nature genetics

Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes

Tami D Lieberman[1,12], Jean-Baptiste Michel[1,2,12], Mythili Aingaran[3], Gail Potter-Bynoe[4], Damien Roux[5], Michael R Davis Jr[6], David Skurnik[5], Nicholas Leiby[1], John J LiPuma[7,8], Joanna B Goldberg[6], Alexander J McAdam[9], Gregory P Priebe[3,5,10] & Roy Kishony[1,11]

Bacterial pathogens evolve during the infection of their human host[1–8], but separating adaptive and neutral mutations remains challenging[9–11]. Here we identify bacterial genes under adaptive evolution by tracking recurrent patterns of mutations in the same pathogenic strain during the infection of multiple individuals. We conducted a retrospective study of a *Burkholderia dolosa* outbreak among subjects with cystic fibrosis, sequencing the genomes of 112 isolates collected from 14 individuals over 16 years. We find that 17 bacterial genes acquired nonsynonymous mutations in multiple individuals, which indicates parallel adaptive evolution. Mutations in these genes affect important pathogenic phenotypes, including antibiotic resistance and bacterial membrane composition and implicate oxygen-dependent regulation as paramount in lung infections. Several genes have not previously been implicated in pathogenesis and may represent new therapeutic targets. The identification of parallel molecular evolution as a pathogen spreads among multiple individuals points to the key selection forces it experiences within human hosts.

During acute and chronic infections, bacterial pathogens can accumulate mutations that allow them to better adapt to their human hosts[1,2], evade the immune response[12,13] and become more resistant to antibiotic therapy[3,4]. The spectrum of beneficial mutations that arises during the course of a bacterial infection is likely to indicate genetic pathways critical for bacterial pathogenesis *in vivo* and may thus inspire new therapeutic directions. Recent advances in high-throughput sequencing make it possible to follow the genomic evolution of bacterial pathogens[7–9,14–17], but it is still difficult to tease apart the adaptive, driver mutations from neutral, passenger mutations that have been fixed by chance[9–11]. In the laboratory, this difficulty has been addressed by following several populations grown in parallel cultures under identical conditions; the adaptive nature of mutations is indicated by their recurrence in replicate experiments[17–19]. In natural and clinical environments, such studies are more difficult and have not yet been systematically performed on a genome-wide scale. As a result, the global patterns of adaptive evolution that underlie bacterial pathogenesis in humans are not well characterized.

Here we systematically identify recurrent patterns of evolution implicated in pathogenesis by comparing the genetic adaptation of a single bacterial strain in multiple human subjects during the spread of an epidemic. The airways of individuals with cystic fibrosis (a lethal genetic disorder) are particularly prone to long-term bacterial infections. Most individuals with cystic fibrosis become colonized by a dominant bacterial strain that persists for many years[20], allowing significant time for genetic adaption[2]. In the 1990s, a small epidemic of *Burkholderia dolosa*—a rare pathogen associated with cystic fibrosis[21,22] that can be transmitted from person to person[23]—broke out among individuals with cystic fibrosis in Boston[24,25]. A total of 39 individuals were infected (**Fig. 1a**), and all were followed in a Boston hospital where bacteria isolated during normal care were routinely frozen.

We conducted a retrospective study of 112 *B. dolosa* isolates from 14 individuals with cystic fibrosis who were infected in this epidemic outbreak—including the first infected patient in the Boston area, patient zero—over the course of 16 years (**Fig. 1b** and **Supplementary Table 1**). During this period, five of these individuals received a lung transplant, and eight died. Most of the 112 *B. dolosa* isolates were recovered from the subjects' airways, and a few were isolated from the blood of subjects with bacteremia. This collection covers the epidemic with high temporal resolution and enables the study of parallel evolution of the same strain in multiple individuals (**Supplementary Fig. 1**).

We sequenced the whole genome of the 112 *B. dolosa* isolates on an Illumina Genome Analyzer IIx sequencer (75-bp, single-end

[1]Department of Systems Biology, Harvard Medical School, Boston, Massachusetts, USA. [2]Program for Evolutionary Dynamics, Harvard University, Cambridge, Massachusetts, USA. [3]Department of Medicine, Division of Infectious Diseases, Children's Hospital Boston, Boston, Massachusetts, USA. [4]Infection Prevention & Control, Children's Hospital Boston, Boston, Massachusetts, USA. [5]Channing Laboratory, Department of Medicine, Brigham and Women's Hospital, Boston, Massachusetts, USA. [6]Department of Microbiology, University of Virginia Health System, Charlottesville, Virginia, USA. [7]Department of Pediatrics, University of Michigan Medical School, Ann Arbor, Michigan, USA. [8]Department of Epidemiology, University of Michigan School of Public Health, Ann Arbor, Michigan, USA. [9]Department of Laboratory Medicine, Children's Hospital Boston, Boston, Massachusetts, USA. [10]Department of Anesthesia, Division of Critical Care Medicine, Children's Hospital Boston, Boston, Massachusetts, USA. [11]School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts, USA. [12]These authors contributed equally to this work. Correspondence should be addressed to R.K. (roy_kishony@hms.harvard.edu), A.J.M. (alexander.mcadam@childrens.harvard.edu) or G.P.P. (gregory_priebe@childrens.harvard.edu).

Received 6 April; accepted 5 October; published online 13 November 2011; doi:10.1038/ng.997

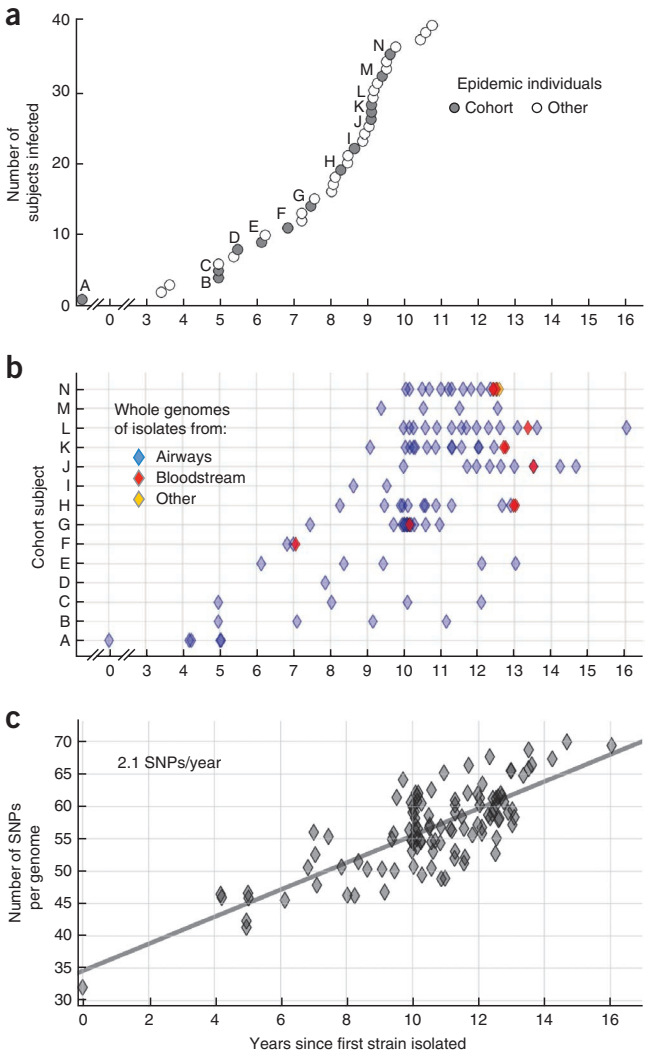

Figure 1 Whole-genome sequencing of 112 epidemic *B. dolosa* isolates recovered from 14 subjects shows the steady accumulation of mutations over years. (**a**) An epidemic of *B. dolosa* spread to 39 individuals with cystic fibrosis (circles) over decades. Time of first attested infection for each subject is indicated in years after the collection of an isolate from patient zero (labeled A). A cohort of 14 individuals from this epidemic (gray circles) was studied retrospectively. (**b**) The genomes of 112 bacterial isolates were sequenced (diamonds; each horizontal line corresponds to a subject). Isolates were recovered over time from the subjects' airways (blue), bloodstreams (red) or other body compartments (yellow; for instance, tissue obtained during surgery). (**c**) The number of SNPs between each isolate and the outgroup is plotted as a function of time (years since first strain isolated). Linear fit is plotted (slope = 2.1 mutations fixed per year).

reads, average read depth of 37×; **Supplementary Fig. 2**) and aligned the reads onto a *B. dolosa* reference genome (*Burkholderia dolosa* Sequencing Project, see URLs). We focused our analysis on SNPs; although structural variants and mobile elements may also be important, they are beyond the scope of this study. Our analysis identified 492 polymorphic loci (**Supplementary Table 2**). The mutations we identified accumulated at a steady rate of ~2 SNPs per year ($r = 0.79$) (**Fig. 1c**), with no discernible difference between subjects (**Supplementary Fig. 3**). This rate of mutation accumulation in the presence of selective forces within the human body is consistent with bacterial mutation-fixation rates reported in long-term human infections[2,26]. The steady accumulation of mutations generated enough genetic diversity to resolve evolutionary relationships between isolates, which were investigated through the creation of a maximum-likelihood phylogenetic tree (**Fig. 2a**).

At the epidemic level, the phylogeny suggested a network of transmission between subjects. Isolates from the same subject tended to form genetically related clusters in the phylogeny (**Fig. 2a** and **Supplementary Fig. 3**). These clusters define subject-specific genetic fingerprints, from which transmission history can be inferred. We constructed the last common ancestor (LCA) for the set of bacterial isolates from each subject having subject-specific fingerprints. The phylogenetic relationships between these inferred strains indicated the likely network of transmission among the 14 subjects (**Fig. 2b**). Because these data account for only 14 of the 39 infected individuals in this epidemic, we cannot determine whether transmission occurred directly from one subject to another or indirectly through an individual not in our study, via a healthcare worker or through a medical device. Nevertheless, this analysis shows that this specific epidemic was transmitted through several people during its spread, indicating the strength of this approach for the identification of the infection network of an epidemic.

At the level of individual subjects, we saw evidence in the phylogenetic analysis of the transfer of multiple *B. dolosa* clones to the subject's bloodstream during bacteremia. We examined isolates from the three subjects from whom we obtained more than one blood isolate (subjects H, K and N). In two of these individuals, we found pairs of blood isolates that evolved from distinct lung isolates (**Fig. 2c**), which is inconsistent with the transmission of a single clone from lungs to blood (**Supplementary Fig. 4a**). This evidence for multiclonal transmission is instead consistent with either a punctuated transmission of multiple clones from the genetically diverse lung[27–30] (**Supplementary Fig. 4b**) or multiple transmissions occurring over time. These different possibilities would lead to recommendations for distinct therapeutic actions: whereas a lung transplant might be effective in preventing the continuous leak of bacteria through lesions of the lung, it would not block the proliferation of bacteria already within the bloodstream. This analysis thus brings into focus unresolved questions about the mechanistic basis of bacteremia.

Finally, we investigated the pathogen's evolution at the gene level. We looked for genetic correlates of known pathogenic phenotypes. We first assayed resistance to ciprofloxacin, a fluoroquinolone frequently prescribed to individuals with cystic fibrosis (**Fig. 3a**). Resistance among the 112 isolates varied over two orders of magnitude (**Supplementary Fig. 5a**). We scored each gene for correlation between the presence of mutations and drug resistance (**Fig. 3b**). This genome-wide association study implicated a single gene in the phenotype, BDAG_02180, a homolog of *Escherichia coli gyrA*. All the genotypes associated with resistance had non-synonymous mutations causing alterations in amino acids T83 or D87, which are known for their role in fluoroquinolone resistance[4,31,32]. Mutations in these residues occurred in six subjects, and in each case, phylogenetic analysis indicated that mutations were independently acquired within the individual after the initial infection event (**Supplementary Fig. 5b).** In some cases, we even found mutations causing alteration of both amino acids in *B. dolosa* strains from the same subject, each carried by a different isolate. These findings support the existence of a strong selective pressure from fluoroquinolones but suggest that there are only a few genetic paths to drug resistance *in vivo*.

We then focused on a second pathogenic phenotype, the presentation of O-antigen repeats in the lipopolysaccharide (LPS) of the bacterial outer membrane, known for its important role in virulence in related species[33–35]. We found that some of our isolates

presented the O-antigen, whereas others did not (**Fig. 3c**). A single nucleotide in the glycosyltransferase gene BDAG_02317 correlated exactly with the presentation of O-antigen repeats (**Fig. 3d**). The ancestral genotype encodes a stop codon at this locus, corresponding to the absence of O-antigen repeats; two different mutations affecting the same amino acid—each restoring a full-length protein—were associated with the presence of the repeats. We confirmed this association experimentally by showing that complementation with

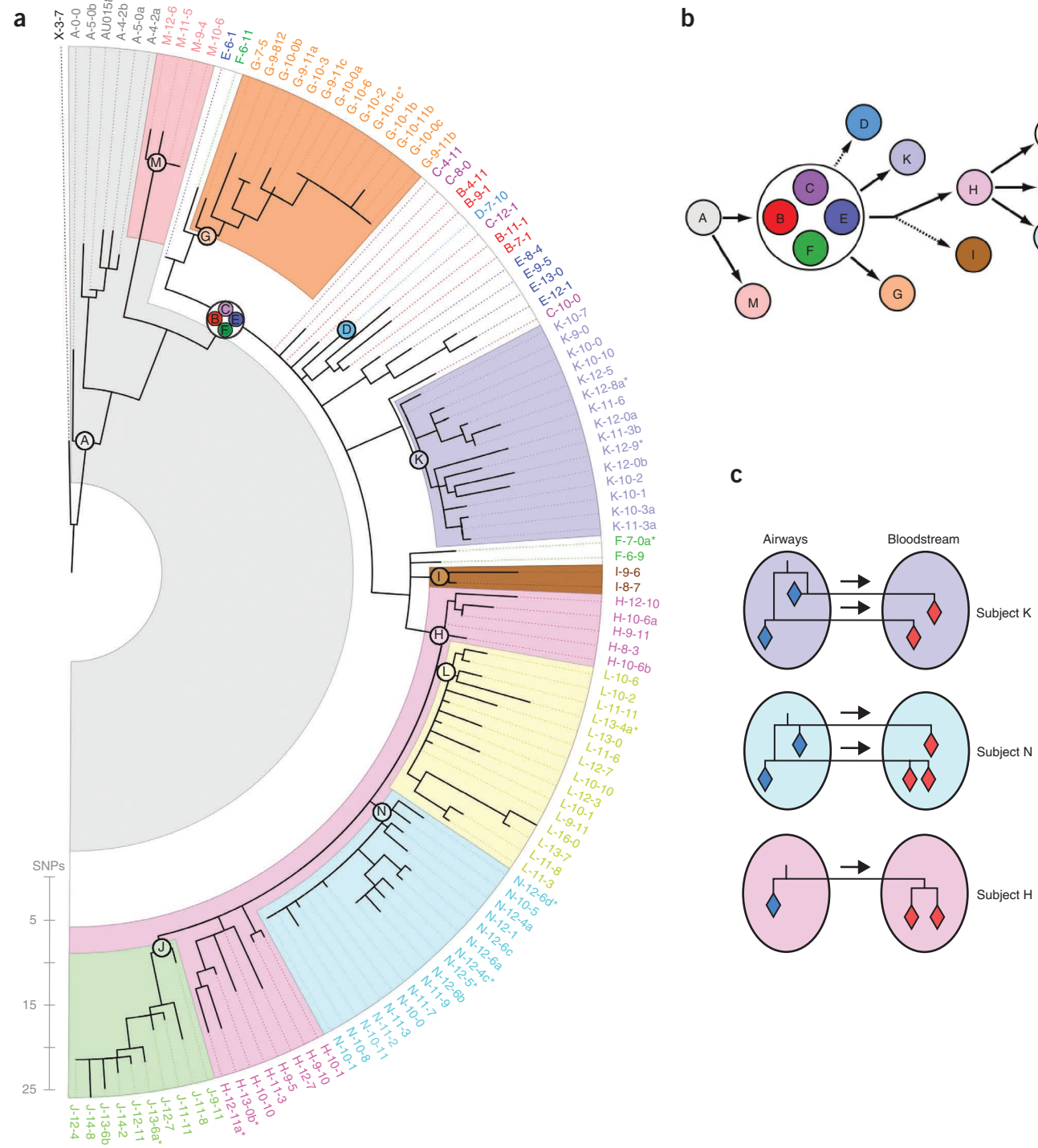

Figure 2 Bacterial phylogeny reveals a likely network of transmission between individuals and between organs. (**a**) Maximum-likelihood phylogenetic tree (SNP scale). The 112 isolates are indicated by thin dashed lines colored according to subject and labeled according to subject and time (for example, C:14-5 was recovered from subject C, 14 years and 5 months after isolation of the first strain). Blood isolates are indicated by an asterisk, and isolates with the same subject ID and date are distinguished by letters. The LCA of isolates from the same subject is represented as a circle of the appropriate color and label. Colored backgrounds indicate subject-specific genetic fingerprints. Subjects B, C, E and F share the same LCA (white background). (**b**) Phylogeny between the inferred LCAs suggests a likely network of infection between subjects (arrows). Dashed arrows indicate less certainty (fewer than three isolates). (**c**) Phylogeny between blood and lung isolates recovered from the same subject shows the transmission of multiple clones to the bloodstream during bacteremia (multiple arrows, subjects K and N).

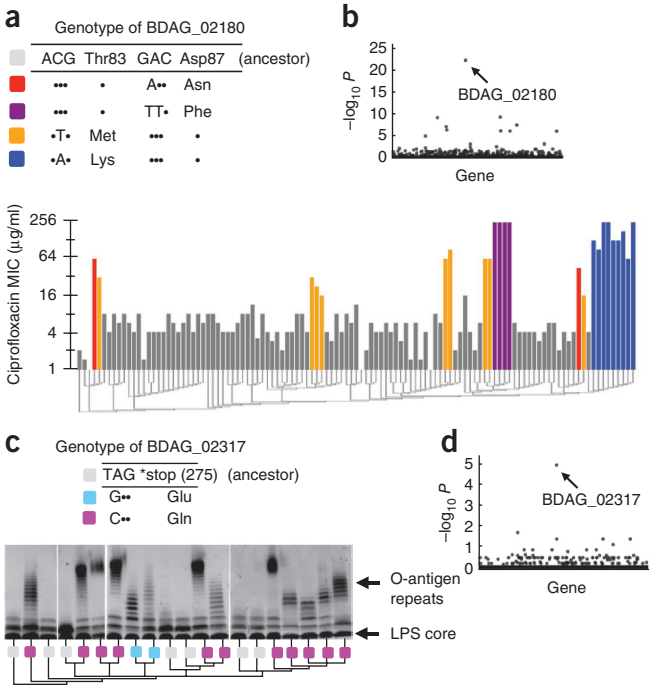

Figure 3 Pathogenic phenotypes are associated with point mutations in key genes. (**a**) Minimal inhibitory concentration (MIC) of ciprofloxacin for each isolate (vertical bars) is correlated with genotypic changes in BDAG_02180, a homolog of *E. coli gyrA*, which result in coding changes in amino acids 83 and 87. Phylogeny is indicated below as a dendrogram, and genotypes at BDAG_02180 are shown in the legend. (**b**) *P* values for correlation between the presence of mutations in each gene and drug resistance levels (Kendall's tau coefficient, τ). (**c**) Silver-stained gels showing the presence of O-antigen repeats (banded pattern) in the LPS of 20 isolates. Phylogeny is indicated below as a dendrogram, and genotypes at BDAG_02317, a homolog of the *wbaD* glycosyltransferase in *E. coli*, are shown in the legend. The presentation of O-antigen repeats corresponds to a recurrent gain-of-function mutation. (**d**) *P* values for correlation between the presence of mutations in each gene and O-antigen presentation (Fisher's exact test).

the full-length glycosyltransferase gene could restore O-antigen presentation (**Supplementary Fig. 6a** and **Supplementary Note**). Harnessing the phylogenetic information, we determined that the LCA of strains from each subject presented the truncated genotype. Thus, the gain-of-function mutations occurred independently in nine subjects (**Supplementary Fig. 6b**), highlighting the strength of the selective pressure for O-antigen presentation during infection. These results identify a previously uncharacterized genetic mechanism for O-antigen switching and hint at the occurrence of a trade-off during person-to-person transmission.

We recognize that the human body challenges bacteria with many selective pressures beyond those discussed above. We therefore developed a systematic approach for identifying genes under positive selection without prior knowledge of the phenotypes being selected. At the genome level, we found no evidence of selection in coding regions (ratio of the mutation rates at non-synonymous and synonymous sites (dN/dS) = ~1) and no significant intragenic bias (**Supplementary Note**). However, we reasoned that genes under selection would be mutated independently in different subjects[17–19]. We leveraged the phylogeny to calculate the number of mutations each gene received, distinguishing genes mutated multiple times from those mutated once but appearing in several subjects through the expansion of the lineage that carried them. We counted 561 independent mutational events in 304 genes (**Supplementary Table 3**). Assuming neutral evolution, we would expect that these mutations would be observed with a random distribution among the 5,014 *B. dolosa* genes and that genes would rarely acquire more than a single mutation. Instead, we observed that many more genes than expected contained multiple mutations (**Fig. 4a**). Seventeen genes were found to have three or more different mutations (neutral expected: ~1 gene, see Online Methods), and four genes had more than ten different mutations (expected: 0 genes).

To determine whether genes that acquired multiple mutations were under positive selection or were merely sites of mutational bias, we calculated the canonical measure for selection, dN/dS (**Fig. 4b**). The large subset of 247 genes that contained only one mutation showed a weak but significant signal for purifying (negative) selection (dN/dS = 0.63,

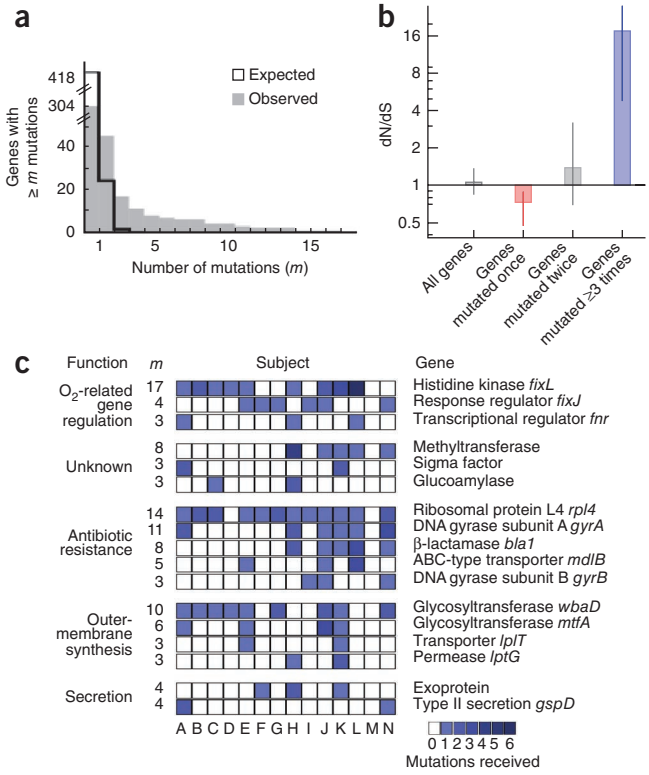

Figure 4 Parallel evolution identifies a set of genes under strong selection during pathogenesis. (**a**) The number of genes that acquired at least *m* mutations during the epidemic is plotted as a function of *m* (gray bars). This distribution contrasts sharply with the distribution expected for neutral evolution (black line). Under neutral evolution, the expectation is that only one gene would receive three or more mutations ($m \geq 3$); instead, 17 such genes were observed. (**b**) The canonical signal for selection (dN/dS) was calculated for the 17 genes with three or more mutations ($m \geq 3$; 109 mutations in all 17 genes), the 28 genes with $m = 2$ and the 247 genes with $m = 1$ (**Supplementary Fig. 3** presents the contribution of these mutations to the molecular clock). dN/dS >1 indicates positive selection (blue), and dN/dS <1 indicates purifying selection (red). Error bars indicate 95% CIs. Calculated over all genes without regard to *m*, this analysis would not show selection. (**c**) Each of the 17 genes (rows) under positive selection contained an acquired mutation in several subjects, as signified by squares (color intensity indicates the number of mutations observed within this subject). The total number of mutations observed within that gene, *m*, is indicated left. Genes are grouped by biological function and labeled with the annotations of close homologs and, when available, with the names of these homologs.

$P < 1 \times 10^{-3}$; see Online Methods). The set of genes with two mutations did not show evidence of selection (dN/dS = 1.4, confidence interval (CI) = 0.7–3.1); it is possible that this set may include a combination of genes that are under some selection and genes that fixed two mutations by chance (22 genes expected under neutral drift, 28 observed). By contrast, the 17 genes that acquired three or more mutations had 18 times as many nonsynonymous mutations as expected by neutral drift and were under strong positive selection (dN/dS = 18, CI = 4.9–152.7). This finding suggests that these 17 genes are not neutral mutational hotspots but underwent adaptive evolution under the pressure of natural selection.

The 17 genes under positive selection (**Fig. 4c** and **Supplementary Table 4**), which are mostly conserved across the *Burkholderia* genus (**Supplementary Fig. 7**), indicate genetic pathways that may be involved in pathogenesis. The inclusion within this group of the two genes previously identified in connection with antibiotic resistance and O-antigen presentation (*gyrA*, 11 mutations; glycosyltransferase *wbaD*, 10 mutations) further connects these genes to pathogenic phenotypes arising under selection. Eleven of the 17 genes encode proteins that belong to functional categories related to pathogenicity, including membrane synthesis (4 genes, including 2 involved in LPS biosynthesis), secretion (2 genes) and antibiotic resistance (5 genes). The presence of six mutations in a second glycosyltransferase in the O-antigen cluster stresses the importance of this pathway for disease development. Notably, the other six genes had not previously been implicated in the pathogenesis of lung infections. Three of these—encoding a glucoamylase, a methyltransferase and a sigma factor—have no well-annotated close homolog, and their roles in pathogenicity are thus unclear. Another gene trio (homologs of *fnr*, *fixL* and *fixJ*), including the most-mutated gene (BDAG_01161, a homolog of *fixL*, that had 17 non-synonymous mutations), can be linked through homology to oxygen-dependent gene regulation[36]. The large number of mutations in this pathway resonates with reports of lowered oxygen tension in the mucus of individuals with cystic fibrosis[37] and of ties between oxygen sensing and virulence modulation in the gastrointestinal tract[38]. Homologs of these three genes have been implicated in diverse regulatory processes[36,38], but their function and the genes they regulate in *B. dolosa* are currently unknown. The identification of 17 *B. dolosa* genes that underwent selective pressure during infection in subjects with cystic fibrosis highlights key pathways involved in pathogenesis and may suggest new therapeutic targets for this and other lung infections.

Tracking the genomic evolution of bacterial pathogens during the infection of their human hosts provides a direct method for observing evolutionary mechanisms *in vivo* and allows the identification of genes central to pathogenesis. This study, which harnesses the combination of high-throughput sequencing and parallel evolution in the clinical setting, is a step toward a comprehensive understanding of genetic adaptation during pathogenesis. Systematically identifying selective pressures acting on pathogens within their hosts may help identify new therapeutic directions.

URLs. *Burkholderia dolosa* Sequencing Project, http://www.broadinstitute.org/; PHYLIP, http://evolution.genetics.washington.edu/phylip.html.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

Accession numbers. Consensus sequences for all 113 *B. dolosa* isolates have been deposited in the NCBI Sequence Read Archive. Accession numbers are listed in **Supplementary Table 1**.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We are grateful to M. Caimano, M. Cendron, P. Kokorowski, S. Lory, C. Marx, N. Delany, S. Walker, M. Waldor and R. Ward for insightful discussions and comments, to O. Iartchouk, A. Brown, M. Light and their team at Partners HealthCare Center for Personalized Genetic Medicine (PCPGM) for Illumina sequencing, to J. Deane and L. Williams for technical assistance, to S. Vargas for assistance with IRB protocols and to M. Baym, M. Ernebjerg, A. Palmer, E. Toprak, K. Vetsigian, Z. Yao and all of the Kishony lab members for helpful discussions and general support. J.B.M. was supported by the Foundational Questions in Evolutionary Biology Prize Fellowship and the Systems Biology PhD Program (Harvard Medical School). G.P.P. was supported in part by The Mannion Fund for Research of the Center for the Critically Ill Child of Children's Hospital Boston. J.J.L. was supported by the Cystic Fibrosis Foundation. This work was supported in part by US National Institutes of Health grants (GM080177 to the Systems Biology Department, Harvard Medical School and GM081617 to R.K.), by a grant from the New England Regional Center of Excellence for Biodefense and Emerging Infectious Diseases (NERCE; AI057159 to R.K.) and by a Harvard Catalyst grant (to R.K., A.J.M. and M. Cendron).

AUTHOR CONTRIBUTIONS

J.-B.M., A.J.M. and R.K. conceived of the study. J.J.L., A.J.M. and G.P.P. collected the clinical samples. T.D.L. and N.L. performed resistance phenotyping. J.B.G., D.R., M.R.D., D.S. and G.P.P. performed LPS phenotyping and complementation. M.A., G.P.-B., A.J.M. and G.P.P. conducted chart review and provided medical information. T.D.L., J.-B.M. and R.K. performed whole-genome sequencing and data analysis. T.D.L., J.-B.M., J.J.L., A.J.M., G.P.P. and R.K. interpreted the results and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interest.

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

1. Suerbaum, S. & Josenhans, C. *Helicobacter pylori* evolution and phenotypic diversification in a changing host. *Nat. Rev. Microbiol.* **5**, 441–452 (2007).
2. Smith, E.E. *et al.* Genetic adaptation by *Pseudomonas aeruginosa* to the airways of cystic fibrosis patients. *Proc. Natl. Acad. Sci. USA* **103**, 8487–8492 (2006).
3. Musher, D.M. *et al.* Emergence of macrolide resistance during treatment of pneumococcal pneumonia. *N. Engl. J. Med.* **346**, 630–631 (2002).
4. Wong, A. & Kassen, R. Parallel evolution and local differentiation in quinolone resistance in *Pseudomonas aeruginosa*. *Microbiology* **157**, 937–944 (2011).
5. Zdziarski, J. *et al.* Host imprints on bacterial genomes—rapid divergent evolution in individual patients. *PLoS Pathog.* **6**, e1001078 (2010).
6. Yang, L. *et al.* Evolutionary dynamics of a bacteria in a human host environment. *Proc. Natl. Acad. Sci. USA* **108**, 7481–7486 (2011).
7. Kennemann, L. *et al. Helicobacter pylori* genome evolution during human infection. *Proc. Natl. Acad. Sci. USA* **108**, 5033–5038 (2011).
8. Mwangi, M.M. *et al.* Tracking the *in vivo* evolution of multidrug resistance in *Staphylococcus aureus* by whole-genome sequencing. *Proc. Natl. Acad. Sci. USA* **104**, 9451–9456 (2007).
9. Harris, S.R. *et al.* Evolution of MRSA during hospital transmission and intercontinental spread. *Science* **327**, 469–474 (2010).
10. Goodarzi, H., Hottes, A.K. & Tavazoie, S. Global discovery of adaptive mutations. *Nat. Methods* **6**, 581–583 (2009).
11. Pleasance, E.D. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**, 191–196 (2010).
12. Moxon, E.R., Rainey, P.B., Nowak, M.A. & Lenski, R.E. Adaptive evolution of highly mutable loci in pathogenic bacteria. *Curr. Biol.* **4**, 24–33 (1994).
13. van der Woude, M.W. & Bäumler, A.J. Phase and antigenic variation in bacteria. *Clin. Microbiol. Rev.* **17**, 581–611 (2004).
14. Croucher, N.J. *et al.* Rapid pneumococcal evolution in response to clinical interventions. *Science* **331**, 430–434 (2011).
15. Holt, K.E. *et al.* High-throughput sequencing provides insights into genome variation and evolution in *Salmonella* Typhi. *Nat. Genet.* **40**, 987–993 (2008).
16. Pallen, M.J. & Wren, B.W. Bacterial pathogenomics. *Nature* **449**, 835–842 (2007).
17. Elena, S.F. & Lenski, R.E. Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. *Nat. Rev. Genet.* **4**, 457–469 (2003).
18. Woods, R. *et al.* Tests of parallel molecular evolution in long-term experiment with *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* **103**, 9107–9112 (2006).
19. Barrick, J.E. *et al.* Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. *Nature* **461**, 1243–1247 (2009).
20. Lipuma, J.J. The changing microbial epidemiology in cystic fibrosis. *Clin. Microbiol. Rev.* **23**, 299–323 (2010).

21. Vermis, K. *et al.* Proposal to accommodate *Burkholderia cepacia* genomovar VI as *Burkholderia dolosa* sp. nov. *Int. J. Syst. Evol. Microbiol.* **54**, 689–691 (2004).
22. Lipuma, J.J. Update on the *Burkholderia cepacia* complex. *Curr. Opin. Pulm. Med.* **11**, 528–533 (2005).
23. LiPuma, J.J., Dasen, S.E., Nielson, D.W., Stern, R.C. & Stull, T.L. Person-to-person transmission of *Pseudomonas cepacia* between patients with cystic fibrosis. *Lancet* **336**, 1094–1096 (1990).
24. Biddick, R., Spilker, T., Martin, A. & LiPuma, J.J. Evidence of transmission of *Burkholderia cepacia, Burkholderia multivorans* and *Burkholderia dolosa* among persons with cystic fibrosis. *FEMS Microbiol. Lett.* **228**, 57–62 (2003).
25. Kalish, L.A. *et al.* Impact of *Burkholderia dolosa* on lung function and survival in cystic fibrosis. *Am. J. Respir. Crit. Care Med.* **173**, 421–425 (2006).
26. Morelli, G. *et al.* Microevolution of *Helicobacter pylori* during prolonged infection of single hosts and within families. *PLoS Genet.* **6**, e1001036 (2010).
27. Sibley, C.D. *et al.* A polymicrobial perspective of pulmonary infections exposes an enigmatic pathogen in cystic fibrosis patients. *Proc. Natl. Acad. Sci. USA* **105**, 15070–15075 (2008).
28. Guss, A.M. *et al.* Phylogenetic and metabolic diversity of bacteria associated with cystic fibrosis. *ISME J.* **5**, 20–29 (2011).
29. Mowat, E. *et al. Psuedomonas aeruginosa* population diversity and turnover in cystic fibrosis infections. *Am. J. Respir. Crit. Care Med.* **183**, 1674–1679 (2011).
30. Wilder, C.N., Allada, G. & Schuster, M. Instantaneous within-patient diversity of *Psuedomonas aeruginosa* quorum-sensing populations from cystic fibrosis lung infections. *Infect. Immun.* **77**, 5631–5639 (2009).
31. Weigel, L.M., Steward, C.D. & Tenover, F.C. *gyrA* mutations associated with fluoroquinolone resistance in eight species of *Enterobacteriaceae*. *Antimicrob. Agents Chemother.* **42**, 2661–2667 (1998).
32. Reyna, F., Huesca, M., Gonzalez, V. & Fuchs, L.Y. *Salmonella typhimurium gyrA* mutations associated with fluoroquinolone resistance. *Antimicrob. Agents Chemother.* **39**, 1621–1623 (1995).
33. Silhavy, T.J., Kahne, D. & Walker, S. The bacterial cell envelope. *Cold Spring Harb. Perspect. Biol.* **2**, a000414 (2010).
34. Vinion-Dubiel, A.D. & Goldberg, J.B. Lipopolysaccharide of *Burkholderia cepacia* complex. *J. Endotoxin Res.* **9**, 201–213 (2003).
35. Ortega, X. *et al.* Reconstitution of O-specific lipopolysaccharide expression in *Burkholderia cenocepacia* strain J2315, which is associated with transmissible infections in patients with cystic fibrosis. *J. Bacteriol.* **187**, 1324–1333 (2005).
36. Crosson, S., McGrath, P.T., Stephens, C., McAdams, H.H. & Shapiro, L. Conserved modular design of an oxygen sensory/signaling network with species-specific output. *Proc. Natl. Acad. Sci. USA* **102**, 8018–8023 (2005).
37. Worlitzsch, D. *et al.* Effects of reduced mucus oxygen concentration in airway *Pseudomonas* infections of cystic fibrosis patients. *J. Clin. Invest.* **109**, 317–325 (2002).
38. Marteyn, B. *et al.* Modulation of *Shigella* virulence in response to available oxygen *in vivo*. *Nature* **465**, 355–358 (2010).

ONLINE METHODS

Study cohort and bacterial isolates. An epidemic of *B. dolosa* affected 39 individuals in the Boston area over a 20-year period (see URLs)[25]. Our cohort includes patient zero, the seven subjects from whom bacterial isolates that were recovered from the bloodstream were available and six subjects chosen at random (**Supplementary Note**). Samples from these individuals were collected during normal care (**Supplementary Table 1**) and frozen. Frozen clinical stocks were streaked on solid medium. A single colony from each plate was chosen at random and frozen in 15% glycerol to create a working library. Time of isolation is reported relative to the collection of an isolate from patient zero (isolate A:0-0). The use of discarded samples for this study was approved by the institutional review boards at Children's Hospital Boston and Harvard Medical School.

Genome sequencing and SNP calling. DNA was extracted from single colonies using standard procedures, and multiplexed genomic libraries were constructed using the Illumina-compatible Nextera DNA Sample Prep kit. Sequencing was performed with 75-bp, single-end reads at a mean read depth of 37×. Reads were aligned using the *B. dolosa* genome AU0158 as a reference, which was isolated from patient zero. We used SAMtools 0.1.12a to manipulate consensus sequences and find SNPs in multiple isolates. Details can be found in the **Supplementary Note**, **Supplementary Figures 1** and **2** and **Supplementary Table 1**.

Rate of evolution. We estimated the number of SNPs accumulated between each isolate and the outgroup, normalizing the SNPs called with confidence to the portion of the genome covered with equally high confidence (**Supplementary Note**). These SNPs were found to accumulate at the constant rate of 2.1 SNPs/year (Pearson's $r = 0.79$), corresponding to a mutation fixation rate of 3×10^{-7} mutations per base pair per year. Within subjects, mutations accumulated at a similar rate (**Supplementary Fig. 3**).

Phylogeny. SNPs were used to construct a maximum-likelihood phylogenetic tree between the 112 isolates (**Supplementary Note**). Our model assumed independent evolution at each site and vertical inheritance. We used the software implementation Dnaml (PHYLIP v3.69 (see URLs). Different transition-to-transversion ratios produced remarkably similar phylogenies. We therefore chose a default model in which all mutations were equally likely. The LCA for strains from each subject was estimated as the most outward node from which all isolates from that subject were descended. Each LCA is an inferred genome that contains all polymorphisms shared among the isolates from that subject.

Ciprofloxacin resistance assays. Bacterial isolates were grown at 37 °C shaking for 24 h in microtiter plates containing LB with logarithmically decreasing concentrations of ciprofloxacin (Sigma-Aldrich; concentrations of 128, 64, 32, 16, 8, 4, 2, 1, 0.5 and 0 mg/l). We used 10 mM hydrochloric acid to solubilize the drug and create a stock solution at 640 mg/l. The MIC of an isolate was estimated as the lowest concentration of ciprofloxacin at which growth was <10% of the maximum growth of that isolate in the absence of drug, as determined by optical density. The reported value for each isolate is the logarithmic average of two replicate experiments performed on different days.

O-antigen repeat assays. Twenty isolates from our library (all taken from the airways, including five sets of isolates taken from the same subject; see **Supplementary Note**) were assayed for O-antigen presence. LPS was extracted from each sample as described previously[29]. Extracts were run on SDS-PAGE gels and visualized using Pro-Q Emerald staining. The presence of low-molecular-weight bands on the gel indicated O-antigen repeats. See **Supplementary Note**, **Supplementary Figure 6a** and **Supplementary Table 5** for details on O-antigen complementation and **Supplementary Figure 8** for details on mucoidy assays.

Genome-wide association study for assayed pathogenic phenotypes. For each phenotypic assay (MIC of ciprofloxacin and presence of O-antigen repeats) and for each gene, we used our 112-strain library to calculate the correlation between the value of the phenotype and the presence of SNPs (with respect to the LCA in the epidemic). Significance was assessed using Kendall's τ coefficient in the case of antibiotic resistance and with Fisher's exact test in the case of the presentation of O-antigen repeats.

Number of mutations. The presence of the same SNP in two distinct isolates can signify one of two scenarios: that the mutation occurred once in an ancestral isolate and was passed on to its descendants or that the mutation occurred twice. We resolved this ambiguity by using the phylogenetic tree derived above, inferring the genotypes of all internal nodes using parsimonious assumptions. This method showed 20 nucleotides that were mutated more than once, including 8 nucleotides that were mutated 3 or more times. We found 8 sites that were mutated to 2 different nucleotides and 1 site where all 4 nucleotides were observed. Overall, we counted 561 mutations. For each gene mutated multiple times, we report the number of mutations that each subject had in **Figure 4c**. For each nucleotide position within the gene, a mutation was counted if there was a polymorphism within that subject or if a mutation first appeared in that subject. In this way, we did not count inherited mutations that arose earlier in the epidemic.

Distribution of mutations expected by random drift. We randomly selected 561 positions in the *B. dolosa* reference genome and counted the distribution of mutations per gene obtained. We repeated this procedure 1,000 times and averaged the results.

Estimating the strength of selective pressures. All but 14 of the 304 genes with polymorphisms correspond to regions of the reference genome with no frameshift mutations. For these 290 genes, we assessed whether mutations changed the protein sequence (nonsynonymous, N) or had no such effect (synonymous, S). For any subset of the 290 genes, it is possible to count the observed N and S mutations and compare these numbers with those expected under random drift (expected N/S of 2.97, robust to transition-to-transversion ratio). The corresponding dN/dS indicated whether selection might act on the group of genes under study. Confidence was estimated with Clopper-Pearson binomial confidence intervals (95% confidence), and one-tailed *P* values were computed by simulation.

Manual annotation of genes. We manually annotated the 17 genes found to be under strong positive selection (**Supplementary Table 4**). Their translated coding sequences were compared to the RefSeq database using the BLASTP algorithm on the NCBI website. The corresponding gene name for well-annotated proteins (three- or four-letter gene names and a link to an NCBI gene) with high homology (*E* value $< 1 \times 10^{-20}$) was assigned to the gene in our study. If homology covered <95% of the protein sequence or if there was no such well-annotated protein, the gene name corresponding to the protein with the highest homology (lowest *E* value) was used. In an exception, the *Shigella flexneri* fnr gene (not in RefSeq) was included in the annotation list based on 89% homology coverage and the presence of two other oxygen-associated genes (*fixJ* and *fixL*) among the 17 genes under strong positive selection.