Historical contingency, rather than higher transmissibility, behind the long-term success of a *Mycobacterium tuberculosis* outbreak strain

Mariana G. López[1], Mª Isolina Campos-Herrero[2], Manuela Torres-Puente[1], Fernando Cañas[3], Jessica Comín[4], Rodolfo Copado[5], Penelope Wintringer[6], Zamin Iqbal[6], Eduardo Lagarejos[2], Miguel Moreno-Molina[1], Laura Pérez-Lago[7], Berta Pino[8], Laura Sante[9], Darío García de Viedma[7,10,‡], Sofía Samper[4,10,‡,*], Iñaki Comas[1,11*]

[1]Tuberculosis Genomics Unit, Instituto de Biomedicina de Valencia (IBV-CSIC), Valencia 46010, Spain
[2]Servicio de Microbiología, Hospital Universitario de Gran Canaria Dr. Negrín, Las Palmas de Gran Canaria 35010, Spain
[3]Hospital Universitario Insular de Gran Canaria, Las Palmas de Gran Canaria 35016, Spain
[4]Instituto Aragonés de Ciencias de la Salud, Fundación IIS Aragón, Zaragoza 50009, Spain
[5]Hospital José Molina Orosa, Las Palmas de Gran Canaria 35500, Spain
[6]European Molecular Biology Laboratory – European Bioinformatics Institute, Hinxton CB10 1SD, UK
[7]Servicio Microbiología Clínica y Enfermedades Infecciosas, Hospital General Universitario Gregorio Marañón, Instituto de Investigación Sanitaria Gregorio Marañón, Madrid 28007, Spain
[8]Hospital Nuestra Señora de la Candelaria, Santa Cruz de Tenerife 38010, Spain
[9]Hospital Universitario de Canarias, Santa Cruz de Tenerife 38320, Spain
[10]CIBER Enfermedades Respiratorias, Instituto de Salud Carlos III, Madrid 28029, Spain
[11]CIBER Epidemiología y Salud Pública, Instituto de Salud Carlos III, Madrid 28029, Spain

[‡] these authors equally contributed
[*] corresponding authors

Abstract

Infectious disease outbreaks are a major public health problem. Tracking their spread and understanding the main factors associated with it are key to assist in disease control. Whole genome sequencing (WGS) of pathogens has been demonstrated to be the best tool to investigate outbreaks, since it helps to answer central questions about their origin and dynamics, and also factors associated with strain success and outbreak evolution. Identification of highly transmissible strains in *Mycobacterium tuberculosis* is a main area of research. Long-lived outbreak strains have been proposed as candidates with high intrinsic transmissibility. However, limited resolution of available typing methods have prevented testing this hypothesis. Here we study an outbreak strain that started to circulate in Gran Canaria (GC, Canary Islands) in 1993 and has caused around 20% of TB cases since then. By using WGS, and phylodynamics analysis we could trace the origin of the strain to 700 years ago in China and infer its introduction in Gran Canaria in the early 90s'. Once established, we document the topology of the outbreak within the Canary Islands. Importantly, we were able to trace the spread of the outbreak beyond the studied area by searching for related sequences within the whole sequence repository, revealing a high success of GC strain in the Canarian archipelago but a limited expansion into the continent. The phylodynamics analyses suggest that determinants associated with the host, instead of higher transmissibility or virulence of the bacteria, are responsible for this long-lived outbreak. We show that historical contingency in the form of a founder effect rather than intrinsic transmissibility explains better the success of the GC strain and probably others of the same nature. Like the GC outbreak strain, many other similar strains in the literature had their origin in the late 20th century suggesting that past epidemics driven by extrinsic factors like HIV are still contributing to the present TB epidemic in local settings.

Introduction

Tuberculosis (TB) is the leading cause of death from a single infectious agent with 1.4 million deaths in 2019 [1], and very much remains a major global health problem. TB

outbreaks, defined as the concentration of an abnormal number of disease cases in space and time, are very common in TB. While outbreaks have generally been assumed to be short-lived, genotyping has been able to identify outbreak strains that became highly prevalent in specific regions and end up transmitting over decades. These outbreak strains can be found worldwide and across MTBC lineages and are usually drivers of local TB burden [2–7]. However, outbreak strains usually have very limited geographical dispersal, raising the question of whether those strains have any intrinsic transmissibility advantage or if their success is associated with random population processes like founder effects or ecological drivers of transmission.

In this work we investigate in detail one of those outbreak strains and compare them to others, previously published, to learn about the origin and epidemiology of long-lived strains. In 2001 a TB outbreak was identified in the Canary Islands. 651 strains were retrospectively analyzed between 1991-1996 in Gran Canaria (GC), the most populated island of the archipelago, using RFLP IS*6110*. A big cluster of 75 isolates was recognized with the first 10 cases diagnosed in 1993. The likely index case was a Liberian refugee who arrived on the island 6 months before diagnosis, in July of 1993[8]. The strain belongs to the Beijing genotype, corroborated by spoligotyping, and was named GC1237. More recently, three GC-outbreak single nucleotide polymorphisms (SNPs) were selected and used to design a specific PCR to rapidly identify secondary cases belonging to GC-outbreak [9]. The fast spread of the GC strain to the other islands of the Canarian archipelago was evidenced by using different molecular typing methods, and also new cases outside the archipelago were identified up to 2014 [9–11]. Until now, GC-outbreak has never been studied by exploring the information obtained from whole genome sequencing (WGS).

In this work, we combine epidemiological and sequencing data, and apply a phylodynamics analysis to trace the origin of the GC outbreak strain, to track its spread, to understand its dynamics and define the factors that underpin its expansion.

Results

Genomic delineation of the GC outbreak

We retrieved 86 samples from outbreak patients previously typified by molecular methods, plus 3 additional samples from the likely index case collected in different years. Out of those, we discarded 24 samples which did not achieve sufficient DNA quality for sequencing, leaving us with a total of 65 usable samples (62 patients, ~10% of the outbreak, Table S1). The GC outbreak was delineated following the diagram detailed in Figure 1.

First, all sequences were queried for the 3 SNP previously defined as GC outbreak markers (partial SNPs profile)[9]. We observed that 2 isolates (GC466, GC515), previously included in the outbreak based on molecular typing methods, did not actually harbor those marker SNPs, thus both were excluded from further analysis. In addition, we queried our collection of sequences and found 7 additional samples, 2 from Valencia Region and 5 from Liberia (Table S2). We then constructed a maximum likelihood (ML) tree with all the samples with the marker SNPs, and observed a differentiated and well-supported monophyletic clade (bootstrap = 99, Figure 2A) including the index case, most of the isolates previously assigned to outbreak and the Valencian cases (60 single TB cases, Table S1). The phylogeny also revealed that the SNPs initially considered as GC outbreak-specific, were not exclusive, since additional global strains also harbour them.

Pairwise distance was also estimated considering the phylogenetic outbreak definition based on WGS data (Table 1), in order to evaluate if distances were in agreement with the accepted recent transmission thresholds of 0-10 SNPs. Mean and median distance within-outbreak were 4.3 and 4 SNPs (range 0-17) respectively, while distance between outbreak and non-outbreak samples were 33 and 30 SNPs (range 18-63).

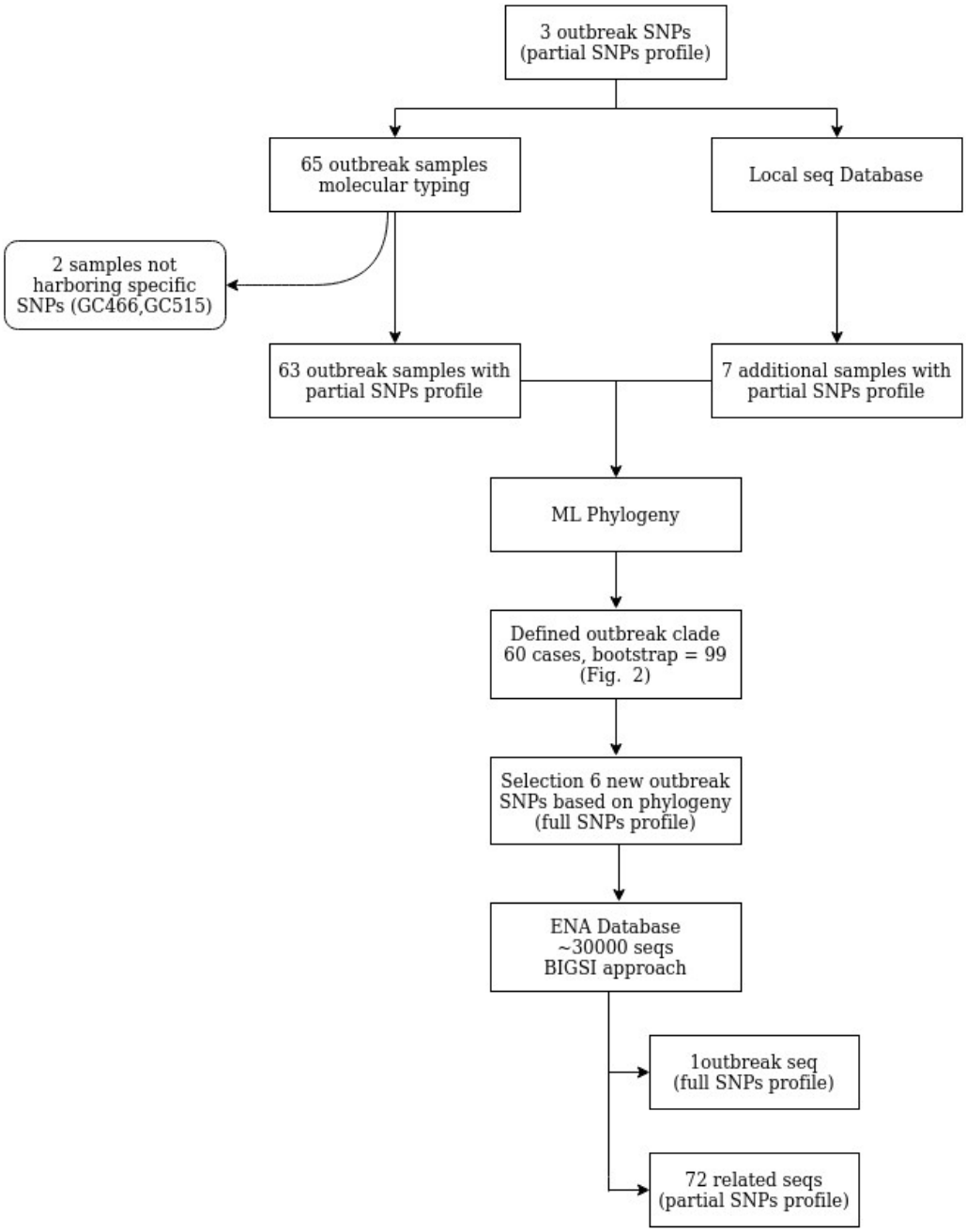

Figure 1. Workflow detailing the GC outbreak delineation procedure

	SNP pairwise distance					
Outbreak	**Min**	**Max**	**Q1**	**Median**	**Mean**	**Q2**
GC	0	17	2	4	4.3	6
BA	0	31	5	7	8.2	10
Bern	0	18	3	5	5.4	7
Denmark	0	33	6	8	9.9	13
Thailand	0	25	9	12	11.8	15

Table 1. Pairwise distance comparison among different outbreaks. Minimum (Min), maximum (Max), mean, median, 1st (Q1) and 2nd (Q2) quartile are provided

When compared to other known outbreak strains, the GC one displayed the lowest within-outbreak mean and median pairwise distance, though similar to the Bern outbreak (Table 1, Figure 2B-C). GC and Bern outbreaks exhibited a unimodal right-skewed pairwise genetic distance distribution, pointing out that most of the samples have low pairwise SNP distance among them. On the contrary, Denmark and Buenos Aires (BA) had a slightly bimodal distribution, suggesting the existence of more than one distinctive clade within both outbreaks, as proposed elsewhere[3]. Thailand outbreak displayed a unimodal normal distribution indicating a high pairwise SNP dispersion.

Thus, distance values obtained for GC also support the phylogenetic circumscription of the outbreak. With this delimitation, two additional samples from Madrid (AB1440701, 48956), belonging to recent Guinean immigrants, and previously identified as infected by the GC1237 strain [9], were excluded (Table S1).

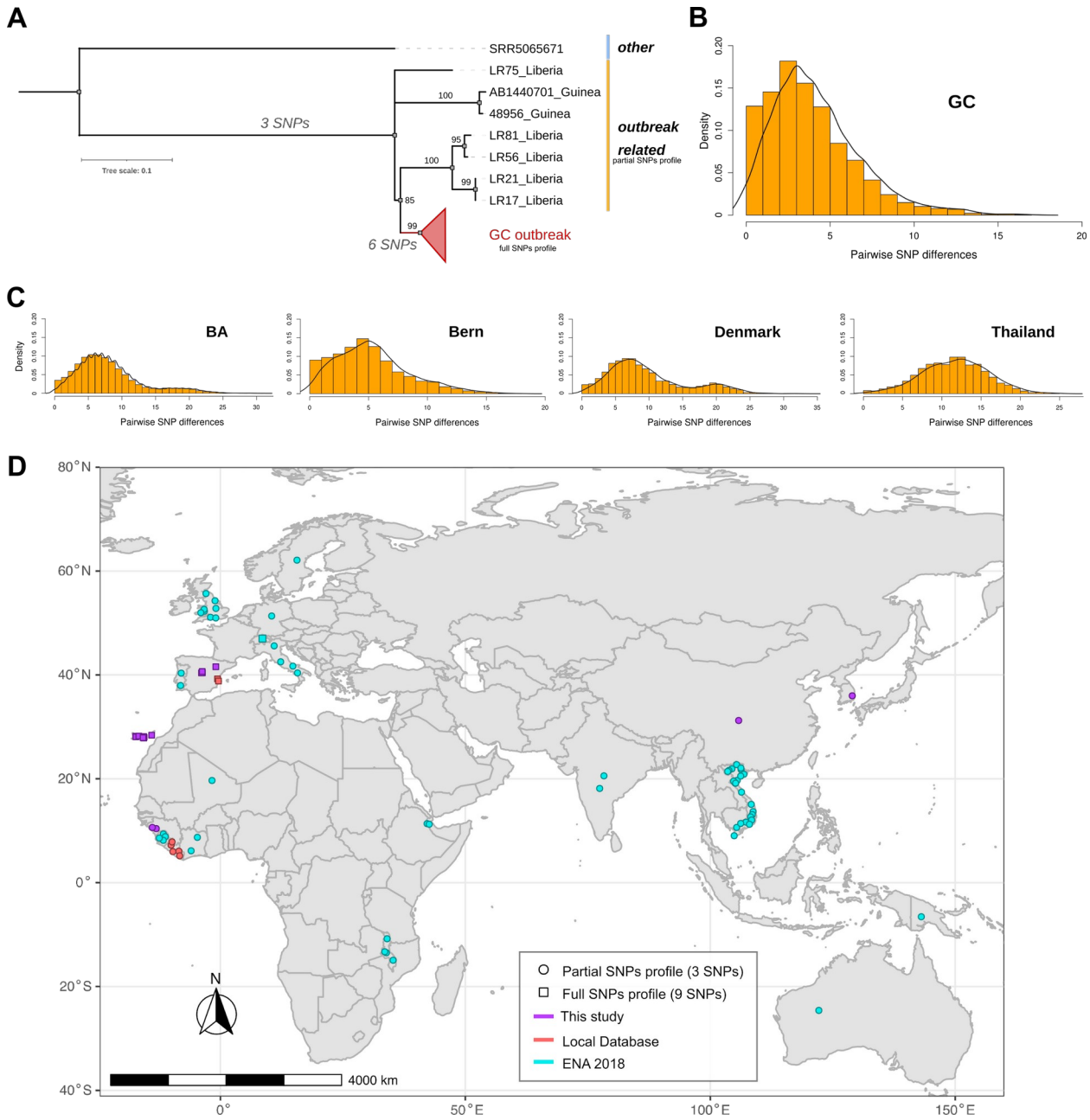

Figure 2. Genomic delineation of the outbreak. **A.** ML tree highlighting the phylogenetic circumscription of the outbreak and related strains identified with the partial SNPs profile. **B-C.** Density graphics of pairwise number of SNPs between samples of GC outbreak and Buenos Aires (BA)[4]; Bern[5]; Denmark [3]; Thailand [2]. **D.** Study of the success of GC outbreak outside the studied area, maps indicating the origin of the sequences identified by the different sources and their particular SNPs profiles; this study includes all samples sequenced as considered part of the outbreak by different molecular typing methods.

Evaluation of success outside the outbreak GC setting

In order to know if the GC strain can be found outside the Canary Islands, and if so if it has a similar epidemiological success, we designed an approach to evaluate all the MTB sequences available in public databases. This gives us the opportunity to study the spread of the outbreak, not only to the study area, but inspect its broader extension. First, the phylogenetic circumscription allows us to identify the six additional markers defining the outbreak clade, most of which were missense mutations located in genes involved in metabolic processes and respiration, with effects likely unrelated with virulence or transmission (Table S3, FIgure 2A). Thus, the full SNPs profile (9 SNPs) was used to query every *M. tuberculosis* sample deposited in ENA by July 2018, using the BIGSI index[12] (see methods). We identified 1 sequence meeting the full SNPs profile, from Switzerland, thus being part of the outbreak, and 72 harboring the partial SNPs profile, thus being related to the outbreak (Table S2, Figure 2D). We consider these sequences in further analysis, as they can shed light on the remote origin of the outbreak strain.

GC outbreak topology

To study the epidemiological characteristics of the GC strain in the Canary Island and abroad, we constructed a median joining network with all samples, including the three additional samples from the likely index case (Table S1, Figure 3). The genomic network displays a star-like structure. The likely index case is located in the center (including the oldest sample and two additional samples collected in 1994, 1995) along with 9 other isolates; most of them from 1993, and all collected from Gran Canaria. In depth investigation of central node (node A) revealed one heterogenous SNP (hSNP), at genomic position 3190007, for which the first sample of the likely index case (GC077) harbor both the reference (T: 82.7%) allele, shared by the rest of samples of the outbreak, and the alternative allele (C: 17.3%), which is fixed in sample GC092 from 1994 (Figure 3). The other hSNPs identified were uninformative since only one allele was fixed, either reference or alternative.

Connected with node A, three additional smaller star-like structures were observed,

corresponding to nodes B, C and E (Figure 3), and resembling secondary outbreaks. Node B represents the later spreading (around 2008) from Gran Canaria to the other islands in the archipelago; with an additional independent introduction in Tenerife. Node C, the smallest secondary outbreak, also occurred in Gran Canaria. Notably, the biggest secondary outbreak, node E, started early and includes samples from the peninsula (Madrid and Zaragoza). Furthermore, the network reveals that the index case initiated a new transmission chain in 1997 (isolate GC1441), also in Gran Canaria. Surprisingly, the two samples from Valencia were directly linked to the origin of the outbreak, without connection between them. Epidemiological data agree with the absence of link between both cases, but there is also no evidence of trip to Canary Islands for either patient, suggesting an additional missing case (or cases) linking them. For the three isolates collected in Madrid, there is some link with Gran Canaria; case Bjn8687 stayed there in prison two years before diagnosis and the others (PP126613, 5171825) also visited the island at some point before diagnosis (Table S1). For the isolate identified from Switzerland, no epidemiological information was obtained. Overall, the extent of secondary transmissions after exportation events seems very limited, since no additional sequences from other places were retrieved from ENA.

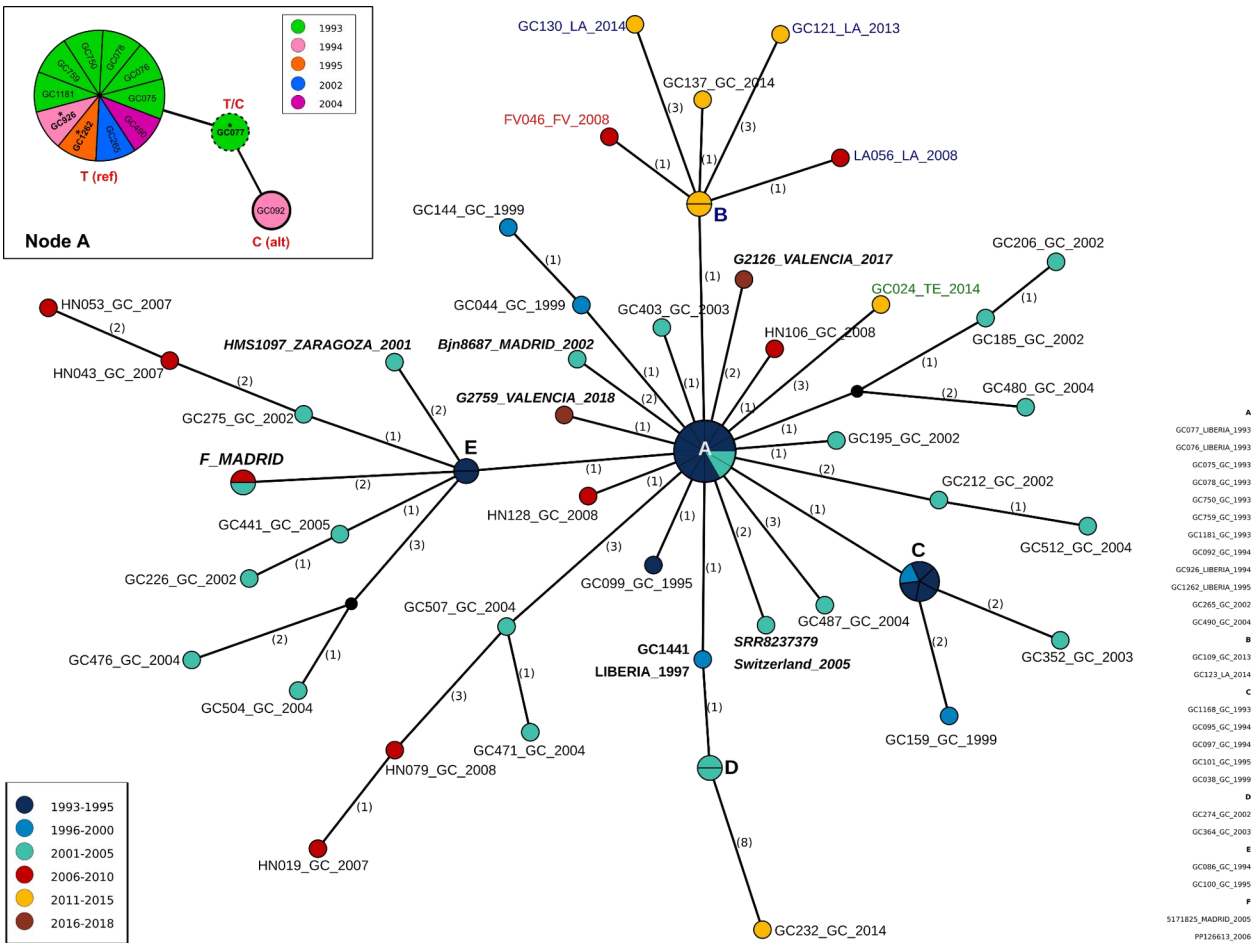

Figure 3. Median joining network analysis. Numbers in parentheses indicate the number of SNPs between nodes. Node size indicates the number of samples with the same genome, node color denotes sampling time; name color indicates different islands (blue: Lanzarote; red: Fuerteventura; green: Tenerife; black: Gran Canaria). Samples from the continent are indicated in italic and bold letters. GC1441 is an additional sample from the index case. Node A resolution with hSNPs pos: 3910007 is detailed; reference (ref) and alternative (alt) alleles distribution among samples is indicated. Names with asterisks indicate samples of the index case.

Phylodynamics analysis

We reasoned that if the GC strain had any transmission advantage this could be reflected in its natural history. Since we found evidence of temporal structure in our dataset (see Methods), it is suitable for applying a Birth-death skyline (BDSKY) model to estimate the epidemiological parameters of the GC outbreak (Supplementary Notes).

The becoming uninfectious or recovery rate (δ) resulted in a median value of 0.49 [0.28 - 0.75, 95% HPD] suggesting an infection period of 2 years [1.3 - 3.6 y] in agreement with the global estimation of 1-3 years previously proposed [13]. The effective reproduction number varies through the period of the outbreak, displaying a particular profile with peaks of expansion (R_e>1) and reduction (R_e<1). The peaks were observed approximately every ten years, and have an extension of three years (Figure 4A). Peaks matched with the secondary outbreak nodes observed in the network, and the periods of outbreak reduction coincide with chains instead of star-like transmission patterns (Figure 3). As observed in the histogram, peaks do not reflect the sampling effort, since periods 2005-2010 and 2010-2015 display a similar number of samples (Figure 4B). While in the first period isolates appeared in terminal nodes (Figure 3), in the second the cases are part of a secondary outbreak (Figure 3, node B). On average, the outbreak is decreasing, since the last period showed a R_e close to 1. Besides its particular profile, R_e never exceeds a value of 12 [6, 0.01 - 11 95% HPD], the maximum number of secondary cases caused by an infected person per year estimated for TB[13]. Overall, there is no indication from phylodynamics analysis that the GC strain has a transmission advantage, at least in terms of secondary cases generated.

On the contrary, if ecological drivers like host or social determinants were behind the initial success of the GC strain, it will be mostly expected among the vulnerable population. In this sense, from all 61 patients belonging to the GC outbreak, 64% presented at least one risk factor including; PDU (Parenteral Drug Users), NPDU (Non-Parenteral Drug Users), imprisonment, HIV infection, alcohol abuse, indigence and no TB treatment adherence (Table S1). All of them are known risk factors for TB infection in low burden countries. Furthermore, considering only patients with available information, the value increases to 76%.

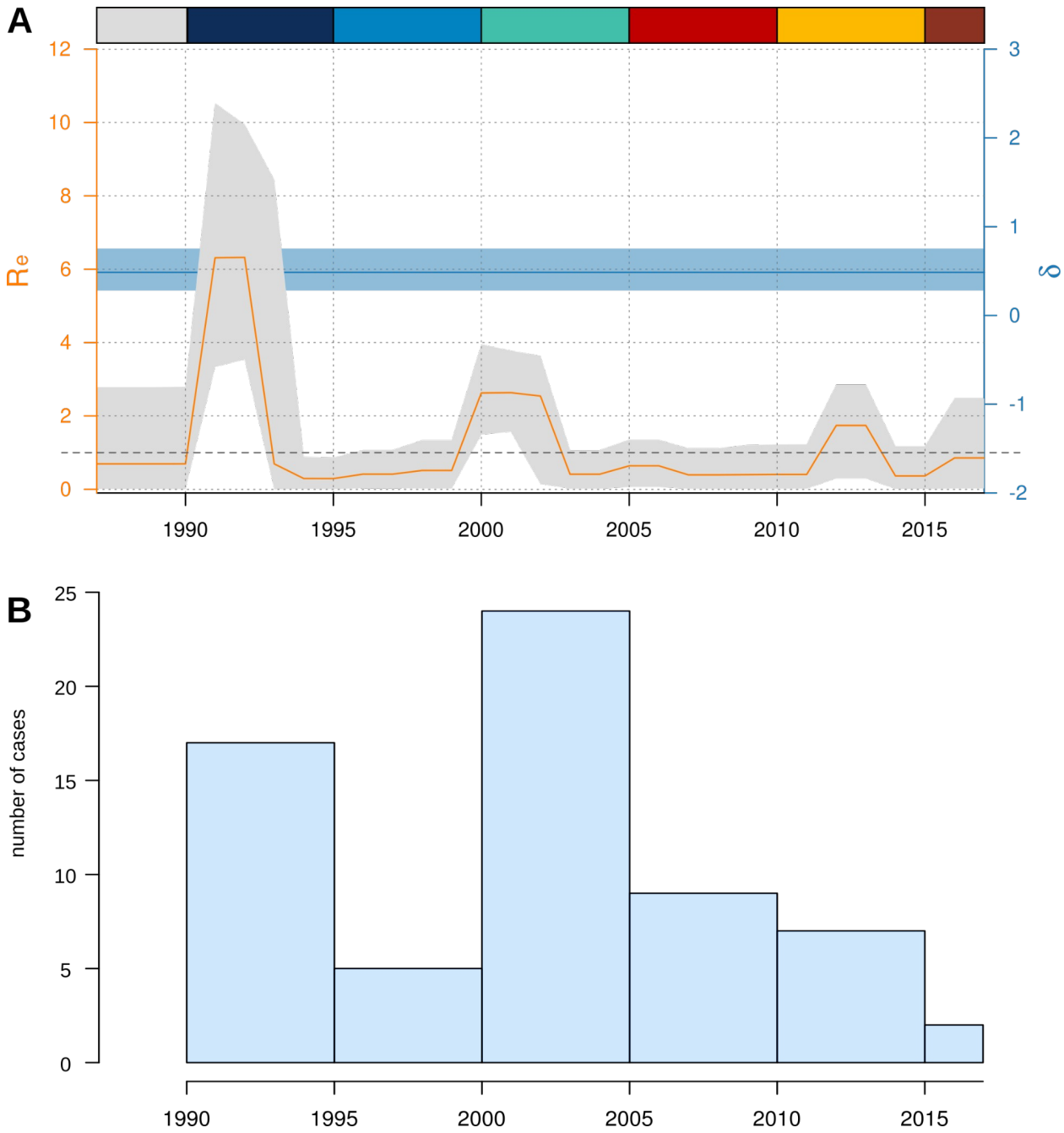

Figure 4. Phylodynamics analysis of GC outbreak. **A.** Birth-death serial skyline results showing reproductive number (R_e) variation across outbreak period and recovery rate (δ). Rectangles in the top indicate periods of time as colored in Figure 3. **B.** Histogram of number of cases sequenced in the different time periods.

Tracking the GC outbreak strain over centuries

The GC outbreak strain belongs to the MTBC lineage 2. A global ML phylogeny was constructed including 740 L2 global strains (MSA with 43401 concatenated SNPs) from different studies (Table S4). Placed in the context of L2 diversity and considering the classification previously proposed [14,15], the GC outbreak strain belongs to L2.2.3 (Figure 5A). The closest clade is mainly composed of African strains (Figure 5B, orange clade) followed by a mixed African-Vietnamese clade (Figure 5B, yellow clade) and more distantly related to China (Figure 5B). The emergence of the outbreak was estimated around 1984 AD [1973-1993 AD, 95% HPD] in agreement with the arrival time of the likely index case to the Archipelago (Figure S1). The MRCA of the African clade was estimated at 1930 AD [1913-1947 AD, 95% HPD] and the MRCA of the African-Vietnamese within the middle of 19th century (Figure 5B). Origin of L2.2 was estimated between 559-819 AD in agreement with previous estimates using the same approach [16]. Phylogeographic and dispersal reconstruction of the outbreak ancestors, placed its oldest ancestor in China around the middle 1300 AD; from there it spread to Vietnam between 1500-1700 AD. From Vietnam it dispersed to Africa between 1930-1940 AD, then to Canary Islands between 1960-1993 AD to later reach Europe between 1993-2002 (Figure 6, File S1).

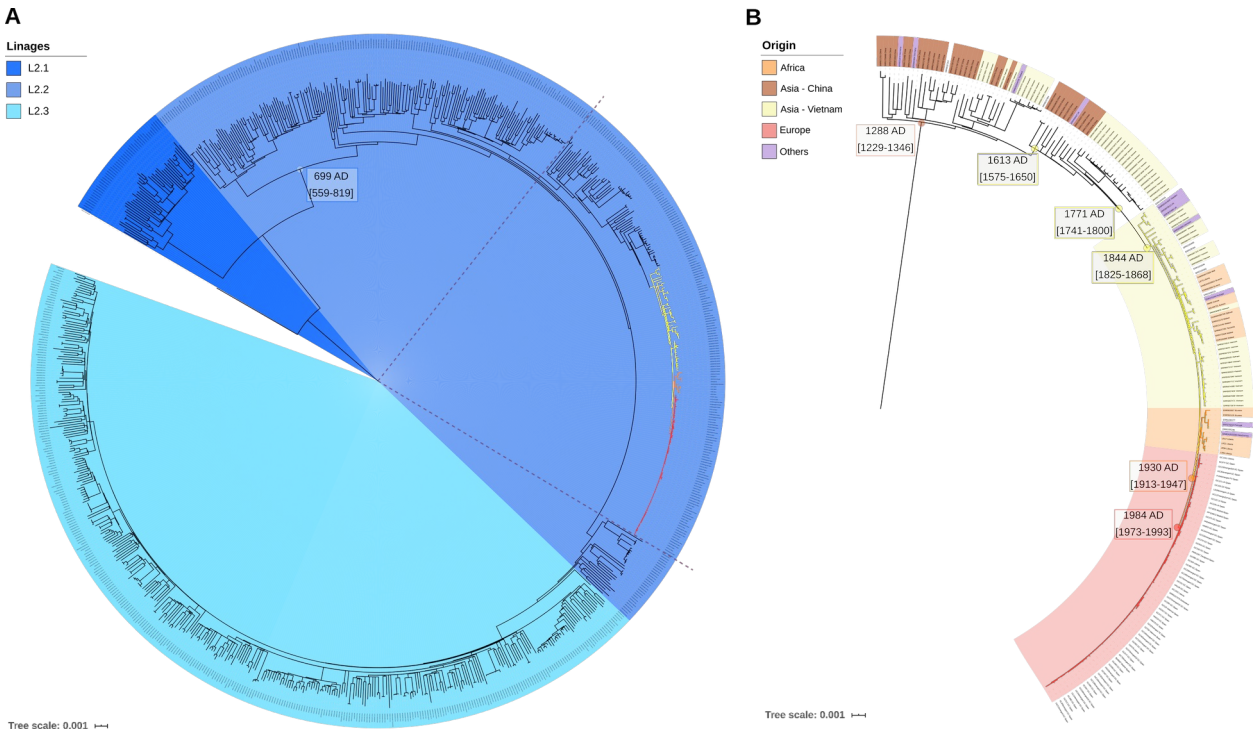

Figure 5. Dating analysis. **A.** Time of the most recent common ancestor (tMRCA) of L2.2. **B.** tMRCAs of GC outbreak (red) and closest nodes. Samples included in red clade (GC) are all from Europe. Colors indicate origin of samples (labels) and MRCA of clades. Branch colors are the same in panels A and B.

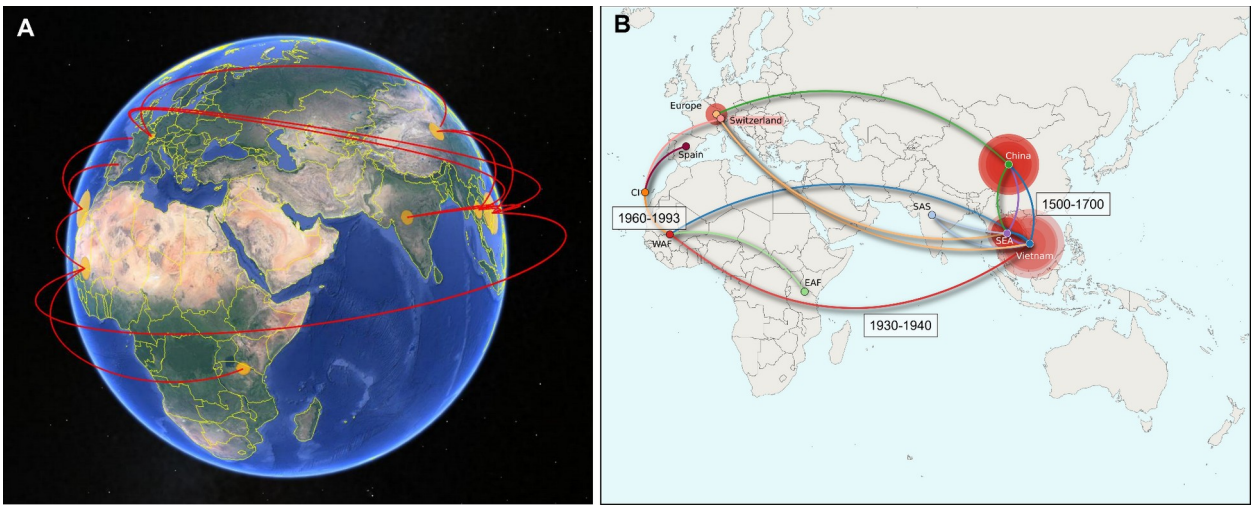

Figure 6. Phylogeographic and dispersal analysis. **A.** Results of the phylogeographic analysis in Google earth view. **B.** Routes of dispersion and time, lines are colored following the destination of the migration, and times from the oldest origin in China, to Vietnam and Canary Islands are denoted.

Discussion

Mycobacterium tuberculosis is an obligate human pathogen and its success in the population deeply depends on human movement. By combining WGS and phylodynamics analyses, here we describe the dynamics of a 25 years outbreak in detail. After the index case arrival to Gran Canaria, the strain spread quickly, since many cases -with identical sequences- were observed in 1993. Two close secondary outbreaks appeared soon, one connected with the samples in the archipelago, and the other extended in the same island, probably as an early independent transmission chain. Lately, a differentiated secondary outbreak spread to the rest of the islands, reaching high representativeness[11]. By evaluating serial samples from different years, we could identify secondary outbreaks associated with the likely index case, probably due to poor treatment adherence. The deep analysis of hSNPs resolved the relations among samples in the central core of the outbreak, adding molecular evidence to support GC077 as the index case (first sample). A similar approach has been described in Lee et al.[17] showing the potential of using hSNP for further genetic discrimination within outbreaks. This approach also provided evidence that most of the observed diversity within the outbreak was already present in the index case samples. Furthermore, this approach allowed us to circumscribe the outbreak, discarding distantly related isolates linked by molecular methods that only query a restricted portion of the genome.

With the added value of querying a large collection of WGS [12,18], our comprehensive approach allowed us to link the GC strain to cases in continental Europe and thus determine the extent of an outbreak beyond the geographic limits where it was studied. Thus, while the GC outbreak strain had a very high local success, reaching a frequency of 27% of all isolates in Gran Canaria[11], it had limited expansion outside the

archipelago; with multiple introductions in the continent not resulting in secondary cases[9].

By combining the finding of a wide SNP profile and by querying a large dataset, we were able to shed light on the remote origin of the outbreak strain. Our analysis places the ancestors of the GC strain hundreds of years ago, in China around 1300 AD. Thanks to the vast amount of MTBC genomic data available we have traced its initial movements from China to Vietnam, later to Liberia and finally to Gran Canaria by the end of the 20th century. Once in Gran Canaria the strain generated a large number of secondary cases there, but was rarely seen outside the archipelago. Thus, here we were able to document how past epidemiological events impact current TB epidemics.

In addition to inferring its remote origin, we had the unique opportunity to study its proximal origin using a phylodynamic approach. Proximal origins of large TB outbreaks are rarely identified by epidemiological information, mainly because of the difficulty in identifying the index patient in an area with many local cases and, in the case of tuberculosis, because of latency periods. Since the GC outbreak was undoubtedly linked to the migration of an infected person from Liberia to Gran Canaria in 1993 [8], this represents a unique opportunity to validate the commonly applied Bayesian approach for tuberculosis. Our results (1984 AD [1973-1993 AD, 95% HPD]) largely agreed with the available epidemiological information.

The main limitations of our study is the low proportion of outbreak cases sequenced from Canary Islands, it accounts for approximately 10% of total cases, however we show that the dataset has enough phylogenetic signal for the analysis presented. In accordance, our phylodynamics results do not correlate to sampling effort (Figure 4). Spain also lacks a systemic WGS program, however in our study we use systematic typing data from three of the largest regions in Spain: Madrid, Aragón and Valencia Region to confirm that the strain had no success beyond the archipelago. Finally, by querying the whole ENA database we could retrieve additional cases sequenced elsewhere and could evaluate the expansion of the outbreak.

A common theme when it comes to MTBC strains is whether some are more transmissible than others or if its success is driven by ecological factors, or both.

Outbreak strains are good candidates to look for intrinsic transmissibility as they are responsible for a large number of cases with sustained transmission over decades. For the GC strain, we estimate an infection period close to two years and less than 12 secondary cases per infected individual, same values proposed for TB in general [13], suggesting that there is no transmissibility advantage of the GC strain, at least in terms of shorter latency periods or higher contagion rate. In addition, it was demonstrated that GC strain did not display higher virulence, nor accumulated mutations in sequential samples or acquired resistance mutations[10], as proposed for other strains belonging to Beijing lineage. Similarly, our extensive query of outbreak sequences in public repositories did not identify that this strain is responsible for outbreaks outside the Canary Islands. For example, an in-depth genotyping effort did not identify secondary cases associated with a case in Madrid that had prolonged disease [10] or in Aragón where systematic genotyping is applied to all TB positive cases. In addition, mutations associated with the GC strain are unlikely to have a role in transmission. For other outbreak strains, notably the Danish C2[19] and the Toronto[20] strains, mutations with a functional role have been proposed but no experimental evidence is available. On the contrary, all those outbreak strains share the fact that they thrive in populations with significant risk factors, particularly during the ‘80s and ‘90s of the past century. Overall, in accordance with Cáceres et al[21] data suggests that the success of the long-lived GC outbreak strain, and probably others, is related to ecological factors associated with founder effects linked to the host and social determinants of the disease rather than an intrinsic transmissibility difference.

Materials and method

Study population

M. tuberculosis DNA samples from 80 different patients, 3 additional ones from the index case, all from Canary Islands, and 1 from Zaragoza were supplied by the Instituto de Investigación Sanitaria de Aragón. In addition, DNA samples from 5 different patients from the peninsula (Madrid) were supplied by the Hospital Universitario Gregorio Marañón, 2 of which belonged to recent immigrants from Guinea (Table S1). A

representative number of samples from the Canary Islands, ranging the defined period of the outbreak (1993-2014), was selected. All samples were previously identified as part of the Gran Canaria (GC) outbreak by genotyping methods.

Epidemiological information was obtained from Servicio de Microbiología (Hospital Universitario de Gran Canaria Dr. Negrín) under the Ethics Committee Application CEIm H.U.G.C. Dr. Negrín 2019-502-1.

Study design

The whole genome of *M. tuberculosis* from the first set of 89 samples (86 patients) was sequenced. Additional samples were looked for by querying a large database, containing more than 6000 sequences [18] and some from Liberia [22], with the three GC outbreak specific SNPs [9] previously selected; Rv2524 (C1398T, SNP2847935), Rv3869 (G1347C, SNP4346385) and Rv0926c (G162A, SNP1033625). Redefinition of SNPs specific primers was conducted based on phylogeny and distance analysis. New samples were identified by inspecting the whole ENA repository and samples from ongoing projects. Phylogeographic and phylodynamics analysis were performed in order to evaluate the outbreak. Comparison with other outbreaks was performed; sequences corresponding to Denmark (PRJEB20214)[3]; Thailand (PRJN244659)[2]; Bern (PRJEB5925)[5]; and Buenos Aires (BA, PRJEB7669)[4] outbreaks were downloaded from ENA and analyzed with the bioinformatics pipeline detailed below.

Whole-Genome Sequencing and bioinformatics analysis

DNA samples were used to prepare sequencing libraries with a Nextera XT DNA library preparation kit (Illumina), following the manufacturer's instructions. Sequencing was performed on an Illumina MiSeq instrument, applying a 2 x 300bp paired-end chemistry. General bioinformatics analysis is described in https://gitlab.com/tbgenomicsunit/ThePipeline/-/tree/master/. Briefly, read files were trimmed and filtered with fastp [23]. Kraken software was used to remove nonMTBC (non-Mycobacterium tuberculosis complex) reads [24]. Sequences were then mapped to an inferred MTBC common ancestor genome (https://doi.org/10.5281/zenodo.3497110)

using bwa [25]. SNPs were called with SAMtools [26] and VarScan2 [27]. GATK HaplotypeCaller [28] was used for InDels calling. SNPs with a minimum of 20 reads (20X) in both strands and quality 20 were selected. InDels with less than 20X were discarded. SnpEff was used for SNP annotation using the H37Rv annotation reference (AL123456.2). Finally, SNPs falling in genes annotated as PE/PPE/PGRS, 'maturase', 'phage', '13E12 repeat family protein'; those located in insertion sequences; those within InDels or in higher density regions (>3 SNPs in 10 bp) were removed due to the uncertainty of mapping. Heterogenous SNPs (hSNPs) were obtained from filtered SNP files, they were classified as positions were > 5% and < 90% of the reads were the alternative allele [17]; we looked for hSNPs only for positions for which at least one sample harbor the SNP fixed, i.e were > 90% of reads were the alternative allele. Lineages were determined by comparing called SNPs with specific phylogenetic positions established [29,30]. An in-house R script was used to detect mixed infections based on the frequency of lineage- and sublineage-specific positions [22]. Pairwise distances were computed with the R *ape* package, statistical analysis and graphics were also performed with R. Raw sequencing data are available under the accession number PRJEB50491 (European Nucleotide Archive)

All WGS Illumina sequencing runs stored in the ENA with metadata identifying them as *M. tuberculosis* complex as of July 2018 were downloaded (N=38075). A de Bruijn graph (k=31) was built from each with Cortex v1.0.5.21 (https://github.com/iqbal-lab/cortex/releases/tag/v1.0.5.21), and sequencing errors were removed by excluding low coverage unitigs from the graph (threshold is sample dependent and automatically chosen by the software). The remaining kmers from these graphs were then used to build a BItsliced Genomic Signature Index (BIGSI)[12]. The index was used to query these ENA samples for the outbreak-related SNPs, by first creating two 61 base-pair probes for each SNP (30bp flanking each side of the SNP, from the reference genome, and one probe for each SNP allele), returning binary information as to which sequencing runs contained all the kmers in the probes.

Phylogenetic analysis

Multisequence alignment (MSA) files were constructed with concatenated SNPs

discarding well-known drug resistance and invariant positions. Maximum likelihood trees were inferred with RAxML v8.2.11 [31] using GTRCATI model and 1000 fast-bootstrap replicates. Tree visualization and editing were conducted in ITOL (https://itol.embl.de/). Specific SNPs were identified using likelihood ancestral reconstruction of Mesquite software v3.61 (http://www.mesquiteproject.org). Genomic network was constructed with the MSA files with a median joining network inference method implemented in PopArt Software [32].

Time and geographic origin of the outbreak

A set of 200 samples including the outbreak, closest clades and a representative subset of global samples selected with Treemer [33] was used to estimate the time of the most common ancestor (tMRCA) of the outbreak and deeper nodes. The evolutionary history of the outbreak was reconstructed with BEAST2 v2.5.1 [34] using GTR (gamma 4) as site model, coalescent constant population as tree prior and relaxed clock log-normal with gamma distribution as prior. We use a tip dating method and also set the clock rate (mean = 4.6×10^{-8} substitutions per site per year with 95% HPD: 3.3×10^{-8} to 6.2×10^{-8}) based on previous publications [35]. MCMC's chain length of 10M with sampling every 1000 steps for the posterior distribution. Three independent runs were performed to reach convergence. Log and tree files were combined with LogCombiner tool discarding 10% of burn-in. Combined files were inspected with Tracer v1.7.1 [36], all parameters reached effective sample size (ESS) > 200 and well mixing. Tree was annotated with TreeAnnotator and visualized with FigTree v1.4.3.

Spatial dispersal dynamics of the outbreak and closest clades was reconstructed using a phylogeographic diffusion in discrete space approach implemented in BEAST v2.5.1 (Beast-classic package, [37]. A set of 142 samples, including the whole outbreak and closest clades, and those global strains with available information of year and location were used. Same clock and tree model, as well as priors, as in dating analysis were used. Same MCMC chains and further analysis as dating were conducted. In addition, SPREAD3 tool was used for geographic visualization of outbreak evolutionary history [38], only dispersion patterns with posterior values higher than 0.9 were plotted.

Phylodynamics of the outbreak

The dynamics of the outbreak was studied with the bayesian Birth-Death Skyline model implemented in BEAST2 v2.5.1. This model allows estimating parameters related to the evolution of the outbreak [39]. The analysis was conducted with all the samples included in the outbreak (64 samples, Table S1). We use a GTR (gamma 4) as site model, strict clock with LogNormal prior with the mean previously published [35] as prior. A Birth death skyline serial tree model was used. Specific parameters were set as follow (i) *becoming uninfectious* (LogNormal; M=0, S=1.25), considering the global estimation of 1 year as TB infectiousness period and not taking into account latency [13] (ii) *origin* (LogNormal; M= 30; S=0.04) it was taking into account that the likely index case has arrived to the archipelago in the beginning of 1993 starting immediately the outbreak; (ii) *reproductive number* (LogNormal; M=0; S= 1.25; dimension = 13) it was considered that in settings with low TB burden each person can generate a maximum of 12 secondary cases [13]; (iv) *sampling proportion* (beta; alpha = 3; beta = 50) it was calculated as the total number of TB cases in Canary Islands between between 1993-2017 (https://www3.gobiernodecanarias.org/sanidad/scs/listaImagenes.jsp?idDocument=fe9e3e94-feee-11e0-ab85-376c664a882a&idCarpeta=0f67aaf7-9d88-11e0-b0dc-e55e53ccc42c) and considering that 20-30% of cases belong to the outbreak resulting in a sampling range of 8-13%. MCMC's chain length of 10M was run with sampling every 1000 steps for the posterior distribution. Log file was inspected with Tracer v1.7.1 [36], all parameters reached ESS>200 and well mixing. Results were inspected and plotted with R package bdskytools (https://github.com/laduplessis/bdskytools).

Clocklike structure of the outbreak dataset was first evaluated with a linear regression analysis of tip dates vs root-to-tip distances with TempEst [40] and by applying a date randomization test (DRT) with 100 randomized replicates of the BDSKY. The clocklike structure was evaluated with the analysis proposed by [41] considering that DRT is passed when the clock rate estimate for the observed data does not overlap with the range of estimates obtained from the randomized sets; intermediate DRT is passed

when the clock rate estimate for the observed data does not overlap with the confidence intervals of the estimates obtained from the randomized sets, and stringent DRT is passed when the confidence interval of the clock rate estimate for the observed data does not overlap with the confidence intervals of the estimates obtained from the randomized sets.

Data and code availability

All sequences have been deposited on the ENA repository under the project number PRJEB50491. All the scripts, tools and reference sequences are included in the repository section of the laboratory webpage (http://tgu.ibv.csic.es/?page_id=1794) and the gitlab page (https://gitlab.com/tbgenomicsunit/ThePipeline). Additional sequences from other studies are listed in Tables S2 and S4. Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

Acknowledgements

Instituto de Salud Carlos III (FIS18/0336), Gobierno de Aragón/Fondo Social Europeo “Construyendo Europa desde Aragón" to SS. This project has been funded by the European Research Council (101001038-TB-RECONNECT), the Ministerio de Economía, Industria y Competividad (PID2019-104477RB-I00) to IC.

Author contribution. M.G.L and I.C conceived the study; M.I.C.H, F.C, R.C, L.S, B.P, L.P.L, D.G.V, S.S, J.C., E.L. collected the samples, obtained and curated the epidemiological data; M.T.P. processed and sequenced the samples; P.W., Z.I performed BIGSI analysis; M.M.M. performed the scripts for DST analysis; M.G.L performed all analyses; M.G.L and I.C wrote the first draft of the manuscript, with revisions from all co-authors.

Declaration of interests

IC received consultancy fees from Foundation for Innovative New Diagnostics. The author has no other competing interests to declare. The remaining authors declare no competing interests.

References

1. Global tuberculosis report 2020 https://www.who.int/publications/i/item/9789240013131.

2. Coscolla, M., Barry, P.M., Oeltmann, J.E., Koshinsky, H., Shaw, T., Cilnis, M., Posey, J., Rose, J., Weber, T., Fofanov, V.Y., et al. (2015). Genomic epidemiology of multidrug-resistant Mycobacterium tuberculosis during transcontinental spread. J. Infect. Dis. *212*, 302–310.

3. Folkvardsen, D.B., Norman, A., Andersen, Å.B., Michael Rasmussen, E., Jelsbak, L., and Lillebaek, T. (2017). Genomic Epidemiology of a Major Mycobacterium tuberculosis Outbreak: Retrospective Cohort Study in a Low-Incidence Setting Using Sparse Time-Series Sampling. J. Infect. Dis. *216*, 366–374.

4. Eldholm, V., Monteserin, J., Rieux, A., Lopez, B., Sobkowiak, B., Ritacco, V., and Balloux, F. (2015). Four decades of transmission of a multidrug-resistant Mycobacterium tuberculosis outbreak strain. Nat. Commun. *6*, 7119.

5. Stucki, D., Ballif, M., Bodmer, T., Coscolla, M., Maurer, A.-M., Droz, S., Butz, C., Borrell, S., Längle, C., Feldmann, J., et al. (2015). Tracking a tuberculosis outbreak over 21 years: strain-specific single-nucleotide polymorphism typing combined with targeted whole-genome sequencing. J. Infect. Dis. *211*, 1306–1316.

6. Casali, N., Broda, A., Harris, S.R., Parkhill, J., Brown, T., and Drobniewski, F. (2016). Whole Genome Sequence Analysis of a Large Isoniazid-Resistant Tuberculosis Outbreak in London: A Retrospective Observational Study. PLoS Med. *13*, e1002137.

7. Roetzer, A., Diel, R., Kohl, T.A., Rückert, C., Nübel, U., Blom, J., Wirth, T., Jaenicke, S., Schuback, S., Rüsch-Gerdes, S., et al. (2013). Whole genome sequencing versus traditional genotyping for investigation of a Mycobacterium tuberculosis outbreak: a longitudinal molecular epidemiological study. PLoS Med. *10*, e1001387.

8. Caminero, J.A., Pena, M.J., Campos-Herrero, M.I., Rodríguez, J.C., García, I., Cabrera, P., Lafoz, C., Samper, S., Takiff, H., Afonso, O., et al. (2001). Epidemiological evidence of the spread of a Mycobacterium tuberculosis strain of

the Beijing genotype on Gran Canaria Island. Am. J. Respir. Crit. Care Med. *164*, 1165–1170.

9. Pérez-Lago, L., Herranz, M., Comas, I., Ruiz-Serrano, M.J., Roa, P.L., Bouza, E., and García-de-Viedma, D. (2016). Ultrafast Assessment of the Presence of a High-Risk Mycobacterium tuberculosis Strain in a Population. J. Clin. Microbiol. *54*, 779–781.

10. Pérez-Lago, L., Navarro, Y., Montilla, P., Comas, I., Herranz, M., Rodríguez-Gallego, C., Serrano, M.J.R., Bouza, E., and de Viedma, D.G. (2015). Persistent Infection by a Mycobacterium tuberculosis Strain That Was Theorized To Have Advantageous Properties, as It Was Responsible for a Massive Outbreak. J. Clin. Microbiol. *53*, 3423–3429.

11. Pérez-Lago, L., Campos-Herrero, M.I., Cañas, F., Copado, R., Sante, L., Pino, B., Lecuona, M., Gil, Ó.D., Martín, C., Muñoz, P., et al. (2019). A Mycobacterium tuberculosis Beijing strain persists at high rates and extends its geographic boundaries 20 years after importation. Sci. Rep. *9*, 1–6.

12. Bradley, P., den Bakker, H.C., Rocha, E.P.C., McVean, G., and Iqbal, Z. (2019). Ultrafast search of all deposited bacterial and viral genomic data. Nat. Biotechnol. *37*, 152–159.

13. Dowdy, D.W., Dye, C., and Cohen, T. (2013). Data needs for evidence-based decisions: a tuberculosis modeler's "wish list" [Review article].

14. Shitikov, E., Kolchenko, S., Mokrousov, I., Bespyatykh, J., Ischenko, D., Ilina, E., and Govorun, V. (2017). Evolutionary pathway analysis and unified classification of East Asian lineage of Mycobacterium tuberculosis. Sci. Rep. *7*, 1–10.

15. Rutaihwa, L.K., Menardo, F., Stucki, D., Gygli, S.M., Ley, S.D., Malla, B., Feldmann, J., Borrell, S., Beisel, C., Middelkoop, K., et al. (2019). Multiple Introductions of Mycobacterium tuberculosis Lineage 2–Beijing Into Africa Over Centuries. https://www.frontiersin.org/article/10.3389/fevo.2019.00112/full.

16. Liu, Q., Ma, A., Wei, L., Pang, Y., Wu, B., Luo, T., Zhou, Y., Zheng, H.-X., Jiang, Q., Gan, M., et al. (2018). China's tuberculosis epidemic stems from historical expansion of four strains of Mycobacterium tuberculosis. Nature Ecology & Evolution *2*, 1982–1992.

17. Lee, R.S., Proulx, J.-F., McIntosh, F., Behr, M.A., and Hanage, W.P. (2020). Previously undetected super-spreading of revealed by deep sequencing. Elife *9*.

18. Chiner-Oms, Á., Sánchez-Busó, L., Corander, J., Gagneux, S., Harris, S.R., Young, D., González-Candelas, F., and Comas, I. (2019). Genomic determinants

of speciation and spread of the complex. Sci Adv *5*, eaaw3307.

19. Folkvardsen, D.B., Norman, A., Andersen, Å.B., Rasmussen, E.M., Lillebaek, T., and Jelsbak, L. (2018). A Major Mycobacterium tuberculosis outbreak caused by one specific genotype in a low-incidence country: Exploring gene profile virulence explanations. Sci. Rep. *8*, 1–8.

20. Mehaffy, C., Guthrie, J.L., Alexander, D.C., Stuart, R., Rea, E., and Jamieson, F.B. (2014). Marked microevolution of a unique Mycobacterium tuberculosis strain in 17 years of ongoing transmission in a high risk population. PLoS One *9*, e112928.

21. Caceres, N., Llopis, I., Marzo, E., Prats, C., Vilaplana, C., de Viedma, D.G., Samper, S., Lopez, D., and Cardona, P.-J. (2012). Low Dose Aerosol Fitness at the Innate Phase of Murine Infection Better Predicts Virulence amongst Clinical Strains of Mycobacterium tuberculosis. PLoS One *7*, e29010.

22. López, M.G., Dogba, J.B., Torres-Puente, M., Goig, G.A., Moreno-Molina, M., Villamayor, L.M., Cadmus, S., and Comas, I. (2020). Tuberculosis in Liberia: high multidrug-resistance burden, transmission and diversity modelled by multiple importation events. Microb Genom *6*.

23. Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics *34*, i884–i890.

24. Wood, D.E., and Salzberg, S.L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biology *15*, R46.

25. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

26. Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics *27*, 2987–2993.

27. Koboldt, D.C., Zhang, Q., Larson, D.E., Shen, D., McLellan, M.D., Lin, L., Miller, C.A., Mardis, E.R., Ding, L., and Wilson, R.K. (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. *22*, 568–576.

28. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303.

29. Coll, F., McNerney, R., Guerra-Assunção, J.A., Glynn, J.R., Perdigão, J., Viveiros,

M., Portugal, I., Pain, A., Martin, N., and Clark, T.G. (2014). A robust SNP barcode for typing Mycobacterium tuberculosis complex strains. Nat. Commun. *5*, 4812.

30. Stucki, D., Brites, D., Jeljeli, L., Coscolla, M., Liu, Q., Trauner, A., Fenner, L., Rutaihwa, L., Borrell, S., Luo, T., et al. (2016). Mycobacterium tuberculosis lineage 4 comprises globally distributed and geographically restricted sublineages. Nat. Genet. *48*, 1535–1543.

31. Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics *30*, 1312–1313.

32. Leigh, J.W., and Bryant, D. (2015). popart : full-feature software for haplotype network construction. Methods in Ecology and Evolution *6*, 1110–1116.

33. Menardo, F., Loiseau, C., Brites, D., Coscolla, M., Gygli, S.M., Rutaihwa, L.K., Trauner, A., Beisel, C., Borrell, S., and Gagneux, S. (2018). Treemmer: a tool to reduce large phylogenetic datasets with minimal loss of diversity. BMC Bioinformatics *19*.

34. Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M.A., Rambaut, A., and Drummond, A.J. (2014). BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. PLoS Computational Biology *10*, e1003537.

35. Bos, K.I., Harkins, K.M., Herbig, A., Coscolla, M., Weber, N., Comas, I., Forrest, S.A., Bryant, J.M., Harris, S.R., Schuenemann, V.J., et al. (2014). Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. Nature *514*, 494–497.

36. Rambaut, A., Drummond, A.J., Xie, D., Baele, G., and Suchard, M.A. (2018). Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. Syst. Biol. *67*, 901–904.

37. Lemey, P., Rambaut, A., Drummond, A.J., and Suchard, M.A. (2009). Bayesian Phylogeography Finds Its Roots. PLoS Comput. Biol. *5*, e1000520.

38. Bielejec, F., Baele, G., Vrancken, B., Suchard, M.A., Rambaut, A., and Lemey, P. (2016). SpreaD3: Interactive Visualization of Spatiotemporal History and Trait Evolutionary Processes. Mol. Biol. Evol. *33*.

39. Stadler, T., Kühnert, D., Bonhoeffer, S., and Drummond, A.J. (2013). Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). Proc. Natl. Acad. Sci. U. S. A. *110*, 228–233.

40. Rambaut, A., Lam, T.T., Max Carvalho, L., and Pybus, O.G. (2016). Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-

Gen). Virus Evol *2*, vew007.

41. Menardo, F., Duchêne, S., Brites, D., and Gagneux, S. (2019). The molecular clock of Mycobacterium tuberculosis. PLoS Pathog. *15*, e1008067.