Whole Genome Sequencing Analysis of Intrapatient Microevolution in *Mycobacterium tuberculosis*: Potential Impact on the Inference of Tuberculosis Transmission

Laura Pérez-Lago,[1,2,6,a] Iñaki Comas,[3,4,a] Yurena Navarro,[1,2,5] Fernando González-Candelas,[3,4] Marta Herranz,[1,2,6] Emilio Bouza,[1,2,6,7] and Darío García-de-Viedma[1,2,5,6]

[1]Servicio Microbiología Clínica y Enfermedades Infecciosas, Hospital General Universitario Gregorio Marañón, Madrid, Spain; [2]Instituto de Investigación Sanitaria Gregorio Marañón, Madrid, Spain; [3]Unidad Mixta Genómica y Salud, Centro Superior de Investigación en Salud Pública (FISABIO)-Universitat de València, Valencia, Spain; [4]CIBER en Epidemiología y Salud Pública CIBERESP, Spain; [5]CEI Campus Moncloa, UCM-UPM, Madrid, Spain; [6]CIBER Enfermedades respiratorias, CIBERES, and [7]Departamento de Medicina, Facultad de Medicina, Universidad Complutense de Madrid, Spain

Background. It has been accepted that the infection by *Mycobacterium tuberculosis* (*M. tuberculosis*) can be more heterogeneous than considered. The emergence of clonal variants caused by microevolution events leading to population heterogeneity is a phenomenon largely unexplored. Until now, we could only superficially analyze this phenomenon by standard fingerprinting (RFLP and VNTR).

Methods. In this study we applied whole genome sequencing for a more in-depth analysis of the scale of microevolution both at the intrapatient and interpatient scenarios.

Results. We found that the amount of variation accumulated within a patient can be as high as that observed between patients along a chain of transmission. Intrapatient diversity was found both at the extrapulmonary and respiratory sites, meaning that this variability can be transmitted and impact on the inference of transmission events. One of the events studied allowed us to track for a single strain the complete process of (i) interpatient microevolution, (ii) intrapatient respiratory variation, and (iii) isolation of different variants at different infected sites of this patient.

Conclusions. Our study adds new data to the understanding of variability in *M. tuberculosis* in a wide clinical scenario and alerts about the difficulties of establishing thresholds to differentiate relatedness in *M. tuberculosis* with epidemiological purposes.

Keywords. tuberculosis; microevolution; whole genome sequencing; intrapatient; interpatient; transmission events.

The application of molecular tools [1] has led to accept that the infection by *Mycobacterium tuberculosis* (*M. tuberculosis*) can be more heterogeneous than traditionally considered. Exogenous reinfections, mixed and compartmentalized infections, and microevolution processes leading to the emergence of clonal variants are versions of this clonal complexity [2–5]. Among these, infections involving clonal variants of *M. tuberculosis* have received little attention. However, the analysis of the mechanisms causing the microevolution—the genetic variability of *M. tuberculosis* at short time scales [4]—of a parental strain into clonal variants is a relevant issue to be addressed. Microevolution is considered when clonal *M. tuberculosis* variants are detected within a single patient simultaneously [6–8] or between patients, generated along the sequential infection of hosts involved in the same transmission cluster [9, 10].

Clonal variants are detected by the observation of subtle modifications in the distribution of IS*6110* copies, which cause RFLP variations and/or in the number of

Received 30 April 2013; accepted 30 July 2013; electronically published 14 August 2013.

[a]Both authors contributed equally to this study.

Correspondence: Darío García de Viedma, PhD, Servicio de Microbiología Clínica y Enfermedades Infecciosas, Hospital General Universitario Gregorio Marañón, Instituto de Investigación Sanitaria Gregorio Marañón, C/Dr Esquerdo, 46, 28007 Madrid, Spain (dgviedma2@gmail.com).

The Journal of Infectious Diseases 2014;209:98–108

DOI: 10.1093/infdis/jit439

repeats in some of the VNTR loci [4, 6, 8, 9, 11]. These subtle genotypic modifications can have functional implications; IS*6110* is able to interrupt genes or modulate the expression of adjacent genes [9, 12], and variations in the number of repeats can modify the transcription of neighboring genes [13, 14].

Until now, we have been almost completely blind to the microevolution occurring outside the genetic elements targeted by the aforementioned fingerprinting strategies. The recent advent of whole genome sequencing (WGS) for the analysis of *M. tuberculosis* can reveal information about the true dimension of microevolution and transform our currently limited knowledge about it. Whole genome has been only recently started to be evaluated as an epidemiological marker for the study of recent transmissions, leading to a new genomic epidemiology [15–17], which is revealing limitations of standard molecular epidemiology. Some of these studies have even defined thresholds for the number of SNPs found between independent isolates to infer whether the respective cases are involved or not in the same recent transmission cluster [17].

In this study, our aim has been to apply WGS to the study of microevolution in *M. tuberculosis*, an issue in which this methodology has not been applied specifically yet. The study has involved the analysis of representatives on intra- and interpatient microevolution. Our first goal was to describe the degree of variation, measured at the level of SNPs, which can be expected to have accumulated in a natural context, which is likely to be especially stressful as is the transmission of consecutive hosts. Second, we aimed at comparing these results with the SNP variability observed between isolates from the same patient, either from the respiratory site or from different body compartments. Finally, we show how inter- and intrapatient variability are part of a continuum in genomic variation, illustrated by a case with several isolates obtained from different body sites and also involved in a transmission cluster.

METHODS

Samples

We selected strains involved in recent transmission, epidemiologically supported clusters (RFLP-IS*6110* and MIRU24) in which at least one of the cases showed a clonal variant [9]. We selected clonal variants differing only in the RFLP pattern, only in the MIRU-VNTR pattern, and in both. We tolerated differences in one band in the RFLP-type or/and in one repeat unit in a single locus of the MIRU-type with respect to the patterns defining the cluster.

We also selected strains that had undergone microevolution within a patient [8]. In these cases, the genotypic differences accepted to define clonal variants were slightly more relaxed, as their close phylogenetic relationships can be assumed because they were simultaneously detected in the same patient. Independent respiratory specimens from a single patient were obtained the same day or ±1 day apart, whereas the maximum differences in sampling times between the respiratory and extra-respiratory specimens were 2, 2, and 8 days, respectively. We found up to 3 repeats in 1 of the 24 loci or 1 repeat in 2 loci of the MIRU type and kept the same 1-band difference criterion for the RFLP pattern.

All intrapatient and interpatient isolates were pan-susceptible.

Genotyping Methods

(i) **IS6110-based RFLP typing:** performed as in [18].

(ii) **MIRU-VNTR typing.** The 24 loci set was applied [19, 20].

Genome Sequencing of Isolates

Sequencing was carried out at GATC Biotech (Konstanz, Germany) for all isolates except those from cluster B (BGI [China]). Illumina TruSeq DNA sample preparation recommendations were followed. Illumina-HiSeq was used (read lengths ranged between 51 and 101 bp, and the percentage of reference positions covered by the reads was always higher than 98%). Fastq files with the raw data for the 36 isolates are deposited (http://www.ebi.ac.uk) under accession number ERP002297.

Read Mapping and Initial SNP Calling

We mapped the reads of each strain against a reference genome using as a main aligner BWA [21] and later MAQ [22] to corroborate the single-nucleotide polymorphism (SNPs) found. As a reference genome we used the most recent common ancestor of the *M. tuberculosis* complex (MTBC) as defined in [23, 24]. For SNP calling with BWA we used the Samtools suite [25] that allows one to process the mapped reads and look for differences between the reads and the reference genome. The initial calls for each strain were further filtered using different parameters to take into account reads mapping to more than 1 position, as well as SNP calls of low quality (minimum coverage 10, minimum mapping quality of the SNP 20).

Identification of Intra- and Intercluster SNPs

Individual lists of SNPs against the reference genome were generated independently for each isolate. The lists corresponding to isolates from the same group were merged and a nonredundant list of variable positions was generated. We used a series of filters to control for false positive and false negative SNP calls (see Supplementary Figure 1). Final SNP lists were annotated using ANNOVAR [26], and custom Perl scripts were used to identify the genome region (coding or intergenic) where each SNP fell into and, in the case of coding genes, whether the mutation was synonymous or nonsynonymous.

Analysis of Heterozygous Calls

Positions showing only heterozygous calls were removed from the analysis as they are difficult to distinguish from mapping errors. However, for some variable positions in some strains within a group, we detected a homozygous SNP call and for others a heterozygous SNP call. As heterozygous SNP calls in this context can mean coexistence of strains with 2 different alleles we performed a deeper analysis of the reads covering these positions. We used the program LoFreq [27] to determine the frequency of the different nucleotides for each position and to evaluate the reliability of the heterozygous SNP calls. Only cases in which the less frequent nucleotide was supported by at least 5 reads in each of the isolates were kept as highly likely polymorphic positions. Finally, the most likely heterozygous SNP calls were confirmed by allele-specific PCR.

Median-joining Networks

With the SNP matrix obtained for each group we constructed a median-joining network using the program NETWORK (www.fluxus-engineering.com). A median joining network is usually used to infer intraspecific phylogenies where small genetic distances are expected. The median joining network resolves all possible evolutionary paths connecting the considered taxa and postulates new nodes. In an epidemiological context these postulated nodes likely indicate incomplete taxon sampling.

Table 1. Single-Nucleotide Polymorphisms (SNPs) Corresponding to Interpatient Microevolution Events

Cluster D, SNPs:1				
CD-R1M1 (34)	CD-R1M1 (35)	CD-R1**M2** (37)	SNP meaning	
T	C	C	S	
2008	2008	2008		
Cluster C, SNPs:19				
CC-R1M1 (30)	CC-R1M1 (31)	CC-R1M1 (32)	CC-**R2**M1 (33)	SNP meaning
T	T	T	**G**	N
C	C	C	**G**	S
C	C	C	**T**	N
C	C	C	**T**	S
T	T	T	**C**	S
T	T	T	**G**	N
G	**C**	G	G	S
A	A	A	**C**	N
T	T	T	**G**	N
G	G	G	**A**	S
G	G	G	**A**	S
A	A	A	**G**	S
G	G	**C**	G	N
C	C	C	**T**	S
G	**C**	G	G	S
A	**G**	A	A	N
G	**A**	G	G	N
G	G	G	**A**	S
C	C	C	**T**	S
2008	2008	2006	2007	
Cluster F, SNPs:8				
CF-R1M1 (8)	CF-R1**M2** (9)	CF-**R2M3** (10)	SNP meaning	
G	C	C	N	
A	**G**	A	N	
C	C	**T**	N	
C	C	**T**	N	
T	T	**C**	I	
A	A	**C**	N	
A	G	G	N	
C	**T**	C	Stop	
2008	2007	2007		

The letter code indicates whether variants differ in the RFLP pattern (R in bold) or in the MIRU pattern (M in bold). The no. identifying each isolate is in brackets. Total no. of SNPs in the event is indicated in the head of the table. The SNPs for each of the variants are highlighted in bold. (S: synonymous, N: nonsynonymous and I: intergenic). The years of isolation are indicated in the bottom line.

RESULTS

We selected cases of interpatient and intrapatient microevolution events showing subtle differences in their fingerprints (Supplementary Figure 2). A panel of 36 isolates was selected for WGS. On average, each nucleotide position in the data set was read 473.34 times, with strain coverage 105X-3,885X.

After mapping and SNP calling, in silico analysis of diagnostic SNPs revealed that all sequences belonged to lineage 4 and a maximum-likelihood phylogenetic tree together with 36 genomes of unrelated lineage 4 strains confirmed the independence of the epidemiological clusters (Supplementary Figure 3).

Inter-patient Microevolution

We first compared the amount of sequence diversity accumulated in a transmission cluster with identical fingerprints to that of the microevolved clusters in which subtle changes in the genotyping patterns were observed. We selected 3 clusters representing differences only in MIRU, only in RFLP, and in both (Supplementary Figure 2).

We characterized 2 isolates from 2 independent cases, JP7 (2009) and JP8 (2010), as representative of a homogeneous cluster without microevolution (Cluster JP), which involved 5 epi-linked cases. Fully identical sequences were obtained by WGS analysis with no SNPs between the 2 isolates. Low SNP-variability has been found occasionally in linked clusters [17, 28].

Cluster D involved three patients (2008). Two isolates were identical by RFLP and MIRU (CD-R1M1), and the other presented a single-locus variant in MIRU (CD-R1M2). One SNP was found between the 2 CD-R1M1 isolates (Table 1), but no SNPs were found between the clonal variants.

Cluster C involved 5 cases occurring between 2003 and 2008, but only 4 isolates (2006–2008) were available. Three isolates (2 years apart) were indistinguishable by MIRU/RFLP (CC-R1M1), and 1 clonal variant (2007) differed in 1 IS*6110* copy (CC-R2M1). A total of 19 SNPs were found among them. Five SNPs were identified among the CC-R1M1 isolates, and the remaining 14 SNPs accumulated in the clonal variant (Table 1).

Cluster F (2007–2008) involved 6 cases. Four were identical by RFLP and MIRU (CF-R1M1), one differed by MIRU (CF-R1M2), and the remaining one by RFLP and MIRU (CF-R2M3). One representative of each of the 3 clonal variants was analyzed. A total of 8 SNPs were distributed among the 3 different representatives (2, 2, and 4 SNPs; Table 1).

A median-joining (MJ) network was generated to represent the genetic changes of the interpatient microevolution events. The analysis of MJ networks for the interpatient microevolution revealed a star-like pattern that is more compatible with a single patient infecting multiple secondary cases (a "super-spreader") (Figure 1). Another application of MJ network analysis is the identification of putative unnoticed cases of active

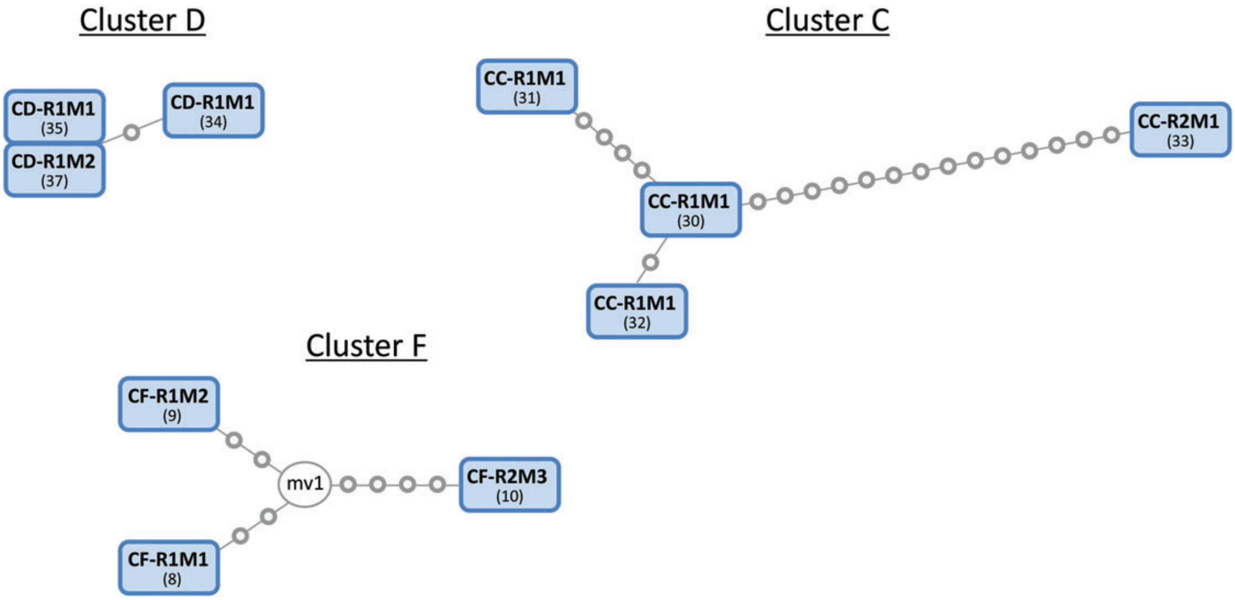

Figure 1. Median-joining networks for the 3 epidemiologically confirmed transmission clusters showing microevolution events. The nodes of the networks correspond to the *Mycobacterium tuberculosis* (*M. tuberculosis*) respiratory isolates showing differences in the RFLP (R1, R2) and/or MIRU-VNTR (M1, M2, M3) patterns. Each dot along the lines linking the nodes corresponds to a single-nucleotide polymorphism (SNP) detected between the variants in the connected nodes. In the case of cluster F a hypothetical, unsampled variant, is postulated by the algorithm (mv1). The no. identifying each isolate is in brackets. Abbreviations: CC, Cluster C; CD, Cluster D; CF, Cluster F.

tuberculosis involved in transmission events, as reflected in the inference of an unsampled haplotype in cluster F (mv1; Figure 1),

Intrapatient Microevolution

After measuring the SNP variability accumulated during the infection of sequential hosts, we evaluated the intrapatient variability in 4 patients.

The identification of intrapatient microevolution was based on MIRU-VNTR screening because this is the only way for detecting clonal complexity directly from *M. tuberculosis* cultures without subculturing and analyzing single colonies. Therefore, intrapatient microevolution events were assigned to 2 categories depending on whether the MIRU variants that had defined the microevolution event were different or not also by RFLP analysis. In order to evaluate potential SNP variability associated to the infection of independent sites which was not revealed by standard genotyping, all additional isolates from extra-respiratory sites for the cases with microevolution were included, regardless whether they shared fingerprinting pattern with the respiratory isolates or not.

Patients 10 and I were representatives of intrapatient microevolution involving clonal variants differing in the MIRU

Table 2. Single-Nucleotide Polymorphisms (SNPs) Corresponding to Intrapatient Microevolution Events

Patient 10, SNPs:7						
P10-R1M1 (12)	P10-R1**M2** (13)	SNP meaning				
A	**G**	N				
G	**A**	N				
A	**G**	N				
T	**C**	S				
G	**T**	S				
T	**C**	N				
G	**A**	S				
Respiratory	Respiratory					
Patient I, SNPs:3						
PI-R1M1 (26)	PI-R1M1 (29)	PI-R1M1 (28)	PI-R1**M2** (27)	SNP meaning		
C	**G**	**G**	**G**	S		
T	**C**	T	T	N		
C	C	**G**	C	N		
Respiratory	Urine	Lymph node	Lymph node			
Patient 8, SNPs:6						
P8-R1M1 (22)	P8-**R2**M1 (25)	P8-**R2M2** (21)	P8-**R2M3** (24)	P8-**R3M3** (23)	SNP meaning	
G	G	G	**A**	G	N	
A	G	G	G	G	S	
G	G	G	**A**	**A**	N	
C	C	C	**T**	**T**	S	
G	G	G	**A**	**A**	N	
A	**A**	G	G	G	N	
Respiratory	Respiratory	Respiratory	Respiratory	Respiratory		
Patient H, SNPs:8						
PH-R1M1 (20)	PH-R1M1 (14)	PH-**R2M2** (15)	PH-**R2M2** (19)	PH-**R2M3** (18)	PH-**R3**M1 (16)	SNP meaning
C	C	**A**	C	C	C	N
G	**C**	G	G	G	G	S
T	**A**	T	T	T	**A**	N
C	C	**G**	C	C	C	I
G	G	G	**T**	G	G	S
T	T	**A**	**A**	**A**	T	I
C	**G**	C	C	C	C	N
A	A	A	A	**T**	A	N
Respiratory	Respiratory	Respiratory	Blood	Blood	Blood	

The letter code indicates whether variants differ in the RFLP pattern (R in bold) or in the MIRU pattern (M in bold). The no. identifying each isolate is in brackets. The type of specimen from which each variant was isolated is indicated at the bottom of each column. Total no. of SNPs in the event is indicated in the head of the table. Each of the SNPs for each of the variants in highlighted in bold. (S, synonymous; N, nonsynonymous; and I, intergenic).

patterns but sharing identical RFLP types (Table 2). In patient 10, the 2 respiratory clonal variants (double locus variants, DLV) differed in 7 SNPs. In patient I, the 2 clonal variants (–SLV), one respiratory and the other isolated from a lymph node (PI-R1M1 and PI-R1M2), differed in one SNP. In this patient, 2 additional isolates isolated from urine and a lymph node and which shared identical genotyping patterns with the respiratory variant showed the same SNP variant than the one observed in the lymph node plus one additional SNP each (Table 2).

Patients 8 and H were representative of intrapatient microevolution involving clonal variants differing in both MIRU and RFLP patterns. In patient 8, 5 clonal variants (showing 3 different RFLP and 3 different MIRU microevolved patterns) were isolated from the respiratory site. Six SNPs were identified among them, 2 specific for one of the variants each, and the remaining ones shared by 2 variants (Table 2). In patient H we analyzed 2 respiratory clonal variants (PH-R1M1 and PH-R2M2, with another representative isolate of each of them, from sputum and blood). Two additional variants (PH-R2M3 and PH-R3M1) were found in blood. A total of 8 SNPs were found among all the variants, 6 among the respiratory isolates (3 between the 2 respiratory isolates of the same PH-R1M1 variant) and the remaining 2 SNPs in the blood variants. Three SNPs were found between the 2 PH-R2M2 variants isolated from sputum and blood samples (Table 2).

The topology of the MJ networks for intrapatient microevolution events was different and more complex than those for interpatient evolution (Figure 1). For patient 8, in which microevolution occurred at the respiratory level, the MJ network showed a linear topology, with a sequential accumulation of SNPs leading to the serial emergence of variants in 2 divergent branches from the likely infecting strain when compared with the MTBC ancestor (Figure 2). Networks for cases with respiratory and extra-respiratory isolates were especially complex. For instance, the MJ networks for patients H and I revealed evolution through several independent branches (Figure 2). In patient H, 2 blood isolates had no direct link with any respiratory sample, thus suggesting the presence of an unsampled variant in the lung or the extinction of the original infecting strain.

Simultaneous Analysis of Intra-patient and Inter-patient Variability

Cluster B represented a unique case, as isolates with both within- and among-host diversity were available. This scenario corresponded to an epidemiologically supported transmission

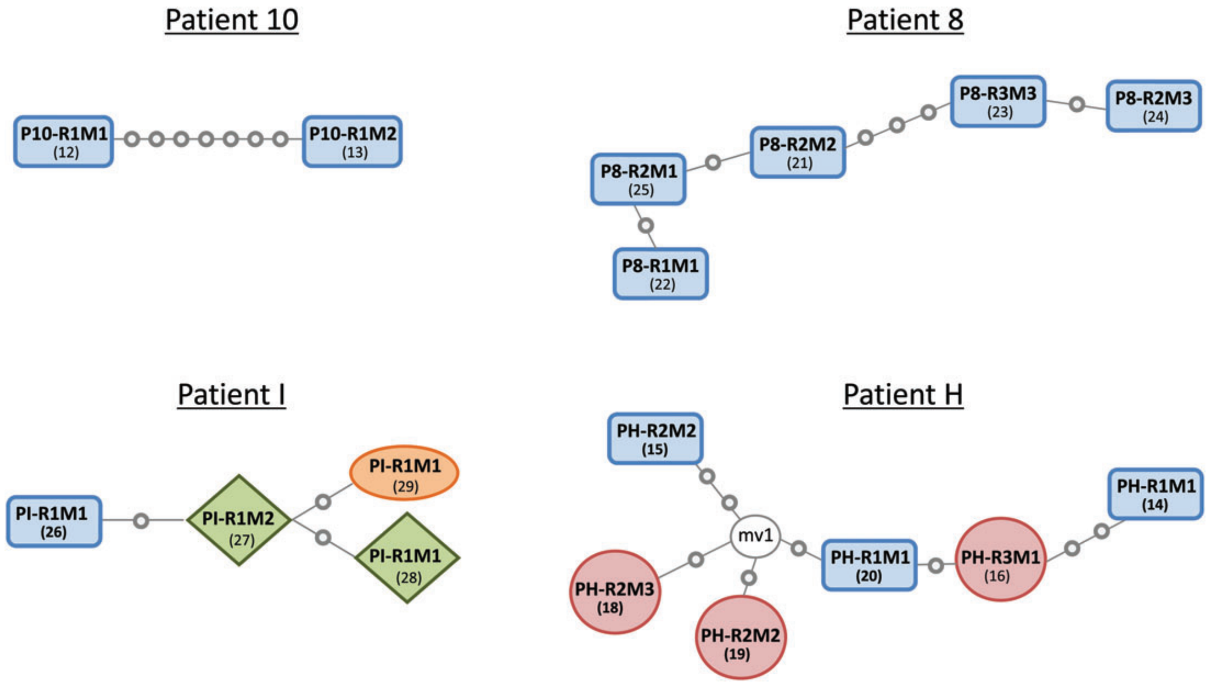

Figure 2. Median-joining networks for the *Mycobacterium tuberculosis* (*M. tuberculosis*) isolates sampled simultaneously from the same patient showing microevolution. The nodes of the networks correspond to the *M. tuberculosis* isolates showing differences in the RFLP (R1, R2, R3) and/or MIRU-VNTR (M1, M2, M3) patterns. Each dot along the lines linking the nodes corresponds to a single-nucleotide polymorphism (SNP) detected between the variants in the connected nodes. In the case of cluster H a hypothetical, unsampled variant, is postulated by the algorithm (mv1). Rectangle boxes correspond to respiratory isolates, diamond boxes to lymph node isolates, circle boxes to blood isolates, and ellipse boxes to urine isolates. The no. identifying each isolate is in brackets.

Table 3. Single-Nucleotide Polymorphisms (SNPs) Corresponding to the Inter/Intrapatient Microevolution Event

Cluster B/Patient J, SNPs:10							
CB-R1M1(38)	CB-**R2**M1(39)	CB-**R3**M1(40)	CB-**R3**M1(41)	CB-**R3**M1(42)	CB-**R3**M1(43)	CB-**R3M2**(44)	SNP meaning
G	G	**T**	G	G	G	G	I
T	T	T	**C**	**C**	**C**	**C**	S
A	A	**C**	A	A	A	A	S
C	C	C	C	**G**	**G**	**G**	N
T	**C**	T	T	T	T	T	N
T	T	T	T	**A**	T	T	N
A	C	C	C	C	C	C	Stop
G	G	G	**A**	**A**	**A**	**A**	N
G	G	G	G/**A**	**A**	**A**	**A**	N
C	C	C	C/**T**	**T**	**T**	**T**	N
Respiratory	Respiratory	Respiratory	Respiratory	Blood	Urine	Urine	
2005	2008	2005	2008				
Cluster B							
			Patient J				

The letter code indicates whether variants differ in the RFLP pattern (R in bold) or in the MIRU pattern (M in bold). The no. identifying each isolate is in brackets. The type of specimen from which each variant was isolated is indicated at the bottom of each column. Total no. of SNPs in the event is indicated in the head of the table. The SNPs for each of the variants are highlighted in bold. (S, synonymous; N, nonsynonymous; and I, intergenic). Variants corresponding to the patients in cluster B and to patient J are indicated at the bottom. The years of isolation for the interpatient event are indicated in the bottom file.

cluster involving 9 cases (2003–2008) in which clonal variants appeared both along the transmission chain and at different infected sites of one of the cases.

We analyzed 4 isolates from 4 cluster members with 3 clonal variants CB-R1M1, CB-R2M1 and CB-R3M1 (differing in RFLP but sharing MIRU types). Six SNPs were observed among them, 4 between the 2 isolates of the same variant CB-R3M1 (Table 3). Of note, in one of the cluster cases (patient J) infected by one of these clonal variants (CB-R3M1), 2 heterozygous SNPs, G/A and C/T, were found at the respiratory site. In the same positions the other 3 cases in the cluster showed G and C, respectively, which suggested that a new variant appeared from the transmitted variant in the heterozygotic case and that both coexisted at the time of sampling (Figure 3). The frequency of the alternative allele in both cases (15% and 77%) was markedly different, suggesting that they appeared at different time points.

Analyses of these heterozygous calls in isolates of the same patient from extra-pulmonary sites revealed that by the time the bacteria was out of the lungs the alternative allele was already fixed in the population (Figure 3*B*, 100% relative frequency in all extra-pulmonary isolates). The CB-R3M1 variant was isolated from urine and blood of the same patient. It presented A and T in the heterozygous positions, indicating that the heterozygous variant which appeared at the respiratory site was incorporated shortly before or after the extra-respiratory infection. Two additional SNPs were found in extra-respiratory sites (Table 3). Finally, a new MIRU variant (CBR3-M2) was found in urine, but no SNP was found between urine variants. Altogether, 10 SNPs were accumulated in the interpatient-intrapatient complete event, 6 and 4 of which corresponded to the inter- and intrapatient comparisons, respectively.

In this cluster B (Figure 3), as well as in the case of cluster F (Figure 1) the existence of unsampled, central haplotypes was inferred using the MJ analysis. Again, a star-like pattern compatible with a "super-spreader" was observed.

Comparison of Interpatient and Intrapatient Variability

In order to evaluate the degree of SNPs variability observed for the inter- and intrapatient events we performed a pairwise comparison of the number of SNPs identified in each group. We found that the genetic diversity within a patient can be as high as that found between patients (Figure 4, Mann-Whitney test, $P = .0523$, $P = .334$ when the CC-R2M1 outlier from cluster C was excluded).

DISCUSSION

The development of WGS strategies holds the promise to be the ultimate tool to analyze the transmission and epidemiology of pathogens. In the case of *M. tuberculosis* [15–17, 28–30] WGS-based approaches have been applied only recently, many of them analyzing transmission clusters in search of a SNP threshold as a new criterion to refine the definition of transmission [17]. The precise definition of SNP thresholds requires the evaluation of the variability, especially in those circumstances in which variability might be higher than the average. In this work, we have applied WGS to several sets of epidemiologically related isolates in which microevolution events had been

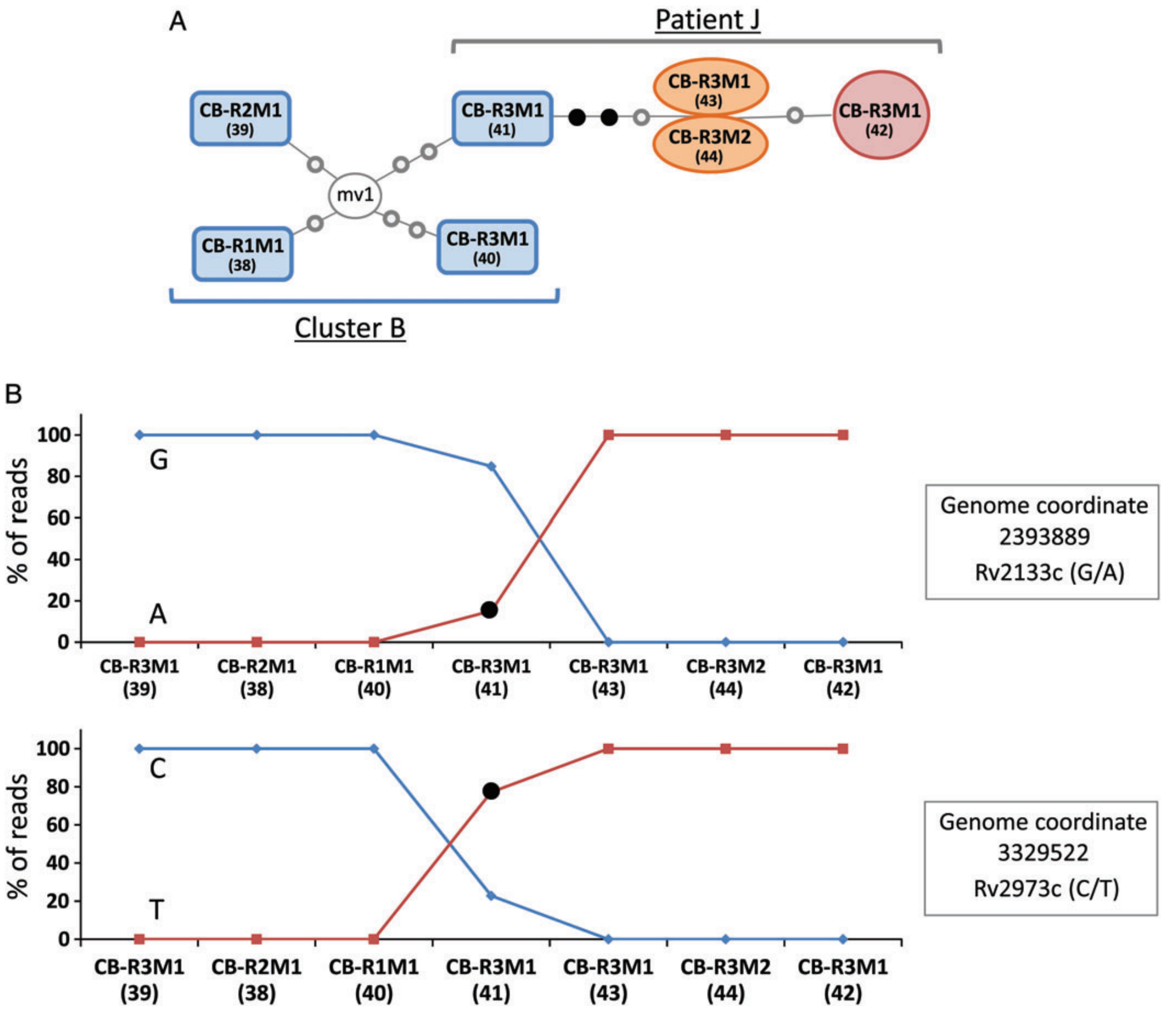

Figure 3. *A*, Median joining network for the *Mycobacterium tuberculosis* (*M. tuberculosis*) isolates involved in the interpatient/intrapatient microevolution event. The figure shows the respiratory isolates from the 4 clustered cases (cluster B: *rectangle boxes*) together with 3 additional extra-respiratory isolates from one of them (patient J): 2 from urine (*ellipse boxes*) and one from blood (*circle box*). Each dot along the lines linking the nodes corresponds to a single-nucleotide polymorphism (SNP) detected between the variants in the connected nodes. Black dots correspond to the frequency of the alternative allele of the polymorphic SNP in the respiratory isolate from patient J (CB-R3M1 (41)) that eventually become fixed in the urine and blood extrapulmonary isolates. The no. identifying each isolate is in brackets. *B*, The percentage of reads with the majority and the minority alleles in the 2 polymorphic sites for all isolates are shown. The order of the variants in the graphic does not try to follow the true chronology.

observed by standard typing, suggesting larger than average genomic diversity that might also be reflected in higher SNP diversity. The detection of clonal variants likely due to microevolution phenomena is not anecdotal. We found them in 7/703 (1%) of the respiratory cases, in 5/71 (7%) of the cases with both respiratory and extra-respiratory tuberculosis and in 9/74 (12%) of the transmission clusters analyzed [8, 9].

Although we must be cautious when directly comparing results that might have applied different SNP-calling pipelines, our results were generally in agreement with the thresholds proposed [17]. Nevertheless, our data also show the difficulties in setting strict universal thresholds for delimiting transmission. Although the existence of missing links between the cases in cluster C accumulating 14 SNPs cannot be ruled out, we have also documented transmission chains with isolates differing in >5 SNPs. Similar bursts of mutations in a single isolate have been found [31].

From the host-to-host transmission point of view the use of complete genome sequences brings a different perspective to tuberculosis molecular epidemiology. Now it is possible to know the precise definition of the topology and dynamics of the transmission event [15–17, 31]. This is because backward mutations are very rare in *M. tuberculosis*, and the relatively rare homoplasy [24] allows defining the position of every case in the transmission topology according to the pattern of SNP acquisition. The interpatient networks analyzed here revealed that, instead of a linear stepwise case-to-case transmission, branched transmissions with one case causing more than one secondary case are

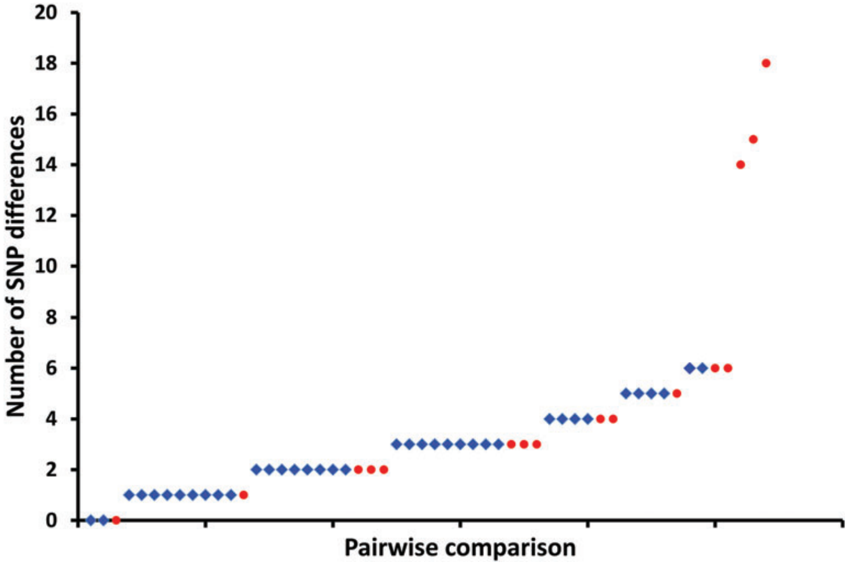

Figure 4. Overlapping pairwise single-nucleotide polymorphism (SNP) distances between isolates belonging to the same patient (*diamond dots*) or clusters of transmission (*circle dots*).

common. Similar topologies in other studies have allowed defining super-spreader cases and understanding that they are more frequent than assumed [16, 17]. In this context, the introduction of analytical methods capable of inferring putative missing nodes, such as median-joining networks, to reveal the connections between variants has the advantage of postulating these undiagnosed cases when a node in the network is not occupied by any of the analyzed isolates [32].

Little attention has been paid to the degree of genetic variation that can be observed during the course of an infection within a patient [17, 33]. Our study design, analyzing both inter- and intrapatient isolates, has allowed us to establish that the genetic diversity within a patient can be as high as that found between patients. Our results highlight that *M. tuberculosis* can accumulate an unexpected level of SNP variation in a general situation, out of processes of acquisition of resistance where a high intrapatient SNP variability is expected [33, 34]. We have also provided evidence that intrapatient diversity is not exclusive of extra-pulmonary sites, which do not impact on transmission, but it was also observed in the respiratory site of several patients (patients 8 and 10), suggesting that substantial diversity can exist in the lung and be eventually transmitted.

We could evaluate the host-to-host and within-patient variability for the same *M. tuberculosis* strain that infected several cases, and in one of them pulmonary and extra-respiratory variants emerged. This analysis confirmed the equivalent interpatient/intrapatient variability results and also allowed us to monitor the fixation of polymorphisms from an initial heterogeneous population in the lungs to a homogeneous population in extrapulmonary sites.

From a public health perspective, it is of paramount importance to evaluate the genetic diversity accumulated, transmitted, and able to infect a new individual. Intrapulmonary variants, such as those revealed here in patients 8, 10, and H, can have a distorting effect on the interpretation of transmission networks and epidemiological investigations. The coexistence of variants within a patient after its initial infection can blur the inference of transmission links based on strict SNP cutoffs, given that the transmitted strain may have accumulated many SNPs with respect to other isolates of the same cluster of transmission. In this context, the analysis of low-frequency variants in sputum samples, usually discarded in WGS approaches but suitable using deep-sequencing approaches, can be critical to infer transmission links from whole genome sequences.

In summary, the findings in this study add complexity to the picture of variability in *M. tuberculosis*. Our results suggest that a constant rate of SNP acquisition may not be valid for all situations/strains, and it is unclear whether strict SNP thresholds to infer transmissions are universally valid or can be applied in only certain situations. Here we present several strains that had shown variability in standard fingerprinting patterns despite proven epidemiological links. The genome sequence of these isolates shows that SNP microevolution within a host can be unexpectedly high and equivalent to the levels of microevolution observed in a host-to-host transmission situation. Additionally, microevolution occurs both at the extra-respiratory level and at respiratory sites, which could impact on the diversity expected for transmission chains. The coexistence of microevolved variants within a host, along with other environmental and/or clinical factors known to foster the accumulation of mutations, could fully transform our expectations of what must be considered as related or unrelated in *M. tuberculosis* epidemiology.

Supplementary Data

Supplementary materials are available at *The Journal of Infectious Diseases* online (http://jid.oxfordjournals.org/). Supplementary materials consist of data provided by the author that are published to benefit the reader. The posted materials are not copyedited. The contents of all supplementary data are the sole responsibility of the authors. Questions or messages regarding errors should be addressed to the author.

Notes

Acknowledgments. We are grateful to Juanjo Palacios, Miguel Martínez Lirola, and María Jesús Ruiz Serrano for allowing us to select some of the MTB strains isolated in their respective institutions as a source for this study. We acknowledge Milagros González Rivera, in charge of the Sequencing Unit from Instituto de Investigación Sanitaria Gregorio Marañón, for her support in the introduction of WGS strategies in our institution.

Financial support. This work was supported by Fondo de Investigaciones Sanitarias (FIS S09/02205, PI12/02080). L. P. held a Juan de la Cierva contract from the Ministerio de Ciencia e Innovación (JCI-2009-05713) and now holds a contract from CIBERES (Ref CP13/27/01). I. C. is supported by European Union funding from the Marie Curie FP7 actions (project number 272086). Research by Y. N. was partly supported by a PICATA predoctoral fellowship (BE55/11) from the Moncloa Campus of International Excellence (UCM-UPM, Instituto de Investigación Sanitaria Gregorio Marañón). F. G. C. was supported by Project BFU2011-24112 from Ministerio de Economía y Competitividad (Spain).

Potential conflicts of interest. All authors: No reported conflicts.

All authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Conflicts that the editors consider relevant to the content of the manuscript have been disclosed.

References

1. Cohen T, van Helden PD, Wilson D, et al. Mixed-strain *mycobacterium tuberculosis* infections and the implications for tuberculosis treatment and control. Clin Microbiol Rev **2012**; 25:708–19.
2. Martín A, Herranz M, Navarro Y, et al. Evaluation of the inaccurate assignment of mixed infections by *Mycobacterium tuberculosis* as exogenous reinfection and analysis of the potential role of bacterial factors in reinfection. J Clin Microbiol **2011**; 49:1331–8.
3. Stavrum R, Mphahlele M, Ovreas K, et al. High diversity of *Mycobacterium tuberculosis* genotypes in South Africa and preponderance of mixed infections among ST53 isolates. J Clin Microbiol **2009**; 47:1848–56.
4. Al-Hajoj SA, Akkerman O, Parwati I, et al. Microevolution of *Mycobacterium tuberculosis* in a tuberculosis patient. J Clin Microbiol **2010**; 48:3813–6.
5. Shamputa IC, Jugheli L, Sadradze N, et al. Mixed infection and clonal representativeness of a single sputum sample in tuberculosis patients from a penitentiary hospital in Georgia. Respir Res **2006**; 7:99.
6. Andrade MK, Machado SM, Leite ML, Saad MH. Phenotypic and genotypic variant of MDR-*Mycobacterium tuberculosis* multiple isolates in the same tuberculosis episode, Rio de Janeiro, Brazil. Braz J Med Biol Res **2009**; 42:433–7.
7. García de Viedma D, Marin M, Andres S, Lorenzo G, Ruiz-Serrano MJ, Bouza E. Complex clonal features in an *Mycobacterium tuberculosis* infection in a two-year-old child. Pediatr Infect Dis J **2006**; 25:457–9.
8. Navarro Y, Herranz M, Pérez-Lago L, et al. Systematic survey of clonal complexity in tuberculosis at a populational level and detailed characterization of the isolates involved. J Clin Microbiol **2011**; 49:4131–7.
9. Pérez-Lago L, Herranz M, Lirola MM, Bouza E, Garcia de Viedma D. Characterization of microevolution events in *Mycobacterium tuberculosis* strains involved in recent transmission clusters. J Clin Microbiol **2011**; 49:3771–6.
10. Ijaz K, Yang Z, Matthews HS, Bates JH, Cave MD. *Mycobacterium tuberculosis* transmission between cluster members with similar fingerprint patterns. Emerg Infect Dis **2002**; 8:1257–9.
11. Hawkey PM, Smith EG, Evans JT, et al. Mycobacterial interspersed repetitive unit typing of *Mycobacterium tuberculosis* compared to IS6110-based restriction fragment length polymorphism analysis for investigation of apparently clustered cases of tuberculosis. J Clin Microbiol **2003**; 41:3514–20.
12. McEvoy CR, Falmer AA, Gey van Pittius NC, Victor TC, van Helden PDWarren RM. The role of IS*6110* in the evolution of *Mycobacterium tuberculosis*. Tuberculosis (Edinb) **2007**; 87:393–404.
13. Tantivitayakul P, Panapruksachat S, Billamas P, Palittapongarnpim P. Variable number of tandem repeat sequences act as regulatory elements in *Mycobacterium tuberculosis*. Tuberculosis (Edinb) **2010**; 90:311–8.
14. Akhtar P, Singh S, Bifani P, et al. Variable-number tandem repeat 3690 polymorphism in Indian clinical isolates of *Mycobacterium tuberculosis* and its influence on transcription. J Med Microbiol **2009**; 58:798–805.
15. Schurch AC, Kremer K, Daviena O, et al. High-resolution typing by integration of genome sequencing data in a large tuberculosis cluster. J Clin Microbiol **2010**; 48:3403–6.
16. Gardy JL, Johnston JC, Ho Sui SJ, et al. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. N Engl J Med **2011**; 364:730–9.
17. Walker TM, Ip CL, Harrell RH, et al. Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. Lancet Infect Dis **2013**; 13:137–46.
18. van Embden JD, Cave MD, Crawford JT, et al. Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. J Clin Microbiol **1993**; 31:406–9.
19. Supply P, Allix C, Lesjean S, et al. Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of *Mycobacterium tuberculosis*. J Clin Microbiol **2006**; 44:4498–510.
20. Oelemann MC, Diel R, Vatin V, et al. Assessment of an optimized mycobacterial interspersed repetitive- unit-variable-number tandem-repeat typing system combined with spoligotyping for population-based molecular epidemiology studies of tuberculosis. J Clin Microbiol **2007**; 45:691–7.
21. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics **2010**; 26:589–95.
22. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res **2008**; 18:1851–8.
23. Comas I, Chakravartti J, Small PM, et al. Human T cell epitopes of *Mycobacterium tuberculosis* are evolutionarily hyperconserved. Nature Genet **2010**; 42:498–503.
24. Comas I, Coscolla M, Luo T, et al. Out-of-Africa migration and Neolithic co-expansion of tuberculosis with modern humans. Nature Genet **2013**; Published online 1 September 2013. doi:10.1038/ng.2744.
25. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics **2009**; 25:2078–9.
26. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res **2010**; 38:e164.
27. Wilm A, Aw PP, Bertrand D, et al. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucleic Acids Res **2012**; 40:11189–201.
28. Roetzer A, Diel R, Kohl TA, et al. Whole genome sequencing versus traditional genotyping for investigation of a *Mycobacterium tuberculosis* outbreak: a longitudinal molecular epidemiological study. PLoS Med **2013**; 10:e1001387.
29. Schurch AC, van Soolingen D. DNA fingerprinting of *Mycobacterium tuberculosis*: from phage typing to whole-genome sequencing. Infect Genet Evol **2012**; 12:602–9.

30. Bryant JM, Schurch AC, van Deutekom H, et al. Inferring patient to patient transmission of *Mycobacterium tuberculosis* from whole genome sequencing data. BMC Infect Dis **2013**; 13:110.
31. Schurch AC, Kremer K, Kiers A, et al. The tempo and mode of molecular evolution of *Mycobacterium tuberculosis* at patient-to-patient scale. Infect Genet Evol **2010**; 10:108–14.
32. Woolley SM, Posada D, Crandall KA. A comparison of phylogenetic network methods using computer simulation. PloS One **2008**; 3:e1913.
33. Sun G, Luo T, Yang C, et al. Dynamic population changes in *Mycobacterium tuberculosis* during acquisition and fixation of drug resistance in patients. J Infect Dis **2012**; 206:1724–33.
34. Comas I, Borrell S, Roetzer A, et al. Whole-genome sequencing of rifampicin-resistant *Mycobacterium tuberculosis* strains identifies compensatory mutations in RNA polymerase genes. Nature Genet **2011**; 44:106–10.
35. Casas-Fischer R, Penedo-Pallares A, Palacios-Gutierrez JJ, Moreno-Torrico A. Outbreak or coincidental cases of tuberculosis? Genotyping provides the clue. Am J Infect Control **2012**; 40:9–10.
36. Darriba D, Taboada GL, Doallo R, Posada D. jModelTest 2: more models, new heuristics and parallel computing. Nat Methods **2012**; 9:772.
37. Akaike H. A new look at the statistical model identification. IEEE Trans Inf Technol Biomed **1976**; 19:5.
38. Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol **2003**; 52:696–704.