Dispersal of *Mycobacterium tuberculosis* to indigenous populations driven by historical European trade in the South Pacific

Claire V. Mulholland[1,2], Abigail C. Shockey[3], Htin L. Aung[2,4], Ray T. Cursons[1], Ronan F. O'Toole[5,6], Sanjay S. Gautam[5], Daniela Brites[7,8], Sebastien Gagneux[7,8], Sally A. Roberts[9], Noel Karalus[10], Gregory M. Cook[2,4], Caitlin S. Pepperell[3,11]*, and Vickery L. Arcus[1,2]*

[1] School of Science, University of Waikato, Hamilton, New Zealand

[2] Maurice Wilkins Centre for Molecular Biodiscovery, University of Auckland, Auckland, New Zealand

[3] Department of Medical Microbiology and Immunology, School of Medicine and Public Health, University of Wisconsin-Madison, Madison, USA

[4] Department of Microbiology and Immunology, School of Biomedical Sciences, University of Otago, Dunedin, New Zealand

[5] School of Medicine, University of Tasmania, Hobart, Australia

[6] School of Molecular Science, La Trobe University, Victoria, Australia

[7] Swiss Tropical and Public Health Institute, Basel, Switzerland

[8] University of Basel, Basel, Switzerland

[9] LabPLUS, Auckland City Hospital, Auckland, New Zealand

[10] Waikato Hospital, Hamilton, New Zealand

[11] Department of Medicine, Division of Infectious Diseases, School of Medicine and Public Health, University of Wisconsin-Madison, Madison, USA

* Corresponding Authors

Email: vic.arcus@waikato.ac.nz (VLA), cspepper@medicine.wisc.edu (CSP)

Abstract

The *Mycobacterium tuberculosis* complex lineage 4 (L4), also known as the "Euro-American" lineage, is the most widely dispersed of the seven human adapted lineages. L4 is comprised of ten sublineages including L4.4, which has a moderate global distribution and is the most common L4 sublineage in New Zealand. We have used a phylodynamics approach and a dataset of 236 global *M. tuberculosis* genomes to trace the origins and dispersal of L4.4 strains in New Zealand that are predominantly found in Māori and Pacific people. We identify an L4.4.1.1 sublineage clade of European origin, likely French, that is prevalent in indigenous populations in both New Zealand and Canada. Molecular dating suggests that expansion of European trade networks in the early 19th century led to dispersal of this clade to the South Pacific. We also identify historical and social factors within the region that have contributed to the local spread and expansion of these strains, including recent Pacific migrations to New Zealand and the rapid urbanization of Māori in the 20th century. Our results offer new insight into the dispersal of *M. tuberculosis* in the South Pacific region and provide a striking example of the role of historical European migrations in the dispersal of *M. tuberculosis*.

Author Summary

Tuberculosis kills more people worldwide than any other infectious disease and indigenous populations are disproportionately affected by the disease. Here, we have used a large global dataset of *Mycobacterium tuberculosis* bacterial genomes to trace the historical origins of tuberculosis strains in New Zealand that are most frequently found in Māori and Pacific people. These strains are locally known as the 'Rangipo' and 'Otara' strains (both Māori place names) and belong to the "Euro-American" lineage of *M. tuberculosis*. Via genome analysis, we find that these strains are closely related to *M. tuberculosis* strains found in indigenous populations in Canada that have a European origin. We used a molecular dating approach (a molecular clock) to infer the ages of these strains and date divergence events. The timing we infer corresponds to the introduction of these strains

to Polynesia via expanding European trade networks in the South Pacific in the early 19th century and suggests that the Otara strain has migrated to New Zealand from the Pacific Islands multiple times. Our results provide insight into human social phenomena underlying the expansion and dispersal of *M. tuberculosis* and reassert the important role of European colonial migrations in the global dispersal of the *M. tuberculosis* Euro-American lineage. This work also highlights the pejorative and stigmatizing mislabelling of the New Zealand strains with indigenous Māori place names, suggesting that these strains should be renamed.

Introduction

Tuberculosis (TB) is caused by the bacterial pathogen *Mycobacterium tuberculosis* (*Mtb*) and other members of the *Mtb* complex (MTBC). TB kills more people globally than any other infectious disease. There are however considerable regional variations in TB incidence rates, and indigenous people are generally found to have higher rates of disease than non-indigenous people [1]. 'Proximate determinants' of TB such as smoking and food insecurity are generally more prevalent in indigenous people compared to non-indigenous people, and may be substantial contributors to the high burden of TB in these communities [2]. Understanding how the pathogen was dispersed and is maintained among indigenous populations is important for designing improved strategies for TB control in these often disproportionally affected populations.

The MTBC comprises seven human adapted lineages, which show strong phylogeographic structure and vary in the extent of their global distribution [3-5]. The most widely globally dispersed MTBC lineage is lineage 4 (L4), also known as the "Euro-American" lineage [6, 7]. Ten sublineages of L4 have been described, which vary in their global distributions from broad to highly localized [8]. Spatial and temporal patterns of L4 dispersal suggest that it was spread through European colonial migrations to Africa and the Americas [3, 4, 6, 7, 9, 10].

Oceania followed the Americas as the last major region to be reached by Europeans. The Oceanic region includes the > 1000 islands of Polynesia scattered across the central and southern Pacific Ocean. Little is known about the origins and dispersal of *Mtb* in this region. It is commonly assumed that TB was introduced to Polynesia with the arrival of European sailors and settlers, however TB-like lesions in skeletons predating European arrival challenge this view [11]. New Zealand is the largest country in Polynesia and is home to the local indigenous Māori people and the largest diaspora of communities of Polynesian people globally [12], providing a unique setting for the investigation of *Mtb* dispersal and transmission in indigenous Polynesian populations. Present-day *Mtb* genotypes in New Zealand Europeans, Māori and Pacific people in New Zealand are dominated by L4 strains [13], consistent with introduction of modern-day strains by Europeans.

Collectively, Māori and Pacific ethnicities account for ~70% of New Zealand born TB cases, and experience much higher notification rates than New Zealand Europeans (average notification rates for 2011–2015: Māori 4.7/100,000; Pacific peoples 16.2/100,000; Europeans 0.8/100,000) [14]. Molecular typing of New Zealand *Mtb* isolates by MIRU-VNTR show Māori and Pacific people also have a high proportion of cases with shared molecular types, i.e. are "clustered" (74.1% and 80.1% of isolates, respectively) [14]. The largest cluster identified by related 24-loci MIRU-VNTR typing patterns is known as the 'Rangipo' cluster. This strain predominantly occurs in Māori (~90% of cases), accounting for around one-quarter of TB cases in this population (82/333 culture positive cases, 2005-2014) (J Sherwood, ESR, personal communication), and has been responsible for numerous TB outbreaks for over the last 30 years [15-17]. Two other large clusters known as the 'Southern Cross' and 'Otara' clusters are predominately found in Pacific people (>90% cases) (J Sherwood, ESR, personal communication).

The most common L4 sublineage in New Zealand is L4.4, which accounts for 43% of New Zealand born L4 cases [8]. In this study, we have used a genomic dataset of 236 global *Mtb* L4.4 isolates from 19 different countries (including 23 recent and newly sequenced New Zealand clinical isolates belonging to the Rangipo and Otara clusters) and a phylodynamics approach to investigate the dispersal of this sublineage to the South Pacific. We show that the Rangipo and Otara clusters belong to a L4.4.1.1 sublineage clade that is frequently found in indigenous populations in Canada and New Zealand. This clade includes the DS6^{Quebec} lineage that was dispersed to Western Aboriginal Canadians by French-Canadian fur traders in the 18th–19th centuries [10]. We suggest that migration of this clade into the South Pacific was likely driven by the expansion of European trade networks in the 19th century, with the whaling trade serving as a route for dispersal to indigenous Polynesian populations. We trace dispersal of this clade to two separate migration events and infer distinct demographic trajectories following establishment of these *Mtb* populations in Polynesia. Dispersal and subsequent expansion of these pathogen populations is related to historical and social drivers of TB transmission. Our results demonstrate the power of phylogeographic approaches to study *Mtb* dispersal at both the global and local scale providing valuable new insights into human social phenomena underlying the dispersal of this globally successful bacterial pathogen.

Results

Phylogeny of the New Zealand *Mtb* strains

Single nucleotide polymorphism (SNP)-based lineage assignment using Illumina whole genome sequencing (WGS) data assigned Rangipo and Otara isolates to the L4.4.1.1 (S-type) sublineage and Southern Cross to L4.3.3 (LAM). A total of 23 New Zealand L4.4.1.1 genomes (16 Rangipo, 7 Otara) spanning a 22-year period (1991–2013) met quality criteria for inclusion in downstream analyses. A maximum likelihood phylogeny of these shows the Rangipo and Otara strains form two well-differentiated monophyletic clades with differing phylogenetic structures (S1 Fig). The

126 Rangipo cluster is characterized by short terminal branches and low genetic diversity (pairwise SNPs
127 0–12, median 4) suggesting recent clonal expansion and temporally short transmission chains,
128 consistent with its association with local outbreaks. Conversely, Otara isolates have long terminal
129 branches and higher genetic diversity (pairwise SNPs 1–102, median 91). This is consistent with
130 predominant reactivation disease due to a locally endemic strain rather than a recent transmission
131 cluster as previously thought based on MIRU-VNTR typing.
132
133 **Global phylogeny of the L4.4 sublineage**
134 To investigate the origins and dispersal of the New Zealand L4.4 strains, we compiled a dataset
135 comprising our 23 New Zealand Rangipo and Otara strain genomes and 213 L4.4 genomes from 18
136 different countries representing all five major global regions (S2 Fig). WGS reads were mapped to the
137 H37Rv reference genome and repetitive genomic regions were removed prior to alignment. High
138 quality variant sites were extracted producing a 9024 bp SNP alignment used to infer a global L4.4
139 maximum likelihood phylogeny (Fig 1A and S3 Fig). This shows L4.4 comprises three
140 well-differentiated sublineages, L4.4.1.1, L4.4.1.2 and L4.4.2 (pairwise F_{ST} values between lineages
141 0.50–0.57), consistent with the Coll classification system [18].
142
143 L4.4 has previously been observed at high proportions in parts of Asia and Africa [8].
144 Consistent with this observation, a large proportion of isolates in our dataset come from these regions
145 (S2 Fig). Our results reveal differing global distributions and population structures of the L4.4
146 sublineages. We find L4.4.2 is essentially restricted to Eastern Asia consistent with *in situ* growth and
147 diversification of this sublineage. Conversely, both L4.4.1.1 and L4.4.1.2 are relatively well
148 distributed globally, indicative of high rates of migration and efficient dispersal.
149

Within L4.4.1.1, we identified a clade comprised predominantly of isolates from New Zealand and Canada. The Canadian isolates belong the DS6^{Quebec} lineage, which is endemic in French Canadians in Quebec and Western Aboriginal Canadian populations, and is characterized by the presence of the DS6^{Quebec} deletion [10, 19]. We therefore termed this clade "DS6Q". All Rangipo and Otara isolates belong to the DS6Q clade. Examination of mapped sequencing reads found that both of these clusters carry the DS6^{Quebec} deletion and the presence of the deletion was further confirmed by PCR and Sanger sequencing. The Rangipo and Otara clusters are not monophyletic within the DS6Q clade, consistent with at least two separate introductions of this clade into New Zealand.

Phylogenetic placement of the DS6^{Quebec} deletion

The DS6^{Quebec} deletion removes an approximately 11.4 kb region truncating or removing the genes from Rv1755c/*plcD* to Rv1765c (between positions 1987457 to 1998849 in H37Rv) (Fig 1B) [19]. Examination of mapped reads found that all L4.4.1.1 and L4.4.1.2, but not L4.4.2 genomes, harboured the DS6^{Quebec} deletion, showing that this is a characteristic deletion of L4.4.1. One L4.4.1.1 genome had an earlier start to the deletion (position 1987142) indicating a subsequent small deletion event. This same region is also removed by a similar but evolutionarily independent ~12 kb RD152 deletion (positions 1986636 to 1998621) in L2/Beijing strains (Fig 1B) [20]. The genomic region affected by RD152 and DS6^{Quebec} is highly variable and is associated with frequent insertion of IS*6110* elements [21], suggesting homologous recombination between adjacent IS*6110* elements as the likely mechanism responsible for these similar but evolutionarily distinct deletion events.

Temporal evolution of the L4.4.1.1 sublineage and the DS6Q clade

The polytomy at the root of the DS6Q clade and the polyphyletic nature of the New Zealand and Canadian isolates implies dispersal of several closely related strains from a common origin. It is likely that the DS6^{Quebec} lineage was introduced to Canada from France [10], suggesting a similar

European origin for the New Zealand strains. French whalers had a notable presence in New Zealand and Polynesia during the South Pacific whaling era (1790–1860) [22] and the arrival of whalers and other traders in the region is associated with the introduction of new diseases, including TB [23, 24]. We hypothesized the DS6Q clade may have been introduced to Polynesia via this route. To further explore this hypothesis, the temporal evolution and dispersal of the L4.4.1.1 sublineage was investigated by Bayesian evolutionary analysis using BEAST2 [25] and an alignment of 3161 variable nucleotide positions from 117 L4.4.1.1 genomes, which included all New Zealand and global L4.4.1.1 isolates with known year of isolation at the time of analysis. Both root-to-tip regression (R^2 0.229) and date randomization tests detected sufficient temporal signal in the data set for calibration of the molecular clock by tip-dating (S4 Fig).

The L4.4.1.1 phylogeny, mutation rate and node ages were inferred using strict and relaxed (UCLD) molecular clocks with different coalescent demographic models. Nucleotide substitution rate was modelled using the general-time reversible model (GTR). All models produced similar rate and date estimates (median 6.15×10^{-8}–6.64×10^{-8} substitutions/site/year (s/s/y); widest 95% highest posterior density (HPD) intervals over all models, 4.23×10^{-8}–9.08×10^{-8}) (Table 1). Model comparison using path sampling determined that the strict clock with the Bayesian skyline demographic model provided the best fit to the data (S1 Table). Under this model we estimated a substitution rate of 6.28×10^{-8} s/s/y (95% HPD, 4.54×10^{-8}–8.10×10^{-8}), resulting in a time to most recent common ancestor (TMRCA) estimate of 1492 for L4.4.1.1 (95% HPD, 1325–1629). Our substitution rate estimate is similar to the results from other studies using contemporary L4 and mixed lineage MTBC genomes, all of which produced median rate estimates of ~7×10^{-8}–1×10^{-7} s/s/y [26-30].

Table 1. *Mycobacterium tuberculosis* complex L4.4.1.1 sublineage substitution rate and time to most recent common ancestor (TMRCA) estimates.

Clock model	Demographic model	Substitution rate (x 10^{-8} s/s/y)[1]	L4.4.1.1 TMRCA	DS6Q TMRCA	Rangipo TMRCA	Otara TMRCA
Strict	Constant	6.63 (4.34–9.06)	1513 (1301–1673)	1671 (1529–1777)	1978 (1965–1987)	1827 (1744–1886)
Strict	Exponential	6.63 (4.37–9.08)	1513 (1299–1672)	1672 (1530–1781)	1978 (1965–1986)	1827 (1745–1887)
Strict	Skyline	6.28 (4.54–8.10)	1492 (1325–1629)	1652 (1535–1741)	1980 (1969–1988)	1813 (1746–1868)
UCLD	Constant	6.49 (4.23–8.93)	1500 (1277–1667)	1665 (1518–1774)	1977 (1964–1986)	1824 (1741–1887)
UCLD	Exponential	6.64 (4.37–8.98)	1511 (1294–1667)	1673 (1528–1772)	1977 (1965–1986)	1828 (1748–1888)
UCLD	Skyline	6.15 (4.39–7.98)	1480 (1300–1624)	1645 (1524–1743)	1980 (1968–1988)	1809 (1739–1867)

Median values are reported and 95% HPD interval shown in brackets. Estimates reported in the text are shown in bold. The GTR model of substitution was used for all analyses.

[1] Substitution rate is in substitutions per site per year (s/s/y).

The Bayesian skyline plot suggests the L4.4.1.1 sublineage underwent a rapid population expansion following its emergence, and corresponding migration analyses show a spike in migration at this time (Fig 2A, 2B). This was followed by a period where the population size remained consistent until another period of population growth in the 19th century, during which time migration tapers off. Our phylogeographic reconstruction is indicative of migration from Africa to Southeast Asia, as well as Europe to Canada and Oceania (Fig 2) . These patterns of connectivity and the dispersal of L4.4.1.1 through Africa and Europe are consistent with previous reconstructions of the migratory history of L4 [6].

Our estimated TMRCA of the DS6Q clade is 1652 (95% HPD, 1535–1741) and the TMRCA of Rangipo and the closest Canadian clade was 1691 (95% HPD, 1588–1776). This is coincident with the French migration to Quebec between 1608–1760 [31], and is thus consistent with a European,

likely French, origin of the DS6Q clade (Fig 3). The TMRCA estimate for the Otara strain is 1813 (95% HPD, 1746–1868), which coincides with arrival of European whalers and other traders, including sealers, bêche-de-mer and sandalwood traders, to the Pacific region the early 19th century [32, 33]. Our TMRCA estimate of the Rangipo strain is 1980 (95% HPD, 1969–1988), indicating this strain is either a relatively recent introduction or clonal expansion from a previously introduced unsampled DS6Q strain.

Discussion

Using a phylodynamics inference from WGS data, we have performed a global characterization of the L4.4 sublineage and identify a L4.4.1.1 sublineage clade that is common in indigenous populations in Canada and Polynesia. Molecular dating estimated the TMRCA for this clade, termed the "DS6Q" clade, to have existed in the mid-17th century, which is coincident with the French migration to Quebec and thus consistent with a French origin as previously reported for the $DS6^{Quebec}$ lineage [10] (Fig 3). Our results indicate multiple migrations of closely related DS6Q strains out of Europe to Canada and to the South Pacific, providing a striking example of the role of European colonial and trade migrations in driving the global spread of L4.

The early 19th century TMRCA estimate for the Otara strain fits with an introduction to Polynesia by French/European whalers or other traders (Fig 3). To date, the Otara strain has predominantly been identified in Pacific people living in New Zealand. Considering the recent history of migration to New Zealand from the Pacific Islands, our results suggest this strain was initially dispersed to the Pacific Islands from Europe and subsequently migrated to New Zealand. Although limited data are available on early migration from the Pacific Islands, only very small numbers of Pacific people began to settle in New Zealand in the 1800s and in 1916 there were only 151 Pacific Island Polynesians in New Zealand [34]. The Pacific population in New Zealand began to slowly

increase in the first half of the 20th century until an upsurge in migration in the 1950s–1970s [35], and by 1976 the Polynesian population in New Zealand had increased to 61,354 [34]. The TMRCA of the Otara strain is therefore consistent with initial dispersal to the Pacific Islands by Europeans in the early 1800s. Long internal branches in the phylogeny stretching back to the 19th century suggest multiple subsequent introductions to New Zealand that may have accompanied more recent Pacific migrations.

Unlike New Zealand, which began to receive large influxes of European, predominantly British and Irish, migrants following British annexation in 1840 [34], other Polynesian islands did not experience the same *en masse* arrival of European emigrants, lending further support to this being a trade-associated introduction. The whaling trade was also the only French economic activity of any scale in the South Pacific during the early 19th century [36], the most significant years of which were 1832–1846 [37]. As with the Canadian fur trade, commercial success of the whaling industry depended on establishing productive social and economic relationships with the local people. Intermarriage played a central role in industry establishment and success in both the Canadian fur trade and South Pacific whaling [38]. During the whaling era, large numbers of Polynesians relocated from villages to harbour settlements for trade and employment opportunities, Polynesian men were often recruited as crew on whaling ships accounting for up to one-fifth of European whaling crews [24, 33]. Such interactions would have established strong social ties conducive for the dispersal of *Mtb.* Accordingly, contact with European trade vessels and ports have been implicated in the introduction of TB and other infectious diseases into Polynesia [23, 24].

The L4.4.1.1 sublineage defined by SNP genotyping corresponds to the S lineage, also known as the ‘S-type’, classified by spoligotyping [18, 39]. Molecular typing has shown that the S lineage also has a notable presence in French Polynesia, accounting for over one-third of *Mtb* isolates in Tahiti

(10/27, 37%) [40]. Tahiti was made a French protectorate in 1842 and a colony in 1880, and was an important commerce hub provisioning European whaling and trade vessels in the early 19th century. Although no WGS data were available for inclusion in phylogenetic analyses, we speculate that this lineage may have been introduced to Tahiti via the same historical migrations that introduced the New Zealand DS6Q strains to Polynesia. In addition to DS6Q strains in Canada and New Zealand, the DS6Q clade also contains isolates from Russia. Unlike New Zealand and Canada where DS6Q strains occur at relatively high frequencies in indigenous populations, L4.4 is rare in Russia [8, 41]. Historically, Western Europe and Russia have been culturally and politically more connected and trade between them dates back to ancient times [42]. Russia was also engaged the colonial fur trade [43], providing possible avenues for dispersal of DS6Q strains.

Unlike the older endemic Otara strain, our results indicate that the Rangipo cluster arose from a relatively recent clonal expansion, due to either a more recent introduction or emergence from a previously introduced DS6Q strain. Between 1840–1843 the majority of French whaling voyages included New Zealand (70/81, 86.4%) [37] and French whaling provides a conceivable route for historical introduction of DS6Q strains into New Zealand from France. Alternatively, Rangipo may be a more recent introduction. The TMRCA follows a period of mass migration to New Zealand from the Pacific Islands in the 1950s–1970s offering another plausible route. Although it is evident that the Rangipo cluster has ultimately emerged from a strain of European origin, more in-depth sampling of L4.4.1.1 isolates from both New Zealand and the Pacific may provide a clearer picture of the route this strain took from Europe to New Zealand and will shed additional light on the dispersal of this sublineage in this region.

The Rangipo strain was named for its association with a large TB outbreak in the late-1990s involving cases who had spent time in the Rangipo prison [15]. Prior to this, health professionals

were aware of clusters of infection caused by this strain first appearing in the early-1990s (N Karalus, personal communication). Our TMRCA estimate for the Rangipo cluster predates this outbreak, although its introduction into the prison environment has presumably helped contribute to its further spread. The TMRCA of the Rangipo strain coincides with major demographic changes in the Māori population that occurred in the mid-20th century. Māori TB mortality rates declined sharply in the mid-1900s [44], which presumably would have imposed a bottleneck on the *Mtb* population. Along with falling TB rates, between 1945–1980 Māori also experienced one of the fastest rates of urbanisation of any population in the world [45]. This was accompanied by significant environmental changes including overcrowded housing and increased prison incarceration rates, both of which are TB risk factors [46, 47]. The temporal association between emergence of the Rangipo cluster and the urbanisation of Māori suggests that human social phenomena are important contributors to the expansion and dispersal of *Mtb*.

Both Rangipo and Otara are Māori place names and Otara is a city that is home to large populations of Pacific people, associating these names with Māori and Pacific people more generally. Our results show that these strains are a product of European contact and colonization and highlight the pejorative naming of these strains with Māori names. Naming diseases by place of origin stigmatizes the associated population and the name 'Rangipo' also further perpetuates the stigma attached to the disease by associating it with prison and criminality. Stigma increases the emotional suffering of TB patients and has implications for TB control efforts, for example by affecting health-seeking behaviours and adherence to treatment [48]. The findings of this work point to the appropriateness of renaming these clusters to refrain from further stigmatizing communities where TB is present and perpetuating stigma associated with the disease, and further work will seek to formally rename these clusters in consultation with Māori.

317 Recently, Brynildsrud et al. [9] reconstructed the migratory history of the L4 sublineage,
318 including isolates from Europe, Africa, the Americas and Southeast Asia, but not the South Pacific.
319 Global dispersal of L4 was found to be dominated by historical migrations out of Europe and dispersal
320 of L4 to Africa and the Americas occurred concomitant with European colonial migrations [9]. We
321 observe the same scenario with the introduction of L4.4 to the South Pacific and the DS6Q clade
322 provides a striking example of the role of European expansion in the global dispersal of *Mtb*. Our
323 analyses reveal the migration of several closely related DS6Q strains out of Europe in the 17th–19th
324 centuries to remote and unconnected populations driven by European colonial migrations and
325 expanding trade networks. In a separate study by O'Neill et al. [6] (preprint), the evolutionary history
326 of L4 was found to be characterized by rapid diffusion and high rates of migration, with range
327 expansion contributing to the growth of this lineage. Consistent with this, our results suggest efficient
328 dispersal of L4.4 and a more extensive demographic analysis of the L4.4.1.1 sublineage revealed
329 increased population growth concurrent with a spike in migration in the 16th century following
330 emergence of this lineage. This timing is coincident with the European age of exploration, providing
331 a plausible factor that may have contributed to the growth and dispersal of this sublineage. A similar
332 pattern of increased population growth and migration during this era has also been detected for L4 as
333 a whole [6]. We detect L4.4.1.1 population growth in the 19th century that could be attributable to
334 various colonial activities around this time involving countries represented in our sample; the French-
335 Canadian fur trade (1710–1870) [49], the South Pacific whaling trade (1790-1860) [22], and the rapid
336 occupation and colonisation of much of the African continent during the New Imperialism period
337 (1876–1912) [50]. While the later population decline in the late 20th century coincides with the
338 dramatic decline in TB incidence in the developed world over the last century.
339
340 The phylogeography of bacterial pathogens can provide valuable insights into the migratory
341 history of their human hosts. Most notably, *Helicobacter pylori* has been identified as reliable marker

342 to deduce human population movements, providing valuable insights into ancient human migrations
343 in the Pacific and globally [51, 52]. In this study, we identify multiple migrations of several closely
344 related *Mtb* strains to geographically distant and unconnected indigenous populations driven by
345 European colonial and trade expansion in the 17–19th centuries. The presence of the DS6Q clade in
346 indigenous Polynesian populations provides a potential marker of these historical migrations and
347 reasserts the role of European migrations in the global dispersal of L4 *Mtb*. Our results highlight the
348 power of phylodynamic methods and the utilization of public WGS data repositories to trace recent
349 migrations of *Mtb* in high resolution, uncovering human movements and social changes that have
350 contributed to the dispersal and success of *Mtb* in indigenous Polynesian populations.
351
352

353 ## Materials and Methods

354 ### Genomic data

355 We have recently sequenced eighteen, five and seven *Mtb* isolates from the New Zealand
356 Rangipo, Southern Cross and Otara clusters, respectively, on the Illumina MiSeq platform [53-55]. In
357 this study, we sequenced a further four Rangipo genomes on Illumina MiSeq as previously described
358 [56]. Sequencing data were submitted to the National Centre for Biotechnology Information (NCBI)
359 Nucleotide Archive (PRJNAXXXXXX). MTBC lineage was determined in KvarQ [57].
360

361 Global L4.4 genomes included 23 unpublished genomes from Canada and 190 publicly
362 available genomes from recently published studies [9, 29, 41, 58-65] and Broad Institute sequencing
363 initiatives (broadinstitute.org) (S1 File). Canadian genomes were sequenced on the Illumina platform
364 as previously described [66] and sequencing data were submitted to the National Centre for
365 Biotechnology Information (NCBI) Nucleotide Archive (PRJNAXXXXXX). Publicly available
366 genomes were assembled from a list of L4 genomes [8] from which those belonging to L4.4 were

367 selected after being identified with KvarQ [57]. Additional L4.4 genomes were identified through
368 literature searches and screening with KvarQ. Genomes identified as low or mixed coverage were
369 excluded, and if more than one genome sequence was available for a sample only the first listed was
370 used. Country and year of isolation were obtained from the NCBI BioSample database.
371
372 **Reference guided assembly and variant calling**
373 Raw reads were trimmed with TrimGalore!
374 (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) using quality threshold of 15 and
375 reads less than 20 bp long were discarded. Trimmed reads were mapped to the H37Rv reference
376 genome (NC_000962.3) [67] using BWA-MEM [68]. Duplicates were removed using Picard tools
377 (https://broadinstitute.github.io/picard/) and local realignment was performed using GATK [69].
378 Mapping quality was assessed using Qualimap [70] and genomes were excluded if the depth of
379 coverage was <25X or if <75% of trimmed reads mapped to the reference genome. Variants were
380 called using Pilon [71] using a minimum depth threshold of 10, base quality threshold of 20 and
381 mapping quality threshold of 40. VCF files generated by Pilon were converted to FASTA format using
382 in house scripts that treat ambiguous calls and deletions as missing data (https://github.com/pepperell-
383 lab/RGAPepPipe). Bases in repetitive regions including genes annotated as PE/PPE, PGRS,
384 REP13E12, transposable and phage elements were removed from FASTA sequences prior to
385 alignment. Variant sites were extracted from concatenated whole genome alignments using SNP-sites
386 [72]. Genomes with missing data at > 10% of sites were excluded from further analyses and only sites
387 where at least 90% of isolates had high quality base calls were included in phylogenetic and molecular
388 dating analyses. VCF and bam files were manually examined for the presence of the DS6^{Quebec} deletion
389 (positions 1987457 to 1998849) [19].
390
391 **Maximum likelihood phylogenetic inference**

Maximum likelihood trees were inferred using PhyML 3.1 [73] with x1000 bootstrap replicates using the general time reversible (GTR) substitution model as this was the best fitting model based on the Bayesian information criterion in jmodeltest2 [74]. A phylogeny of 23 New Zealand L4.4.1.1 sublineage genomes was inferred from a 345 bp whole-genome SNP alignment. Pairwise SNP distances were calculated from these variant sites using the poppr::bitwise.dist package in R [75] using the 'missing_match = T' option to count sites with missing data as matching. A 9024 bp SNP alignment was used to infer a global L4.4 phylogeny, this included all high-quality New Zealand and global L4.4 genomes (n 236, S1 File, Dataset 1) and H37Rv. The R-package PopGenome [76] was used to calculate pairwise fixation indices (F_{ST}) from variant sites to estimate population separation between lineages, specifying groups by lineage as determined by Kvarq.

Bayesian phylogenetic analysis

Bayesian evolutionary analysis of the L4.4.1.1 sublineage was performed in BEAST2 [25] using 3161 variant sites extracted from a 3949977 bp alignment of 117 L4.4.1.1 genomes with known year of isolation (S1 File, Dataset 2). XML-input files were manually modified to specify the number of invariant sites as calculated by scaling the number of non-SNP sites in the full alignment by the frequency of each base.

Assessment of temporal signal for tip-based calibration. The molecular clock was calibrated using tip dates covering a 26-year period (1987–2013). To determine if the temporal signal was sufficient for accurate molecular dating, the dataset was assessed using root-to-tip regression and date randomization (S4 Fig). A maximum likelihood tree was constructed in PhyML and Tempest was used to determine root-to-tip distance for regression analysis against tip date, revealing a modest temporal signal in the data (R^2 0.229). The DS6Q clade sample subset (n 47) showed weaker temporal signal (R^2 0.139) but similar slope (4.6×10^{-4}) to the full L4.4.1.1 dataset (2.3×10^{-4}). To further

validate the temporal signal, sampling dates were randomized 20 times and analysed with BEAST2 using a strict clock and constant demographic model with the same parameters for the random and real dates. Estimates of the substitution rate and TMRCA showed no overlap in the 95% HPD between the real and randomized dates, indicating that the data contains sufficient temporal signal for tip-based calibration.

Molecular dating. Mutation rates and divergence times were estimated using MCMC sampling in BEAST2 with the BEAGLE library [77]. Analyses were performed using the GTR substitution model, strict and relaxed molecular clocks (uncorrelated relaxed clock with a log-normal distribution (UCLD)) [78], and coalescent constant, exponential and Bayesian skyline [79] demographic models. Two monophyletic taxon sets were created to ensure the root was correctly placed (as determined with high confidence bootstrap support in the maximum likelihood phylogeny). Uniform prior distributions were defined for the substitution rate (1×10^{-10}–1×10^{-6} s/s/y) and effective population size (upper bound of 1×10^{10}). For the Bayesian skyline model, the Jeffrey's (1/X) prior was deselected for the population size parameter as this an improper prior and therefore unsuitable for model evaluation using path sampling. Default priors were used for the remaining parameters. To estimate posterior distributions, three independent chains were run for 100–350 million states sampling every 10000 states. The first 10% of states were discarded as burn-in and chains were assessed for convergence and sufficient mixing (effective sample size > 200 for all parameters) (S5 Fig). Samples from the three independent chains were combined and parameter estimation based on the combined chain. Median estimates are reported unless otherwise specified. The maximum clade credibility (MCC) tree was estimated from combined tree samples in TreeAnnotator.

The performance of various clock and demographic models was evaluated by path sampling analysis [80]. For each model, 100 path steps were specified using the proportions of a $\beta(0.3, 1.0)$

distribution and two separate runs were performed per model to check for consistency. The MCMC was also run in the absence of data to sample prior distributions for each model. Comparison of marginal posterior and prior distributions showed a strong a strong signal from the data indicating our results are just not an artefact reflecting the prior. The effect of the prior on parameter estimation was also examined by using different upper bounds and the default 1/X prior for the effective population size. Congruent rate and date estimates were obtained when the varying prior parameters on population size demonstrating the robustness of our estimates to this prior specification (S6 Fig).

Phylogeographic inference. Ancestral reconstruction was performed using BEAST2, with UN region for each isolate modelled as a discrete trait. Analyses were performed using the GTR model of nucleotide substitution, a strict molecular clock with the estimated substitution rate of 6.28×10^{-8} s/s/y and BSP demographic models. Migration rates over time were inferred from an MCC tree. As described in O'Neill et al. [6], migration events were defined as a change in the most probable reconstructed state from parent to child node. Only nodes with a posterior probability > 80% were considered. Median heights of the parent and child nodes were treated as the range of time in which a migration event could occur. Migration rates through time were inferred by summing the number of migration events during each year of the phylogeny, divided by the total number of branches in existence during each year of the phylogeny. The Bayesian stochastic search variable selection method (BSSVS) [81] implemented in BEAST2 was used to identify well-supported migration rates between UN regions in the phylogeographic analyses. SpreaD3 [82] was used to calculate Bayes factor for each pairwise rate.

Acknowledgments

The authors would like to thank Dr. Jill Sherwood, ESR (The Institute of Environmental Science and Research, N.Z.), for providing public health data.

References

1. Tollefson D, Bloss E, Fanning A, Redd JT, Barker K, McCray E. Burden of tuberculosis in indigenous peoples globally: a systematic review. Int J Tuberc Lung Dis. 2013;17(9):1139-50.

2. Cormier M, Schwartzman K, N'Diaye DS, Boone CE, dos Santos AM, Gaspar J, et al. Proximate determinants of tuberculosis in Indigenous peoples worldwide: a systematic review. Lancet Glob Health. 2019;7(1):e68-e80.

3. Hershberg R, Lipatov M, Small PM, Sheffer H, Niemann S, Homolka S, et al. High functional diversity in Mycobacterium tuberculosis driven by genetic drift and human demography. PLoS Biol. 2008;6(12):e311.

4. Hirsh AE, Tsolaki AG, DeRiemer K, Feldman MW, Small PM. Stable association between strains of Mycobacterium tuberculosis and their human host populations. Proc Natl Acad Sci U S A. 2004;101(14):4871-6.

5. Gagneux S. Ecology and evolution of Mycobacterium tuberculosis. Nat Rev Microbiol. 2018;16(4):202.

6. O'Neill MB, Shockey AC, Zarley A, Aylward W, Eldholm V, Kitchen A, et al. Lineage specific histories of Mycobacterium tuberculosis dispersal in Africa and Eurasia. bioRxiv. 2018; Preprint. DOI:10.1101/210161.

7. Gagneux S, DeRiemer K, Van T, Kato-Maeda M, de Jong BC, Narayanan S, et al. Variable host-pathogen compatibility in Mycobacterium tuberculosis. Proc Natl Acad Sci U S A. 2006;103(8):2869-73.

8. Stucki D, Brites D, Jeljeli L, Coscolla M, Liu Q, Trauner A, et al. Mycobacterium tuberculosis lineage 4 comprises globally distributed and geographically restricted sublineages. Nat Genet. 2016;48(12):1535-43.

490 9. Brynildsrud OB, Pepperell CS, Suffys P, Grandjean L, Monteserin J, Debech N, et al. Global
491 expansion of Mycobacterium tuberculosis lineage 4 shaped by colonial migration and local
492 adaptation. Sci Adv. 2018;4(10):eaat5869.
493 10. Pepperell CS, Granka JM, Alexander DC, Behr MA, Chui L, Gordon J, et al. Dispersal of
494 Mycobacterium tuberculosis via the Canadian fur trade. Proc Natl Acad Sci U S A. 2011;108(16):5.
495 11. Buckley HR, Tayles N, Halcrow SE, Robb K, Fyfe R. The people of Wairau Bar: a re-
496 examination. J Pac Archaeol. 2010;1(1):1-20.
497 12. Spickard P, Rondilla JL, Hippolite Wright D, editors. Pacific Diaspora: Island Peoples in the
498 United States and Across the Pacific. Honolulu, USA: University of Hawai'I Press; 2002.
499 13. Yen S, Bower JE, Freeman JT, Basu I, O'Toole RF. Phylogenetic lineages of tuberculosis
500 isolates in New Zealand and their association with patient demographics. Int J Tuberc Lung Dis.
501 2013;17(7):892-7.
502 14. ESR. Tuberculosis in New Zealand: Annual Report 2015. Institute of Environmental Science
503 and Research Ltd (ESR), Porirua, N.Z.: 2018.
504 15. De Zoysa R, Shoemack P, Vaughan R, Vaughan A. A prolonged outbreak of tuberculosis in
505 the North Island. N Z Public Health Rep. 2001;8(1):1-3.
506 16. Colangeli R, Arcus VL, Cursons RT, Ruthe A, Karalus N, Coley K, et al. Whole genome
507 sequencing of Mycobacterium tuberculosis reveals slow growth and low mutation rates during latent
508 infections in humans. PLoS One. 2014;9(3):e91024.
509 17. McElnay C, Thornley C, Armstrong R. A community and workplace outbreak of tuberculosis
510 in Hawke's Bay in 2002. N Z Med J. 2004;117(1200):U1019.
511 18. Coll F, McNerney R, Guerra-Assuncao JA, Glynn JR, Perdigao J, Viveiros M, et al. A robust
512 SNP barcode for typing Mycobacterium tuberculosis complex strains. Nat Commun. 2014;5:4812.

513 19. Nguyen D, Brassard P, Menzies D, Thibert L, Warren R, Mostowy S, et al. Genomic
514 characterization of an endemic Mycobacterium tuberculosis strain: evolutionary and epidemiologic
515 implications. J Clin Microbiol. 2004;42(6):2573-80.
516 20. Tsolaki AG, Gagneux S, Pym AS, Goguet de la Salmoniere YO, Kreiswirth BN, Van
517 Soolingen D, et al. Genomic deletions classify the Beijing/W strains as a distinct genetic lineage of
518 Mycobacterium tuberculosis. J Clin Microbiol. 2005;43(7):3185-91.
519 21. Ho TB, Robertson BD, Taylor GM, Shaw RJ, Young DB. Comparison of Mycobacterium
520 tuberculosis genomes reveals frequent deletions in a 20 kb variable region in clinical isolates. Yeast.
521 2000;17(4):272-82.
522 22. Haines D. Lighting up the World? Empires and Islanders in the Pacific Whaling Industry,
523 1790-1860. In: Fusaro M, Polónia A, editors. Maritime history as global history. Oxford: Liverpool
524 University Press; 2010. p. 159–76.
525 23. Lange R. Plagues and Pestilence in Polynesia - the 19th-Century Cook Islands Experience.
526 Bull Hist Med. 1984;58(3):325-46.
527 24. Chappell DA. Double ghosts: Oceanian voyagers on Euroamerican ships. Armonk, New York,
528 U.S.A.: M.E. Sharpe, Inc.; 1997.
529 25. Bouckaert R, Heled J, Kuhnert D, Vaughan T, Wu CH, Xie D, et al. BEAST 2: a software
530 platform for Bayesian evolutionary analysis. PLoS Comput Biol. 2014;10(4):e1003537.
531 26. Eldholm V, Monteserin J, Rieux A, Lopez B, Sobkowiak B, Ritacco V, et al. Four decades of
532 transmission of a multidrug-resistant Mycobacterium tuberculosis outbreak strain. Nat Commun.
533 2015;6:7119.
534 27. Roetzer A, Diel R, Kohl TA, Ruckert C, Nubel U, Blom J, et al. Whole genome sequencing
535 versus traditional genotyping for investigation of a Mycobacterium tuberculosis outbreak: a
536 longitudinal molecular epidemiological study. PLoS Med. 2013;10(2):e1001387.

537 28. Ford CB, Shah RR, Maeda MK, Gagneux S, Murray MB, Cohen T, et al. Mycobacterium
538 tuberculosis mutation rate estimates from different lineages predict substantial differences in the
539 emergence of drug-resistant tuberculosis. Nat Genet. 2013;45(7):784-90.
540 29. Walker TM, Ip CL, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, et al. Whole-genome
541 sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study.
542 Lancet Infect Dis. 2013;13(2):137-46.
543 30. Pepperell CS, Casto AM, Kitchen A, Granka JM, Cornejo OE, Holmes EC, et al. The role of
544 selection in shaping diversity of natural M. tuberculosis populations. PLoS Pathog.
545 2013;9(8):e1003543.
546 31. Charbonneau H, Boleda M, Bates RL. The First French Canadians: Pioneers in the St.
547 Lawrence Valley. Newark, NJ: University of Delaware Press; 1993.
548 32. Campbell IC. Worlds apart: a history of the Pacific Islands. Christchurch, N.Z.: Canterbury
549 University Press; 2011.
550 33. Fischer SR. A history of the Pacific Islands. 2nd ed. Basingstoke, Hampshire: Palgrave
551 Macmillan; 2013.
552 34. United Nations, Economic Social Commission for Asia and the Pacific. The Population of
553 New Zealand: Country Monograph Series, No. 12. Bangkok, Thailand: United Nations; 1985.
554 35. Dunsford D, Park J, Littleton J, Friesen W, Herda P, Neuwelt P, et al. Better lives: the struggle
555 for health of transnational pacific peoples in New Zealand, 1950-2000. Department of Anthropology,
556 University of Auckland; 2011.
557 36. Maclellan N. After Moruroa: France in the South Pacific. Chesneaux J, editor. Melbourne:
558 Ocean Press; 1998.
559 37. Foucrier A. The French and the Pacific world, 17th-19th centuries: explorations, migrations
560 and cultural exchanges. Aldershot, Hampshire: Ashgate Variorum; 2005.

561 38. Stevens K, Wanhalla A. Intimate Relations: Kinship and the Economics of Shore Whaling in
562 Southern New Zealand, 1820-1860. J Pac Hist. 2017;52(2):135-55.
563 39. Brudey K, Driscoll JR, Rigouts L, Prodinger WM, Gori A, Al-Hajoj SA, et al. Mycobacterium
564 tuberculosis complex genetic diversity: mining the fourth international spoligotyping database
565 (SpolDB4) for classification, population genetics and epidemiology. BMC Microbiol. 2006;6:23.
566 40. Osman DA, Phelippeau M, Drancourt M, Musso D. Diversity of Mycobacterium tuberculosis
567 lineages in French Polynesia. J Microbiol Immunol Infect. 2017;50(2):199-206.
568 41. Casali N, Nikolayevskyy V, Balabanova Y, Harris SR, Ignatyeva O, Kontsevaya I, et al.
569 Evolution and transmission of drug-resistant tuberculosis in a Russian population. Nat Genet.
570 2014;46(3):279-86.
571 42. Öhberg A. Russia and the World Market in the Seventeenth Century: A Discussion of the
572 Connection between Prices and Trade Routes. Scand Econ Hist Rev. 1955;3(2):154.
573 43. Rich EE. Russia and the colonial fur trade. Econ Hist Rev. 1955;7(3):307-28.
574 44. MacLean FS. Challenge for health: a history of public health in New Zealand. Wellington,
575 N.Z.: Government Printer; 1964.
576 45. Pool I. Te Iwi Maori: A New Zealand Population, Past, Present and Projected. Auckland, NZ:
577 Auckland University Press; 1991.
578 46. Clark M, Riben P, Nowgesic E. The association of housing density, isolation and tuberculosis
579 in Canadian First Nations communities. Int J Epidemiol. 2002;31(5):940-5.
580 47. Baussano I, Williams BG, Nunn P, Beggiato M, Fedeli U, Scano F. Tuberculosis incidence in
581 prisons: a systematic review. PLoS Med. 2010;7(12):e1000381.
582 48. Craig GM, Daftary A, Engel N, O'Driscoll S, Ioannaki A. Tuberculosis stigma as a social
583 determinant of health: a systematic mapping review of research in low incidence countries. Int J Infect
584 Dis. 2017;56:90-100.

585 49. Innis HA. The fur trade in Canada: An introduction to Canadian economic history: University
586 of Toronto Press; 1999.
587 50. Pakenham T. The scramble for Africa: Hachette, UK.; 2015.
588 51. Moodley Y, Linz B, Yamaoka Y, Windsor HM, Breurec S, Wu JY, et al. The peopling of the
589 Pacific from a bacterial perspective. Science. 2009;323(5913):527-30.
590 52. Falush D, Wirth T, Linz B, Pritchard JK, Stephens M, Kidd M, et al. Traces of human
591 migrations in Helicobacter pylori populations. Science. 2003;299(5612):1582-5.
592 53. Mulholland CV, Ruthe A, Cursons RT, Durrant R, Karalus N, Coley K, et al. Rapid molecular
593 diagnosis of the Mycobacterium tuberculosis Rangipo strain responsible for the largest recurring TB
594 cluster in New Zealand. Diagn Microbiol Infect Dis. 2017;88(2):138-40.
595 54. Gautam SS, Mac Aogain M, Bower JE, Basu I, O'Toole RF. Differential carriage of virulence-
596 associated loci in the New Zealand Rangipo outbreak strain of Mycobacterium tuberculosis. Infect
597 Dis (Lond). 2017;49(9):680-8.
598 55. Mac Aogáin M, Gautam SS, Bower JE, Basu I, O'Toole RF. Draft genome sequence of a New
599 Zealand Rangipo strain of Mycobacterium tuberculosis. Genome Announc. 2016;4(4):e00657-16.
600 56. Aung HL, Tun T, Moradigaravand D, Koser CU, Nyunt WW, Aung ST, et al. Whole-genome
601 sequencing of multidrug-resistant Mycobacterium tuberculosis isolates from Myanmar. J Glob
602 Antimicrob Resist. 2016;6:113-7.
603 57. Steiner A, Stucki D, Coscolla M, Borrell S, Gagneux S. KvarQ: targeted and direct variant
604 calling from fastq reads of bacterial genomes. BMC Genomics. 2014;15:881.
605 58. Bryant JM, Harris SR, Parkhill J, Dawson R, Diacon AH, van Helden P, et al. Whole-genome
606 sequencing to establish relapse or re-infection with Mycobacterium tuberculosis: a retrospective
607 observational study. Lancet Respir Med. 2013;1(10):786-92.

608 59. Bryant JM, Schurch AC, van Deutekom H, Harris SR, de Beer JL, de Jager V, et al. Inferring
609 patient to patient transmission of Mycobacterium tuberculosis from whole genome sequencing data.
610 BMC Infect Dis. 2013;13:110.
611 60. Clark TG, Mallard K, Coll F, Preston M, Assefa S, Harris D, et al. Elucidating emergence and
612 transmission of multidrug-resistant tuberculosis in treatment experienced patients by whole genome
613 sequencing. PLoS One. 2013;8(12):e83012.
614 61. Guerra-Assunção J, Crampin A, Houben R, Mzembe T, Mallard K, Coll F, et al. Large-scale
615 whole genome sequencing of M. tuberculosis provides insights into transmission in a high prevalence
616 area. Elife. 2015;4:e05166.
617 62. Guerra-Assunção JA, Houben RM, Crampin AC, Mzembe T, Mallard K, Coll F, et al.
618 Recurrence due to relapse or reinfection with Mycobacterium tuberculosis: a whole-genome
619 sequencing approach in a large, population-based cohort with a high HIV infection prevalence and
620 active follow-up. J Infect Dis. 2014;211(7):1154-63.
621 63. Holt KE, McAdam P, Thai PVK, Thuong NTT, Ha DTM, Lan NN, et al. Frequent transmission
622 of the Mycobacterium tuberculosis Beijing lineage and positive selection for the EsxW Beijing variant
623 in Vietnam. Nat Genet. 2018;50(6):849–56.
624 64. Zhang H, Li D, Zhao L, Fleming J, Lin N, Wang T, et al. Genome sequencing of 161
625 Mycobacterium tuberculosis isolates from China identifies genes and intergenic regions associated
626 with drug resistance. Nat Genet. 2013;45(10):1255.
627 65. Bjorn-Mortensen K, Soborg B, Koch A, Ladefoged K, Merker M, Lillebaek T, et al. Tracing
628 Mycobacterium tuberculosis transmission by whole genome sequencing in a high incidence setting: a
629 retrospective population-based study in East Greenland. Sci Rep. 2016;6:33180.
630 66. Doroshenko A, Pepperell CS, Heffernan C, Egedahl ML, Mortimer TD, Smith TM, et al.
631 Epidemiological and genomic determinants of tuberculosis outbreaks in First Nations communities in
632 Canada. BMC Med. 2018;16(1):128.

633 67. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, et al. Deciphering the biology
634 of Mycobacterium tuberculosis from the complete genome sequence. Nature. 1998;393(6685):537-
635 44.
636 68. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.
637 arXiv. 2013; Preprint. Available from: arXiv:1303.3997
638 69. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for
639 variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet.
640 2011;43(5):491-8.
641 70. Garcia-Alcalde F, Okonechnikov K, Carbonell J, Cruz LM, Gotz S, Tarazona S, et al.
642 Qualimap: evaluating next-generation sequencing alignment data. Bioinformatics. 2012;28(20):2678-
643 9.
644 71. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated
645 tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One.
646 2014;9(11):e112963.
647 72. Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, et al. SNP-sites: rapid
648 efficient extraction of SNPs from multi-FASTA alignments. Microb Genom. 2016;2(4):e000056.
649 73. Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies
650 by maximum likelihood. Syst Biol. 2003;52(5):696-704.
651 74. Darriba D, Taboada GL, Doallo R, Posada D. jModelTest 2: more models, new heuristics and
652 parallel computing. Nat Methods. 2012;9(8):772.
653 75. Kamvar ZN, Tabima JF, Grunwald NJ. Poppr: an R package for genetic analysis of populations
654 with clonal, partially clonal, and/or sexual reproduction. PeerJ. 2014;2:e281.
655 76. Pfeifer B, Wittelsburger U, Ramos-Onsins SE, Lercher MJ. PopGenome: an efficient Swiss
656 army knife for population genomic analyses in R. Mol Biol Evol. 2014;31(7):1929-36.

657 77. Ayres DL, Darling A, Zwickl DJ, Beerli P, Holder MT, Lewis PO, et al. BEAGLE: an
658 application programming interface and high-performance computing library for statistical
659 phylogenetics. Syst Biol. 2012;61(1):170-3.
660 78. Drummond AJ, Ho SY, Phillips MJ, Rambaut A. Relaxed phylogenetics and dating with
661 confidence. PLoS Biol. 2006;4(5):e88.
662 79. Drummond AJ, Rambaut A, Shapiro B, Pybus OG. Bayesian coalescent inference of past
663 population dynamics from molecular sequences. Mol Biol Evol. 2005;22(5):1185-92.
664 80. Lartillot N, Philippe H. Computing Bayes factors using thermodynamic integration. Syst Biol.
665 2006;55(2):195-207.
666 81. Lemey P, Rambaut A, Drummond AJ, Suchard MA. Bayesian phylogeography finds its roots.
667 PLoS Comput Biol. 2009;5(9):e1000520.
668 82. Vrancken B, Baele G, Lemey P, Bielejec F, Suchard MA, Rambaut A. SpreaD3: Interactive
669 Visualization of Spatiotemporal History and Trait Evolutionary Processes. Mol Biol Evol.
670 2016;33(8):2167-9.
671

672 **Fig 1. Global phylogeny of the *Mycobacterium tuberculosis* complex L4.4 sublineage and**
673 **genomic structure and phylogenetic placement of the DS6^{Quebec} deletion.** (A) Whole genome SNP
674 maximum likelihood phylogeny of L4.4 comprised of 236 isolates from 19 different countries,
675 including 23 isolates from the New Zealand Rangipo and Otara clusters. A black asterisk indicates the
676 DS6^{Quebec} deletion. Tips and terminal branches are coloured by global region and lineages labelled
677 according to the nomenclature of Coll et al. [18]. A black circle indicates the node position of the most
678 recent common ancestor of L4.4 (rooted to H37Rv, not shown). (B) Schematic of the DS6^{Quebec}
679 deletion. Genomic regions in H37Rv that are deleted by the DS6^{Quebec} deletion and the RD152 deletion
680 in the Beijing/W lineage are shown.
681
682 **Fig 2. Demographic analysis of the *Mycobacterium tuberculosis* complex L4.4.1.1 sublineage.**
683 (A) Effective population size (N_e) through time of L4.4.1.1. Median N_e and 95% highest posterior
684 density pictured as black line and grey shading, respectively. X-axis in calendar years. (B) Migration
685 events through time of L4.4.1.1. Black line depicts the rate of migration through time, calculated as
686 the sum of migration events occurring across every year of the phylogeny divided by the total number
687 of branches during each year of the phylogeny. Grey shading depicts the rates inferred after the
688 addition or subtraction of a single migration event. X-axis in calendar years. (C) Migration matrices
689 of L4.4.1.1. Heatmap of pairwise relative migration rates between UN regions. Only relative rates
690 with Bayes factor > 5 shown. (D) MCC phylogeny of L4.4.1.1. Tips and terminal branches coloured
691 according to UN region of isolation. Pie charts on nodes coloured according to geographic state
692 probabilities. X-axis in calendar years.
693
694 **Fig 3. Dated Bayesian phylogeny of the DS6Q clade and historical timeline.** Events shown include
695 the French migration to Quebec (1608–1760), the South Pacific whaling era (l790–1860), the rapid

696 urbanization of Māori in the 20th century (1945–1980), and the surge in Pacific migration into New
697 Zealand in the 1950s–1970s.

Supporting information

S1 Fig. Maximum likelihood phylogeny of 23 New Zealand *Mycobacterium tuberculosis* Rangipo and Otara strain isolates.

S2 Fig. Global distribution of isolates included in this study. (A) Dataset 1; (B) Dataset 2. Pie charts show the proportion of isolates from each of the L4.4 sublineages and circle sizes correspond to the number of isolates from each country.

S3 Fig. Maximum likelihood phylogeny of the *Mycobacterium tuberculosis* complex L4.4 sublineage. Whole genome SNP phylogeny of 236 L4.4 genomes from 19 different countries, including 23 isolates from the New Zealand Rangipo and Otara clusters. Rooted to H37Rv. Tips and terminal branches are coloured by global region and lineages labelled according to the nomenclature of Coll et al. [18]. A black asterisk indicates the DS6^{Quebec} deletion and the DS6Q clade is highlighted in grey.

S4 Fig. Assessment of temporal signal for molecular dating analyses. Regression analysis of root-to-tip distance and year of isolation for (A) the L4.4.1.1 molecular dating dataset, and (B) the DS6Q sample subset. Tip date randomization median estimates and 95% HPD intervals for (C) substitution rate in substitutions/site/year (s/s/y), and (D) tree height in years (since 2013), after tip randomization and for real dates for the full L4.4.1.1 sample (n 117).

S5 Fig. Assessment of MCMC chain convergence. Trace outputs for key parameters from three independent chains for the best model as determined by path sampling (GTR, strict clock, Bayesian skyline demographic). (A) Posterior probability; (B) substitution rate in substitutions/site/year (s/s/y); (C) tree height (years since 2013); and (D-F) TMRCAs for nodes of key interest in calendar years.

S6 Fig. Effect of the population size prior on parameter estimation. Comparison of parameter estimates for (A) substitution rate in substitutions/site/year (s/s/y), and (B) tree height (years since 2013), using the Jeffreys (1/X) and uniform population size priors with different upper bounds (GTR, strict clock, constant population demographic).

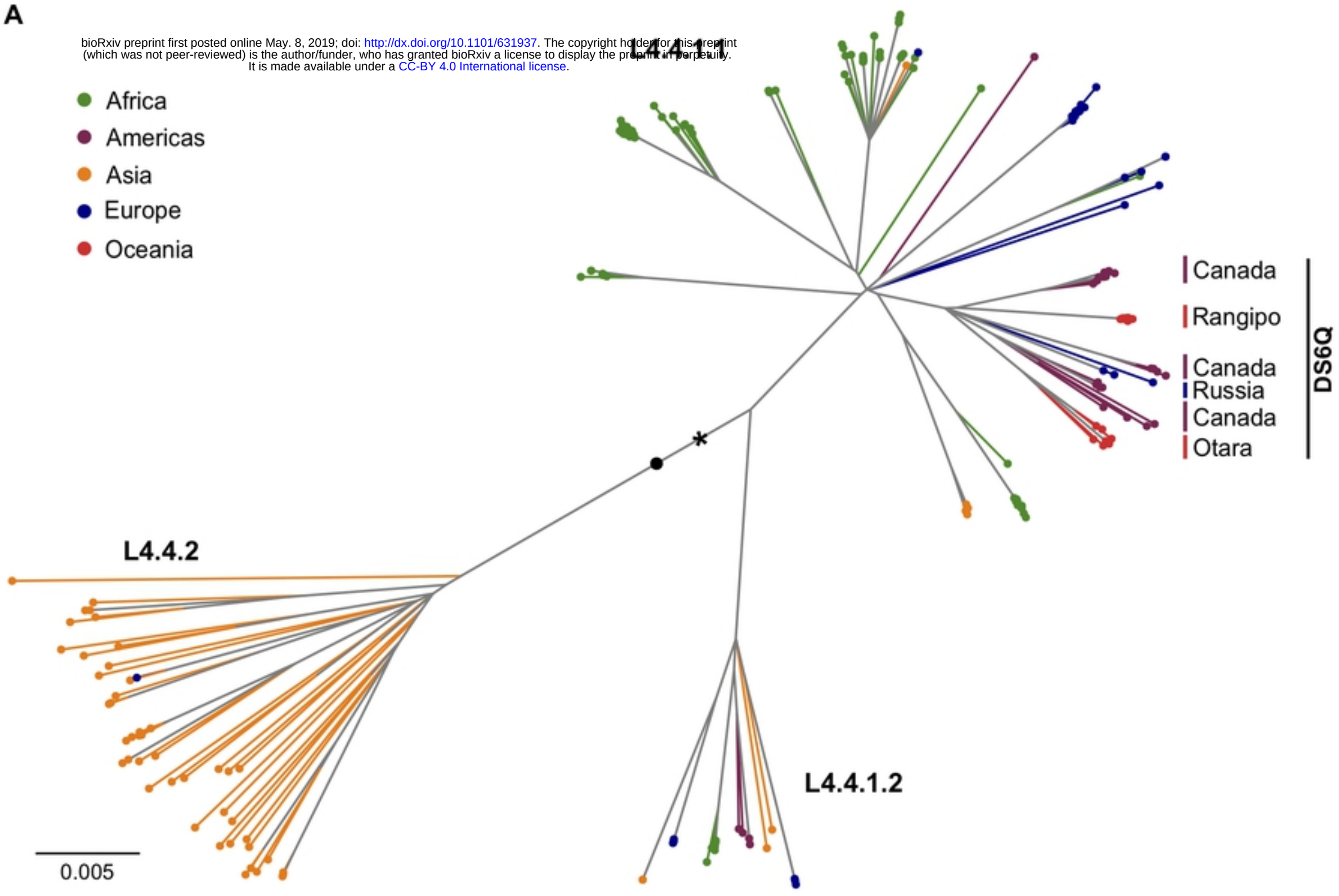
S7 Fig. Dated Bayesian phylogeny of the L4.4.1.1 sublineage showing individual node ages. Bayesian MCC tree of 117 *Mycobacterium tuberculosis* L4.4.1.1 genomes (GTR, strict clock, Bayesian skyline demographic). Median node ages (years since 2013) are shown. A grey box highlights the DS6Q and the New Zealand Rangipo and Otara clusters are labelled.

S8 Fig. Dated Bayesian phylogeny of the L4.4.1.1 sublineage showing posterior probabilities of individual nodes. Bayesian MCC tree of 117 *Mycobacterium tuberculosis* L4.4.1.1 genomes (GTR, strict clock, Bayesian skyline demographic). A grey box highlights the DS6Q and the New Zealand Rangipo and Otara clusters are labelled.

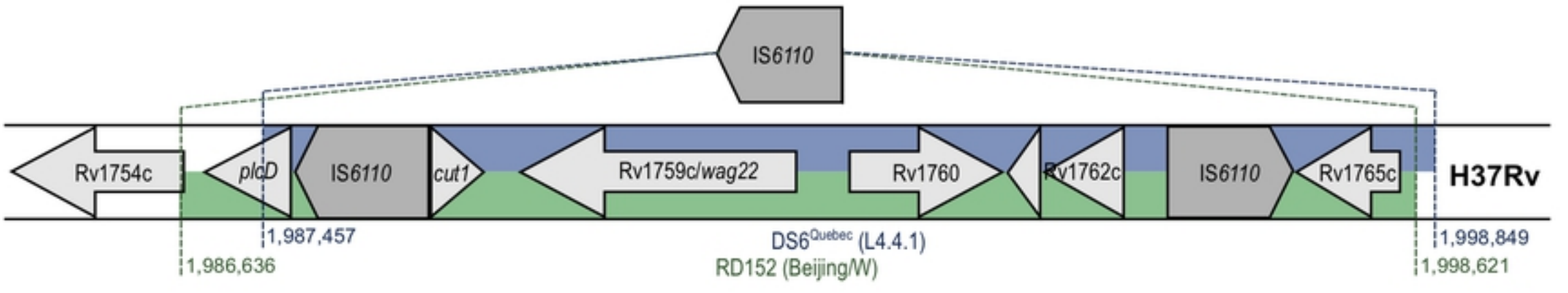
S1 Table. BEAST2 model evaluation by path sampling analysis. Marginal likelihood estimation (MLE) for different clock and population demographic models was determined using path-sampling analysis in BEAST2. The GTR nucleotide substitution model was used for all analyses. Mean log-MLEs are reported for two replicate runs performed to check for consistency. Bayes factors calculated relative to the top ranked model.

S1 File. Genomic datasets used in this study. Dataset 1, L4.4 dataset used for maximum likelihood phylogenetic analysis. Dataset 2, L4.4.1.1 dataset used for molecular dating and demographic analyses.

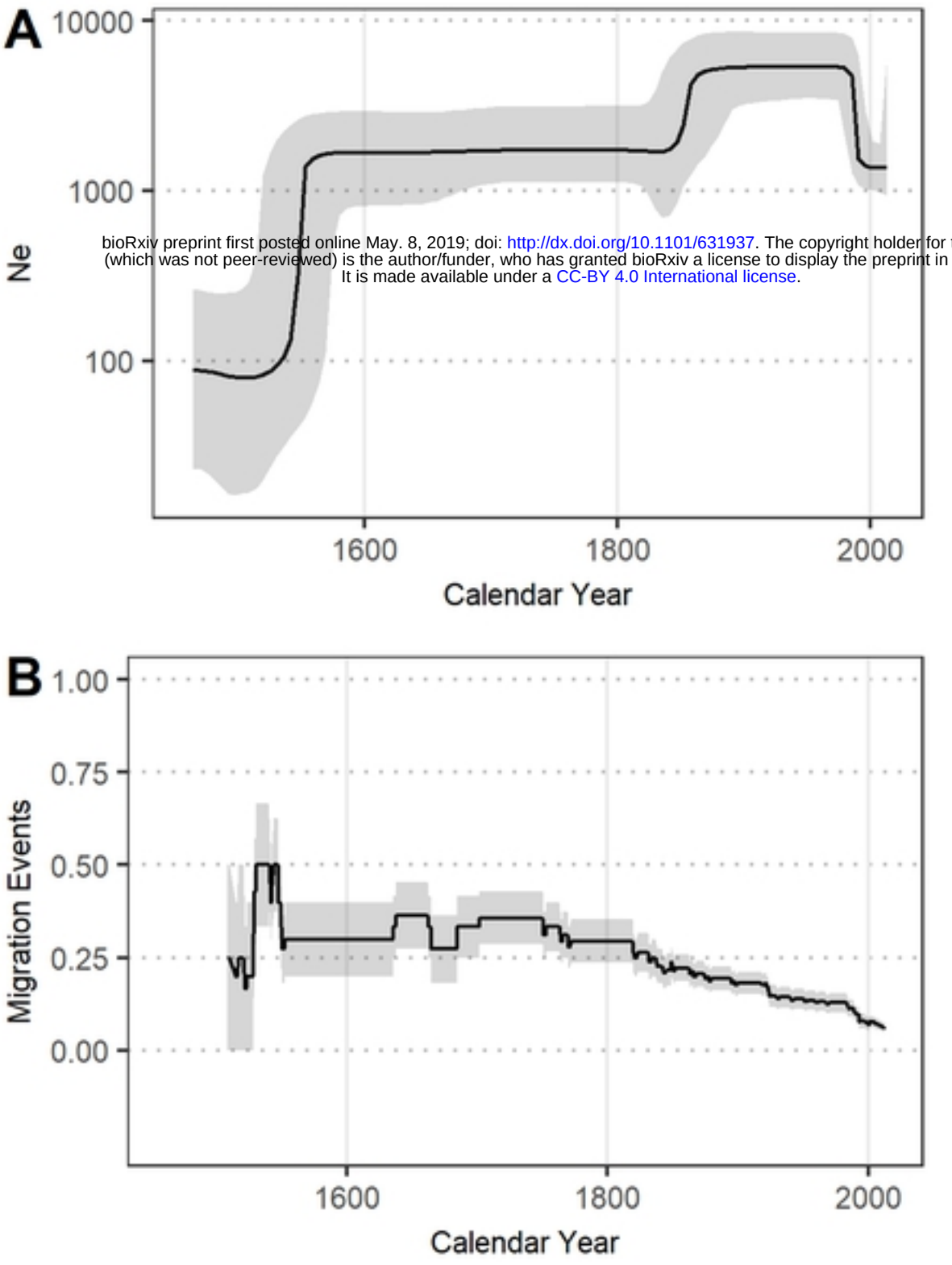
A



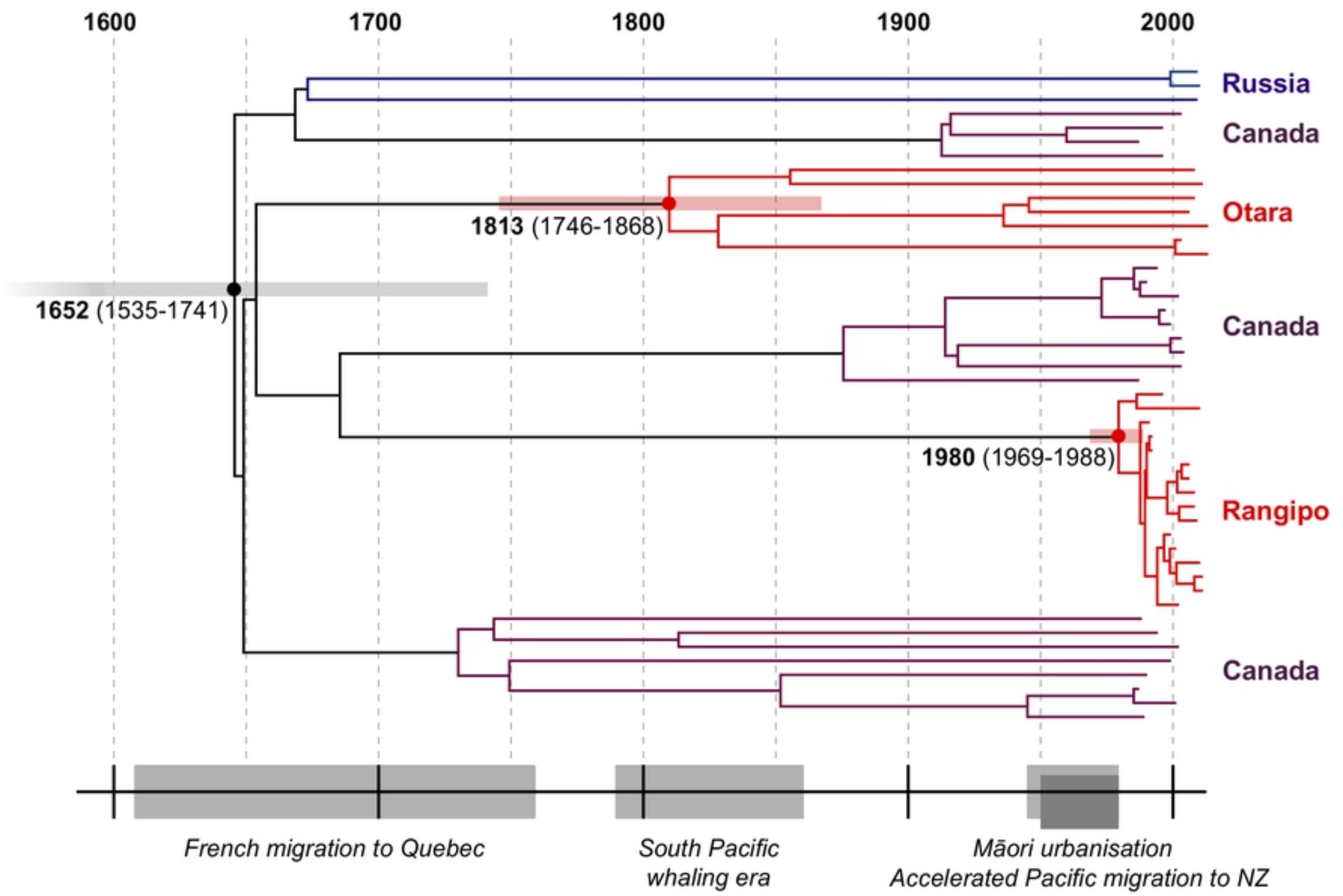
B



Main text figure 1



Main text figure 2



Main text figure 3