 1). The genomes of all 111 isolates were whole-genome sequenced on the Illumina platform, yielding a median coverage depth of 117X (S1 Fig). Multi-locus sequence typing was performed *in silico* by extracting seven loci from the whole genome sequence data (*atpD*, *gltB*, *gyrB*, *recA*, *lepA*, *phaC*, *trpB*) and comparing them to the *Burkholderia cepacia* complex MLST Databases in pubMLST. This analysis revealed that all isolates were clonally related and of the sequence type ST-783 [24].

Genomic diversity and phylogenetic analysis suggest underlying population structure. The *de novo* genome assembly of a single isolate recovered from the third chronic infection sputum sample was used as the reference for the mapping assembly of all other isolates. This particular isolate was chosen as the reference since it had the best overall *de novo* assembly metrics. The reference assembly consisted of 6,444,123 bases across 26 contigs, which were pseudo-scaffolded against the complete genome of *B. multivorans* ATCC 17616 (as ordered in Fig 2A). Through a conservative variant calling pipeline [25], a total of 1,892 SNPs and 328 indels segregating among the 111 isolates were identified, with 1,039, 672, and 180 SNPs being found on chromosomes, 1, 2, and 3 respectively. Only a single SNP was found in a contig which did not map to the ATCC 17616 genome. Overall, 740 (39.1%) SNPs and 163 (49.9%)

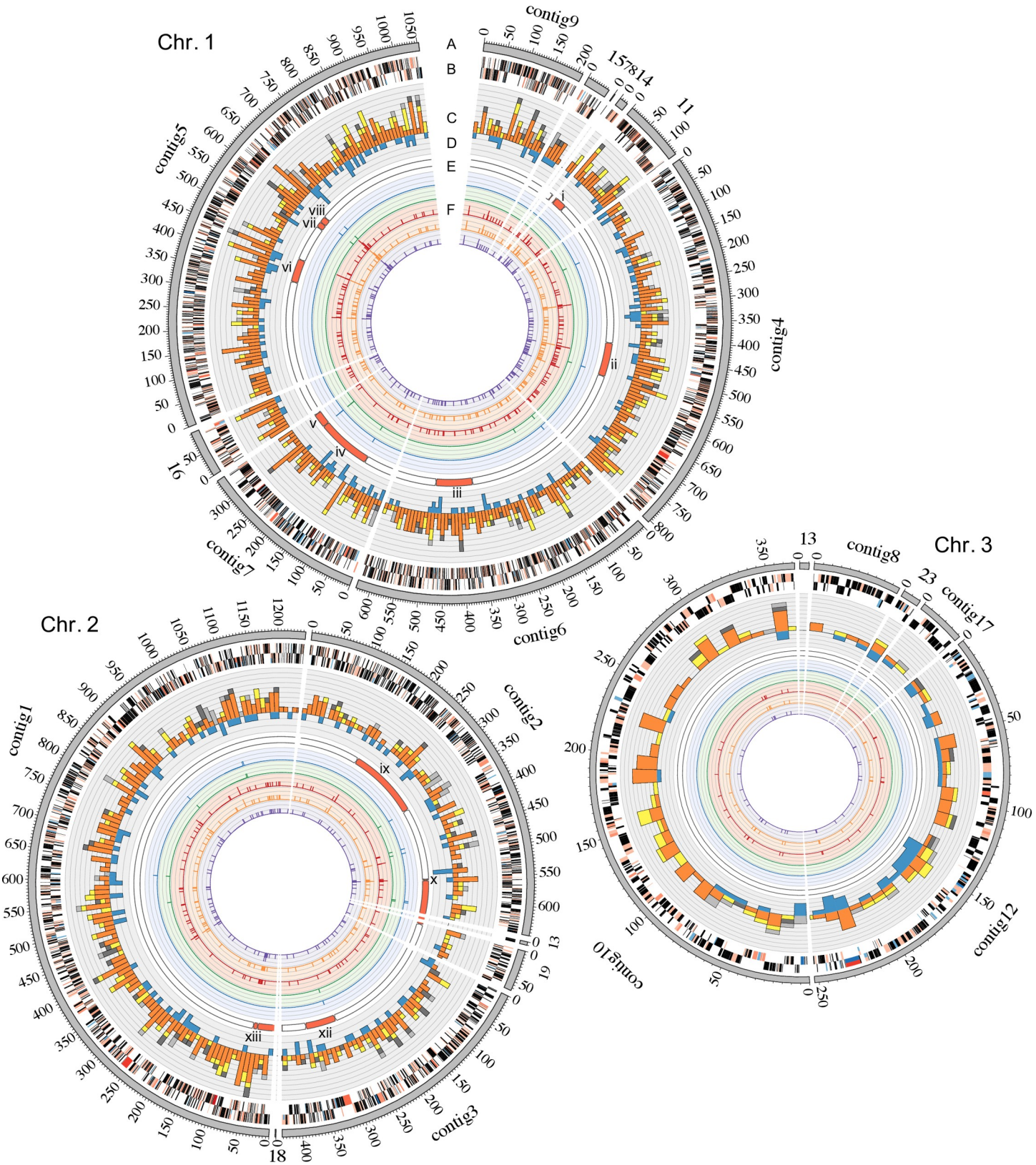

Fig 2. Genomic Characterization of 111 *B multivorans* isolates. (A) Contigs (gray outer ring) of the *de novo* reference were arranged according to the three chromosomes of the complete genome of *B. multivorans* ATCC 17616. This genome was obtained from expectorated sputum collected in the third chronic infection

sample. (B) Genome annotation according to RAST. (C) SNP count per 10 Kb as a function of their location in the contigs. Non-synonymous (orange), synonymous (yellow), putative regulatory (dark grey) and intergenic (light grey). (D) Indel (blue) count per 10 Kb. (E) Recombinogenic regions, as predicted by DnaSP Hudson-Kaplan four gamete test, are shown as red blocks. (F) Variants Associated with Antibiotic Resistance. From outermost to innermost ring: aztreonam and ceftazidime (β-lactam), amikacin and tobramycin (aminoglycoside), and ciprofloxacin (quinolone). This figure was prepared with circus v. 0.69 [90].

https://doi.org/10.1371/journal.ppat.1007453.g002

indels were parsimonious informative (PI, i.e. non-singleton), and 226 (11.9%) SNPs and 99 (30.2%) indels segregated in at least two sampling time points. From the 1,892 SNPs, 70.4%, 15.7%, and 13.9% were non-synonymous, synonymous, and intragenic substitutions respectively. 51.3% of the intergenic SNPs were found in putative regulatory regions (defined as the intergenic region within 150 bases from the start codon of any gene). The population showed a genetic diversity average of 123.62 ± 120.98 (number of SNP differences, mean ± standard deviation) pairwise differences. The distribution of these difference suggested an underlying population structure since genetic diversity was not uniform even among isolates from the same specimen (S2 Fig).

We reconstructed the core genome phylogenetic relationships among all isolates using an alignment of the 1,892 SNPs and Bayesian, maximum likelihood, and maximum parsimony approaches (Fig 3A and S3B and S3C Fig). All three methods gave consistent results. The root of the tree was identified by including *B. multivorans* ATCC 17616, *B. multivorans* BAA 247, *B. multivorans* AU1185, *B. multivorans* DDS 15A-1, and *B. mallei* ATCC 23344 in the phylogenetic analysis (S3A Fig). Additionally, these strains and our 111 isolates were placed in the phylogenetic context of other bacteria in the *Burkholderia* genus (S4 Fig). The tree topology indicates that the incident isolate diverged from the chronic and post-transplant isolates at the base of the tree. The ten isolates from the post-transplant sample are highly divergent (relative to the total diversity) and form a basally branching, monophyletic clade. The chronic infection isolates form a less divergent monophyletic clade, which diversified into subgroups. The same general structure is also observed in a network-based (i.e. neighbor-net) phylogenetic approach (S5 Fig), where two groups of chronic infection isolates cluster in a star-like phylogeny. Star phylogenies are characterized by roughly equal divergence from the common ancestor, and are associated with recent purges in genetic variation due to selective or demographic processes [26].

Population structure analysis clusters the isolates into five groups. We used the Monte Carlo Markov Chain analysis of SNPs and indels implemented in STRUCTURE to infer population structure among the 111 isolates [27]. We identified the lowest number of subpopulations that maximized the likelihood of data; hence determining the underlying population structure in the data without overestimating the number of subpopulations [28]. There were three subpopulations that arose from single common ancestors, which we labelled groups R, B, and G, comprising 54, 26, and 10 isolates, respectively (Fig 3C and 3D). The ancestral composition of the incident isolate and seven of the chronic infection isolates, recovered at collection points T1, T2 and T10, resembled a combination of the three identified subpopulations. This group of isolates was labeled RBG. Another group labeled RB (13 isolates) has an admixed ancestry from the ancestral subpopulations of R and B.

Isolates from groups RBG and RB were found in low frequencies through different samples from the chronic infection period (Fig 3B). In contrast, isolates from group R or B were more dominant in this same period. The isolates from group R were first observed at the third time point of the chronic infection samples, and they remained the most abundant group in subsequent chronic samples (Fig 4). In contrast, the abundance of group B isolates decreased over time. The genetic diversity, measured as number of SNPs, significantly differed between these groups (one-way ANOVA: $F(4,1902) = 1{,}426.133$, p-value < 0.0001), with group G (those

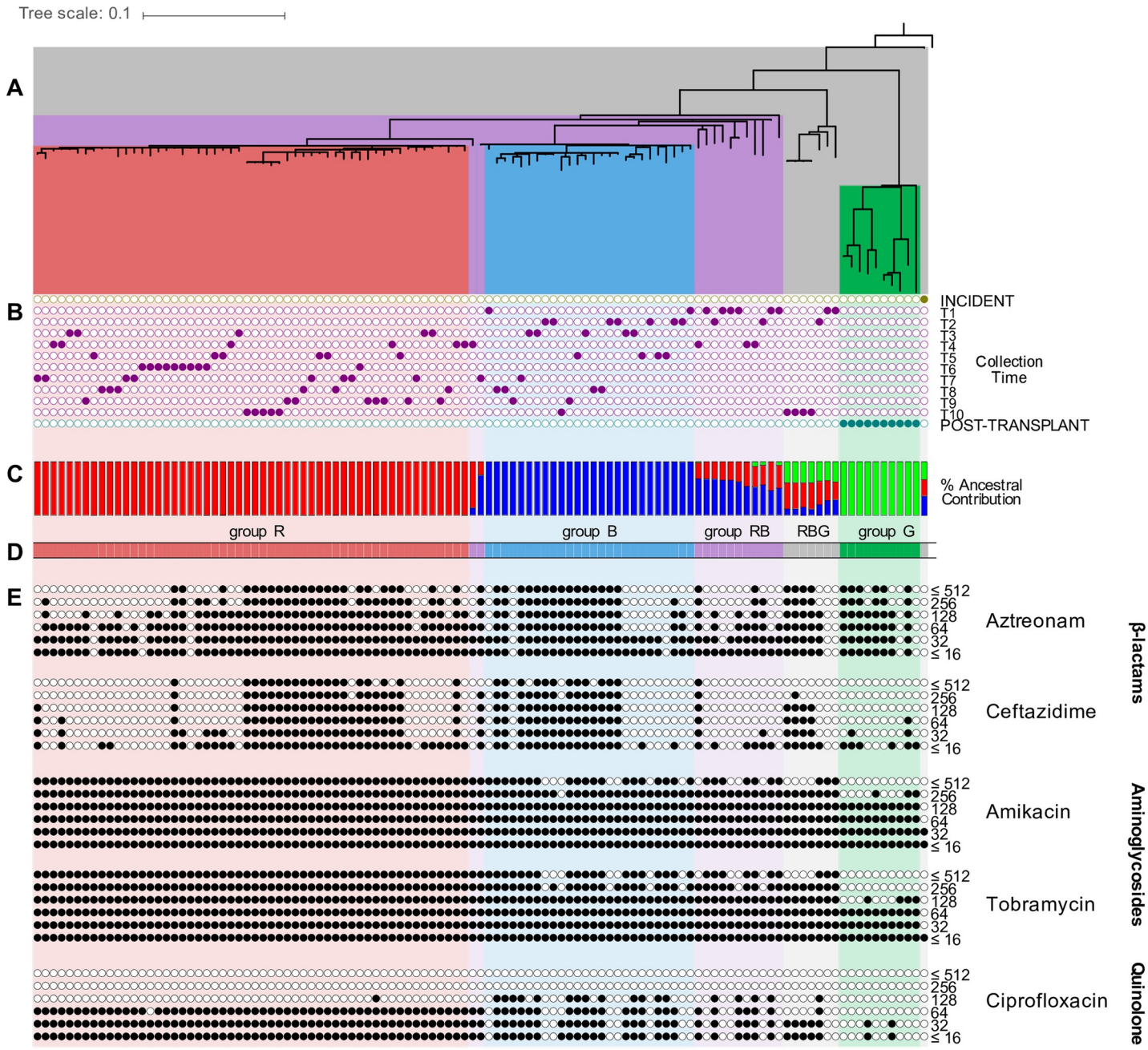

Fig 3. Population structure and antibiotic resistance profiles. (A) Phylogenetic relationships of the 111 *B. multivorans* isolates were estimated employing a Bayesian approach based on genome-wide single nucleotide polymorphisms (SNPs). (B) Time of collection for each isolate. (C) Population structure analysis as assessed by Structure v2.3.4 with three expected ancestral subpopulations. Ancestral subpopulations are coded as red (R), blue (B), and green (G). (D) Isolates are grouped based on their ancestral composition. Group R, B, G, RB, and RBG are shaded in red, blue, green, purple, and grey respectively. (E) Antibiotic susceptibility for each isolate, the highest black circle represents the MIC (μg/mL), to the β-lactams: aztreonam and ceftazidime, the aminoglycosides: amikacin and tobramycin, and the quinolone: ciprofloxacin are shown as filled circles at six different concentration thresholds. This figure was elaborated at the interactive tree of life (iTOL) website v. 3 [91].

https://doi.org/10.1371/journal.ppat.1007453.g003

recovered exclusively post-transplant) being the most diverse, followed by groups RBG and RB, then groups R and B (S6A Fig).

The time to the most recent common ancestor (tMRCA) calculated as days before the last sample for all isolates and the various STRUCTURE-defined groups is shown in Table 1. This

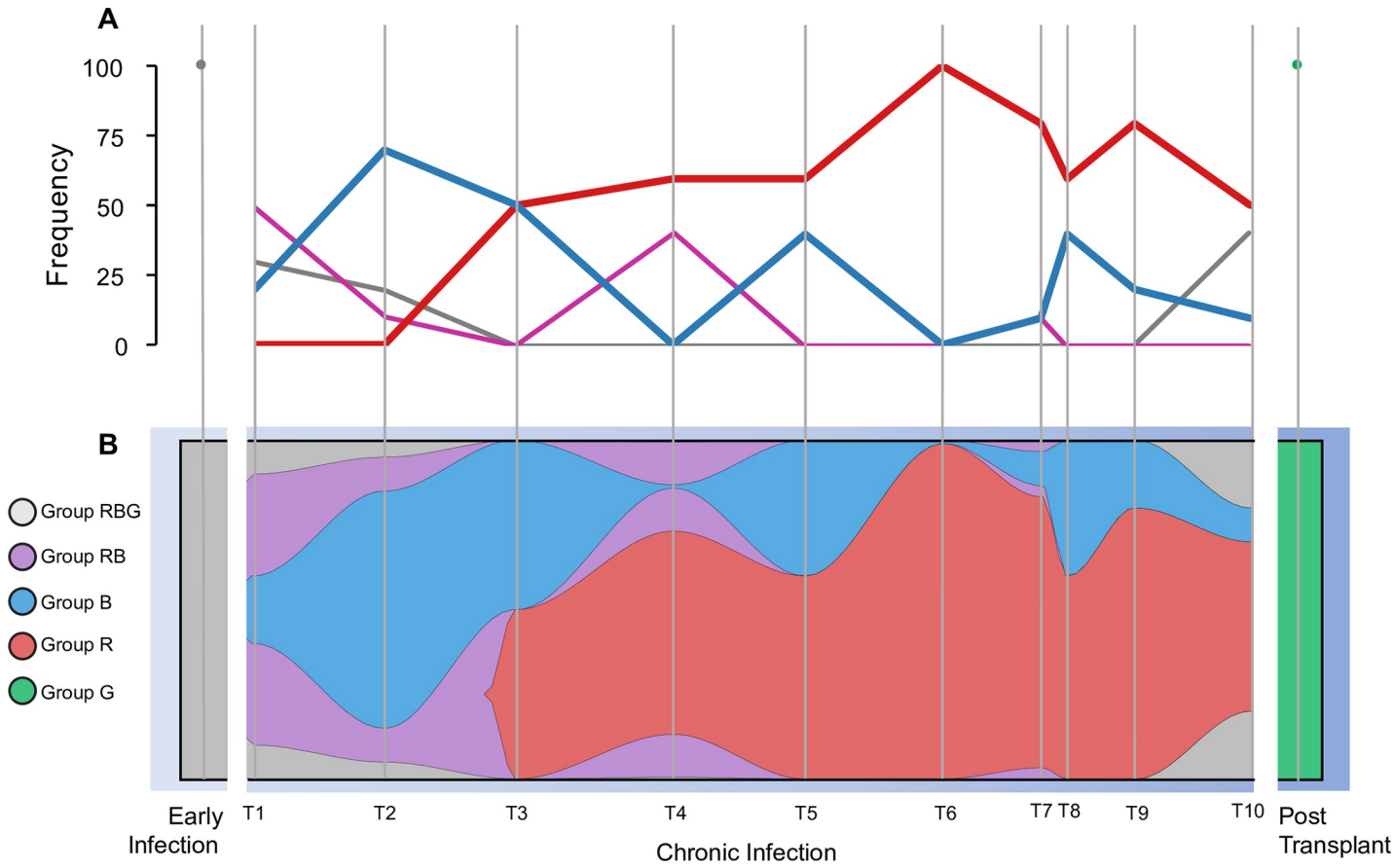

Fig 4. Population genomics of the community over time. Groups R, B, G, RB, and RBG are coloured in red, blue, green, purple, and grey respectively. (A) Frequency of each group over time. (B) The clonal graph was created with the assumption that RBG is the group of isolates resembling the ancestor of all the isolates, and RB is the group of isolates resembling the ancestor of group R and B. The distance between sample times is relative to the actual number of days between them. This plot was created using fishplot v. 0.3 [92].

https://doi.org/10.1371/journal.ppat.1007453.g004

analysis shows that the RGB group, which includes all of the chronic infection isolates as well as the post-transplant isolates, coalesced to a common ancestor at roughly the same time as the full isolate collection, including the incident infection (11.18 and 11.57 years before the final sample, respectively). This result supports the hypothesis that the infection of the transplanted lung originated from the same source as the incident isolate despite being separated by approximately ten years, as opposed to being derived from the chronic infection population. Additionally, groups R and B diverged at approximately the same time (3.38 and 3.61 years before the final sample, respectively). Unfortunately, we are unable to determine if these were

Table 1. Time to most recent common ancestry.

Group	tMRCA (years)	95% HDP Interval
All Strains	11.57	9.73–15.50
Group RBG	11.18	9.73–14.11
Group RB	4.86	4.01–5.93
Group G	2.53	1.90–2.99
Group B	3.61	3.45–3.80
Group R	3.38	3.25–3.53

https://doi.org/10.1371/journal.ppat.1007453.t001

allopatric populations colonizing distinct regions in the lung, or sympatric populations coexisting within the same compartment due to our sampling of expectorated sputum. The post-transplant reinfection population (group G) had a most recent common ancestor of 2.53 years before the final samples, which is consistent with the fact that patient CF170 underwent lung transplantation approximately three years before the end of the study (i.e. the final sample).

d_N/d_S estimates support positive selection in the population

We determined the ratio of non-synonymous to synonymous substitutions (d_N/d_S) as an estimate of selection. Since we expect that the importance of natural selection and/or genetic drift will be more accurately reflected on those SNP segregating in the population over multiple sampling time-points than on variants that segregate only in a single sample, we determined the d_N/d_S ratios both for all SNPs as well as for only those that segregate in at two or more time-points–'multi-time' SNPs (S6B Fig). The d_N/d_S for the overall population was 1.35 (95% confidence interval, CI = 1.19–1.53) and 1.34 for multi-time SNPs (CI = 0.94–1.96), which may indicate weak positive selection, or simply the segregation of mildly deleterious variants. Only groups R and RB multi-time SNPs showed d_N/d_S above the neutral expectation of 1.0 (group R d_N/d_S = 2.05, CI = 0.57–11.15, group RB d_N/d_S = 2.38, CI = 1.08–6.18), although the confidence intervals for the group R are quite large. All other groups had d_N/d_S ratios only slightly elevated (ranging from 1.04–1.63), although the differences between groups were not statistically significant.

Further support for positive selection comes from a significantly negative Tajima's D test (D = -2.21, $P < 0.01$) and Fu and Li's tests ($D^* = -6.11$, $P < 0.02$; $F^* = -5.20$, $P < 0.02$). While all three of these results can be explained by both positive selection and recent population expansion, the combination of these results with the high nucleotide diversity and $d_N/d_S > 1.0$ is most consistent with positive selection.

GWAS identification of variants associated with antibiotic resistance

We assumed that the intensive antibiotic exposure during the chronic infection sampling period would result in strong selection for resistance-associated genotypes in *B. multivorans*. Minimum inhibitory concentrations (MICs) for two β-lactams (aztreonam, ceftazidime), two aminoglycosides (tobramycin and amikacin), and the fluoroquinolone ciprofloxacin were determined for all isolates by agar dilution using Clinical and Laboratory Standards Institute procedures [29]. Isolates from the three phases of infection had distinct susceptibility profiles. The incident isolate had MICs of 8 μg/mL or less for all agents tested, while all chronic infection and post-transplant isolates had significantly higher MICs for both of the aminoglycosides tested (t-test $p < 0.0001$, Fig 3E), but variable MICs for the β-lactams and fluoroquinolone tested (range: ≤8 to >512 μg/mL).

The 1,892 SNP positions segregating among the 111 isolates were grouped in 150 distinct mutational profiles (i.e. one or more SNP positions that share the same pattern of reference vs. alternative base among the strain collection, S7 Fig). Prior to population control, each of these mutational profiles was examined for a statistical association to the five tested antibiotics at six different levels of resistance and these associations were Bonferroni corrected for multiple testing. Five mutational profiles (comprising 17 SNPs) were associated with resistance to both β-lactam antibiotics, and one mutational profile (comprising 2 SNPs) was associated specifically with ceftazidime (S8 and S9 Figs). Ten mutational profiles (comprising 250 SNPs) were associated with resistance to amikacin, tobramycin, and ciprofloxacin. Additionally, two mutational profiles (comprising 31 SNPs) were associated with resistance to both aminoglycosides, and four mutational profiles (comprising 33 SNPs) were associated specifically with ceftazidime.

Next, we tested these variants against population structure controls, counting only those associated variants that were observed in multiple subpopulation groups as determined by the population structure analysis. This criterion could be satisfied by one of three mechanisms: 1) the mutations arose in the subpopulations through multiple independent mutational events, 2) they arose in a common ancestor of multiple subpopulations and have been maintained in multiple lineages while being lost in others, or 3) the variants arose in one lineage, but were transmitted to another via recombination. Out of all mutational profiles associated with elevated MICs for both β-lactams, one (comprising a single SNP) passed the population structure control (S8B Fig). This SNP was found in 20.4% of isolates in group R, and 50% of isolates in group RBG. This variant leads to a non-synonymous amino acid substitution (P39S) in AmpD (BMUL_2790), a protein extensively studied for its role in resistance to β-lactams [30, 31]. This mutation was predicted to have a deleterious effect on the protein function of AmpD by PROVEAN analysis (score = -8.0, S10A Fig). The *ampD* locus appears to be an important selective target since it was independently mutated a total of five times within our collection. A second SNP in *ampD* (leading to the non-synonymous amino acid mutation F52S) was found in a mutational profile that was similarly associated with β-lactam MICs; nevertheless, it failed to pass the population structure control. Additionally, two mutational profiles associated to the aminoglycosides and ciprofloxacin passed the population structure control (S8E Fig). One of these mutational profiles, was defined by a non-synonymous amino acid substitution (P211L) in an *araC* family transcriptional regulator locus (BMUL_3951; KEGG orthology group K18991). PROVEAN analysis indicates that this mutation is unlikely to have a deleterious effect on the protein function (score = 6.906). The second mutational profile was defined by a non-synonymous substitution (P304S) in an outer member protein or porin (BMUL_3342; KEGG orthology group K03285). While this mutation is not expected to have a deleterious effect on protein function (PROVEAN score = 3.273), the BMUL_3342 locus was independently mutated two additional times.

Additional variants potentially associated with pathoadaptation can be detected by identifying multi-mutated loci

Pathoadaptation is the process of selective enhancing bacterial virulence via mutational changes that lead to the modulation or loss of function of pre-existing genes [32]. Genes that are independently mutated multiple times provide strong evidence of parallel adaptation [33]. While these mutational patterns are typically associated with pathoadaptation towards virulence and / or resistance, they may also reflect more general adaptation to both the biotic and abiotic lung environment. The former may include adaptation driven by host derived pressures as well as microbiological pressures from both conspecific and heterospecific strains. The latter may include adaptation driven by simple environmental variables such as temperature, moisture, pH, etc.

We observed 328 loci with two or more polymorphisms at distinct positions along the gene in our collection (Table 2). Given the genome size and the total number of polymorphisms (both SNPs and indels), we only consider the 62 loci carrying three or more independent mutations to be statistically significant (p-value $< 0.05/[1{,}892$ SNPs + 328 indels = 2220 polymorphisms]). 184 SNPs (9.7%) and 26 indels (7.9%) were found in these 62 loci. No individual nucleotide site was mutated more than one time. In other words, the mutations were clustered by locus rather than by specific nucleotide position, reducing the likelihood that this pattern was due to mutational hotspots. We further excluded the possibility that multi-mutated loci showed excess polymorphism simply due to an increased mutational rate by examining the mutational class spectrum for the multi-mutated loci relative to the genome-wide average.

Table 2. Parallel pathoadapted loci with multiple independent mutations.

Locus	Encoded Protein	No. of SNPs / Indels	Probability [a]	Biological Relevance	Annotated homologs: organism (query coverage)
BMUL_0641	LysR family transcriptional regulator	7/0	1.65×10^{-23}	Antibiotic Resistance	*bpeT*: *Paraburkholderia xenovorans* (93.8)
BCEN2424_5592 [c]	Glycosyltransferase 36	4/2	1.03×10^{-19}	?	*chvB*: *Burkholderia oklahomensis* (99.2)
BMUL_4010	NAD-glutamate dehydrogenase	5/0	6.48×10^{-16}	Amino acid metabolism	*gdh2*: *Burkholderia vietnamiensis* (99.4)
BMUL_0487	Hypothetical protein	5/0	6.48×10^{-16}	Lipopolysaccharide biosynthesis	*pagL*: *Pseudomonas aeruginosa* (99.9)
BMUL_4327	Porin	3/2	6.48×10^{-16}	Antibiotic Resistance	*opcP1*: *Burkholderia pseudomallei* (99.3)
BMUL_2790	N-acetyl-anhydromuranmyl-L-alanine amidase	5/0	6.48×10^{-16}	Antibiotic Resistance	*ampD*
BMUL_1598	Amino acid adenylation domain-containing protein	4/0	4.06×10^{-12}	Antibiotic Biosynthesis	*lgrC*: *Brevibacillus brevis* (99.3)
BMUL_0353	YD repeat-containing protein	3/1	4.06×10^{-12}	Secretion	*VgrG*: *Aggregatibacter aphrophilus* (77.7)
BMUL_0449	Preprotein translocase subunit	4/0	4.06×10^{-12}	Quorum Sensing	*SecB*
BMUL_2632	Chaperone protein	4/0	4.06×10^{-12}	Protein Folding	*dnaJ*
BMUL_4942	Signal transduction histidine kinase	3/1	4.06×10^{-12}	Biofilm Formation	*wspE*: *Ralstonia solanacearum* (98.8)
BMUL_2775	UDP-N-acetylmuramate—L-alanyl-gamma-D-glutamyl- meso-diaminopimelate ligase	4/0	4.06×10^{-12}	Antibiotic Resistance	*mpl*: *Burkholderia mallei* (99.8)
BMUL_1444	Transcription termination factor	4/0	4.06×10^{-12}	Transcription Machinery	rho
BMUL_0954	Glycoside hydrolase 15-like protein	4/0	4.06×10^{-12}	Nutrient Metabolism	*cga*: *Burkholderia mallei* (97.9)
BMUL_4115	Outer membrane autotransporter	4/0	4.06×10^{-12}	Secretion	*ssp*: *Stenotrophomonas maltophilia* (81.4)
BMUL_0250	50S ribosomal protein L4	3/0	2.55×10^{-8}	Translation	*rplD*
BMUL_5547	Conjugation protein	2/1	2.55×10^{-8}	Quorum Sensing	*trbI*: *Rhodoferax ferrireducens* (59)
BMUL_2931	TPR repeat-containing protein	3/0	2.55×10^{-8}	Antibiotic Resistance	*bamD*: *Ralstonia solanacearum* (99.1)
BMUL_3678	Integral membrane sensor signal transduction histidine kinase	3/0	2.55×10^{-8}	Signal Transduction	*rstB*: *Burkholderia mallei* (97.6)
BMUL_3503	L-serine dehydratase 1	3/0	2.55×10^{-8}	Antibiotic Biosynthesis	*sdaA*: *Ralstonia solanacearum* (100)
BMUL_0690	RND efflux system outer membrane lipoprotein	2/1	2.55×10^{-8}	Antibiotic Resistance	*oprM*: *Burkholderia mallei* (97.4)
BMUL_0663	Alpha/beta hydrolase fold protein	3/0	2.55×10^{-8}	?	PA0368: *Pseudomonas aeruginosa* (87.6)
BMUL_0431	Histidine kinase	1/2	2.55×10^{-8}	Signal Transduction	*dctB*: *Paraburkholderia xenovorans* (96.1)
BMUL_4510	Signal transduction histidine kinase	2/1	2.55×10^{-8}	Chemotaxis	*cheA*: *Paraburkholderia xenovorans* (96.7)
BMUL_1970	Major facilitator transporter	3/0	2.55×10^{-8}	Transport across the Membrane	RPA4808: *Rhodopseudomonas palustris* (99.5)
BMUL_2008	Major facilitator transporter	2/1	2.55×10^{-8}	Transport across the Membrane	*oxlT6*: *Paraburkholderia xenovorans* (99.8)
BMUL_2621	DNA mismatch repair protein	1/2	2.55×10^{-8}	DNA Repair	*mutL*
BMUL_4037	Esterase	3/0	2.55×10^{-8}	?	PA3628: *Pseudomonas aeruginosa* (82.7)
BMUL_3977	Metallophosphoesterase	2/1	2.55×10^{-8}	?	BMAA1343: *Burkholderia mallei* (99.7)
BMUL_4949	Aldehyde dehydrogenase	2/1	2.55×10^{-8}	?	gabD: Burkholderia mallei (1)

(*Continued*)

Table 2. (Continued)

Locus	Encoded Protein	No. of SNPs / Indels	Probability [a]	Biological Relevance	Annotated homologs: organism (query coverage)
BMUL_3951	AraC family Transcriptional regulator	3/0	2.55 X 10^{-8}	Antibiotic Resistance	*mtrA*: *Neisseria gonorrhoeae* (99.7)
BMUL_6019	Cytosine/purines uracil thiamine allantoin permease	2/1	2.55 X 10^{-8}	Transport across the Membrane	BMAA0417: *Burkholderia mallei* (96.9)
BMUL_0307	Amino acid carrier protein	3/0	2.55 X 10^{-8}	Transport across the Membrane	*alsT*: *Burkholderia mallei* (96.6)
BMUL_5501	Cytochrome c oxidase subunit I	3/0	2.55 X 10^{-8}	Nutrient Metabolism	*coxAC*: *Burkholderia pseudomallei* (72.5)
BMUL_5087	Short-chain dehydrogenase/reductase SDR	3/0	2.55 X 10^{-8}	Nutrient Metabolism	*fabG*: *Paraburkholderia xenovorans* (93.9)
BMUL_4813	RNA polymerase sigma factor	3/0	2.55 X 10^{-8}	Translation	*rpoD*
BMUL_3197	Beta-galactosidase	3/0	2.55 X 10^{-8}	Nutrient Metabolism	*bgaB*: *Burkholderia thailandensis* (99.8)
BMUL_3212	Feruloyl-CoA synthase	3/0	2.55 X 10^{-8}	Nutrient Metabolism	*fcs*: *Pandorea pnomenusa* (94.5)
BMUL_3315	PA-phosphatase like phosphoesterase	1/2	2.55 X 10^{-8}	Antbiotic Resistance	*bcrC*: *Nitrospirillum amazonense* (99.6)
BMUL_3752	Peptidoglycan-binding LysM	3/0	2.55 X 10^{-8}	?	*RSc3430*: *Ralstonia solanacearum* (5.6)
BMUL_3615	Aldehyde oxidase	3/0	2.55 X 10^{-8}	?	*iorB*: *Pseudomonas aeruginosa* (98.8)
BMUL_1686	Ribonuclease R	3/0	2.55 X 10^{-8}	Translation	*vacB*: *Burkholderia mallei* (94.1)
BMUL_4615 [b]	Amidophosphoribosyltransferase	3/0	2.55 X 10^{-8}	Amino acid metabolism	*purF*: *Burkholderia mallei* (99.8)
BMUL_4605	UTP-glucose-1-phosphate uridylyltransferase	3/0	2.55 X 10^{-8}	Amino acid metabolism	*galU-2*: *Burkholderia mallei* (100)
ABD05_14940 [d]	Isochorismatase	3/0	2.55 X 10^{-8}	Quorum Sensing	*entB*: *Burkholderia ambifaria* (100)
BMUL_1431	GAF modulated sigma54 specific transcriptional regulator	2/1	2.55 X 10^{-8}	Transcription Machinery	*acoR*: *Paraburkholderia xenovorans* (97.6)
BMUL_1377	N-acetyltransferase GCN5	3/0	2.55 X 10^{-8}	?	BMA1429: *Burkholderia mallei* (96.6)
BMUL_0964	DNA polymerase III subunit alpha	3/0	2.55 X 10^{-8}	DNA Repair	*dnaE*: *Burkholderia mallei* (100)
BMUL_0692	Carbohydrate kinase FGGY	2/1	2.55 X 10^{-8}	Nutrient Metabolism	*xylB*: *Paraburkholderia xenovorans* (99.4)
BMUL_0477	Error-prone DNA polymerase (DnaE2)	3/0	2.55 X 10^{-8}	DNA Repair	*dnaE2*: *Burkholderia mallei* (99.9)
BMUL_0443	Phosphoenolpyruvate-protein phosphotransferase	3/0	2.55 X 10^{-8}	Signal Transduction	*ptsI*: *Burkholderia mallei* (94.1)
BMUL_3068	Aldehyde dehydrogenase	3/0	2.55 X 10^{-8}	?	BMA3273: *Burkholderia mallei* (100)
BMUL_4835	Hypothetical protein	2/1	2.55 X 10^{-8}	?	STY4627: *Salmonella enterica* (99)
BMUL_1873	UvrD/REP helicase	3/0	2.55 X 10^{-8}	DNA Repair	*uvrD*: *Burkholderia mallei* (99.9)
BMUL_2536	Hypothetical protein	3/0	2.55 X 10^{-8}	?	RSp0803: *Ralstonia solanacearum* (35.5)
BMUL_2710	Outer membrane autotransporter	3/0	2.55 X 10^{-8}	Transport across the Membrane	aidA-I: Enterobacter sp. 638 (57.5)
BMUL_0123	Heavy metal translocating P-type ATPase	3/0	2.55 X 10^{-8}	Transport across the Membrane	*cadA*: *Burkholderia mallei* (82.5)
BMUL_0116	Acyl-CoA dehydrogenase domain-containing protein	3/0	2.55 X 10^{-8}	Lipid Metabolism	*aidB*: *Burkholderia mallei* (99.6)
BMUL_0075	Two component transcriptional regulator	2/1	2.55 X 10^{-8}	Signal Transduction	Bxe_A0008: *Paraburkholderia xenovorans* (100)
BMUL_4226	4-hydroxyphenylpyruvate dioxygenase	3/0	2.55 X 10^{-8}	Amino acid metabolism	*hppD*: *Paraburkholderia xenovorans* (96.3)

(*Continued*)

Table 2. (Continued)

Locus	Encoded Protein	No. of SNPs / Indels	Probability [a]	Biological Relevance	Annotated homologs: organism (query coverage)
BMUL_4749	Amino acid permease	2/1	2.55×10^{-8}	?	PA0789: *Pseudomonas aeruginosa* (98.1)

[a] Probability of resampling with replacement any locus n times, given a genome size of N. P = (1/N)^(n − 1).

[b] A mutation occurred in the intergenic region flanking the start codon of this locus.

[c] This locus is not found in ATCC 17616. The homolog with highest similarity is in *B. cenocepacia* HI2424

[d] This locus is not found in ATCC 17616. The homolog with highest similarity is in *B. pyrrocinia DSM10685*

https://doi.org/10.1371/journal.ppat.1007453.t002

While the rate of non-synonymous, synonymous and intergenic mutations among all 1,892 SNPs is 70.5%, 15.6%, and 13.9% respectively, the mutational class spectrum of the SNPs found among multi-mutated loci is 83.1% non-synonymous, 11.7% synonymous, and 3.2% intergenic substitutions. Therefore, the mutational class distribution of SNPs found in multi-mutated loci is significantly skewed toward an excess of non-synonymous mutations ($P < 0.0001$, chi-square test).

Some of these multi-mutated loci are known to play significant roles in antibiotic resistance. For example, a gene encoding a LysR family transcriptional regulator (BMUL_0641) has seven independently acquired mutations. The probability of any one gene being mutated seven times given our dataset is 1.65×10^{-23}. Homologs of this locus in other *Burkholderia* species are annotated as *bpeT*, which is strongly associated with drug resistance [34–36]. A locus with five multiple mutations ($P = 6.48 \times 10^{-16}$) encodes N-acetylmuramoyl-L-alanine amidase (AmpD, BMUL_2790), which is associated with resistance to β-lactam antibiotics [30].

We performed a functional enrichment analysis on the multi-mutated loci and found that the Gene Ontology (GO) function phosphorelay signal transduction system was overrepresented in multi-mutated genes compared to the whole genome ($P = 0.050$). The phosphorelay signal transduction system has been previously described as a therapeutic target, given that it controls the expression of genes encoding virulence factors [37].

We also found ten genes that had two independent mutations located in the same or adjacent codon (Table 3). The mutational class spectrum of the SNPs associated with this

Table 3. Pairs of mutations occurring in the same or in neighboring codons.

Encoded Protein	Proximity
Regulatory protein GntR, HTH:GntR, C-terminal	Adjacent codon
Oligopeptide ABC transporter, periplasmic oligopeptide-binding protein (OppA)	2 codons away
Citrate-proton symporter	2 codons away
CDP-6-deoxy-delta-3,4-glucoseen reductase-like	2 codons away
RNA polymerase sigma factor (RpoD) [a]	Same codon
Endo-1,4-beta-xylanase Z precursor [b]	Adjacent codon
Isoquinoline 1-oxidoreductase beta subunit [b]	2 codons away
LSU ribosomal protein L4p (L1e) [b]	Same codon
Chaperone protein (DnaJ) [c]	Adjacent codon
LysR family transcriptional regulator [d]	2 codons away

[a] Loci additionally mutated 1 more time. Additional mutation is synonymous.

[b] Loci additionally mutated 1 more time. Additional mutation is non-synonymous.

[c] Locus additionally mutated 2 more times. All non-synonymous mutations.

[d] Locus additionally mutated 5 more times. All non-synonymous mutations.

https://doi.org/10.1371/journal.ppat.1007453.t003

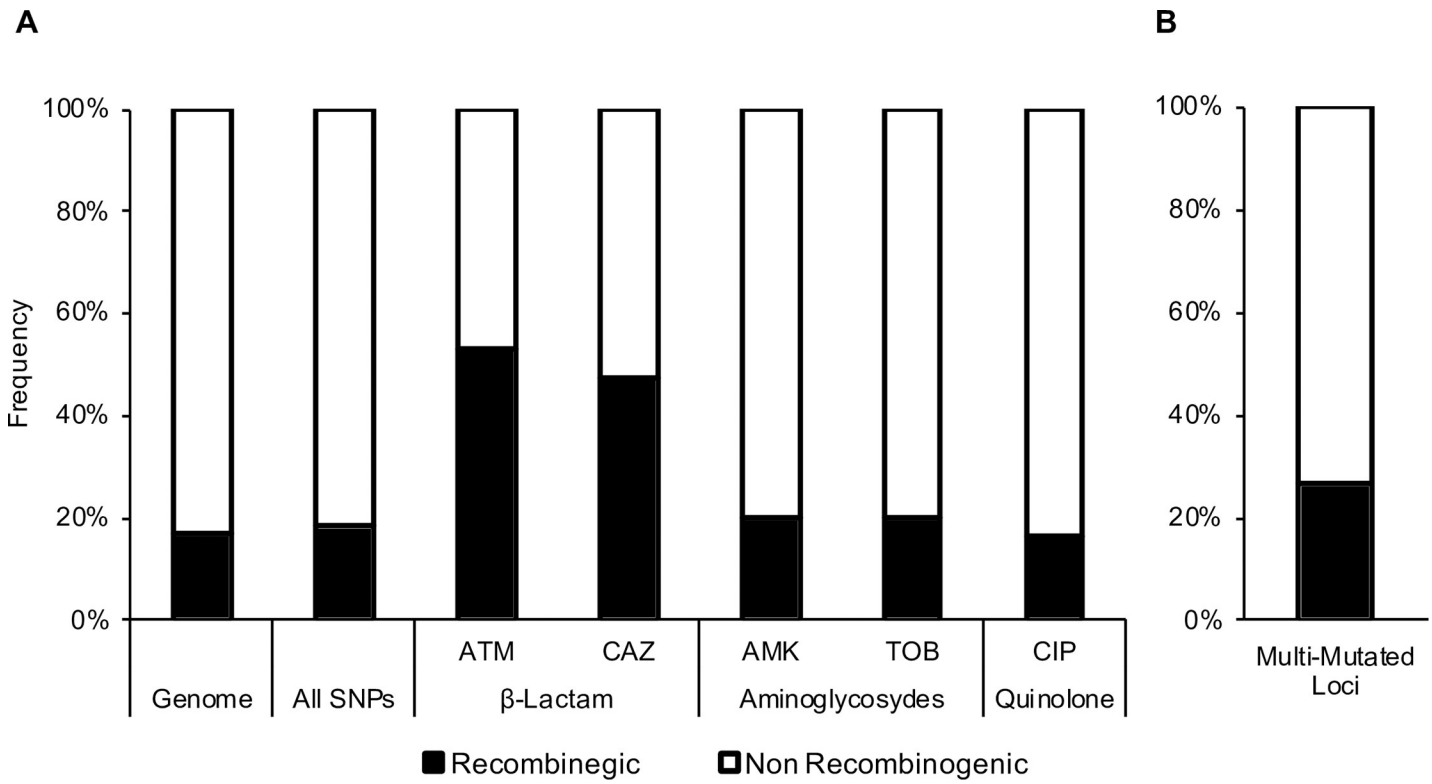

Fig 5. Distribution of pathoadaptive variants in recombinogenic regions of the genome. (A) Distribution of the mutations associated with the tested antibiotics in the identified recombinogenic regions and in the rest of the genome (*** $p < 0.0001$, chi square test with multiple test correction). (B) Distribution of the mutations in multi-mutated loci in the identified recombinogenic regions and in the rest of the genome (*** $p < 0.001$, chi square test with multiple test correction).

https://doi.org/10.1371/journal.ppat.1007453.g005

observation is of 90%, 10% and 0%, non-synonymous, synonymous, and intergenic substitutions, respectively. In this case, the fraction of non-synonymous mutations is significantly higher than the fraction found for both all SNPs, as well as all the SNPs in the multi-mutated loci ($P < 0.00001$, chi-square test). One of the genes with multiple independent mutations in the same codon encodes for RNA polymerase sigma factor (RpoD), which is associated with the expression of housekeeping genes [38]. One of the mutations in this locus is fixed between the post-transplant isolates and the rest of the isolates, and the other mutation is fixed between the isolates in group RBG collected in the tenth sample time and the rest of the isolates.

Parallel pathoadaptive variants are overrepresented in recombinogenic regions

We looked for signals of recombination in our isolates using both the four-gamete tests of Hudson and Kaplan [39] and BratNextGen [40]. We identified a minimum of 15 regions with signatures of recombination in at least one of these methods (Fig 2D). Three of these events were identified between sites in different genome assembly contigs; therefore, they were not considered in downstream recombination analysis. The nucleotide length of this recombinogenic regions ranged from 4,783 bases to 192,532 bases, and these regions account for 15.1% of the assembled genome. 300 (15.9%) out of the total 1,892 SNPs and 47 indels (14.3%) occur in these regions, which is not significantly different than expected given the recombinogenic proportion of the genome.

A recombinogenic region on the first chromosome (281,829–322,435) was involved in genetic exchange between isolates in the groups B and RB. This region contained 16 SNPs, out of which four were statistically associated with resistance to aminoglycosides and to ciprofloxacin prior to population control (Fig 2Ei, Supplementary Table 1). Two of these mutations, which segregated in different isolates, occurred in adjacent bases and led to amino acid substitutions in the same codon of a gene encoding the 50S ribosomal protein L4p (L1e, BMUL_0250). The other two mutations led to two non-synonymous amino acid substitutions in a gene encoding glycerol-3-phosphate transporter ATP-binding subunit (BMUL_0301). Another recombinogenic region in the first chromosome (1,566,898–1,695,617) affected only isolates from the post-transplant sample. 47 SNPs were detected in this region, four of these were associated with resistance to both aminoglycosides and to ciprofloxacin, and one was associated only to aminoglycosides prior to population control (Fig 2Eiii, Supplementary Table 2). These five mutations led to five non-synonymous mutations in the genes encoding ABC transporter-like protein (BMUL_2127), acyl carrier protein (BMUL_2180), malonyl CoA-acyl carrier protein transacylase (BMUL_2182), D-amino acid dehydrogenase small subunit (BMUL_2240), and DL-methionine transporter ATP-binding subunit (BMUL_2245), respectively. We were not able to identify the source of the remaining identified recombination events.

We examined association between the recombinant regions and the polymorphisms associated with antibiotic resistance. 20.1% (56 of 279) of SNPs associated with both aminoglycosides assayed (amikacin & tobramycin), and 16.4% (46 of 281) of SNPs associated with ciprofloxacin were found in recombinogenic regions (Fig 5A). These ratios failed to reject the null hypothesis of random distribution of mutations around the genome. On the other hand, 52.9% (9 of 17) and 47.4% (9 of 19) of the SNPs associated with aztreonam and ceftazidime, respectively, were found in recombinogenic regions, which significantly differs from null expectations ($p < 0.0001$, chi square test). Additionally, while the phylogenies of aminoglycoside and ciprofloxacin associated SNPs resemble the overall phylogeny, the phylogenies of β-lactam associated SNPs have topologies different from the topology of the overall phylogeny (S11 Fig).

Finally, 26.6% (49 of 184) of SNPs and 8.5% (49 of 47) of indels found in multi-mutated loci (those with at least three distinct polymorphic positions) occur in the identified recombinogenic regions (Fig 5B). Intriguingly, while the proportion of SNPs in these multi-mutated loci are overrepresented in recombinogenic regions ($P < 0.0001$, chi square test), the proportion of indels are not.

Discussion

Our study investigated how *B. multivorans* evolves within the lungs of an individual afflicted with CF using a deep longitudinal sampling design (i.e. multiple isolates obtained per sputum sample) to capture both the overall population diversity and the temporal shifts that occurred at different phases of the infection, including the colonization of a new lung allograft. To identify the source of genetic diversity in this *B. multivorans* population, we needed to understand: 1) the genetic relationships between the incident isolate that was recovered from the first BCC-positive sputum culture, the chronic strains that persisted in the population, and the population of strains that re-established an infection post-transplant; 2) whether there were multiple colonization events of the patient by divergent clones; 3) how genetic diversity was generated and dispersed in the population; and 4) how the pathogen adapts and responds to clinical treatment. While we were unable to address all of these questions, we have concluded that the chronic population originated from either the incident isolate, or a clone that shared a recent

common ancestor with the incident isolate. Furthermore, all of the chronic isolates descended from a single common ancestor, ruling out multiple independent colonization events.

One clear signal is that the *B. multivorans* isolates recovered from the post-transplant lung did not originate from the chronic population. In fact, it appears that the post-transplant isolates originated from the same source as the incident isolate. Based on the current literature, the most likely source of these isolates is the upper respiratory tract, although environmental sources cannot be ruled out [41–44]. Upper airway sampling was not performed on this patient, so we have no information on the microbiome of this compartment. While some transplant procedures attempt to clean the nasal reservoir prior to transplant via nasal washing / scraping, we do not know if this procedure was done on this patient. If the upper airway was the source for both the incident isolate and the post-transplant isolates, the latter would have been exposed to ten additional years of antimicrobial treatments than the former, perhaps explaining why these isolates have antibiotic susceptibility pattern more similar to the chronic isolates. We also note that the post-transplant population is much more genetically diverse than any of the chronic populations. This could suggest that this population was rapidly adapting to an environmental change, such as the shift from CF to non-CF conditions, which would include, differences in immune response, the composition of the allograft microbiome, and treatment regimens. Alternatively, it could reflect colonization by a population of related strains. It is possible that given sufficient time this population would eventually be winnowed down to a single surviving clone (as is seen with the incident infection) due to selection and / or genetic drift.

A major motivator for this study was to better understand how pathogens adapt to their hosts over the course of disease progression and treatment; an issue that can be addressed using statistical association tests. Correcting for the genetic structure of the bacterial population poses a challenge to the implementation of these tests. Population structure in this context refers relationships among strains due to descent from a common ancestor and limited recombination. This structure results in the linkage of segregating genetic variation around the genome, which makes it very difficult to distinguish a causal mutation that is responsible for a phenotype of interest from a neutral variant that occurred in the same genetic background. In the absence of recombination, the neutral mutation will have the same population distribution as the causal mutation due to genetic hitchhiking. This issue is particularly prevalent when studying largely isolated and recently evolved populations, such as the case of pathogens evolving within a host.

To overcome these two issues, we imposed a lineage control filter on our GWAS approach, in which we focused only on mutations that occurred in multiple, distinct, genetic lineages. This pattern can best be explained by recombination of polymorphisms between lineages, but formally, could also be due to extensive gene loss. Our analysis showed that linkage disequilibrium was only disrupted in a relatively small number of polymorphisms (those polymorphisms shown as orange circles; S8B–S8E Fig). This reinforces the need for deep sampling since the infrequent recombination signals may have been missed if isolates were only collected from a single sample, or if only single isolates were recovered from each sample. Consequently, the tractability of GWAS in this *B. multivorans* population was greatly enhanced by our sampling schema.

Using the established lineage structure of the *B. multivorans* population as control for our association study, we identified two non-synonymous SNPs associated with resistance to the aminoglycosides amikacin and tobramycin, and to the quinolone ciprofloxacin. One of these SNPs occurs in a locus encoding an AraC family transcriptional regulator, which is homologous to MtrA in *Neisseria gonorrhoeae*, an obligate human pathogen [45]. MtrA is required for the induction of the *mtrCDE*-encoded efflux pump system, which removes macrolide

antibiotics, penicillin, and antimicrobial effectors of the innate defense from the cell [46]. Our PROVEAN analysis predicted that this mutation would not significantly impact the function of the encoding protein, but the appropriate regulation of this efflux pump system could prove crucial for the survival of these bacteria. The second SNP associated with aminoglycoside and ciprofloxacin resistance was found in a locus annotated as a porin. This locus encodes a member of the general bacterial porin family, and shares common ancestry with *Burkholderia pseudomallei*'s OpcP1, which is a subunit of the porin oligomer OpcPO [47]. This family of porins has been associated with the bacterial survival in the airways of the CF lungs by limiting the uptake of small hydrophilic molecules, including ciprofloxacin, into the cell [48, 49]. The function of the encoding protein was not estimated to change because of this SNP; nevertheless, the adequate functioning of these porins in the outer membrane of *Burkholderia multivorans* plays an important role in their survival and resistance [50].

Additionally, we identified a single SNP associated with resistance to the β-lactams aztreonam and ceftazidime. This SNP occurs in the *ampD* gene, which affects the expression of the β-lactamase AmpC and likely also PenB [30] and is expected to have a deleterious effect in the encoding protein. This observation is not unexpected as bacteria treated with β-lactams benefit from the constitutive overproduction of β-lactamase. Overall, AmpD seems to play an important role in the adaptation of this *B. multivorans* population to antimicrobial treatment since four other independent non-synonymous mutations, all of which are expected to have deleterious effects on the protein, occur at this locus (S10A Fig).

Our population structure control criterion, which focuses on those polymorphisms present in multiple lineages, resulted in the exclusion of some variants associated with resistance or virulence, e.g. one of the four mutations in *ampD*, which was statistically associated with β-lactam resistance. A population structure control is critical for distinguishing putatively causative mutations from hitchhiking variants that are carried along by linkage disequilibrium. Filtering in this manner reduces the number of false positives; nevertheless, variants underlying phenotypes of interest could be segregating in linkage disequilibrium blocks, and therefore, may not be identified in our GWAS approach (i.e. false negatives).

We observed that mutations associated with resistance to β-lactams (prior to lineage controls) occur disproportionately in recombinogenic regions (Fig 2F), while variants associated with both aminoglycosides or ciprofloxacin are randomly distributed with respect to recombinogenic regions. Patient CF170 received both long-term maintenance β-lactam and aminoglycoside treatments in addition to multiple short-term β-lactam treatments that included cycles of ceftazidime, piperacillin/tazobactam, meropenem, and cefepime. This more aggressive and varied course of treatment with β-lactams could potentially explain the increased role of recombination in the dissemination of putatively beneficial polymorphisms, similar to what has been observed in other pathogens [51, 52].

Parallel evolution has been shown to be a reliable signal for identifying genes involved in host adaptation, including virulence and resistance to antibiotics, among CF lung pathogens [25, 53–57]. Our analysis identified numerous genes showing a statistical excess of independent mutations (i.e. putative parallel pathoadaptations) [25, 32, 57]. Examining multi-mutated loci can reveal the heterogeneous selective pressures that bacteria must adapt to in order to reside within the lung. For instance, a gene encoding a transcription regulator of multidrug resistance efflux pumps independently accumulated seven different mutations leading to eight unique alleles in our population of 111 *B. multivorans* isolates. We also found seven different alleles of a locus encoding cyclic β-1,2-glucan synthase, which is linked to bacteria's ability to elude host cell defenses [58]. A number of loci underlying virulence-associated traits, such as quorum sensing and biofilm production, also carry multiple independent mutations. Particularly interesting are multi-mutated loci with no characterized function, or with no prior

linkage to resistance or virulence. These loci include a NAD-glutamate dehydrogenase locus BMUL_4010, which was mutated five independent times over the course of the study, and a glycosyl transferase protein (BCEN2424_5592), not previously seen in *B. multivorans* that was mutated six times (4 SNPs and 2 indels) during the course of the study. Examples such as these provide excellent candidates for characterizing the spectrum of ways pathogens adapt to their hosts, including selection for antibiotic resistance, adaptation to the host immune system and physical environment, resource utilization, microbe-microbe competition, and even unknown selective forces. Perhaps the strongest signals of parallel pathoadaptation involve those cases where mutations occur independently in the same or adjacent codon. These observations suggest a specific form of selective pathoadaptation, which identifies the specific residue or region of the locus that potentially plays a role in selection.

While the most frequently found targets of parallel evolution are loci associated with antibiotic resistance other classes of targets have also been identified [25, 53, 57, 59]. For instance, Silva *et al.* reported parallelism in an OmpR-like response regulator, which is involved in the mucoidy phenotype of *B. multivorans*, and later showed its association with persistence in the CF lungs [54, 60]. We found two related multi-mutated genes encoding an OmpR family-sensor histidine kinase (BMUL_3678) and an OmpR family response regulator (BMUL_0075). A study of the within-host evolution of *B. pseudomallei* in seven Australasian CF patients by *Viberg et al.* [61] found multiple independent mutations in genes involved in DNA repair (*mutS*), translation (*rpoD*), protein folding (*dnaK*), and secretion (*vgrG*). Similarly, we observed multiple independent mutations in genes involved in the same processes (*mutL*, BMUL_2621; *rpoD*, BMUL_4813; *dnaJ*, BMUL_2632; and *vgrG*, BMUL_0353). While these examples of parallel evolution are suggestive of the pathoadaptive direction of our *B. multivorans* population, we cannot conclusively determine which mutation or group of mutations are responsible for the pathoadaptation of the bacterial population in the lungs of patient CF170.

Finally, our study highlighted an intriguing role for recombination in the development of antimicrobial resistance in *B. multivorans*. We observed that multi-mutated loci were over-represented within recombinogenic regions, along with an excess of mutations associated with β-lactam resistance. This suggests that while recombination plays an important role in the pathoadaptation of this *B. multivorans* population, its selective benefit may be environment dependent.

Our study illustrates the relevance of deep, longitudinal sampling to the implementation of GWAS approaches in a population under positive selection. We identified the potential genetic basis behind the antibiotic resistance of a *B. multivorans* population in a single host. Moreover, this approach allowed us to study variants associated to antibiotic resistance and revealed that resistance to β-lactams may be passed within the population via recombination. This study is limited to *in silico* predictions of the impact mutations on protein function, and future efforts should include functional validation of these mutants; nevertheless, many of the identified genes are already well-established targets for antibiotic resistance. Additionally, our findings are restricted to a single patient and a single bacterial species; extending this approach in other systems under positive selection will be required to establish the generalizability of the findings. Nevertheless, this study is one of the first examining in depth the fine-scale evolution of *B. multivorans* in the lungs of a CF patient as it transitions from chronic infection to the eventual reinfection of a transplanted allograft.

Materials and methods

Ethics statement

All protocols involving the collection, handling and laboratory use of respiratory specimens were approved by the Research Ethics Boards of St. Michael's Hospital (Protocol #09–289)

(Toronto, Canada) and the University Health Network (Protocol #09-0420-T) (Toronto, Canada). We obtained written informed consent from the study subject prior to specimen collection and sputa were produced voluntarily. All experiments involving clinical specimens were performed in accordance with the *Tri-Council Policy Statement*: *Ethical Conduct for Research Involving Humans*, of the Canadian Institutes of Health Research, the Natural Sciences and Engineering Research Council of Canada, and the Social Sciences and Humanities Research Council of Canada.

Specimen collection and isolation of *B. multivorans*

Sputum specimens were collected by expectoration from a 29-year-old male (CF170), with a homozygous ΔF508 CFTR genotype being followed at the Adult CF Clinic at St. Michael's Hospital (Toronto, Canada). Ten sputum specimens were collected over a 10-month period while the patient was in the advanced stages of CF lung disease (assessed by the forced expiratory volume in 1 second (FEV_1), FEV_1 which was 27–39% predicted throughout the course of the study), and an additional sputum specimen obtained after the patient had undergone double lung transplantation. All specimens were processed for bacterial culture as previously described [62]. After 48h of incubation, cultures were visually inspected, and each distinct colony morphotype was described using eight characteristics of physical appearance (pigmentation, size, surface texture, surface sheen, opacity, mucoidy, autolysis and margin shape). Ten colonies were selected from each sputum culture in relation to the diversity of colony types present. The incident isolate was obtained from the *Burkholderia cepacia* complex repository at St. Michael's Hospital and was recovered from the first BCC positive sputum culture produced by the study patient (Toronto, Canada). Isolates were stored at −80˚C in 20% (v/v) glycerol after a 20h subculture in LB broth (Wisent Inc., QC, CA) and confirmed as *Burkholderia* spp. by a secondary subculture onto both *Burkholderia cepacia* selective (BCSA) (HiMedia Laboratories, Mumbai, IN) and MacConkey (Becton Dickinson, MD, USA) agars, as well as being tested for growth at 42˚C. The *recA* gene was sequenced from each isolate as described by Spilker *et al*. for preliminary speciation [63].

Antimicrobial susceptibility testing

Each isolate confirmed as *B. multivorans* was screened for antimicrobial susceptibility by agar dilution using Clinical and Laboratory Standards Institute procedures [29]. We tested susceptibility to representatives of the β-lactam (aztreonam [ATM], ceftazidime [CAZ]), fluoroquinolone (ciprofloxacin [CIP]) and aminoglycoside (amikacin [AMK], tobramycin [TOB]) (Sigma-Aldrich, ON, Canada) classes. Minimum inhibitory concentrations (MIC), defined as the lowest concentration of each antibiotic to inhibit growth, were reported as the median MIC of three independent experiments. Growth was assessed following 24 to 48 h of incubation on Mueller-Hinton agar (Becton, Dickinson, MD, USA). The *B. multivorans* ATCC 17616 strain was included as a positive control, while *P. aeruginosa* ATCC 27853 and *E. coli* ATCC 25922 were used as quality controls.

Sequencing and quality control

B. multivorans isolates were whole-genome sequenced on the MiSeq and NextSeq Illumina platforms. This sequences can be found in NCBI's BioProject Accession: PRJNA475602. The number of bases sequenced per isolate ranged from 213 to 2,262 million bases, and the median was 1,002 million bases. Trimmomatic v. 0.33 was used to remove adapters and quality trim the sequencing reads from each isolate (parameter settings: PE -phred33 ILLUMINACLIP: adapters.fa:2:30:10 LEADING:5 TRAILING:5 SLIDINGWINDOW:4:25) [64]. Sequencing

reads with guanine homopolymers longer than ten bases were trimmed with cutadapt v. 1.9.1 (parameter settings: -a "G[65]") [66]. Reads bellow 100 bases were removed using Trimmomatic v. 0.33 (parameter settings: PE–phred33 MINLENGTH:100). The resulting quality-controlled sequencing reads yielded a median read depth per position of 117X (range 32-276X).

De novo and reference mapping assembly

Each of the isolates was *de novo* assembled using the CLC Genomics Workbench v. 8.0.1 (Aarhus, Denmark). Contigs with a scaffolding depth lower than 10X and/or with a size smaller than 1 Kb were removed from further analyses. Isolate CF170-3b, which was sequenced with 250 bp-long paired-end reads, yielded the best assembly metrics in 26 contigs with lengths ranging from 1,010 to 1,243,078 bases and an N50 of 654,231. The final assembly length of the CF170-3b isolate was of 6,444,123 bp. These contigs were annotated at the RAST server using the native gene caller and Classic RAST as the annotation scheme [64]. Additionally, each CDS identified by RAST was blasted against the genome of *B. multivorans* ATCC17616 (if no hit found, we blasted against *B. cenocepacia* 22E-1 and *B. cenocepacia* HI2424) [67, 68]. Further, this genome was functionally annotated with blast2go v 4.1.9 [69] including blastx v. 2.6.0+ [67] and the KOALA annotation tool, which enabled KEGG orthology annotation [70]. Statistical results from the functional enrichment analysis were Bonferroni corrected for multiple testing using the number of multi-mutated genes (P-value/62). The contigs of the CF170-3b genome were used as the reference for mapping assembly of each remaining isolate. We performed three different reference-mapping assemblies including BWA v 0.7.12 [71], LAST v 284v [72] and novoalign v 2.08.03 (Novocraft Technologies).

Single Nucleotide Polymorphism (SNP) and indel Calling

SAMtools and BCFtools v 0.1.19 were used to produce the initial set of variants [73]. We implemented a method previously described to detect SNPs among the 111 isolates [25, 53]. First, 1,892 high-confidence polymorphic positions were identified using the following criteria: 1) variant Phred quality score of $\geq$ 30, 2) variants must be found at least 150 bp away from either the edge of the reference contig or an indel, and 3) variants must be called in the three reference mapping experiments. Second, we reviewed each high-confidence polymorphic position in each isolate with a relaxed Phred score threshold of 25. Support for either the reference or the SNP call was verified with a multi-hypothesis correction which required that at least 80% of the sequencing reads endorsed the SNP or the reference. If the data did not support either base, then the position was called as an ambiguous base ('N'). The ambiguous call rate was lower than 0.01%.

Candidate indels detected by BWA and SAMtools were examined by realigning mapped and unmapped sequencing reads to the indel regions using Dindel v. 1.01 [74]. High-confidence indel positions were defined as sites with: 1) variant Phred quality score of $\geq$ 35; 2) at least two forward and two reverse reads; and 3) sequencing coverage $\geq$ 10. These indel positions were reviewed in each isolate. The final indel call required a Phred quality score $\geq$ 25 and an allele frequency $\geq$ 80%. Ambiguous indel calls were defined as those where the allele frequency was $\leq$ 20%.

Population and single genome sequencing evaluation

We performed bulk population sequencing on the post-transplant specimen to confirm that our isolate sampling depth appropriately represented the real *B. multivorans* population diversity (S12 Fig). The sequencing reads from each of the ten isolates from the post-transplant sample were rarified to 1/10th of the number of sequencing reads produced by the population

sequencing experiment. These reads were combined in corresponding paired-end fasta files. Next, population and single isolate sequencing reads were mapped to the *de novo* assembled genome of the CF170-3b isolate using BWA. Mutation allele frequencies for each experiment were estimated as previously described by Lieberman *et al.* [53].

Phylogenetic, population structure, coalescent and recombination analyses

Using the 1,892 SNPs, we created a genome-wide alignment to reconstruct the phylogenetic relationships among the 111 isolates. The phylogeny was calculated using MrBayes v. 3.2.6 [75]. The nucleotide substitution model that best fit our data was the General Time Reversible (GTR) with gamma-distributed rate variation across sites (LnL = -13,152.7810, AIC = 26,832.1306) as calculated with jModelTest v. 2.1.10 [76]. The Bayesian analysis was run through four different chains of 1 million Markov Chain Monte Carlo (MCMC) generations sampled every 100 MCMC generations and the burn-in period was of 250,000 MCMC generations. The final average standard deviation of split frequencies was of $7.3x10^{-3}$, and the potential scale reduction factor (PSRF) of the substitution model parameters ranged from $1–6.66x10^{-5}$ to $1 + 4.83x10^{-4}$. The phylogeny was rooted with *B. multivorans* ATCC 17616 [77]. The network-based phylogenetic analysis was performed using SplitsTree v 4.14.4 [78]. We employed the Jukes-Cantor distance matrix to implement the neighbor-net Network (Fit = 99.804).

The variance among the 111 isolates, including SNPs and indels, was employed to investigate the population structure using the Structure software v 2.3.4 [27]. Structure employs a Bayesian algorithm to detect the number of ancestral populations (K), also known as clusters, which describe the variance and covariance observed in a test population. The number of clusters ranging from 1–10 was tested in triplicates through 1 million MCMC generations sampled every 1,000 MCMC generations and a burn-in period of 250,000 MCMC generations. We used the correlated allele frequencies model, and admixture was allowed in these analyses. We plotted the estimated ln probability of data for the tested levels of K, and identified the smallest stable K as the optimum value since it maximized the global likelihood of the data (S13 Fig) [79]. The estimated ln probability of data plateaus at K = 3, where the variance of ln likelihood ranges from 2,343.0 to 2,353.1. Assuming three ancestral populations, the isolates were classified into five different groups according to their ancestry. Isolates whose ancestry is attributed exclusively (>90%) to either ancestral population one, two, or three are grouped in group red (R), (B), or (G), respectively. Group RB includes isolates with admixed ancestry from clusters one and two (at least 10% of both cluster one and two, and less than 10% of cluster three). Isolates whose ancestral composition is made up from a combination of all three clusters (at least 10% of each cluster) are in group RBG.

We used BEAST v. 1.8.4 to implement a Bayesian approach to inferring the time to the most recent common ancestor (tMRCA) for the entire population and each group individually [80]. Next, we employed the GTR nucleotide substitution model, and estimated the nucleotide substitution frequencies with MEGA7 using the Maximum Likelihood Estimate of the Substitution Matrix tool ([AC] = 0.0091, [AG] = 0.4281, [AT] = 0.0016, [CG] = 0.0260, [GT] = 0.0061, and [CT] = 0.5290) [81]. Preliminary analyses consisting of duplicate 10 million generations and a 10% burn-in were used to estimate the appropriate molecular clock and demographic models. We tested the Bayesian skygrid, constant size and the exponential, logarithmic and expansion growth population size models using three different molecular clock models (strict and the lognormal and exponential uncorrelated relaxed clocks). The exponential relaxed uncorrelated molecular clock and the Bayesian skygrid model was inferred the most appropriate given our data ([AIC] = 26,228.421) [82]. The final analysis was run in duplicate for 1 billion MCMC generations sampled every 1,000 MCMC generation, and the burn-in

period was set at 20% of the MCMC generations. The inferred molecular clock was consistent with the number of mutations observed in our isolates through time (S14 Fig).

Population genetic tests and detection of recombination events in each contig were performed with DnaSP v. 5.10.01 [83] and BratNextGen, which was run with 500 iterations [40]. We calculated the pairwise homoplasy index (Φw, PHI statistic), which considers the minimum number of homoplasies needed to account for the linkage between two sites [84]. This statistic rejected the null hypothesis of no recombination in the regions we had identified as recombinogenic ($p<0.01$). Additionally, a phylogenetic analysis of the identified recombinogenic regions reveal different topologies compared to the overall phylogeny (S15 Fig).

SNP to phenotype association

Each mutational profile was tested for statistical association to each antibiotic. In order to discard mutational profiles specific to a subpopulation, mutations were simulated to occur along the phylogeny through a parsimonious process, so as to identify mutations which occurred independently in more than one subpopulation. Mutations arisen through a single mutational event in a single subpopulation were deemed to be in linkage disequilibrium with the mutations that are fixed in that subpopulation.

We tested the null hypothesis that the presence or absence of each of the 1,892 SNPs, summarized in 150 distinct mutational profiles, is equally likely found in antibiotic resistant isolates using Fisher's exact test. These tests were conducted for each examined antibiotic at six different MIC resistance thresholds ($\leq$16, 32, 64, 128, 256 and $\leq$512 MIC). For each test, we created a contingency table reflecting the distribution of each mutation profile in isolates with lower and greater MIC than each resistance threshold. *P* values were adjusted based on the total number of tests (number of mutational profiles), and only associations with a *P* value $< 3.36 \times 10^{-4}$ (0.05 / 150) were considered significant to control for multiple testing. Next, we simulated gains or losses of these mutational events following a continuous-time Markov chain along a ClonalFrameML v. 1.0–19 phylogeny as implemented in GLOOME v. 01.266 using the default parameters [85, 86]. We defined independent mutational events as those with a probability greater than 0.95 and to control for population structure, we required multiple independent mutational events in at least two STRUCTURE-defined groups.

d_N/d_S calculations

We calculated the expected N/S ratio by simulating all potential mutations in all CDS in the reference genome and recording all the outcomes of the particular mutational spectrum as non-synonymous or synonymous amino acid substitutions. For instance, A>T mutations are 18.9 times more likely to lead to a non-synonymous amino acid substitution than a C>T mutation. The reported d_N/d_S was the ratio between the observed value of N/S and the expected value of N/S given each type of mutation. The confidence intervals were estimated consistent with binomial sampling. This method was first reported by Lieberman *et al.*, in 2014 [87].

In silico mutation impact prediction

To predict the potential impact of non-synonymous SNPs on the biological function of a protein, we employed PROVEAN v. 1.1.3 [88]. These calculations were performed on the GPC supercomputer at the SciNet HPC Consortium [89].

Supporting information

S1 Fig. Sequencing coverage. Whole genome sequencing of 111 isolates of *B. multivorans* in the Illumina platform. (A) Distribution of number of bases sequenced per isolate. (B) Distribution of median read depth per position.
(PDF)

S2 Fig. Genetic diversity over time. (A) Pairwise nucleotide differences between isolates collected from the same collection sample. Incident infection is not included since only one isolate was recovered from that time point. (B) Nucleotide differences between each isolate and the incident infection isolate.
(PDF)

S3 Fig. Additional phylogenetic analysis to support outgroup position and robustness of the phylogenetic topology. A) Maximum likelihood phylogeny including *B. multivorans* ATCC 17616, *B. multivorans* BAA247, *B. multivorans* DDS15A-1, and *B. multivorans* AU1185 from the Burkholderia Genome database [68]. This tree was estimated using the General Time Reversible (GTR) model in MEGA7 with 500 bootstrap iterations, and it was rooted with *B. mallei* ATCC 23344 as the outgroup [81]. B) Maximum likelihood phylogeny rooted using *B. multivorans* ATCC 17616 as the outgroup. This tree was estimated under the GTR model in MEGA7 using 500 bootstrap iterations [81]. C) Maximum parsimony phylogeny rooted with *B. multivorans* ATCC 17616 as the outgroup. This tree was estimated using MEGA7 and 500 bootstrap iterations [81]. D) Hierarchical clustering based on the presence and absence of insertions or deletions among the 111 isolates using Euclidian distances as implemented by the vegan package in R [93]. This dendrogram was rooted with the incident isolate as the outgroup.
(PDF)

S4 Fig. CF170 isolates in the context of other *Burkholderia* genus genomes. The sequences of seven housekeeping genes (atpD, gltB, gyrB, lepA, phaC, recA, and trpB) from *B. xenovorans* LB 400, *B. oklahomensis* C6786, *B. thailandensis* E264, *B. mallei* ATCC 23344, *B. pseudomallei* K96243, *B. vietnamiensis* G4, *B. ambifaria* AMMD, *B. cenocepacia* HI2424, *B. pyrrocinia* DSM 10685, *B. dolosa* AU 0158, *B. multivorans* ATCC 17616, *B. multivorans* 15A-1, *B. multivorans* BAA 247, *B. multivorans* CGD2M, and *B. multivorans* AU1185 were extracted as defined by pubMLST [24]. These sequences were aligned with MUSCLE (default parameters) [94], and the resulting alignment was used to recreate their phylogenetic relationships with a Maximum Likelihood approach (Bootstrap = 1,000).
(PDF)

S5 Fig. Neighbor-Net phylogeny. This network-based phylogeny was calculated in SplitsTree v. 4.14.4. Individual strain names at the tips of each branch have been replaced with pie charts indicating the distribution of dates during which the strains were sampled (indicated by the circular legend).
(PDF)

S6 Fig. Genetic diversity and selection analysis per group. (A) Pairwise nucleotide differences between isolates from the same group based on ancestry. (B) d_N/d_S per group calculated including all SNPs and using only SNPs observed in multiple time points (MTP). d_N/d_S and the respective confidence intervals were calculated as described by Lieberman *et al.* [87].
(PDF)

S7 Fig. SNP positions with identical distribution of reference or alternative bases across the strain collection are grouped into mutational profiles. Here, "0"s and "1"s represent the reference or alternative base, respectively, at each SNP position for each strain. SNP1 is the only position where only Strain1 has a base alternative to the reference. Hence, mutational profile 1, 1-0-0-0, comprises only one SNP. On the other hand, Strain4 is the only strain with a variant base for positions SNP2 and SNP3. Therefore, mutational profile 2, 0-0-0-1, comprises SNP2 and SNP3.
(PDF)

S8 Fig. Mutational profiles associated with antibiotic resistance. (A) Maximum Likelihood phylogeny of 111 *B. multivorans* isolates was elaborated using RaxML v. 7.0.4 with a GTR + gamma model and 1,000 bootstraps [95]. Here, we show all mutation profiles associated with antibiotic resistance prior to lineage control in black and with lineage control in orange. (B) resistance to both β-lactams, (C) to amikacin only, (D) to both aminoglycosides, (E) to both aminoglycosides and to ciprofloxacin, (F) and to ciprofloxacin only. A filled circle represents a SNP call in the corresponding isolate compared to the reference.
(PDF)

S9 Fig. Resistance levels at which genetic associations are statistically significant. Mutational profiles were tested for association against six levels of antibiotic resistance (<16, <32, <64, <128, <256 and <512 MIC) to five antibiotics (amikacin, tobramycin, aztreonam, ceftazidime and ciprofloxacin). Black boxes show the levels of resistance at which the mutational profiles were statistically significant including multi-testing correction. Associations to ciprofloxacin antibiotic resistance are shown up to <128 MIC since no isolate had a MIC of 256 or greater in relation to that antibiotic.
(PDF)

S10 Fig. Mutations in *ampD* locus. (A) Distribution of the PROVEAN scores of all identified non-synonymous substitutions highlighting SNPs in multi-mutated loci (yellow) and in the *ampD* gene (red or blue if associated to β -lactam resistance). Red lines represent thresholds from most specific (highest), to most sensitive (lowest) to determine if a mutation is deleterious to the function of the gene in which it occurs. (B) Crystal structure of protein product of AmpD (PDB ID:2Y2B, [96]) in complex with 1,6-anhydro-N-acetylmuramic acid and L-ala-gamma-D-glu-meso-diaminopimelic acid, which are associated to the cell-wall degradation pathway. Mutations found in our *B. multivorans* population are colored in red or blue (mutations associated with β-lactam resistance).
(PDF)

S11 Fig. Phylogenetic analysis of SNPs associated with antibiotic resistance. Maximum likelihood phylogenies for SNPs associated with resistance to A) Amikacin and Tobramycin, B) Ciprofloxacin, C) Aztreonam, and D) Ceftazidime were recreated in MEGA7 using the GTR model and 500 bootstrap iteration [81]. Each phylogeny was midpoint rooted.
(PDF)

S12 Fig. Population and single isolate sequencing. Sequencing reads from each isolate from the post-transplant sample were rarified to 1/10th of the number of reads in the population sequencing experiment; then they were combined so that the number of reads would be the same for both experiments. Sequencing reads from the population and single isolate experiments were mapped to the same reference as described above. Mutation allele frequencies for both experiments were calculated using the quality thresholds described by Lieberman *et al.* [53]. (A) Grey circles represent mutation allele frequencies in the deep population sequencing

experiment (y axis) versus in single isolate sequencing (x axis). The dashed line represents the $x = y$ function and the solid line is the best fit line taking into account all data points (R^2 = 0.9928, 95% confidence interval = 0.9918–0.9937). Red circles represent alleles found in the single isolate sequencing experiment but not in the deep sequencing one. Fixed mutations between the reference and all the post-transplant isolates are colored blue. (B) Proportion of false positives in the single isolate sequencing experiment.
(PDF)

S13 Fig. Determining the number of ancestral populations that explain the variance and covariance in CF170 *B. multivorans* population. We ran three independent chains for each K between one and ten. The estimated ln probability of data plateaus at K = 3 in all chains.
(PDF)

S14 Fig. Regression analysis of the root-to-tip distance as a function of time of isolation using the TempEst program [97]. Each circle represents the average root-to-tip distance of the isolates from the respective sampling time point. The resulting trend shows that the inferred molecular clock was consistent with the changes seen in our isolates through time (R^2 = 0.97, $P < 0.0001$).
(PDF)

S15 Fig. Phylogenetic analysis of recombinogenic regions. We recreated the phylogenies for each of the identified recombinogenic regions using the Maximum Likelihood method as implemented in MEGA7 with the GTR model and 500 bootstrap iterations [74]. The labels of each phylogeny correspond to the labels in Fig 2E. Each tree was rooted midpoint.
(PDF)

S1 Table. Mutations occurring in the recombinogenic region between RB & B isolates.
(DOCX)

S2 Table. Mutations occurring in the recombinogenic regions among post-transplant isolates.
(DOCX)

S3 Table. NCBI BioProject ID, BioSample IDs, and Genbank Accession IDs for genomes of study isolates.
(DOCX)

Acknowledgments

We would like to thank Dr. Tami Lieberman for her assistance with the estimation of the d_N/d_S rates, and the entire Guttman laboratory for their helpful comments and input.

Author Contributions

Conceptualization: Julio Diaz Caballero, Yvonne C. W. Yau, David M. Hwang, David S. Guttman.

Data curation: Julio Diaz Caballero, Pauline W. Wang.

Formal analysis: Julio Diaz Caballero, Shawn T. Clark, Bryan Coburn.

Funding acquisition: David M. Hwang, David S. Guttman.

Investigation: Julio Diaz Caballero, Shawn T. Clark, Bryan Coburn, David S. Guttman.

Methodology: Julio Diaz Caballero.

Project administration: Sylva L. Donaldson, David S. Guttman.

Resources: D. Elizabeth Tullis, Yvonne C. W. Yau, Valerie J. Waters, David M. Hwang, David S. Guttman.

Supervision: Pauline W. Wang, Bryan Coburn, David M. Hwang, David S. Guttman.

Writing – original draft: Julio Diaz Caballero, David S. Guttman.

Writing – review & editing: Julio Diaz Caballero, Shawn T. Clark, Bryan Coburn, Yvonne C. W. Yau, Valerie J. Waters, David M. Hwang, David S. Guttman.

References

1. Vandamme P, Dawyndt P. Classification and identification of the *Burkholderia cepacia* complex: Past, present and future. Syst Appl Microbiol. 2011; 34(2):87–95. https://doi.org/10.1016/j.syapm.2010.10.002 PMID: 21257278
2. De Smet B, Mayo M, Peeters C, Zlosnik JE, Spilker T, Hird TJ, et al. *Burkholderia stagnalis* sp. nov. and *Burkholderia territorii* sp. nov., two novel *Burkholderia cepacia* complex species from environmental and human sources. Int J Syst Evol Microbiol. 2015; 65(7):2265–71. https://doi.org/10.1099/ijs.0.000251 PMID: 25872960
3. Courtney JM, Bradley J, McCaughan J, O'Connor TM, Shortt C, Bredin CP, et al. Predictors of mortality in adults with cystic fibrosis. Pediatr Pulmonol. 2007; 42(6):525–32. https://doi.org/10.1002/ppul.20619 PMID: 17469153
4. Stephenson AL, Sykes J, Berthiaume Y, Singer LG, Aaron SD, Whitmore GA, et al. Clinical and demographic factors associated with post-lung transplantation survival in individuals with cystic fibrosis. J Heart Lung Transplant. 2015; 34(9):1139–45. https://doi.org/10.1016/j.healun.2015.05.003 PMID: 26087666
5. Drevinek P, Mahenthiralingam E. *Burkholderia cenocepacia* in cystic fibrosis: epidemiology and molecular mechanisms of virulence. Clin Microbiol Infect. 2010; 16(7):821–30. https://doi.org/10.1111/j.1469-0691.2010.03237.x PMID: 20880411
6. Lipuma JJ. The changing microbial epidemiology in cystic fibrosis. Clin Microbiol Rev. 2010; 23(2):299–323. https://doi.org/10.1128/CMR.00068-09 PMID: 20375354
7. Lipuma JJ. Update on the *Burkholderia cepacia* complex. Curr Opin Pulm Med. 2005; 11(6):528–33. PMID: 16217180
8. Jones AM, Dodd ME, Govan JR, Barcus V, Doherty CJ, Morris J, et al. *Burkholderia cenocepacia* and *Burkholderia multivorans*: influence on survival in cystic fibrosis. Thorax. 2004; 59(11):948–51. https://doi.org/10.1136/thx.2003.017210 PMID: 15516469
9. Leitao JH, Sousa SA, Ferreira AS, Ramos CG, Silva IN, Moreira LM. Pathogenicity, virulence factors, and strategies to fight against *Burkholderia cepacia* complex pathogens and related species. Appl Microbiol Biotechnol. 2010; 87(1):31–40. https://doi.org/10.1007/s00253-010-2528-0 PMID: 20390415
10. Zlosnik JE, Zhou G, Brant R, Henry DA, Hird TJ, Mahenthiralingam E, et al. *Burkholderia* species infections in patients with cystic fibrosis in British Columbia, Canada. 30 years' experience. Ann Am Thorac Soc. 2015; 12(1):70–8. https://doi.org/10.1513/AnnalsATS.201408-395OC PMID: 25474359
11. Rhodes KA, Schweizer HP. Antibiotic resistance in *Burkholderia* species. Drug Resist Updat. 2016; 28:82–90. https://doi.org/10.1016/j.drup.2016.07.003 PMID: 27620956
12. Price AL, Spencer CC, Donnelly P. Progress and promise in understanding the genetic basis of common diseases. Proc Biol Sci. 2015; 282(1821):20151684. https://doi.org/10.1098/rspb.2015.1684 PMID: 26702037
13. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet. 2008; 9(5):356–69. https://doi.org/10.1038/nrg2344 PMID: 18398418
14. Power RA, Parkhill J, de Oliveira T. Microbial genome-wide association studies: lessons from human GWAS. Nat Rev Genet. 2017; 18(1):41–50. https://doi.org/10.1038/nrg.2016.132 PMID: 27840430
15. Farhat MR, Shapiro BJ, Kieser KJ, Sultana R, Jacobson KR, Victor TC, et al. Genomic analysis identifies targets of convergent positive selection in drug-resistant *Mycobacterium tuberculosis*. Nat Genet. 2013; 45(10):1183–9. https://doi.org/10.1038/ng.2747 PMID: 23995135

16. Chewapreecha C, Harris SR, Croucher NJ, Turner C, Marttinen P, Cheng L, et al. Dense genomic sampling identifies highways of pneumococcal recombination. Nat Genet. 2014; 46(3):305–9. https://doi.org/10.1038/ng.2895 PMID: 24509479

17. Chewapreecha C, Marttinen P, Croucher NJ, Salter SJ, Harris SR, Mather AE, et al. Comprehensive identification of single nucleotide polymorphisms associated with beta-lactam resistance within Pneumococcal mosaic genes. PLoS Genet. 2014; 10(8):e1004547. https://doi.org/10.1371/journal.pgen.1004547 PMID: 25101644

18. Chen PE, Shapiro BJ. The advent of genome-wide association studies for bacteria. Curr Opin Microbiol. 2015; 25:17–24. https://doi.org/10.1016/j.mib.2015.03.002 PMID: 25835153

19. Sheppard SK, Didelot X, Meric G, Torralbo A, Jolley KA, Kelly DJ, et al. Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in *Campylobacter*. Proc Natl Acad Sci U S A. 2013; 110(29):11923–7. https://doi.org/10.1073/pnas.1305559110 PMID: 23818615

20. Chaston JM, Newell PD, Douglas AE. Metagenome-wide association of microbial determinants of host phenotype in *Drosophila melanogaster*. MBio. 2014; 5(5):e01631–14. https://doi.org/10.1128/mBio.01631-14 PMID: 25271286

21. Earle SG, Wu CH, Charlesworth J, Stoesser N, Gordon NC, Walker TM, et al. Identifying lineage effects when controlling for population structure improves power in bacterial association studies. Nat Microbiol. 2016; 1:16041. https://doi.org/10.1038/nmicrobiol.2016.41 PMID: 27572646

22. Didelot X, Maiden MC. Impact of recombination on bacterial evolution. Trends Microbiol. 2010; 18(7):315–22. https://doi.org/10.1016/j.tim.2010.04.002 PMID: 20452218

23. Hughes D, Andersson DI. Evolutionary consequences of drug resistance: shared principles across diverse targets and organisms. Nat Rev Genet. 2015; 16(8):459–71. https://doi.org/10.1038/nrg3922 PMID: 26149714

24. Jolley KA, Maiden MC. BIGSdb: Scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics. 2010; 11:595. https://doi.org/10.1186/1471-2105-11-595 PMID: 21143983

25. Diaz Caballero J, Clark ST, Coburn B, Zhang Y, Wang PW, Donaldson SL, et al. Selective sweeps and parallel pathoadaptation drive *Pseudomonas aeruginosa* evolution in the cystic fibrosis lung. MBio. 2015; 6(5):e00981–15. https://doi.org/10.1128/mBio.00981-15 PMID: 26330513

26. McVean G. The structure of linkage disequilibrium around a selective sweep. Genetics. 2007; 175(3):1395–406. https://doi.org/10.1534/genetics.106.062828 PMID: 17194788

27. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000; 155(2):945–59. PMID: 10835412

28. Kalinowski ST. The computer program STRUCTURE does not reliably identify the main genetic clusters within species: simulations and implications for human population structure. Heredity (Edinb). 2011; 106(4):625–32. https://doi.org/10.1038/hdy.2010.95 PMID: 20683484

29. CLSI. Methods for Dilution Antimicrobial Susceptibility Tests for Bacteria That Grow Aerobically; Approved Standard M07-A9. Wayne, PA: Clinical and Laboratory Standards Institute; 2012.

30. Hwang J, Kim HS. Cell wall recycling-linked coregulation of AmpC and PenB beta-lactamases through *ampD* mutations in *Burkholderia cenocepacia*. Antimicrob Agents Chemother. 2015; 59(12):7602–10. https://doi.org/10.1128/AAC.01068-15 PMID: 26416862

31. Kong KF, Schneper L, Mathee K. Beta-lactam antibiotics: from antibiosis to resistance and bacteriology. APMIS. 2010; 118(1):1–36. https://doi.org/10.1111/j.1600-0463.2009.02563.x PMID: 20041868

32. Sokurenko EV, Hasty DL, Dykhuizen DE. Pathoadaptive mutations: gene loss and variation in bacterial pathogens. Trends Microbiol. 1999; 7(5):191–5. PMID: 10354593

33. Wood TE, Burke JM, Rieseberg LH. Parallel genotypic adaptation: when evolution repeats itself. Genetica. 2005; 123(1–2):157–70. PMID: 15881688

34. Podnecky NL, Wuthiekanun V, Peacock SJ, Schweizer HP. The BpeEF-OprC efflux pump is responsible for widespread trimethoprim resistance in clinical and environmental *Burkholderia pseudomallei* isolates. Antimicrob Agents Chemother. 2013; 57(9):4381–6. https://doi.org/10.1128/AAC.00660-13 PMID: 23817379

35. Schweizer HP. Mechanisms of antibiotic resistance in *Burkholderia pseudomallei*: implications for treatment of melioidosis. Future Microbiol. 2012; 7(12):1389–99. https://doi.org/10.2217/fmb.12.116 PMID: 23231488

36. Randall LB, Georgi E, Genzel GH, Schweizer HP. Finafloxacin overcomes *Burkholderia pseudomallei* efflux-mediated fluoroquinolone resistance. J Antimicrob Chemother. 2017; 72(4):1258–60. https://doi.org/10.1093/jac/dkw529 PMID: 28039270

37. Stephenson K, Hoch JA. Two-component and phosphorelay signal-transduction systems as therapeutic targets. Curr Opin Pharmacol. 2002; 2(5):507–12. PMID: 12324251

38. Potvin E, Sanschagrin F, Levesque RC. Sigma factors in *Pseudomonas aeruginosa*. FEMS Microbiol Rev. 2008; 32(1):38–55. https://doi.org/10.1111/j.1574-6976.2007.00092.x PMID: 18070067

39. Hudson RR, Kaplan NL. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics. 1985; 111(1):147–64. PMID: 4029609

40. Marttinen P, Hanage WP, Croucher NJ, Connor TR, Harris SR, Bentley SD, et al. Detection of recombination events in bacterial genomes from large population samples. Nucleic Acids Res. 2012; 40(1):e6. https://doi.org/10.1093/nar/gkr928 PMID: 22064866

41. Beaume M, Kohler T, Greub G, Manuel O, Aubert JD, Baerlocher L, et al. Rapid adaptation drives invasion of airway donor microbiota by *Pseudomonas* after lung transplantation. Sci Rep. 2017; 7:40309. https://doi.org/10.1038/srep40309 PMID: 28094327

42. Nunley DR, Grgurich W, Iacono AT, Yousem S, Ohori NP, Keenan RJ, et al. Allograft colonization and infections with *Pseudomonas* in cystic fibrosis lung transplant recipients. Chest. 1998; 113(5):1235–43. PMID: 9596300

43. Walter S, Gudowius P, Bosshammer J, Romling U, Weissbrodt H, Schurmann W, et al. Epidemiology of chronic *Pseudomonas aeruginosa* infections in the airways of lung transplant recipients with cystic fibrosis. Thorax. 1997; 52(4):318–21. PMID: 9196512

44. Folkesson A, Jelsbak L, Yang L, Johansen HK, Ciofu O, Hoiby N, et al. Adaptation of *Pseudomonas aeruginosa* to the cystic fibrosis airway: an evolutionary perspective. Nat Rev Microbiol. 2012; 10 (12):841–51. https://doi.org/10.1038/nrmicro2907 PMID: 23147702

45. Zalucki YM, Dhulipala V, Shafer WM. Dueling regulatory properties of a transcriptional activator (MtrA) and repressor (MtrR) that control efflux pump gene expression in *Neisseria gonorrhoeae*. MBio. 2012; 3 (6):e00446–12. https://doi.org/10.1128/mBio.00446-12 PMID: 23221802

46. Warner DM, Folster JP, Shafer WM, Jerse AE. Regulation of the MtrC-MtrD-MtrE efflux-pump system modulates the in vivo fitness of *Neisseria gonorrhoeae*. J Infect Dis. 2007; 196(12):1804–12. https://doi.org/10.1086/522964 PMID: 18190261

47. Tsujimoto H, Gotoh N, Yamagishi J, Oyamada Y, Nishino T. Cloning and expression of the major porin protein gene opcP of *Burkholderia* (formerly *Pseudomonas*) cepacia in *Escherichia coli*. Gene. 1997; 186(1):113–8. PMID: 9047353

48. Hancock RE. Resistance mechanisms in *Pseudomonas aeruginosa* and other nonfermentative gram-negative bacteria. Clin Infect Dis. 1998; 27 Suppl 1:S93–9.

49. Mira NP, Madeira A, Moreira AS, Coutinho CP, Sa-Correia I. Genomic expression analysis reveals strategies of *Burkholderia cenocepacia* to adapt to cystic fibrosis patients' airways and antimicrobial therapy. PLoS One. 2011; 6(12):e28831. https://doi.org/10.1371/journal.pone.0028831 PMID: 22216120

50. Blair JM, Webber MA, Baylay AJ, Ogbolu DO, Piddock LJ. Molecular mechanisms of antibiotic resistance. Nat Rev Microbiol. 2015; 13(1):42–51. https://doi.org/10.1038/nrmicro3380 PMID: 25435309

51. Garcia-Solache M, Lebreton F, McLaughlin RE, Whiteaker JD, Gilmore MS, Rice LB. Homologous recombination within large chromosomal regions facilitates acquisition of beta-lactam and vancomycin resistance in *Enterococcus faecium*. Antimicrob Agents Chemother. 2016; 60(10):5777–86. https://doi.org/10.1128/AAC.00488-16 PMID: 27431230

52. Aubert D, Naas T, Nordmann P. Integrase-mediated recombination of the *veb1* gene cassette encoding an extended-spectrum beta-lactamase. PLoS One. 2012; 7(12):e51602. https://doi.org/10.1371/journal.pone.0051602 PMID: 23251590

53. Lieberman TD, Michel JB, Aingaran M, Potter-Bynoe G, Roux D, Davis MR Jr., et al. Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes. Nat Genet. 2011; 43 (12):1275–80. https://doi.org/10.1038/ng.997 PMID: 22081229

54. Silva IN, Santos PM, Santos MR, Zlosnik JE, Speert DP, Buskirk SW, et al. Long-term evolution of *Burkholderia multivorans* during a chronic cystic fibrosis infection reveals shifting forces of selection. mSystems. 2016; 1(3)

55. Nunvar J, Capek V, Fiser K, Fila L, Drevinek P. What matters in chronic *Burkholderia cenocepacia* infection in cystic fibrosis: Insights from comparative genomics. PLoS Pathog. 2017; 13(12):e1006762. https://doi.org/10.1371/journal.ppat.1006762 PMID: 29228063

56. Ciofu O, Johansen HK, Aanaes K, Wassermann T, Alhede M, von Buchwald C, et al. *P. aeruginosa* in the paranasal sinuses and transplanted lungs have similar adaptive mutations as isolates from chronically infected CF lungs. J Cyst Fibros. 2013; 12(6):729–36. https://doi.org/10.1016/j.jcf.2013.02.004 PMID: 23478131

57. Marvig RL, Sommer LM, Molin S, Johansen HK. Convergent evolution and adaptation of *Pseudomonas aeruginosa* within patients with cystic fibrosis. Nat Genet. 2015; 47(1):57–64.

58. Arellano-Reynoso B, Lapaque N, Salcedo S, Briones G, Ciocchini AE, Ugalde R, et al. Cyclic beta-1,2-glucan is a *Brucella* virulence factor required for intracellular survival. Nat Immunol. 2005; 6(6):618–25. https://doi.org/10.1038/ni1202 PMID: 15880113

59. Schaefers MM, Liao TL, Boisvert NM, Roux D, Yoder-Himes D, Priebe GP. An Oxygen-Sensing two-component system in the *Burkholderia cepacia* complex regulates biofilm, intracellular invasion, and pathogenicity. PLoS Pathog. 2017; 13(1):e1006116. https://doi.org/10.1371/journal.ppat.1006116 PMID: 28046077

60. Silva IN, Pessoa FD, Ramires MJ, Santos MR, Becker JD, Cooper VS, et al. OmpR regulator of *Burkholderia multivorans* controls mucoid-to-nonmucoid transition and other cell envelope properties associated with persistence in the cystic fibrosis lung. J Bacteriol. 2018;

61. Viberg LT, Sarovich DS, Kidd TJ, Geake JB, Bell SC, Currie BJ, et al. Within-host evolution of *Burkholderia pseudomallei* during chronic infection of seven Australasian cystic fibrosis patients. MBio. 2017; 8 (2)

62. Clark ST, Diaz Caballero J, Cheang M, Coburn B, Wang PW, Donaldson SL, et al. Phenotypic diversity within a *Pseudomonas aeruginosa* population infecting an adult with cystic fibrosis. Sci Rep. 2015; 5:10932. https://doi.org/10.1038/srep10932 PMID: 26047320

63. Spilker T, Baldwin A, Bumford A, Dowson CG, Mahenthiralingam E, LiPuma JJ. Expanded multilocus sequence typing for *Burkholderia* species. J Clin Microbiol. 2009; 47(8):2607–10. https://doi.org/10.1128/JCM.00770-09 PMID: 19494070

64. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014; 30(15):2114–20. https://doi.org/10.1093/bioinformatics/btu170 PMID: 24695404

65. Yu YS, Rambo T, Currie J, Saski C, Kim HR, Collura K, et al. In-depth view of structure, activity, and evolution of rice chromosome 10. Science. 2003; 300(5625):1566–9. https://doi.org/10.1126/science.1083523 PMID: 12791992

66. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. 2011. 2011; 17(1)

67. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997; 25 (17):3389–402. PMID: 9254694

68. Winsor GL, Khaira B, Van Rossum T, Lo R, Whiteside MD, Brinkman FS. The *Burkholderia* Genome Database: facilitating flexible queries and comparative analyses. Bioinformatics. 2008; 24(23):2803–4. https://doi.org/10.1093/bioinformatics/btn524 PMID: 18842600

69. Gotz S, Garcia-Gomez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, et al. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res. 2008; 36(10):3420–35. https://doi.org/10.1093/nar/gkn176 PMID: 18445632

70. Kanehisa M, Sato Y, Morishima K. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. J Mol Biol. 2016; 428(4):726–31. https://doi.org/10.1016/j.jmb.2015.11.006 PMID: 26585406

71. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25(14):1754–60. https://doi.org/10.1093/bioinformatics/btp324 PMID: 19451168

72. Shrestha AM, Frith MC. An approximate Bayesian approach for mapping paired-end DNA reads to a reference genome. Bioinformatics. 2013; 29(8):965–72. https://doi.org/10.1093/bioinformatics/btt073 PMID: 23413433

73. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25(16):2078–9. https://doi.org/10.1093/bioinformatics/btp352 PMID: 19505943

74. Albers CA, Lunter G, MacArthur DG, McVean G, Ouwehand WH, Durbin R. Dindel: accurate indel calls from short-read data. Genome Res. 2011; 21(6):961–73. https://doi.org/10.1101/gr.112326.110 PMID: 20980555

75. Ronquist F, Huelsenbeck JP. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 2003; 19(12):1572–4. PMID: 12912839

76. Darriba D, Taboada GL, Doallo R, Posada D. jModelTest 2: more models, new heuristics and parallel computing. Nat Methods. 2012; 9(8):772.

77. Stanier RY, Palleroni NJ, Doudoroff M. The aerobic pseudomonads: a taxonomic study. J Gen Microbiol. 1966; 43(2):159–271. https://doi.org/10.1099/00221287-43-2-159 PMID: 5963505

78. Huson DH, Bryant D. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol. 2006; 23(2):254–67. https://doi.org/10.1093/molbev/msj030 PMID: 16221896

79. Porras-Hurtado L, Ruiz Y, Santos C, Phillips C, Carracedo A, Lareu MV. An overview of STRUCTURE: applications, parameter settings, and supporting software. Front Genet. 2013; 4:98. https://doi.org/10.3389/fgene.2013.00098 PMID: 23755071

80. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol. 2012; 29(8):1969–73. https://doi.org/10.1093/molbev/mss075 PMID: 22367748

81. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol. 2016; 33(7):1870–4. https://doi.org/10.1093/molbev/msw054 PMID: 27004904

82. Gill MS, Lemey P, Faria NR, Rambaut A, Shapiro B, Suchard MA. Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. Mol Biol Evol. 2013; 30(3):713–24. https://doi.org/10.1093/molbev/mss265 PMID: 23180580

83. Librado P, Rozas J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics. 2009; 25(11):1451–2. https://doi.org/10.1093/bioinformatics/btp187 PMID: 19346325

84. Bruen TC, Philippe H, Bryant D. A simple and robust statistical test for detecting the presence of recombination. Genetics. 2006; 172(4):2665–81. https://doi.org/10.1534/genetics.105.048975 PMID: 16489234

85. Didelot X, Wilson DJ. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. PLoS Comput Biol. 2015; 11(2):e1004041. https://doi.org/10.1371/journal.pcbi.1004041 PMID: 25675341

86. Cohen O, Ashkenazy H, Belinky F, Huchon D, Pupko T. GLOOME: gain loss mapping engine. Bioinformatics. 2010; 26(22):2914–5. https://doi.org/10.1093/bioinformatics/btq549 PMID: 20876605

87. Lieberman TD, Flett KB, Yelin I, Martin TR, McAdam AJ, Priebe GP, et al. Genetic variation of a bacterial pathogen within individuals with cystic fibrosis provides a record of selective pressures. Nat Genet. 2014; 46(1):82–7. https://doi.org/10.1038/ng.2848 PMID: 24316980

88. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. PLoS One. 2012; 7(10):e46688. https://doi.org/10.1371/journal.pone.0046688 PMID: 23056405

89. Chris L, Daniel G, Leslie G, Richard P, Neil B, Michael C, et al. SciNet: Lessons Learned from Building a Power-efficient Top-20 System and Data Centre. Journal of Physics: Conference Series. 2010; 256 (1):012026.

90. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an information aesthetic for comparative genomics. Genome Res. 2009; 19(9):1639–45. https://doi.org/10.1101/gr.092759.109 PMID: 19541911

91. Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res. 2016; 44(W1):W242–5. https://doi.org/10.1093/nar/gkw290 PMID: 27095192

92. Miller CA, McMichael J, Dang HX, Maher CA, Ding L, Ley TJ, et al. Visualizing tumor evolution with the fishplot package for R. BMC Genomics. 2016; 17(1):880. https://doi.org/10.1186/s12864-016-3195-z PMID: 27821060

93. Dixon P. VEGAN, a package of R functions for community ecology. J Veg Sci. 2003; 14(6):927–30.

94. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004; 32(5):1792–7. https://doi.org/10.1093/nar/gkh340 PMID: 15034147

95. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014; 30(9):1312–3. https://doi.org/10.1093/bioinformatics/btu033 PMID: 24451623

96. Carrasco-Lopez C, Rojas-Altuve A, Zhang W, Hesek D, Lee M, Barbe S, et al. Crystal structures of bacterial peptidoglycan amidase AmpD and an unprecedented activation mechanism. J Biol Chem. 2011; 286(36):31714–22. https://doi.org/10.1074/jbc.M111.264366 PMID: 21775432

97. Rambaut A, Lam TT, Max Carvalho L, Pybus OG. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). Virus Evol. 2016; 2(1):vew007. https://doi.org/10.1093/ve/vew007 PMID: 27774300