nature medicine

Genomic diversity in autopsy samples reveals within-host dissemination of HIV-associated *Mycobacterium tuberculosis*

Tami D Lieberman[1,2], Douglas Wilson[3], Reshma Misra[4], Lealia L Xiong[1], Prashini Moodley[4], Ted Cohen[5–7,10] & Roy Kishony[1,8–10]

Mycobacterium tuberculosis remains a leading cause of death worldwide, especially among individuals infected with HIV[1]. Whereas phylogenetic analysis has revealed _M. tuberculosis_ spread throughout history[2–5] and in local outbreaks[6–8], much less is understood about its dissemination within the body. Here we report genomic analysis of 2,693 samples collected post mortem from lung and extrapulmonary biopsies of 44 subjects in KwaZulu-Natal, South Africa, who received minimal antitubercular treatment and most of whom were HIV seropositive. We found that purifying selection occurred within individual patients, without the need for patient-to-patient transmission. Despite negative selection, mycobacteria diversified within individuals to form sublineages that co-existed for years. These sublineages, as well as distinct strains from mixed infections, were differentially distributed throughout the lung, suggesting temporary barriers to pathogen migration. As a consequence, samples taken from the upper airway often captured only a fraction of the population diversity, challenging current methods of outbreak tracing and resistance diagnostics. Phylogenetic analysis indicated that dissemination from the lungs to extrapulmonary sites was as frequent as between lung sites, supporting the idea of similar migration routes within and between organs, at least in subjects with HIV. Genomic diversity therefore provides a record of pathogen diversification and repeated dissemination across the body.

Mutations that pathogens accumulate over time can be used to reconstruct their transmission history. In theory, the same principles that are used to track spread across the globe and between individuals can inform our understanding of how *M. tuberculosis* spreads within the lungs and throughout the body[9,10]. Although *M. tuberculosis* primarily causes lung infections, it can disseminate to extrapulmonary organs, particularly in subjects who are co-infected with HIV. Understanding how *M. tuberculosis* diversifies and spreads within individuals is crucial to outbreak tracing[11], clinical diagnosis of resistance[12–14] and design strategies to minimize *de novo* emergence of antibiotic resistance[15–17].

Here we conducted a postmortem study in KwaZulu-Natal, South Africa, to capture the spatial diversity of *M. tuberculosis* within each person (**Fig. 1**). Inclusion criteria included tuberculosis as cause of death, suspected disseminated infection and fewer than 4 d of antitubercular treatment (Online Methods). Ninety-six of the 100 subjects enrolled were HIV positive, reflecting the high rate of HIV positivity among those dying from tuberculosis in the province. For each subject, a minimally invasive autopsy was performed, and specimens were collected from each of six lung sites (three regions within each lung; **Supplementary Table 1**), as well as from the liver and spleen. Each lung biopsy site was processed individually, whereas multiple biopsies from the other organs were pooled to form a single site per organ. Endotracheal aspirate, a proxy for what might have been detectable from a sputum specimen, was also collected. Specimens were cultured on a solid medium to select for *M. tuberculosis*. We focused on 44 subjects who had positive cultures for at least one lung site and at least one extrapulmonary site. Of these, all but two were HIV positive (**Supplementary Table 2**).

A total of 2,693 *M. tuberculosis* samples from 329 sites were analyzed for single-nucleotide polymorphisms (SNPs) at the whole-genome level (**Supplementary Table 3**). For the first 12 subjects, the colonies that grew on the solid medium were scraped into a single tube, resulting in one DNA sample per site. For the remaining subjects, up to 15 individual colonies from each site were processed as separate samples (in many cases, colonies were found to be formed by multiple cells). Genomic libraries were sequenced on the HiSeq platform to an average of 178× for scrape samples and 59× for colony samples. Reads were aligned to an approximate ancestral genome of the *M. tuberculosis* complex[3], allowing inference of derived and ancestral alleles. Covariation of mutation frequencies across samples from each subject distinguished pre-existing polymorphisms owing to mixed infection from *de novo* mutations (Online Methods; **Supplementary Fig. 1**).

[1]Department of Systems Biology, Harvard Medical School, Boston, Massachusetts, USA. [2]Center for Microbiome Informatics and Therapeutics, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [3]Department of Internal Medicine, Edendale Hospital, University of KwaZulu-Natal, Pietermaritzburg, South Africa. [4]Department of Infection Prevention and Control, Nelson R Mandela School of Medicine, University of KwaZulu-Natal, Durban, South Africa. [5]Department of Epidemiology of Microbial Diseases, Yale School of Public Health, New Haven, Connecticut, USA. [6]Division of Global Health Equity, Brigham and Women's Hospital, Boston, Massachusetts, USA. [7]Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts, USA. [8]Faculty of Biology, Technion–Israel Institute of Technology, Haifa, Israel. [9]Faculty of Computer Science, Technion–Israel Institute of Technology, Haifa, Israel. [10]These authors jointly directed this work. Correspondence should be addressed to T.C. (theodore.cohen@yale.edu) and R.K. (rkishony@technion.ac.il).

Received 28 June; accepted 14 September; published online 31 October 2016; doi:10.1038/nm.4205

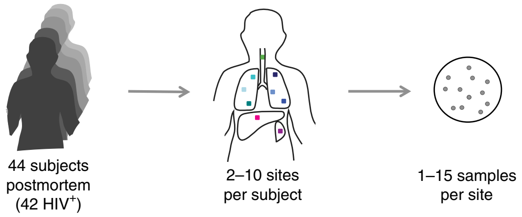

Figure 1 Postmortem analysis of 2,693 *M. tuberculosis* samples from 329 sites across the bodies of 44 subjects. Subjects who died in the emergency, admitting or inpatient wards of Edendale Hospital in Kwazulu-Natal, South Africa, were eligible to be enrolled in our study if they had been on antitubercular therapy for fewer than 4 d before death. For each subject, postmortem biopsies were taken from each of six locations within the lung, with each being treated as a separate site, and from multiple biopsies within the liver and spleen, which were pooled to form one site per organ. Aspirates of endotracheal secretions were taken, as were aspirates from clinically apparent ascitic fluid, pleural fluid or pus collections when detected. Lung sites 1–3 were taken from the right lung, and lung sites 4–6 were taken from the left lung (top to bottom; **Supplementary Table 1**). Each site was cultured separately for the presence of *M. tuberculosis*. DNA from one or more samples (up to 15 per site) was sequenced at the whole-genome level.

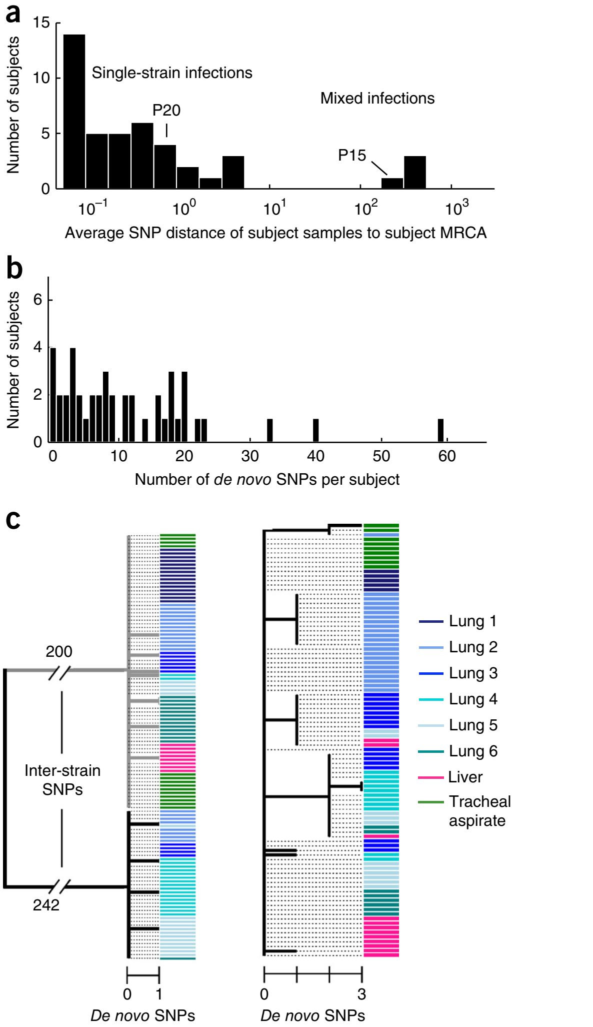

Figure 2 Genomic sequencing reveals variation due to mixed infection and *de novo* mutation. (**a**) A histogram of dMRCA estimates across subjects, the average number of single-nucleotide mutations per *M. tuberculosis* cell in that subject relative to each subject's population's most recent common ancestor (MRCA). Subjects P15 and P20 are indicated on the histogram. (**b**) A histogram of the number of *de novo* mutations within each subject's *M. tuberculosis* population, after removal of polymorphisms arising because of mixed infection (**Supplementary Fig. 1**). (**c**) Within-subject phylogenies created from inferred *M. tuberculosis* genotypes from samples obtained from subjects P15 (left) and P20 (right). Genotypes were called if they were found with at least 20% frequency within a sample (each sample can contain multiple genotypes). Colored bars indicate site.

We then used a constraint-based method to infer individual genotypes in nonclonal samples (**Supplementary Fig. 2**).

Four subjects had mixed infections, providing a prevalence estimate (9%) consistent with high-endemic areas[15,18,19]. These subjects were identified by the average mutational distance of their *M. tuberculosis* population to its most recent common ancestor (referred to as dMRCA; Online Methods; **Fig. 2a**). Given the *M. tuberculosis* molecular clock rate of ~0.3–0.5 SNPs/year[5,8,20], these subjects had dMRCA values consistent with those of mixed infection (201–439 SNPs/cell), whereas all others had values of fewer than 5 SNPs/cell. In two of these four cases, the minority strain was detected at low frequency across samples (5% and 11%), highlighting the importance of deep sampling for the detection of mixed infections.

Mutations appearing *de novo* showed evidence of purifying selection. The number of *de novo* mutations detected varied considerably across subjects (median 8.5 SNPs; range 0–59; **Fig. 2b**) and was somewhat dependent on the number of specimens (**Supplementary Fig. 3a**), but it was generally too low for selection to be analyzed on a subject-by-subject basis. Considering all of the *de novo* mutations together, the relative proportions of different nucleotide-to-nucleotide mutations closely matched those of mutations between subjects (**Supplementary Fig. 4**). *De novo* mutations had fewer amino acid–changing mutations than expected under a model of neutral evolution (normalized ratio of nonsynonymous to synonymous mutations (dN/dS) = 0.78; $P < 0.01$; Online Methods), similar to mutations observed between subjects (dN/dS = 0.80, $P < 0.001$). These data suggest that purifying selection, which has also been identified from between-patient comparisons[21,22], does not require patient-to-patient transmission but can occur within each subject.

De novo mutations led to sublineages that coexisted for many years in some patients. When focusing on single-strain infections, dMRCA estimates varied considerably across subjects (range 0–5 SNPs/cell, consistent with diversification for 0–16 years; **Fig. 2a**) and were not sensitive to the number of samples (**Supplementary Fig. 3c,d**). Some subjects had dMRCA values consistent with many years of within-patient diversification, yet most subjects had much lower dMRCA estimates (median 0.24 SNPs/cell, consistent with diversification for <1 year), consistent with the fact that immunocompromised subjects are likely to have fulminant courses of disease. Subjects who were HIV seronegative or diagnosed with HIV infection only post mortem had higher dMRCA values on average (six subjects; median 1.0 SNP/cell; consistent with diversification for 2.2–3.4 years; $P < 0.02$ by Wilcoxon rank–sum test), consistent with a longer duration of infection in patients who are not immunocompromised[23,24]. For all of the subjects with *de novo* mutations, even when the dMRCA value was small, diversification led to coexisting short-branched sublineages (**Fig. 2c**).

The diversity arising from *de novo* mutation and mixed infections was heterogeneously distributed across lung sites. In subjects with mixed infections, strain distribution varied considerably across sites,

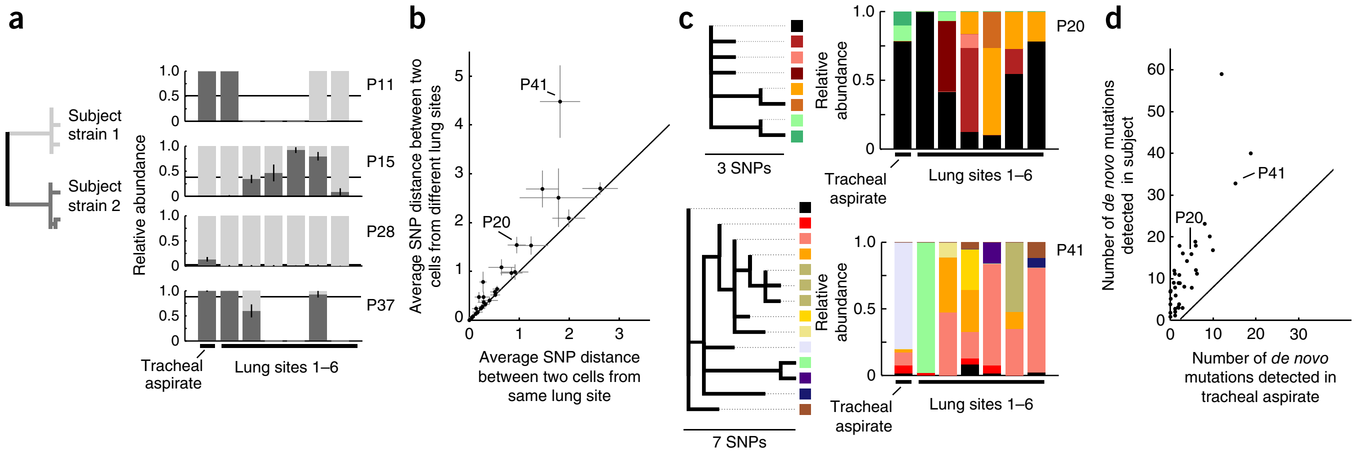

Figure 3 *M. tuberculosis* diversity within the lungs is spatially structured. For each subject, we compared the distribution of strains, mutations and genotypes across the lungs and tracheal aspirate, averaging across samples from each site. (**a**) The relative abundance of both strains in each respiratory site (tracheal aspirate and lung sites 1–6, left to right) is indicated for subjects with mixed infection. Different shades of gray represent the two strains. Error bars indicate the s.e.m. for subjects with multiple samples per site. The horizontal black lines indicate the average across the sites from each subject. (**b**) For each subject with a single-strain infection and multiple samples per site, the average distance between two cells from different lung sites is plotted against the average distance between two cells from the same lung site (averaging across sites). Error bars indicate the s.e.m. across sites or pairs of sites. (**c**) For two representative subjects, the relative abundance of each genotype within each respiratory site is indicated (tracheal aspirate and lung sites 1–6, left to right) (right), and the phylogenetic relationships of these genotypes are shown on the left. Color represents genotype. (**d**) For each subject, the number of *de novo* mutations detected within the subject is plotted against the number of *de novo* mutations detected within only the tracheal aspirate. Low-abundance mutations were the most likely to be missing from the tracheal samples (**Supplementary Fig. 5a**,**b**). A small offset was added to improve visualization (up to 0.5 mutations; same value added in *x* and *y*).

often with only one strain detected per lung site (**Fig. 3a**). Considering only *de novo* mutations, cells from the same site were significantly more similar to one another than to cells from different sites ($P < 0.05$ by one-sided two-sample *t*-test for 17/30 subjects with multiple sites and at least one *de novo* mutation; **Fig. 3b**). Inspection of intrapersonal phylogenies further revealed that minority genotypes can be localized to a single site and that some sites were dominated by minority sublineages not seen elsewhere within the lungs (for example, **Fig. 3c**). This spatial heterogeneity was neither universal nor long-lived; many new mutations were shared across lung sites, and no long branches were unique to a single lung or site. These results are consistent with reports of spatial heterogeneity in other lung infections[9] and in *M. tuberculosis* infections of HIV-negative individuals[25–27]. Whatever the mechanism, this heterogeneity meant that many *de novo* mutations and strains infecting subjects' lungs were undetected in tracheal aspirate samples (**Fig. 3d** and **Supplementary Fig. 5**). Although sputum samples were not available, this suggests that spatial heterogeneity can compromise the reliability of antibiotic-resistance profiling and other assays performed even on the entire diversity within respiratory samples[28].

Unexpectedly, the genetic distance between cells from different sites, both within the lungs and between organs, did not depend on spatial proximity. Cells from sites in different lungs (right versus left) were, on average, no more different than cells taken from sites in the same lung (**Fig. 4a**). Furthermore, two cells from different organs were no more different than two cells taken from different lung sites (**Fig. 4b** and **Supplementary Fig. 6**). Moreover, genotypes from the liver or spleen sites did not cluster phylogenetically, suggesting dissemination of multiple genotypes to each of these organs (**Fig. 4c**). To compare rates of transmission between organs and within the lungs, we estimated the number of transmitted genotypes between the lungs and extrapulmonary organs inferred by their joint phylogeny (**Fig. 4c**; Online Methods). We inferred multiple transmitted genotypes between the lungs and extrapulmonary organs for many subjects (**Fig. 4c**,**d**). These multiple genotypes might result from a single transmission event involving multiple genotypes or from repeated transmission events[29]; we note that the detection of different genotypes in the liver and spleen from the same individual suggests multiple independent transmission events. Notably, the number of genotypes transmitted from the lungs to extrapulmonary sites was as high as the number of transmissions between different lung sites (**Fig. 4d**).

The similar patterns of migration within and between organs suggest that dissemination of *M. tuberculosis* across the lungs of these subjects is no easier than dissemination across organs. These results support the hypothesis that both intra-lung and inter-organ spread are mediated by similar dissemination mechanisms, such as by macrophage trafficking through the bloodstream or by other mechanisms of hematogenous spread[30–32]. It would be interesting to investigate the generality of these results in subjects with less advanced disease or without HIV.

The diversity of genotypes within individuals and their stratification across the body have implications for outbreak tracing. A major impetus for whole-genome sequencing of *M. tuberculosis,* as well as other pathogens, has been to identify transmission links between individuals. Although future approaches will incorporate the pathogen diversity within each person[33–35], current approaches establish an epidemiological linkage between people if the mutational distance between single samples from each patient is below a set threshold[7,8,22,36]. Considering all of the genotypes of *M. tuberculosis* within a subject, 20% of subjects harbored genotypes separated by more than a common threshold of 5 SNPs[8] (**Supplementary Table 2**; 9% emerging from multiple-strain infection), and 11% harbored genotypes separated by more than an alternative threshold of 12 SNPs[37]. These estimates may be even greater when considering longer-duration infections, such as those of HIV-negative individuals. Although the thresholds chosen will depend on the mutation-detection sensitivity unique to each study, the frequency of substantial within-patient distances adds weight to the mounting evidence[11,33,38,39]

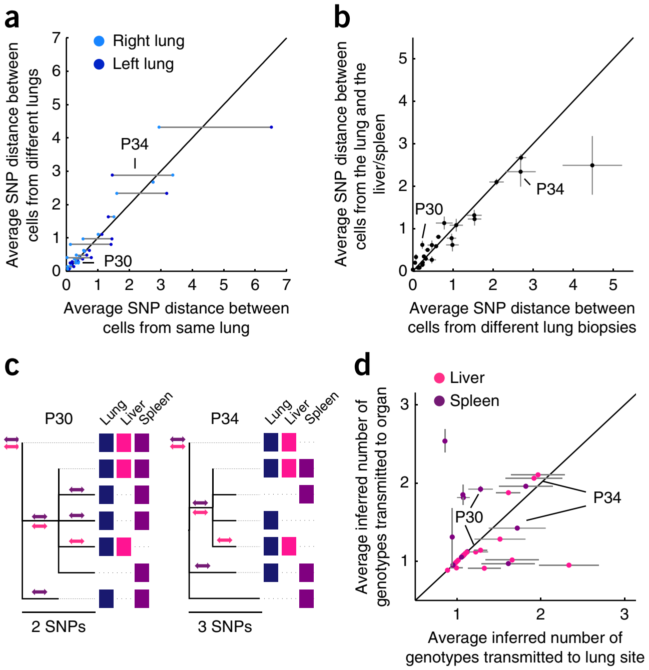

Figure 4 *M. tuberculosis* dissemination within lungs, between lungs and between organs follows similar dynamics. (**a**,**b**) For each subject with a single-strain infection, the average distance between cells from the right lung and left lung is plotted against the average distance between cells from the same lung (right lung, dark blue; left lung, light blue) (**a**) and the average distance between a cell from a lung site and a cell from an extrapulmonary organ is plotted against the average distance between cells from the different lung sites (**b**). Error bars indicate s.e.m. across sites or pairs of sites. (**c**) Examples of intrasubject phylogenies showing shared genotypes. Genotypes are indicated as 'within an organ' with a colored box in the corresponding column if they were found at >20% frequency in at least two samples from that organ. Potential transmission events between the lungs and extrapulmonary organs are identified by mutations shared across organs and indicated with a double-sided arrow of the corresponding color. (**d**) For each subject, the minimum number of transmitted genotypes to the liver and/or spleen from the lungs was inferred (*y* axis) and compared to the mean of the minimum number of transmitted genotypes to each lung site from the rest of the lungs (averaging across sites; *x* axis). Simulations were used to normalize for sampling efforts (Online Methods). Error bars indicate the s.e.m. across lung sites. A small offset was added for visibility (up to 10%; same value in *x* and *y*).

against ruling out epidemiological links on the basis of mutational thresholds between single samples.

This study represents the first large-scale genomic investigation of *M. tuberculosis* diversity across the human body. Extensive sampling from multiple body sites and detection of minor alleles within each sample shows that even the notoriously slowly evolving *M. tuberculosis* presents substantial within-patient diversity. The distribution of this diversity across the body suggests that, at least in people with HIV, dissemination within the lungs follows a similar mechanism as dissemination from the lungs to extrapulmonary organs. In total, these findings highlight the potential of whole-genome sequencing to make inferences about the history of individual bacterial infections from the diversity surveyed at a single time point. We anticipate that future studies will extend these approaches to address open questions about *M. tuberculosis* and other pathogens, including the contribution of spatial heterogeneity to the evolution of antibiotic resistance and to the dynamics of dormancy and reactivation.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Accession codes. NCBI Sequence Read Archive (SRA): sequencing reads have been deposited under BioProject accession code PRJNA323744.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS
We thank L. Maziya, Z. Magcaba, N. Xengxe and M.J. Khumalo for their assistance with enrollment and postmortem analyses, the Edendale Hospital management and the KwaZulu-Natal Department of Health. We are grateful to M. Baym and members of the Kishony lab for helpful discussions, C. Duvallet for comments on the manuscript, and H. Chung for discussions of spatial diversity and comments on the manuscript. We thank the team at Macrogen Clinical Laboratories for their help with Illumina sequencing and I. Comas (University of Valencia) for providing the *M. tuberculosis* reference sequence. This work was funded in part by US National Institutes of Health grants DP2 OD006663 (T.C.) and R01-GM081617 (R.K.), F. Hoffman LaRoche, Inc. (R.K.), and European Research Council FP7 ERC grant 281891 (R.K.).

AUTHOR CONTRIBUTIONS

T.C. and D.W. conceived and coordinated the study; T.D.L., D.W., T.C. and R.K. designed the study; D.W. identified eligible decedents and conducted postmortem biopsies; R.M. and P.M. processed and cultured samples, performed antibiotic sensitivity testing and extracted DNA; T.D.L. and L.L.X. prepared genomic libraries; T.D.L. and R.K. analyzed genomic data; T.D.L., D.W., T.C. and R.K. interpreted results and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

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

1. World Health Organization. Global tuberculosis report 2015 1–204 (World Health Organization, Geneva, Switzerland, 2015).
2. Cohen, K.A. *et al.* Evolution of extensively drug-resistant tuberculosis over four decades: whole-genome sequencing and dating analysis of *Mycobacterium tuberculosis* isolates from KwaZulu-Natal. *PLoS Med.* **12**, e1001880 (2015).
3. Comas, I. *et al.* Out-of-Africa migration and neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. *Nat. Genet.* **45**, 1176–1182 (2013).
4. Kay, G.L. *et al.* Eighteenth-century genomes show that mixed infections were common at time of peak tuberculosis in Europe. *Nat. Commun.* **6**, 6717 (2015).
5. Bos, K.I. *et al.* Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. *Nature* **514**, 494–497 (2014).
6. Gardy, J.L. *et al.* Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *N. Engl. J. Med.* **364**, 730–739 (2011).
7. Walker, T.M. *et al.* Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. *Lancet Infect. Dis.* **13**, 137–146 (2013).
8. Bryant, J.M. *et al.* Inferring patient-to-patient transmission of *Mycobacterium tuberculosis* from whole-genome sequencing data. *BMC Infect. Dis.* **13**, 110 (2013).
9. Jorth, P. *et al.* Regional isolation drives bacterial diversification within cystic fibrosis lungs. *Cell Host Microbe* **18**, 307–319 (2015).
10. Lieberman, T.D. *et al.* Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes. *Nat. Genet.* **43**, 1275–1280 (2011).
11. Pérez-Lago, L. *et al.* Whole-genome sequencing analysis of intrapatient microevolution in *Mycobacterium tuberculosis*: potential impact on the inference of tuberculosis transmission. *J. Infect. Dis.* **209**, 98–108 (2014).
12. Zetola, N.M. *et al.* Clinical outcomes among persons with pulmonary tuberculosis caused by *Mycobacterium tuberculosis* isolates with phenotypic heterogeneity in results of drug-susceptibility tests. *J. Infect. Dis.* **209**, 1754–1763 (2014).
13. Black, P.A. *et al.* Whole-genome sequencing reveals genomic heterogeneity and antibiotic purification in *Mycobacterium tuberculosis* isolates. *BMC Genomics* **16**, 857 (2015).
14. Sun, G. *et al.* Dynamic population changes in *Mycobacterium tuberculosis* during acquisition and fixation of drug resistance in patients. *J. Infect. Dis.* **206**, 1724–1733 (2012).
15. Cohen, T. *et al.* Mixed-strain *Mycobacterium tuberculosis* infections and the implications for tuberculosis treatment and control. *Clin. Microbiol. Rev.* **25**, 708–719 (2012).

16. Colijn, C., Cohen, T., Ganesh, A. & Murray, M. Spontaneous emergence of multiple drug resistance in tuberculosis before and during therapy. *PLoS One* **6**, e18327 (2011).
17. Prideaux, B. *et al.* The association between sterilizing activity and drug distribution into tuberculosis lesions. *Nat. Med.* **21**, 1223–1227 (2015).
18. Mankiewicz, E. & Liivak, M. Phage types of *Mycobacterium tuberculosis* in cultures isolated from Eskimo patients. *Am. Rev. Respir. Dis.* **111**, 307–312 (1975).
19. Warren, R.M. *et al.* Patients with active tuberculosis often have different strains in the same sputum specimen. *Am. J. Respir. Crit. Care Med.* **169**, 610–614 (2004).
20. Ford, C.B. *et al. Mycobacterium tuberculosis* mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. *Nat. Genet.* **45**, 784–790 (2013).
21. Pepperell, C.S. *et al.* The role of selection in shaping diversity of natural *M. tuberculosis* populations. *PLoS Pathog.* **9**, e1003543 (2013).
22. Lee, R.S. *et al.* Population genomics of *Mycobacterium tuberculosis* in the Inuit. *Proc. Natl. Acad. Sci. USA* **112**, 13609–13614 (2015).
23. Tiemersma, E.W., van der Werf, M.J., Borgdorff, M.W., Williams, B.G. & Nagelkerke, N.J.D. Natural history of tuberculosis: duration and fatality of untreated pulmonary tuberculosis in HIV-negative patients: a systematic review. *PLoS One* **6**, e17601 (2011).
24. Eldholm, V. *et al.* Impact of HIV co-infection on the evolution and transmission of multidrug-resistant tuberculosis. *eLife* **5**, 306 (2016).
25. Lin, P.L. *et al.* Sterilization of granulomas is common in active and latent tuberculosis despite within-host variability in bacterial killing. *Nat. Med.* **20**, 75–79 (2014).
26. García de Viedma, D., Marín, M., Ruiz Serrano, M.J., Alcalá, L. & Bouza, E. Polyclonal and compartmentalized infection by *Mycobacterium tuberculosis* in patients with both respiratory and extra-respiratory involvement. *J. Infect. Dis.* **187**, 695–699 (2003).
27. Liu, Q. *et al.* Within-patient microevolution of *Mycobacterium tuberculosis* correlates with heterogeneous responses to treatment. *Sci. Rep.* **5**, 17507 (2015).
28. Ford, C. *et al. Mycobacterium tuberculosis*—heterogeneity revealed through whole-genome sequencing. *Tuberculosis (Edinb.)* **92**, 194–201 (2012).
29. Turajlic, S. & Swanton, C. Metastasis as an evolutionary process. *Science* **352**, 169–175 (2016).
30. Krishnan, N., Robertson, B.D. & Thwaites, G. The mechanisms and consequences of the extra-pulmonary dissemination of *Mycobacterium tuberculosis. Tuberculosis (Edinb.)* **90**, 361–366 (2010).
31. McMurray, D.N. Hematogenous reseeding of the lung in low-dose, aerosol-infected guinea pigs: unique features of the host–pathogen interface in secondary tubercles. *Tuberculosis (Edinb.)* **83**, 131–134 (2003).
32. Ssengooba, W., de Jong, B.C., Joloba, M.L., Cobelens, F.G. & Meehan, C.J. Whole-genome sequencing reveals mycobacterial microevolution among concurrent isolates from sputum and blood in HIV-infected TB patients. *BMC Infect. Dis.* **16**, 371 (2016).
33. Worby, C.J., Lipsitch, M. & Hanage, W.P. Within-host bacterial diversity hinders accurate reconstruction of transmission networks from genomic-distance data. *PLoS Comput. Biol.* **10**, e1003549 (2014).
34. Didelot, X., Gardy, J. & Colijn, C. Bayesian inference of infectious disease transmission from whole-genome sequence data. *Mol. Biol. Evol.* **31**, 1869–1879 (2014).
35. Paterson, G.K. *et al.* Capturing the cloud of diversity reveals complexity and heterogeneity of MRSA carriage, infection and transmission. *Nat. Commun.* **6**, 6560 (2015).
36. Guerra-Assunção, J.A. *et al.* Large-scale whole-genome sequencing of *M. tuberculosis* provides insights into transmission in a high-prevalence area. *eLife* **4**, e05166 (2015).
37. Walker, T.M. *et al.* Assessment of *Mycobacterium tuberculosis* transmission in Oxfordshire, UK, 2007-12, with whole-pathogen genome sequences: an observational study. *Lancet Respir. Med.* **2**, 285–292 (2014).
38. Hatherell, H.-A. *et al.* Interpreting whole-genome sequencing for investigating tuberculosis transmission: a systematic review. *BMC Med.* **14**, 21 (2016).
39. Eldholm, V. *et al.* Evolution of extensively drug-resistant *Mycobacterium tuberculosis* from a susceptible ancestor in a single patient. *Genome Biol.* **15**, 490 (2014).

ONLINE METHODS

Study cohort. Subjects were eligible for inclusion in our study if they died while at Edendale Hospital, were age 20 or over, were on antitubercular treatment for fewer than 4 d before death, and were either smear positive for acid-fast bacilli or had at least one focal process compatible with tuberculosis without an alternative diagnosis. The number of subjects studied was determined on the basis of feasibility. Subjects were de-identified before DNA analysis, and mutation calling was performed by researchers who were blinded to subject information.

Sample collection. Following informed consent by family members under a protocol approved by the institutional review board at University of KwaZulu-Natal, a minimally invasive autopsy was performed for each subject. Briefly, samples from three lung sites were collected from each side—at the second, fourth and sixth intercostal spaces. A new sterile needle was used at each site (see **Supplementary Note** for more detail). Specimens were stored on ice and transported to the University of KwaZulu-Natal in Durban. Participants who were not documented in the case notes to be HIV seropositive were tested post mortem using a lateral-flow ELISA assay. Subject information is listed in **Supplementary Table 2**. Cause of death for all subjects was culture-confirmed disseminated tuberculosis.

Biopsy specimens were cut into small pieces and decontaminated using Petroff's method[40] (to enrich for *M. tuberculosis*). Endotracheal aspirate specimens were decontaminated with a 4% *N*-acetyl-L-cysteine (NALC) in NaOH solution. 500 µl of the decontaminated specimen was inoculated into a mycobacteria growth indicator tube (MGIT), and 100 µl was spread on three separate 7H11 plates[41] to select for *M. tuberculosis*. For the first 12 subjects, the colonies (which represented the microbial diversity of the sample) on these three plates were scraped into a single tube for DNA analysis. For the remaining cultures, up to five single colonies, demonstrating morphological variation, were selected from each plate (for a maximum of 15 colonies per specimen). Each colony was then subcultured onto dedicated 7H11 plates and incubated for 3 weeks at 37 °C under aerobic conditions.

Antibiotic resistance testing. Phenotypic antibiotic susceptibility testing was performed on all decontaminated specimens. The modified proportion method was used for antibiotic susceptibility testing. Positive MGIT tubes from each specimen were used to set up the antibiotic susceptibility tests. Tubes were confirmed to be uncontaminated and were incubated for an additional 24–48 h after becoming instrument positive before the test was performed. 7H11 cultures that had been incubated for 3 weeks were used in the event the MGIT cultures were found to be contaminated or to confirm resistance.

Quadrant 7H10 agar plates that comprised a control (antibiotic-free) quadrant and antibiotic-containing quadrants were used. The following concentrations were used: 1 µg/ml isoniazid, 1 µg/ml rifampicin, 7.5 µg/ml ethambutol, 2 µg/ml streptomycin, 2 µg/ml ofloxacin, 5 µg/ml kanamycin, 5 µg/ml ethionamide and 10 µg/ml capreomycin. Cultures were diluted so as to obtain approximately 100 colony-forming units (CFU) on each quadrant and seeded onto each quadrant. The plates were sealed in CO_2-permeable bags and incubated at 37 °C and 5% CO_2 for 1 week, and then under aerobic conditions for an additional 2 weeks. Cultures with control *M. tuberculosis* strains (H37Rv and A169) were set up in parallel. Antibiotic-containing quadrants with CFU > 1% of the CFU on the control quadrant were interpreted as being antibiotic resistant. Either no growth or CFU < 1% of that of the control quadrant was interpreted as being antibiotic susceptible. Cultures were read in a double-blind manner.

No significant variation was found across specimens from a given subject. Specimens from subjects P4, P9 and P12 were multidrug resistant; specimens from P16 were rifampicin resistant, and all others were susceptible to all drugs tested (**Supplementary Table 2**).

DNA extraction. DNA was extracted using the *N*-acetyl-*N*,*N*,*N*-trimethyl ammonium bromide (CTAB) method[42]. *M. tuberculosis* colonies were scraped from agar, resuspended into 500 µl of water in a microcentrifuge tube and heat-killed for 30 min at 80 °C. 70 µl 10% sodium dodecyl sulfate and 50 µl proteinase K (10 mg/ml stock solution) were added, and the solution was incubated for 1 h at 60 °C using a thermomixer on 'low-mode' shaking. Samples were removed from the thermomixer, and 100 µl 5 M NaCl (preheated to 60 °C) was added and mixed thoroughly by hand inversion. 100 µl 10% CTAB (preheated to 60 °C) was added, and the solution was mixed by hand inversion. Tubes were then incubated at 60 °C for 15 min (thermomixer on low-mode shaking). 700 µl chloroform:isoamyl alcohol (24:1) was added, and the solution was mixed by hand inversion. Tubes were then centrifuged for 10 min at 13,000 r.p.m. The upper (aqueous) phase was transferred to a tube with 700 µl of cold isopropanol and mixed by hand inversion. DNA was precipitated by freezing at −20 °C for at least 30 min and centrifugation for 10 min at 13,000 r.p.m. Tubes were drained, and the pellet was washed with 80% ethanol followed by centrifugation for 5–10 min. These were then drained, and the pellet was allowed to dry. DNA was suspended in 55 µl of 10 mM Tris with 1 mM EDTA (TE buffer).

Illumina sequencing and mutation identification. Genomic libraries were constructed and barcoded using a previously described modified version of the Illumina Nextera protocol[43]. Genomic libraries were sequenced on the Illumina HiSeq 2000 platform by Macrogen using paired-end 100-bp reads. Reads were aligned to a modified H37rv reference genome in which each nucleotide reflects this ancestor of the *M. tuberculosis* complex[3] (provided by Iñaki Comas). The limited mutation distance of all extant *M. tuberculosis* to this ancestor (<0.05% of the genome) and limited recombination in *M. tuberculosis*[3] gives us confidence that each difference from this reference that passed our conservative filters (**Supplementary Note**) represents a newly derived mutation. Standard approaches were used for read-filtering and alignment[44]. Average coverage for each sample is listed **Supplementary Table 4**.

Potential cross-contamination or mislabeling events were identified by the presence of multiple mutations that were not found in any other sample from that subject and/or the multiple mutations found in samples from different subjects processed on the same plate. We identified 68 such potentially contaminated samples. New, repeated genomic libraries were prepared for 40 of these 68 samples. In 31 cases, the sequencing results from the duplicate matched more closely with those of other samples from that subject, and the duplicate was used for analysis. In nine cases, the duplicate also showed evidence of contamination, and the two replicates of the sample were discarded (for a total of 37 discarded samples out of 2,730; 1.4%). In our final data, subjects had distinct strains (**Supplementary Fig. 7**). Because some of these samples may have been discarded in error, some true mixed infections may have been called as simple infections.

Many single-colony samples from subjects with mixed infections had hundreds of mutations at intermediate frequencies—suggesting that cells from different strains were comingled in a single colony (**Supplementary Fig. 1c**; consistent with the cording phenotype of *M. tuberculosis*[45]). We therefore searched for both fixed and polymorphic mutations within each sample, using a series of filters and statistical tests to distinguish false positives (for example, alignment and sequencing errors) from real polymorphisms[44]. Additional filters regarding the proportion of samples in each subject with a polymorphism were also used to remove notoriously problematic sites, such as genes encoding members of the proline–glutamate and proline–proline–glutamate (PE and PPE) families[46] (see **Supplementary Note** for details). Among these filters, polymorphic mutations were required to be supported by at least 10% of the reads and supported by at least four reads on each strand. As a consequence, rare variants may not have been detected, particularly in cases in which few samples were available per site.

For each polymorphic position that met the filtering criteria in at least one sample, raw reads were used to determine the allele frequency across samples from that subject. The 518 *de novo* mutations found, and their distribution across samples, are listed in **Supplementary Table 4**. Mutations and genes were annotated according to the H37rv sequence (NC_000962.3). None of these mutations are known to be associated with antibiotic resistance[47].

dN/dS. Calculations for dN/dS, the ratio of nonsynonymous mutations to synonymous mutations divided by the ratio under a neutrality, were performed normalizing for the spectrum of mutations observed as previously described[44].

P values for depletion of nonsynonymous mutations were calculated according to the binomial cumulative distribution function.

Identification of polymorphisms arising from multiple-strain infection. Mutations carried by each of the infecting strains co-vary across samples with one another and inversely co-vary with mutations from the other strain[48]. Mutations arising from multiple-strain infection are defined by their contribution to the primary principal component (PC1) in a principal component analysis (PCA) of the mutation-frequency matrix. *De novo* mutations are identified as those without a significant contribution to PC1 (absolute value < 0.015, cutoff determined empirically; **Supplementary Fig. 1**). Previously described SNP markers were used to assign each strain to a global lineage[49].

Estimation of dMRCA. The average mutational distance across cells in a subject's *M. tuberculosis* population to its most recent common ancestor, referred to as dMRCA, was calculated as the sum of the mutation frequencies at each polymorphic position called within the subject[44]. To normalize for sampling efforts, the dMRCA value was first calculated across samples from each specimen, and these values were averaged for each subject. For subjects who were previously diagnosed with HIV, the mean value of dMRCA was 0.42 SNPs/cell, and for subjects who were HIV negative or diagnosed post mortem, this mean was 1.8 SNPs/cell (**Fig. 2a** and **Supplementary Table 2**).

Interpretation of dMRCA as 'time of infection' assumes that these subjects were infected with a single genotype; if subjects were infected by multiple closely related genotypes, then these estimates are inflated. Alternatively, these estimates may be underestimations if mutations have swept the subject's diversity since infection due to adaptation or bottlenecks (drift). Estimates of dMRCA for subjects with multiple-strain infection are minimal estimates because the strict mutation caller was optimized for detecting *de novo* mutations. Other potential sources of error include: Poisson error in the number of mutations accumulated in each lineage since each subject's MRCA, underestimation due to limited sensitivity in detecting mutations, overestimation due to false positives, incorrect designation of ancestral versus derived alleles, and underestimation due to incomplete sampling of the diversity within each subject.

Genotype identification and phylogenetic inference. We used a conservative algorithm to infer a minimal set of distinct genotypes within each subject based on the assumptions that genotypes are shared across samples from each subject and that each SNP occurred only once within a subject (strict parsimony). A direct consequence of the strict parsimony assumption is that two mutations at high frequency (majority) in the same sample must coexist on the same genotype[50,51]. For each sample within a subject, starting with the most homogenous and covered samples, high-frequency polymorphisms were matched with previously identified genotypes, or if no such match was made, then the polymorphisms found at high frequency were used to define a new genotype (**Supplementary Fig. 2**). Some *de novo* mutations could not be confidently assigned to a genotype and were omitted; combined with our stringent mutation caller, this approach produced conservative phylogenies in which some genotypes were not identified or called with fewer mutations than they really have. After genotypes were identified, each sample was then assigned to one or more genotypes. We note that our algorithm will not work well in cases where parallel nucleotide evolution is expected or where pure samples are not available, and that the cutoffs in our implementation may need to be adjusted for different use cases (**Supplementary Note**).

Pairwise genetic distances. We calculated the average pairwise distance between cells in different samples, considering the frequency of each mutation f at each confident polymorphic position x on the genome, in samples i and j as:

$$\sum_x f_{xi} + f_{xj} - (2 \times f_{xi} \times f_{xj})$$

Only lung, liver and spleen sites with more than two samples per site were included in the analysis. To normalize for different sampling efforts across sites, we first calculate the mean at each site (x axis; **Fig. 3b**) or pairs of sites and then average across the site or pairs of sites. Error bars indicate the s.e.m. of the means across the sites or pairs of sites.

Transmission analysis. We developed an algorithm based on the number of shared mutations between sites (**Fig. 4c**). Only genotypes found in two separate samples within an organ or lung site were considered to reduce the effects of potential cross-contamination. We iterated through mutations shared between the destination site and the rest of the lung. After each iteration, the genotype closest to the subject's MRCA containing a shared mutation was chosen and counted as a transmission event. All mutations in that genotype were removed from the list of unaccounted-for shared mutations, and the next iteration was performed. In this way, groups of mutations appearing on the same branch of the phylogeny were counted only as a single transmitted genotype. Finally, the presence of the subject's MRCA in both locations suggests a transmission of this ancestral genotype (**Fig. 4c**). Simulations were performed to normalize for different sampling efforts across sites and organs, omitting one site from the lung each trial (for estimations of intra-organ transfer, this omitted site was then treated as destination). Only sites with more than six samples were included, and each site was randomly subsampled, taking the minimum number of samples across sites with >6 samples. For each omitted lung site, 1,000 trials were performed and averaged.

Code availability. Custom MATLAB code used for calling mutations and analyzing the data can be found at https://github.com/kishonylab/TB-diversity-across-organs.

40. Petroff, S.A. A new and rapid method for the isolation and cultivation of tubercle bacilli directly from the sputum and feces. *J. Exp. Med.* **21**, 38–42 (1915).
41. Cohn, M.L., Waggoner, R.F. & McClatchy, J.K. The 7H11 medium for the cultivation of mycobacteria. *Am. Rev. Respir. Dis.* **98**, 295–296 (1968).
42. Somerville, W., Thibert, L., Schwartzman, K. & Behr, M.A. Extraction of *Mycobacterium tuberculosis* DNA: a question of containment. *J. Clin. Microbiol.* **43**, 2996–2997 (2005).
43. Baym, M. *et al.* Inexpensive multiplexed library preparation for megabase-sized genomes. *PLoS One* **10**, e0128036 (2015).
44. Lieberman, T.D. *et al.* Genetic variation of a bacterial pathogen within individuals with cystic fibrosis provides a record of selective pressures. *Nat. Genet.* **46**, 82–87 (2014).
45. Middlebrook, G., Dubos, R.J. & Pierce, C. Virulence and morphological characteristics of mammalian tubercle bacilli. *J. Exp. Med.* **86**, 175–184 (1947).
46. Casali, N. *et al.* Microevolution of extensively drug-resistant tuberculosis in Russia. *Genome Res.* **22**, 735–745 (2012).
47. Coll, F. *et al.* Rapid determination of antituberculosis drug resistance from whole-genome sequences. *Genome Med.* **7**, 51 (2015).
48. Cleary, B. *et al.* Detection of low-abundance bacterial strains in metagenomic data sets by eigengenome partitioning. *Nat. Biotechnol.* **33**, 1053–1060 (2015).
49. Coll, F. *et al.* A robust SNP barcode for typing *Mycobacterium tuberculosis*–complex strains. *Nat. Commun.* **5**, 4812 (2014).
50. Jiao, W., Vembu, S., Deshwar, A.G., Stein, L. & Morris, Q. Inferring clonal evolution of tumors from single-nucleotide somatic mutations. *BMC Bioinformatics* **15**, 35 (2014).
51. Fischer, A., Vázquez-García, I., Illingworth, C.J.R. & Mustonen, V. High-definition reconstruction of clonal composition in cancer. *Cell Rep.* **7**, 1740–1752 (2014).