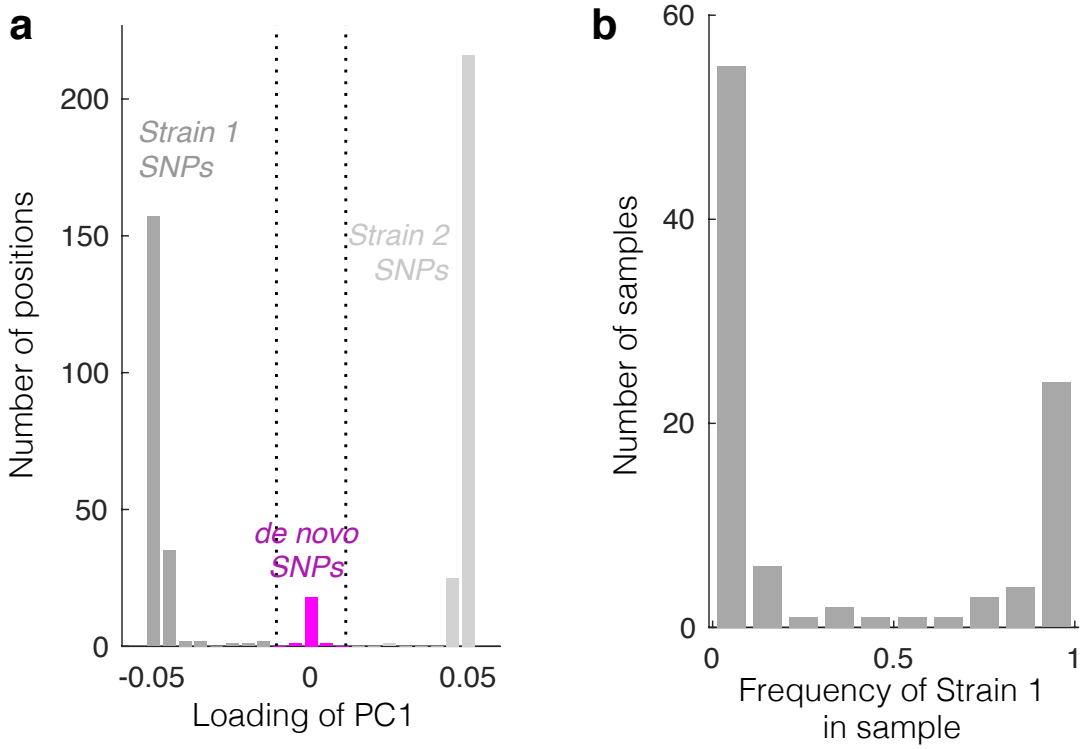
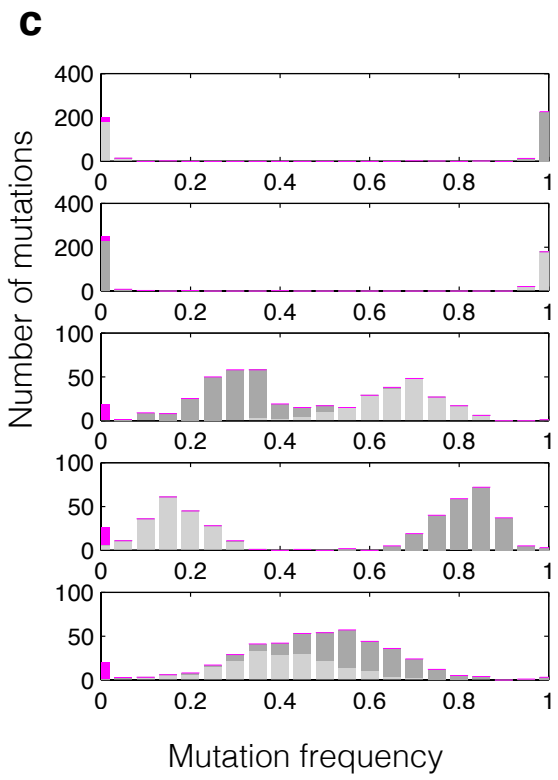
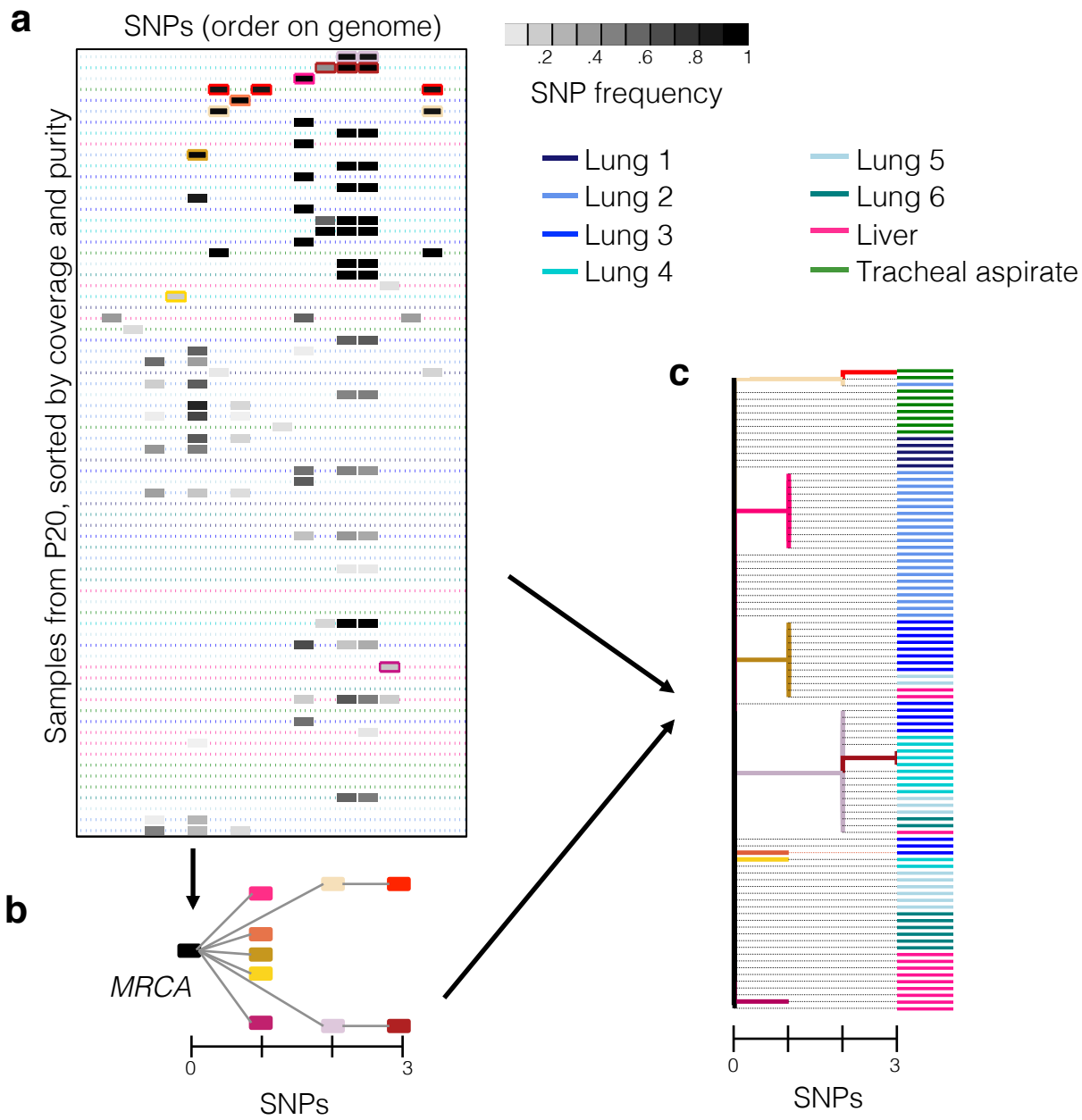
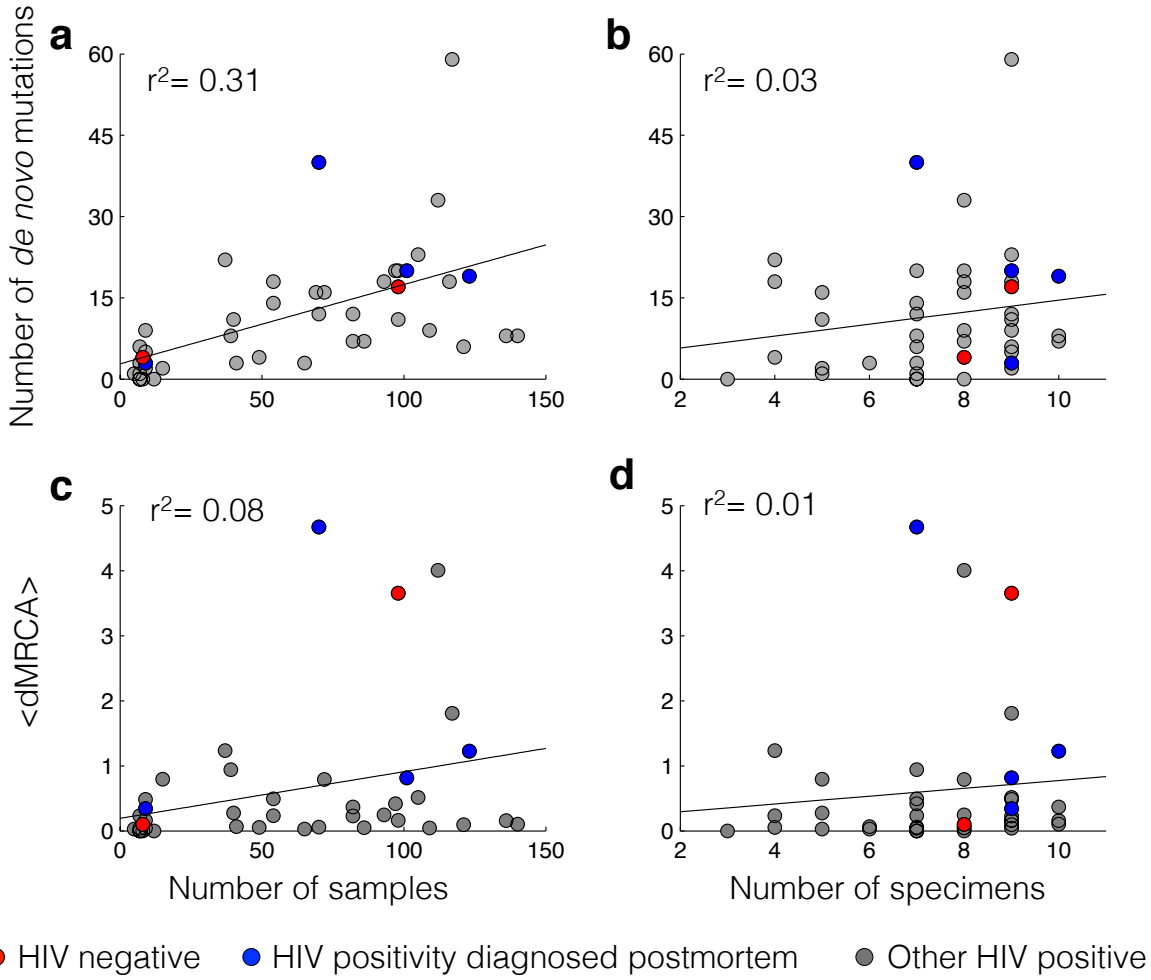




Supplementary Figure 1: Co-variation separates de novo mutations from preexisting mutations arising from mixed infection. Analysis of subject P15 is shown as an example. For each subject, a principal component analysis was performed on the matrix of derived mutation frequencies across samples from that subject. (**a**) A histogram of each mutation's contribution to the primary principal component is shown (PC1, explains > 98% for each subject). *De novo* mutations (magenta) were defined as those without a significant contribution to PC1 (absolute value < 0.011, determined empirically; dotted line) while mutations with significant contributions to PC1 were defined as coming from one strain or the other (dark gray, strain 1, < -0.011; light gray, strain 1, > 0.011). (**b**) A histogram of the weight of PC1 across samples (the frequency of Strain 1 in the sample), shows that some samples contained a single strain while other were mixed. (**c**) A histogram of mutation frequencies is shown for each of 5 samples from this subject, demonstrating that mutations assigned to each strain (by their contribution to PC1, panel a; strain 1, dark gray; strain 2, light gray) co-varied across samples and anti-covaried with mutations assigned to the other strain.

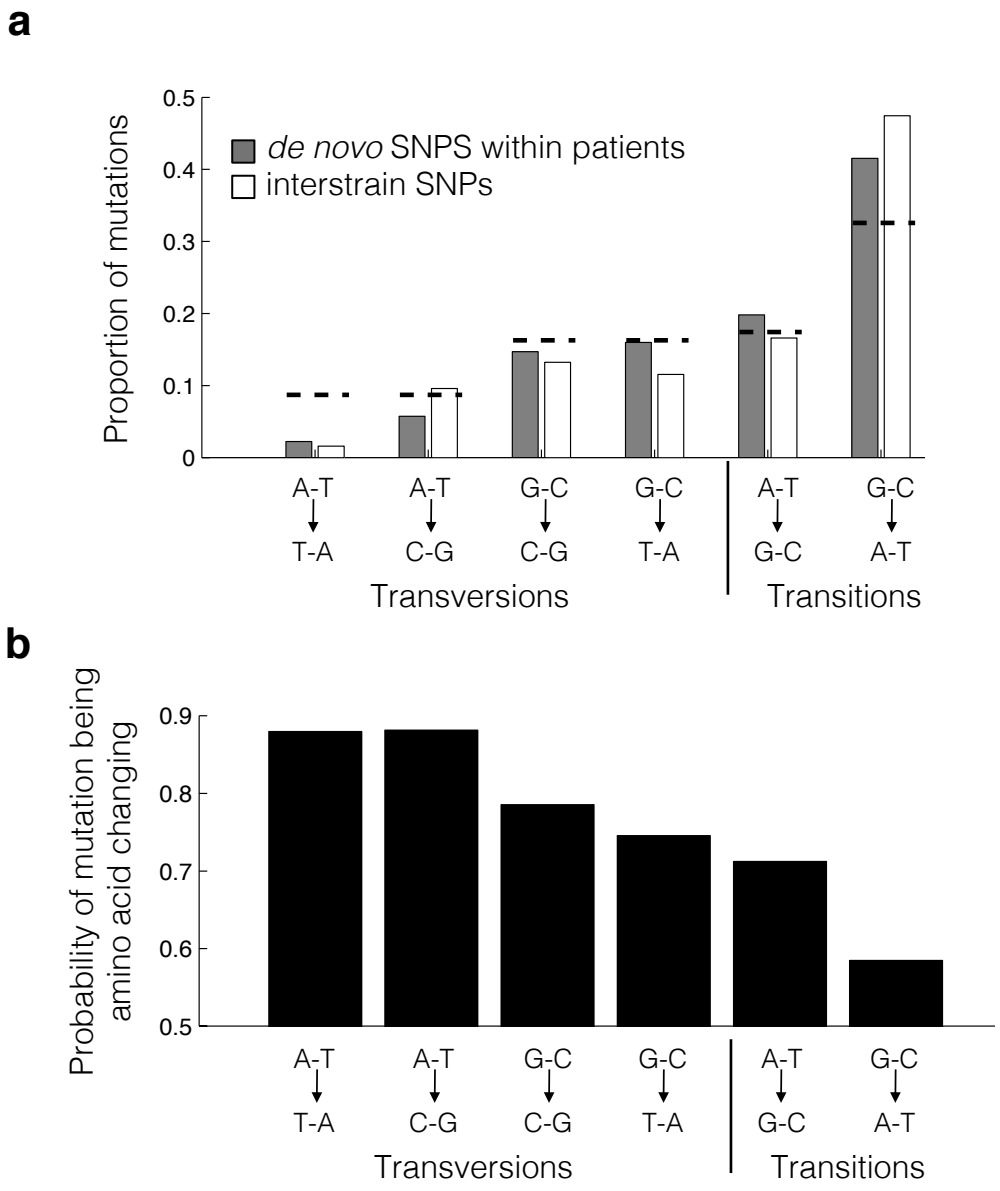


Supplementary Figure 2: A constraint based method defines within-host genotypes despite the non-clonality of samples. Data from Subject P20 is shown as an example. To identify genotypes despite the non-clonality of samples, we developed a conservative algorithm that identifies a minimal set of distinct genotypes whose combination explains the mutation frequency data (Online Methods & Supplementary Note). (**a**) A matrix of samples by positions is shown. Shading indicates the proportion of reads in that sample aligning to that position that support a mutation. Samples are sorted by their input to the algorithm (a function of coverage and average allele frequency across variant positions). Each set of colored outlines in the same horizontal line (with a unique color) indicates a defined genotype and the sample in which it was defined. Not all mutations were assigned to a genotype (columns without an outlined box). Dotted lines in background represent which site each sample was taken from. (**b**) Once genotypes were defined, their relationship to one another and the most recent common ancestor of the subject's population (MRCA) was determined by parsimony. (**c**) Following the identification of genotypes, each sample was defined as a combination of these genotypes, and a detailed evolutionary history across organs was inferred.

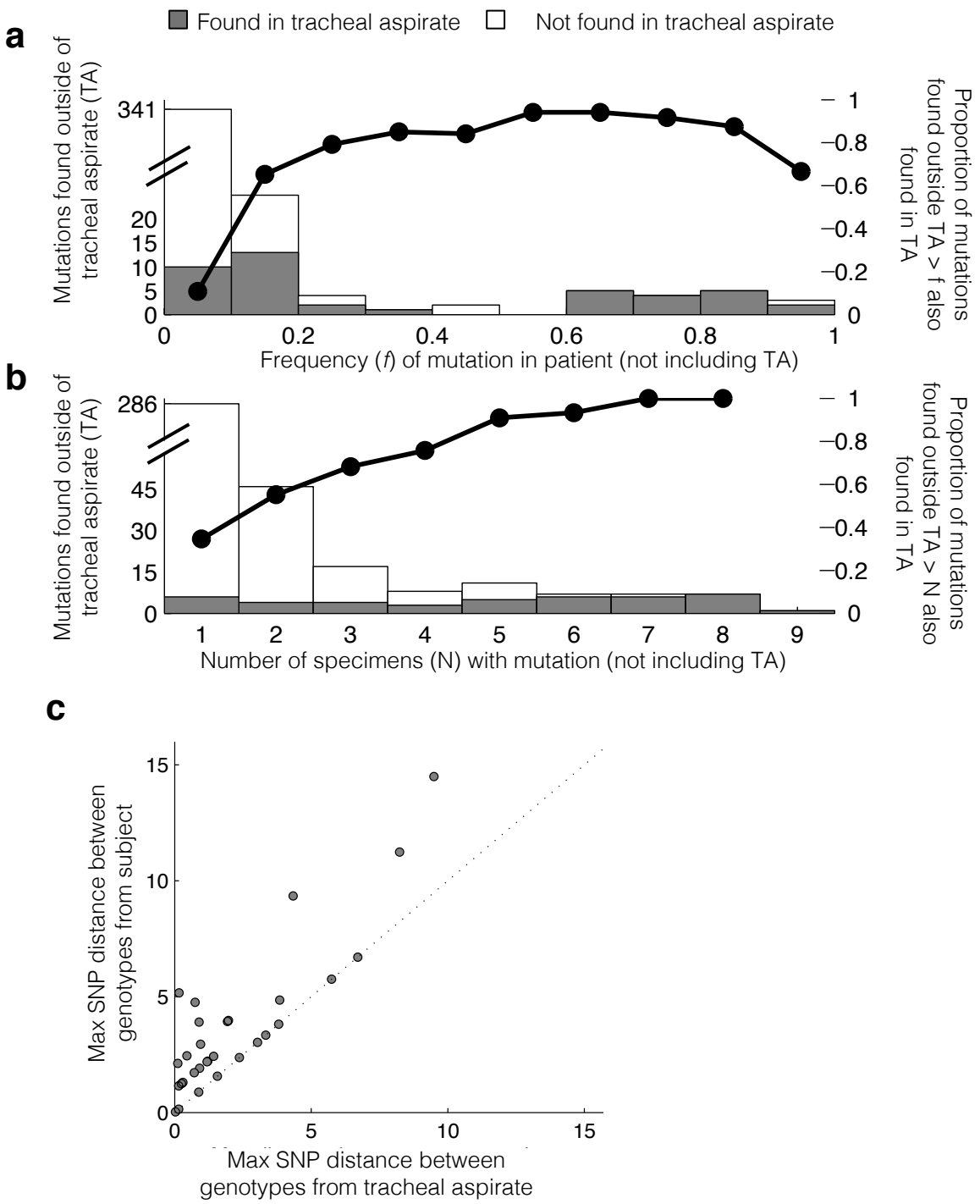


Supplementary Figure 3: Number of *de novo* mutations detected increases with number of samples analyzed. The number of *de novo* mutations detected per subject is shown as a function of (**a**) the number of samples and (**b**) number of specimens collected from that subject. The average mutational distance to most recent comment ancestor (<dMRCA>, see Methods), is also shown as a function of (**c**) the number of samples and (**d**) number of specimens collected from that subject. For panels (**c**) and (**d**), only subjects with single-strain infection are shown.

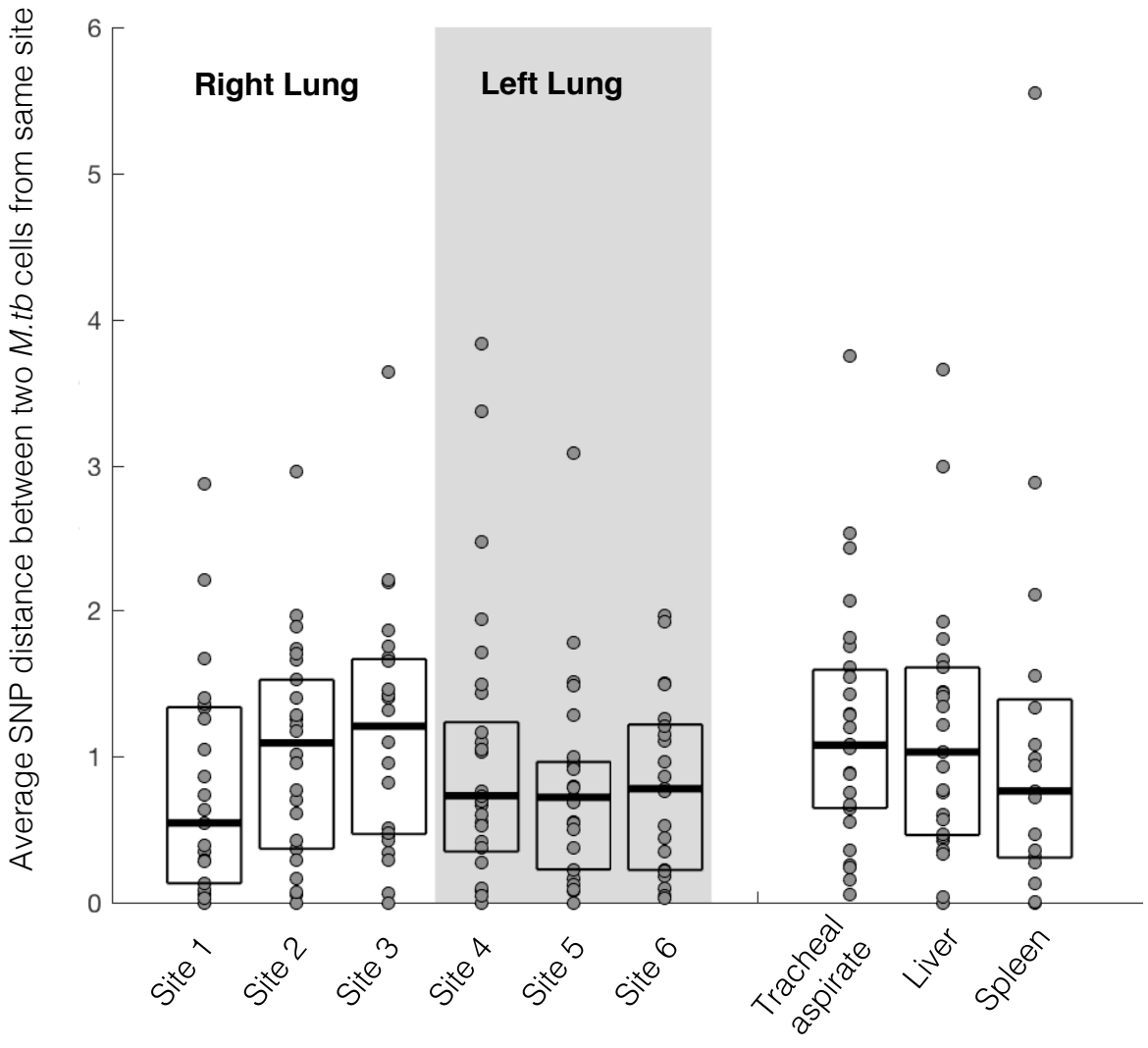
In each panel, r^2 is shown for the best linear fit for the data. Across panels, red circles indicate HIV negative subjects, blue circles indicate subjects diagnosed with HIV postmortem, and gray circles indicate subjects diagnosed with HIV prior to death.



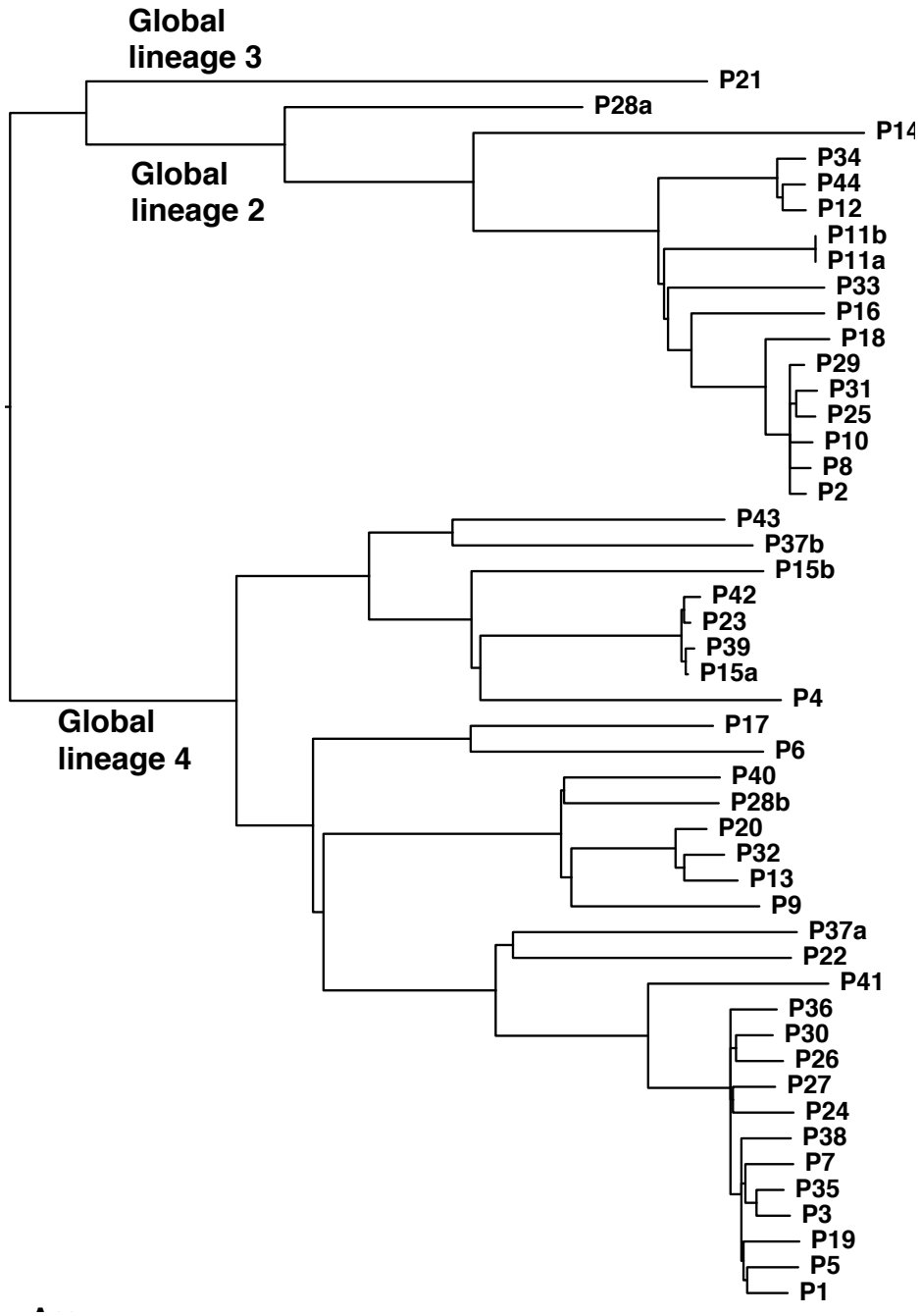
Supplementary Figure 4: Spectrum of mutations similar between *de novo* SNPs and SNPs separating lineages. (**a**) Polymorphic positions emerging because of *de novo* mutations (gray, n=518) and those found between patients (white, n=4094) were classified into 6 categories, without regard for strand (e.g. A->C is equivalent to T->G). Directionality was inferred using the reference strain as the ancestor (Methods). An expectation based on GC content and transitions being twice as likely as transversions is indicated by the black dotted lines. (**b**) The probability of each type of mutation causing an amino acid change across the TB genome is shown (neutral model), showing that the most frequent mutations are those which are most likely to preserve amino acid identity.



Supplementary Figure 5: Rare mutations are more likely to be missing from the endotracheal aspirate. The number of mutations observed outside the tracheal aspirate that are (gray bars) and are not (white bars) also found in the tracheal aspirate above a 5% frequency threshold are shown, as a function (**a**) the frequency of the mutations within the subject (averaging across sites) and (**b**) the number of sites in the subject in which that mutation was observed above the same 5% threshold. In each plot, we also show the probability that an observed mutation is also found in the tracheal aspirate (black lines), as a function of (**a**) intrapatient frequency across non-aspirate samples and (**b**) number of specimens. (**c**) For each subject with single strain infection, the maximum distance between pairs of genotypes in the tracheal aspirate is plotted against the maximum distance between pairs of genotypes within the subject. Jitter added for visibility (Up to 0.5 SNPs).



Supplementary Figure 6: Within-site diversity across sites For each subject with a single strain infection and multiple samples per site, the average genetic distance between two cells taken from the same site (normalized by subject, dividing by the mean intra-site distances across sites) is shown as a function of location. Boxes indicate the first, second (median), and third quartiles.



Supplementary Figure 7: Phylogeny of subjects isolates. A maximum parsimony tree was constructed using a single sample from each subject with a simple infection and two samples from each subject with mixed infection. The tree is rooted by the reference genome, an approximation of the ancestor of all TB. As our mutation caller was optimized for avoiding false positives, mutational distances are not accurate. Global lineages were determined by comparison to mutations in Coll et al[1].

Supplementary Table 1: Lung site locations

Biopsy site anatomical landmark	Right lung biopsy site no.	Left lung biopsy site no.
Second intercostal space, mid-clavicular line	1	4
Fourth intercostal space, anterior axillary line	2	5
Sixth intercostal space, posterior axillary line	3	6

Supplementary Table 2: Subject information

Subject	Year of birth	Sex	HIV Status	On anti-retrovirals at admission	# samples sequenced[a]	Average coverage	Max SNP distance between inferred genotypes	<dMRCA>	Global lineage(s)[b]	# de novo SNPs detected	Antimicrobial resistances[c]
P1	1985	male	positive	Yes	9	184	1	0.04	4	2	
P2	1958	male	positive	No	7	81	1	0.24	2	3	
P3	1969	male	positive**	No	9	192	1	0.34	4	3	
P4	1993	female	positive	Yes	7	152	0	0.00	4	0	Isoniazid, Rifampicin, Ethambutol, Ofloxacin, Ethionamide
P5	1983	female	negative	No	8	206	1	0.10	4	4	
P6	1974	male	positive	Yes	7	157	1	0.04	4	1	
P7	1976	male	positive	No	5	216	0	0.03	4	1	
P8	1977	female	positive	No	9	226	0	0.17	2	5	
P9	1978	male	positive	No	9	182	2	0.48	4	9	Isoniazid, Rifampicin, Ethambutol, Streptomycin, Ofloxacin
P10	1973	male	positive	Yes	7	303	0	0.00	2	0	
P11	1976	male	positive	No	7	73	801*	381*	2 & 4	6	
P12	1985	female	positive	Yes	8	150	0	0.00	2	0	Isoniazid, Rifampicin
P13	1975	female	positive	No	12	97	0	0.00	4	0	
P14	1966	female	positive	No	15	60	2	0.79	2	2	
P15	1970	female	positive	No	98	68	444*	201*	4	20	
P16	1980	male	positive	No	54	86	3	0.50	2	14	Rifampicin
P17	1980	male	positive	No	70	49	2	0.06	4	12	
P18	1974	male	positive	Yes	98	75	2	0.16	2	11	
P19	1963	male	positive	Yes	121	46	3	0.09	4	6	
P20	1971	female	positive	Yes	72	65	6	0.79	4	16	
P21	1965	female	positive	No	86	61	2	0.05	3	7	
P22	1949	male	positive	No	82	51	4	0.37	4	7	
P23	1972	male	positive	No	136	36	2	0.16	4	8	
P24	1977	male	positive	No	140	51	2	0.11	4	8	
P25	1975	male	positive	No	105	66	4	0.51	2	23	
P26	1960	female	positive	Yes	41	62	2	0.07	4	3	
P27	1981	male	positive	Yes	65	67	2	0.03	4	3	
P28	1974	male	positive	Yes	116	54	919*	439*	2 & 4	18	
P29	1977	male	positive	Yes	49	61	2	0.05	2	4	
P30	1982	female	positive**	No	123	59	4	1.23	4	19	
P31	1978	female	positive	Yes	82	52	3	0.23	2	12	
P32	1954	male	positive**	No	70	60	14	4.67	4	40	
P33	1976	female	positive	Yes	37	60	5	1.23	2	22	
P34	1980	male	positive	No	117	59	9	1.80	2	59	
P35	1973	female	negative	No	98	68	7	3.65	4	17	
P36	1968	male	positive	No	109	68	2	0.04	4	9	
P37	1967	female	positive	Yes	69	50	825*	392*	4	16	
P38	1977	female	positive	No	97	58	3	0.42	4	20	
P39	1978	male	positive**	No	101	62	3	0.81	4	20	
P40	1971	male	positive	No	39	68	4	0.94	4	8	
P41	1971	female	positive	Yes	112	56	11	4.00	4	33	
P42	1986	female	positive	Yes	54	63	5	0.23	4	18	
P43	1967	female	positive	Yes	40	57	4	0.28	4	11	
P44	1975	male	positive	Yes	93	58	5	0.25	2	18	

[a] See Supplementary Table 3 for more information.
[b] Determined by investigation of previously described lineage-defining positions[1].
[c] See Online Methods. Resistance phenotyping performed per site (not per sample). No variation found across sites taken from the same patient.
* The mutation caller was optimized to minimize false positives for patients with single strain infection and may underestimate strain-separating SNPs for patients with mixed infection. See Online Methods.
**Indicates HIV diagnosed only postmortem. See Online Methods.

Supplementary Table 3: Number of samples sequenced by site and subject

Subject	Lung 1 (Top Right)	Lung 2 (Middle Right)	Lung 3 (Bottom Right)	Lung 4 (Top Left)	Lung 5 (Middle Left)	Lung 6 (Bottom Left)	Endo-tracheal aspirate	Spleen	Liver	Other1	Other1 ID	Other2	Other2 ID
P1	1	1	1	1	1	1	1	1		1	Serous Fluid		
P2	1	1	1	1		1	1	1					
P3	1	1	1	1	1	1	1		1	1	Serous Fluid 1	1	Serous Fluid 2
P4	1	1	1	1	1	1	1						
P5	1	1	1	1	1	1	1		1				
P6		1	1	1	1	1	1	1					
P7	1	1	1			1	1						
P8	1	1	1	1	1	1	1	1	1				
P9	1	1	1	1	1	1	1	1		1	Lymph		
P10	1			1	1		1	1	1	1	Serous Fluid		
P11	1				1	1	1	1	1	1	Serous Fluid		
P12	1	1	1	1	1		1		1	1	Serous Fluid		
P13		6		1			5						
P14		1	1	4	2		7						
P15	15	14	8	15	11	14	14	7					
P16	12	7		7	13		8	6	1				
P17	13	10	15	5		7	12		8				
P18	14	13	14	8	7	9	6	12	15				
P19	15	14	13	13	14	14	11	15	12				
P20	5	14	10	8	9	7	9	10					
P21	13	14	11	11	13	11	3	10					
P22	1	1	15	8	5	9	1	14	14	14	Serous Fluid		
P23	15	13	15	14	14	14	13	15	15	8	Serous Fluid		
P24	14	15	14	15	15	13	9	15	15	15	Serous Fluid		
P25	15	9	15	10	13	8	12	9	14				
P26	2	8	8		3		15	5					
P27		5	10	9			14	14	13				
P28	15	5	11	15	14	13	14	15	14				
P29				13			12	11	13				
P30	13	8	15	13	10	15	14	14	10	11	Serous Fluid		
P31	15	15	15	11	4	3	10	5		4	Serous Fluid		
P32	15	15	3	15	6	1	15						
P33	6	3		15				13					
P34	15	15	14	15	15	14	6	10	13				
P35	11	12	14	12	14	13	11	6	5				
P36	15	7	14	15	14	15	14	15					
P37	15	15			15		11	13					
P38	15	15	15	14	12			13	13				
P39	11	13	13	13	9	11	15	11	5				
P40		5	1	4	8	9	1		11				
P41	14	15	14	14	13	14	13	15					
P42				13	13	13	15						
P43	14		3	7	9		7						
P44	13	15	11	7		2	15	15	15				

Supplementary Table 4 is a separate csv file containing detailed information for all *de novo* mutations detected in *M. tuberculosis* samples grown from autopsy specimens from 44 subjects

Supplementary Note

Site-specific specimen collection for mycobacterial culture

For each subject, respiratory tract secretions and up to 6 needle core biopsies per site were obtained for mycobacterial culture. Cold abscess collections, pleural effusions and ascetic fluid were aspirated if clinically identified. The following techniques were used:

Endotracheal specimen collection: The cricothyroid space was palpated and a puncture made with a scalpel blade through the cricothyroid membrane into the trachea. The puncture tract was held open with tissue forceps while a feeding tube was advanced into the trachea and then withdrawn while aspirating back into a 50 mL syringe. If no secretions were obtained the tube was advanced again until resistance was felt, and 50 mL of sterile 0.9% saline was lavaged down the tube and suctioned back.

***For all tissue specimens*:** 14-gauge core biopsy needles were used to obtain specimens, with a clean needle for each site. Needles were disinfected between subjects (see below). Specimens were put into 10 mL of sterile 0.9% saline in separate labeled screw top containers using one container for each organ, except for lung parenchyma samples, where each site was processed separately (see Supplementary Table 1).

Lung: Biopsies were obtained from each lung through 1 cm punctures in the second intercostal space (mid-clavicular line), fourth intercostal space (anterior axillary line) and sixth intercostal space (posterior axillary line). Repeated passes of the biopsy needle were made to a depth of 5-8 cm medial to the rib plane towards the center of each lung, until up to 6 cores of tissue were obtained.

Liver: The position of the liver was determined by percussion, and a 1 cm incision made over area of maximal dullness in the mid-axillary line (usually the 5th intercostal space). Repeated passes of the biopsy needle were made to a depth of 5-8 cm medial to the rib plane until up to 6 cores of tissue were obtained.

Spleen: A 1 cm incision was made over the 11th intercostal space in the posterior axillary line. Repeated passes of the biopsy needle were made in an anteriomedial plane to a depth of 5 cm medial to the rib plane until up to 6 cores of tissue were obtained.

Lymph node: Enlarged lymph nodes (>2 x 1 cm) in the cervical, axillary and inguinal regions was determined by palpation and a 1 cm incision made over the node. Repeated passes of the biopsy needle were made into the node until up to 6 cores were obtained. Fluctuant areas (i.e. cold abscess) were aspirated.

Serous fluid: Pleural effusion or ascites were detected clinically or radiologically (for pleural collections), aspirated through a 21 gauge needle and collected into separate screw top containers.

Disinfection of biopsy needles: Needles were soaked in hypochlorite solution for 5 minutes, rinsed in warm soapy water, and then cleaned externally with brushes and internally with dental tape. The 5 minute hypochlorite soak was repeated. After the second rinse the needles were soaked in 2% glutaraldehyde solution for 45 minutes, air dried, and reassembled.

Mutation calling

Processing samples from each subject individually, and looking at the raw data across samples from that subject in an interactive MATLAB environment, we implemented a set of filters and tests to reduce false positives, based on our previous work[2]. For fixed mutations, we used SAMtools 1.2[3] to generate a list of potential mutations, and accepted only those positions for which one pair of samples from that subject was discordant on the called base and both members of the pair had FQ scores (lower values indicate agreement between reads) less than −85. This threshold was set strictly to avoid false positives arising from very deep coverage.

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

- **Minor allele frequency:** More than 10% of reads support a particular minor allele
- **Overall coverage:** More than 10 reads align in both the forward and reverse direction, and the total number of reads aligning is below the 99th percentile of covered positions in that sample.
- **Minor allele coverage:** More than 3 reads per major and minor allele aligning in both the forward and reverse direction (4 thresholds: forward minor, forward major, reverse minor, reverse major)
- **Base quality:** Average base quality (provided by sequencer) of greater than 29 for both the major and minor allele calls on both the forward and reverse strand
- **Mapping quality:** Average mapping quality (provided by aligner) of greater than 39 for reads supporting both the major and minor allele on both the forward and reverse strand
- **Tail distance:** Average tail distance of between 25 and 75 for reads supporting both the major and minor allele on both the forward and reverse strand
- **Indels:** Fewer than 20% of the reads aligning to that position support an indel at any position along that read
- **Strand bias:** A P-value of $> 10^{-3}$ supporting a null hypothesis that the minor allele frequency is the same for reads aligning to both the forward and reverse strand (Fisher's exact test).

We then removed genomic positions that were suspected to have false polymorphisms based on the distribution of allele frequencies across samples in that subject. These filters were necessary because we did not explicitly mask repetitive elements like PE/PPE genes, which are known to cause false-positive polymorphisms when performing alignment-based SNP-calling and cover nearly 10% of the genome[4]. We chose a filter-based approach rather than a masking-based approach (1) to detect true variants in these regions and (2) to ensure that other problematic regions not known a-priori (e.g. duplications not present in the reference) were masked. These filters and particular cutoffs were chosen by empirically looking at the raw data across samples – in particular the number of reads supporting each nucleotide across samples, and the number of these reads that aligned to the forward and reverse strand. These filters preferentially removed polymorphisms from well-known problematic genes but also excluded mutations in other regions well. Masking all polymorphisms genes known to be problematic[5] would remove an additional 22 variable positions and not impact the conclusions of this study.

In particular, we implemented the following filters:

- **Co-variation with nearby mutations:** If nearby positions (within 1000bp) strongly co-varied in frequency across samples, these mutations were discarded. This is based on the assumption

that the few mutations arising from de novo mutation within each subject are unlikely to arise both nearby and at the same time. This filter was not applied when considering strain-separating mutations for subjects inferred to have mixed infection.

- **Most samples have polymorphism at same position**: If at least 90% of samples in that subject, at a given genomic position, had raw reads that supported a polymorphism (major allele frequency < 97%), this position was discarded. Given that samples were taken from colonies, we do not expect to see the same polymorphism at intermediate frequency in every or nearly every sample.
- **Many samples support polymorphism at same position but few confidently or at high frequency**: In particular, we removed positions in which < 15% of the samples with major allele frequency <97% (determined by raw reads) were called confidently. Similarly, we removed positions for which more than 33% of the samples supported a mutation (by raw reads) but no single sample supported the mutation at a frequency above 60%.

All de novo mutations that met our filters and their frequencies are listed in Supplementary 4.

Genotype identification and phylogenetic inference

To identify genotypes despite the non-clonality of samples, we developed a conservative algorithm that identifies a minimal set of distinct genotypes whose combination can explain the mutation frequency data. This algorithm is 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. Similar approaches have been used in recent studies, particularly in the reconstruction of cancer evolution[6,7]. The relatively small number of mutations per subject enabled us to use a simplified version of these algorithms. The input into our algorithm is a matrix of mutation frequencies (of size number of samples by number of polymorphic positions), where each mutation found is a derived mutation. Supplementary Figure 3 illustrates this algorithm, including the input data and inferred genotypes, for one example subject. 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.

For each subject, samples were first sorted by purity and coverage. In particular, samples in which the average mutation frequency (of positions polymorphic in that sample above > 0.1) was greater than 0.8 were considered first, followed by less pure samples. Within each group, samples were sorted by coverage. Samples with an average coverage of less than 12 reads were not used for genotype identification.

Then, for each sample, an attempt was made to see if any previously defined genotype (a binary vector of presence/absence of mutation) was consistent with the mutations in the majority of that sample. A previously defined genotype was said to be consistent the sample if: (1) it contained all mutations found > 0.6 frequency in that sample; (2) to account for measurement error, all the mutations in a less strict majority were sufficiently close to a previously defined genotype ($\frac{\sum_i^n |g_i - o_i|}{n} < 1 - F$, where, for each mutations i, $g_i = 1$ if this mutation is in that defined genotype and 0 otherwise, $o_i = 1$ if this mutation is above 0.4 in the sample and 0 otherwise, and F is the assumed frequency of this genotype – the minimum frequency across positions in the sample above 0.5).

New genotypes were then identified in the three following ways. (1) If the above conditions were not met and the mean frequency of mutations in the majority was > 0.6 (to exclude samples with mutations very close to 0.5, which are hard to deconvolve), a new genotype was defined containing all the mutations above 0.6 in that sample. (2) If only one mutation in the sample was above 0.2 frequency, a genotype containing only this mutation is defined as a new genotype. (3) If there was only one additional mutation in a sample with a genotype already observed at high frequency (F >0.8) that was at high enough frequency to be inferred on the background of this majority genotype (>1.15-F), a new genotype which was defined which contained both the mutations in the majority genotype and this mutation.

The thresholds in this algorithm were tuned such that no mutation was ever inferred to occur two independent times within a subject (an unlikely event given so few mutations per subject). In other words, if two mutations were found within the same genotype, these thresholds were set to ensure that only one was permitted to be in a genotype without the other (one mutation happened first).

Not all mutations were assigned to at least one genotype. In particular, some de novo mutations were only found at intermediate frequency within a sample that also contained other mutations at intermediate frequency; these mutations cannot be confidently phased and were not included in phylogenetic inference.

Finally, each sample was assigned to the found genotypes. Starting with genotypes with the most mutations, each genotype was examined for goodness of fit. If, for each position i in the genotype under examination, $f_i > .20$, this genotype was said to be present at frequency $\min(f)$ in that sample. This genotype was then removed from the observations for that sample ($f_i = f_i - \min(f)$) and the search for additional genotypes present in that sample continued.

Supplementary References

1. Coll, F. *et al*. A robust SNP barcode for typing Mycobacterium tuberculosis complex strains. *Nat Commun* **5,** 4812 (2014).
2. 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).
3. Li, H. *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
4. Casali, N. *et al*. Microevolution of extensively drug-resistant tuberculosis in Russia. *Genome Res.* **22,** 735–745 (2012).
5. Comas, I. *et al*. Human T cell epitopes of Mycobacterium tuberculosis are evolutionarily hyperconserved. *Nat. Genet.* **42,** 498–503 (2010).
6. 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).
7. 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).