

# Human ancient DNA analyses reveal the high burden of tuberculosis in Europeans over the last 2,000 years

Gaspard Kerner,<sup>1,2,3,\*</sup> Guillaume Laval,<sup>1</sup> Etienne Patin,<sup>1</sup> Stéphanie Boisson-Dupuis,<sup>2,3,4</sup> Laurent Abel,<sup>2,3,4</sup> Jean-Laurent Casanova,<sup>2,3,4,5,7</sup> and Lluís Quintana-Murci<sup>1,6,7,\*</sup>

## Summary

Tuberculosis (TB), usually caused by *Mycobacterium tuberculosis* bacteria, is the first cause of death from an infectious disease at the world-wide scale, yet the mode and tempo of TB pressure on humans remain unknown. The recent discovery that homozygotes for the P1104A polymorphism of *TYK2* are at higher risk to develop clinical forms of TB provided the first evidence of a common, monogenic predisposition to TB, offering a unique opportunity to inform on human co-evolution with a deadly pathogen. Here, we investigate the history of human exposure to TB by determining the evolutionary trajectory of the *TYK2* P1104A variant in Europe, where TB is considered to be the deadliest documented infectious disease. Leveraging a large dataset of 1,013 ancient human genomes and using an approximate Bayesian computation approach, we find that the P1104A variant originated in the common ancestors of West Eurasians ~30,000 years ago. Furthermore, we show that, following large-scale population movements of Anatolian Neolithic farmers and Eurasian steppe herders into Europe, P1104A has markedly fluctuated in frequency over the last 10,000 years of European history, with a dramatic decrease in frequency after the Bronze Age. Our analyses indicate that such a frequency drop is attributable to strong negative selection starting ~2,000 years ago, with a relative fitness reduction on homozygotes of 20%, among the highest in the human genome. Together, our results provide genetic evidence that TB has imposed a heavy burden on European health over the last two millennia.

Infectious diseases have been the leading cause of mortality since the origin of modern humans in Africa and throughout their subsequent dispersals around the world.<sup>1–5</sup> Tuberculosis (TB [MIM: 607948]) is considered to be the deadliest infection of the common era, with more than one billion deaths over the last 2,000 years,<sup>6–8</sup> and still responsible for more than 1.5 million deaths annually according to the WHO. The human genetic basis of TB susceptibility has remained elusive until the turn of the 21<sup>st</sup> century, when two rare inborn errors of immunity, autosomal-recessive interleukin-12 receptor b1 (IL-12Rb1) and tyrosine kinase 2 (TYK2) deficiencies, were identified in children with severe TB.<sup>9,10</sup> It was only in 2018 that the first common, monogenic predisposition to TB was identified. Homozygotes for the *TYK2* (MIM: 611521) P1104A polymorphism (rs34536443) were found to be at higher risk of developing clinical forms of TB, due to the selective disruption of IL-23-dependent antimycobacterial IFN- $\gamma$  immunity, underlying a recessive trait.<sup>11</sup> A subsequent study revealed an enrichment in P1104A homozygotes among TB cases of a case-control cohort from the United Kingdom, where the allele is most prevalent today (4%).<sup>7</sup> The frequency of P1104A, together with its high penetrance for TB in the homozygous state (>0.8),<sup>11</sup> suggests that about 1/600 British individuals would develop TB during their lifetime because of the mutation, if TB were still highly endemic in Europe.

Pathogen-imposed selective pressures have been paramount during human evolution.<sup>2,4,5</sup> Over the last decade, population genetic studies have documented strong, distinct selection signatures among host defense genes, helping to delineate immunological mechanisms of major importance,<sup>12</sup> and supporting the notion that microbes have had an overwhelming impact on human genome diversity.<sup>4,5</sup> While several studies have provided insight into the periods when malaria has exerted pressure on humans,<sup>13–17</sup> little is known about the historical burden of other infectious diseases associated with past epidemics. Yet, TB appears to have been more lethal than malaria in the common era,<sup>6</sup> making it a stronger selective pressure in endemic regions. Recent evidence based on mycobacterial ancient DNA (aDNA) suggests a Holocene dispersal of *M. tuberculosis* <6,000 years ago (ya),<sup>18,19</sup> a time frame that coincides with the growth of agricultural communities and anthropogenic environmental changes, which may have favored infectious disease transmission.<sup>20</sup>

To investigate the historical burden of TB in humans, we sought to reconstruct the evolutionary history of the *TYK2* P1104A variant. Indeed, this mutation, in the homozygous state, underlies the only known common, monogenic predisposition to TB.<sup>7,11</sup> Moreover, *TYK2* P1104A does not affect the risk for other infectious diseases except, to a milder degree, rare cases of infection by environmental mycobacteria in otherwise healthy individuals.<sup>11</sup> Whereas

<sup>1</sup>Human Evolutionary Genetics Unit, Institut Pasteur, UMR2000, CNRS, 75015 Paris, France; <sup>2</sup>Laboratory of Human Genetics of Infectious Diseases, Necker Branch, INSERM UMR 1163, Necker Hospital for Sick Children, 75015 Paris, France; <sup>3</sup>Paris University, Imagine Institute, 75015 Paris, France; <sup>4</sup>St. Giles Laboratory of Human Genetics of Infectious Diseases, Rockefeller Branch, Rockefeller University, New York, NY 10065, USA; <sup>5</sup>Howard Hughes Medical Institute, New York, NY 10065, USA; <sup>6</sup>Chair of Human Genomics and Evolution, Collège de France, 75005 Paris, France

<sup>7</sup>These authors contributed equally

\*Correspondence: [gakerner@pasteur.fr](mailto:gakerner@pasteur.fr) (G.K.), [quintana@pasteur.fr](mailto:quintana@pasteur.fr) (L.Q.-M.)

<https://doi.org/10.1016/j.ajhg.2021.02.009>

© 2021 The Author(s). This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).



disease-protective variants may rapidly increase in frequency owing to positive Darwinian selection, disease-risk alleles are expected to evolve under strong negative selection and be gradually purged from the population. Because negatively selected variants have become rare, very rare, or even extinct, they are harder to study using genetic data from modern human populations. However, with the increasing availability of genomes from ancient individuals, direct measurements of the intensity of selection are now possible, as significant increases or decreases of allele frequencies can be captured with aDNA from time transects.<sup>21</sup> Thus, the study of the P1104A variant offers an unprecedented opportunity to shed light on the evolutionary history of a deadly human disease such as TB. Of note, P1104A homozygotes have also been shown to enjoy from a protective effect against various autoimmune and inflammatory diseases.<sup>22,23</sup> While this effect could have provided a fitness advantage opposed to that attributable to TB infection, the general late onset manifestation of autoimmune and inflammatory disorders makes unlikely the occurrence of a large counteractive effect.

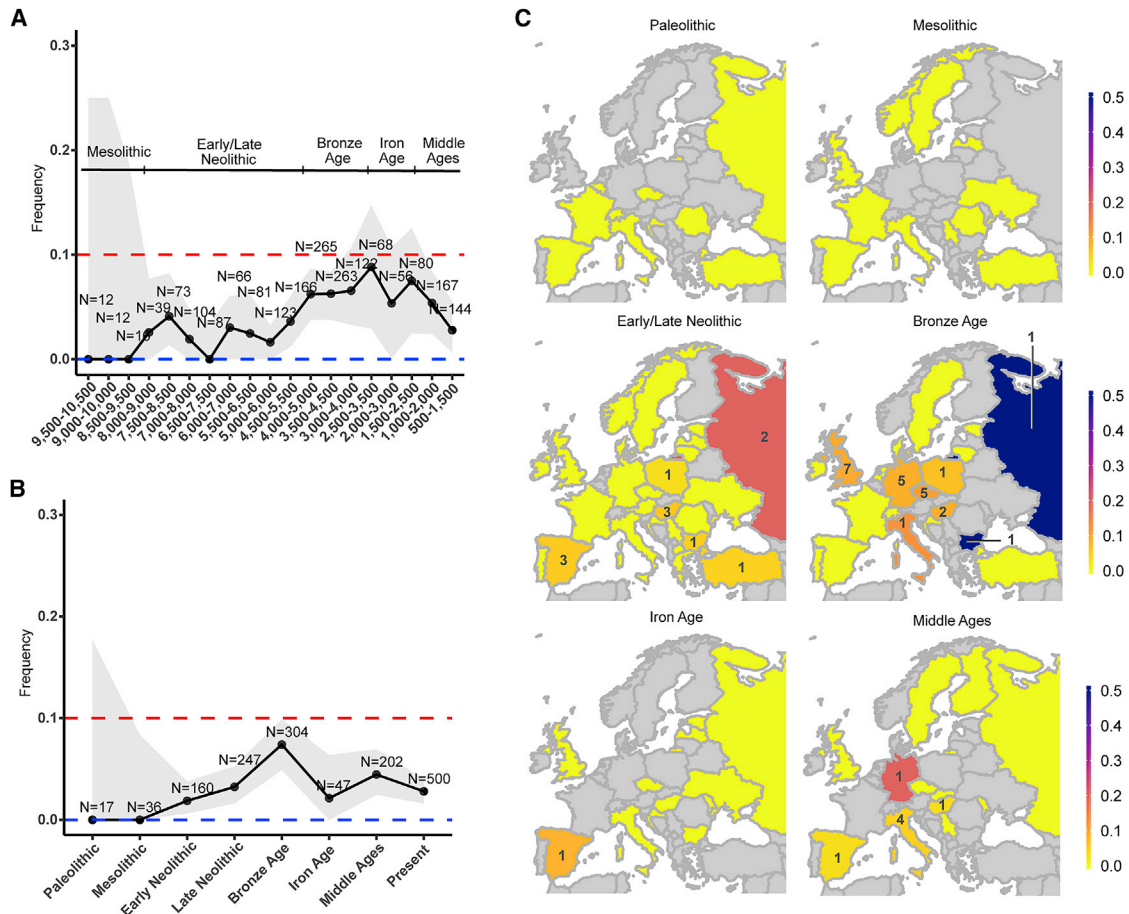
We therefore examined the frequency trajectory of P1104A over the last 10,000 years of European history, by screening a collection of 1,013 genomes that cover a time transect from the Mesolithic period to the Middle Ages (Figure 1A; Table S1). We partitioned the aDNA data into seven epochs and incorporated data from present-day populations (supplemental material and methods). The P1104A variant, which we found to be the result of a single mutational event (Figure S1), appeared for the first time in our dataset during the early Neolithic ~8,500 ya in the Anatolian peninsula, and then spread to Central Europe where it remained at frequencies lower than 3% until ~5,000 ya (Figures 1A–1C). During the Bronze Age, P1104A increased in frequency, reaching its maximum frequency ~3,000 ya at nearly 10%. After the Iron Age, we observed a strong and consistent decrease in frequency of P1104A, resulting in an average frequency of 2.9% among contemporary Europeans.<sup>24</sup>

We estimated the age of the *TYK2* P1104A mutation ( $T_{\text{age}}$ ), tested whether the mutation has been the substrate of natural selection, and inferred the onset ( $T_{\text{onset}}$ ) and strength ( $s$ ) of negative selection acting on homozygotes, using an approximate Bayesian computation (ABC) approach<sup>25</sup> that considers large prior assumptions ( $T_{\text{age}} \sim \mathcal{U}[8.5\text{--}100,000]$  ya,  $T_{\text{onset}} \sim \mathcal{U}[500\text{--}10,000]$  ya and  $s \sim \mathcal{U}[0\text{--}1]$ ; supplemental material and methods). We first determined the extent to which our approach could determine the evolutionary model of P1104A that best explains the observed aDNA data, by comparing the fit of the simulated to the observed data (supplemental material and methods). We assumed a validated demographic model for Europeans,<sup>26</sup> to which we added gene flow from both Near Easterners and Central Asians (Table S2), to account for the large-scale migrations of early farmer populations of the Anatolian plateau and Eurasian steppe populations associated with the Yamnaya culture inferred from

aDNA.<sup>27</sup> In doing so, considering the aforementioned large prior assumptions, we obtained simulated frequency trajectories that closely reproduce that of P1104A, similarly to other genome-wide variants (Figure S2). We also noted a similar, or higher, increase in frequency as that observed for P1104A until the Bronze Age for more than 20% of other aDNA variants within the uncertainty frequency interval of P1104A in the Mesolithic ([0.00–0.10]; Table S3), highlighting the marked impact of the aforementioned migratory events on the frequency of a large fraction of genomic variants, including P1104A. Furthermore, simulated neutral variants closely matched observed frequency distributions of non-coding variants for all epochs (Figure S3), indicating that the demographic model used—present-day Europeans are a mixture of Mesolithic hunter-gatherers, Anatolian Neolithic farmers, and Eastern steppe-related groups<sup>28,29</sup>—well reproduces the neutral patterns of European diversity.

We then estimated the origin of the *TYK2* P1104A mutation, based on its frequency in  $K = 12$  populations sampled at different epochs, including European aDNA data (Paleolithic, Mesolithic, Early Neolithic, Late Neolithic, Bronze Age, Iron Age, and Middle Ages; supplemental material and methods) and present-day Europeans, Middle Easterners, Central Asians (from 1% to 4%), Sub-Saharan Africans (0%), and East Asians (0%) (Figure 2A; Table S3). We found the age of P1104A to be ~30,000 years old (mode = 29,182; 95% CI [20,636–57,285]) (Figure 2B; supplemental material and methods), which is consistent with a previous estimate.<sup>30</sup> Using cross-validation, we found that parameter estimation was accurate across all ages, with 96% of 1,000 estimated 95% CIs including the true simulated value, and also robust to the choice of the summary statistics used (Figures S4A–S4D). While the 95% CI for the age of P1104A overlaps with the divergence time between West and East Eurasians (35–45 kya), the proportion of best-fitting simulated variants originating in the common ancestors of West Eurasians was significantly higher than that of the rest of simulated variants (OR = 7.00, 95% CI [5.70–8.53],  $p < 10^{-10}$ ; Figure 2B; supplemental material and methods). This suggests that P1104A originated in the common ancestors of West Eurasians after the split with East Eurasians, but before the divergence of Europeans, Middle Easterners, and Central Asians. Together, our results provide robust evidence that *TYK2* P1104A appeared during the Upper Paleolithic in West Eurasia, largely predating the estimated emergence of TB in Europe.<sup>18,19,31</sup>

We next investigated the evolutionary forces that could explain the frequency decrease of P1104A since the Bronze Age, where the maximum frequency is observed, by simulating frequency trajectories under neutrality ( $s = 0$ ) or negative selection ( $s > 0$ ) (Figure S5A). We found that simulations matching the estimated frequency of P1104A at the end of the Bronze Age explained both the observed aDNA and modern data only if  $s > 0.1$ . Furthermore, the frequency decrease after the Bronze Age was observed in



**Figure 1. Evolutionary history of the TB-associated *TYK2* P1104A variant**

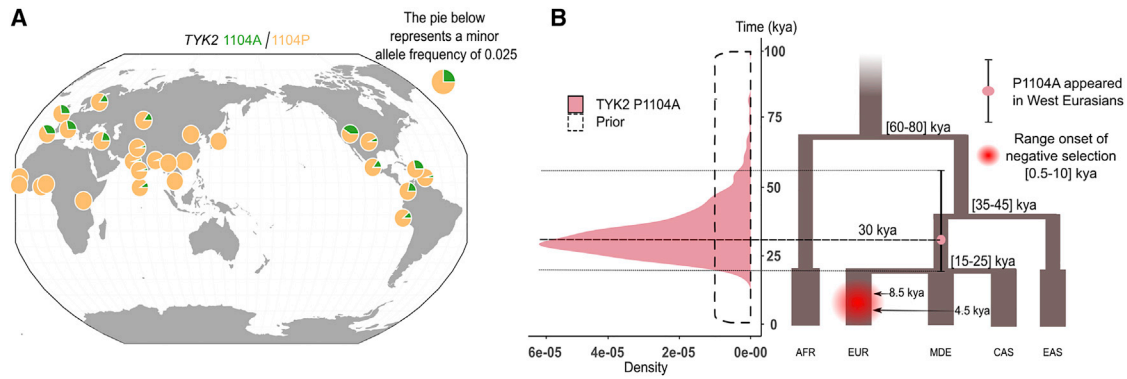
(A and B) European frequency trajectory for the *TYK2* P1104A variant over the last 10,000 years for (A) bins of 1,000 years and sliding windows of 500 years or (B) seven (pre-) historical European epochs and current times. The red and blue horizontal dashed lines indicate a frequency of 10% and 0%, respectively. Uncertainty of the frequency estimation is indicated by a gray colored area, representing the normal approximation of the 95% binomial proportion CI. Large uncertainty for older times is due to small sample sizes. For each bin, at least one carrier was assumed to obtain uncertainty estimates.

(C) Geographical distribution of the *TYK2* P1104A allele by country (using today's political borders), across all defined epochs. Colors indicate frequency estimations by country, from 0 (yellow) to 0.5 (blue). Grey indicates unavailable data. Number of P1104A carriers is indicated with its respective number on each country. Sample sizes for countries with non-zero counts (Table S1) are the following: Early/Late Neolithic: Austria (n = 7), Bulgaria (n = 21), Croatia (n = 10), Czech Republic (n = 8), Denmark (n = 1), Estonia (n = 1), France (n = 4), Germany (n = 13), Greece (n = 9), Hungary (n = 51), Ireland (n = 4), Italy (n = 11), Latvia (n = 20), Lithuania (n = 8), Luxembourg (n = 1), Macedonia (n = 1), Norway (n = 1), Poland (n = 32), Portugal (n = 11), Romania (n = 3), Russia (n = 10), Serbia (n = 14), Spain (n = 57), Sweden (n = 11), Turkey (n = 22), UK (n = 46), Ukraine (n = 27); Bronze Age: Bulgaria (n = 2), Croatia (n = 2), Czech Republic (n = 46), Denmark (n = 2), Estonia (n = 7), France (n = 6), Germany (n = 58), Hungary (n = 17), Ireland (n = 1), Italy (n = 8), Lithuania (n = 4), the Netherlands (n = 10), Poland (n = 15), Portugal (n = 2), Russia (n = 2), Spain (n = 33), Sweden (n = 7), Switzerland (n = 1), Turkey (n = 5), UK (n = 75); Iron Age: Bulgaria (n = 1), Croatia (n = 1), Czech Republic (n = 1), Estonia (n = 3), Hungary (n = 5), Italy (n = 6), Latvia (n = 8), Moldova (n = 4), Russia (n = 2), Spain (n = 12), UK (n = 1); Middle Ages: Czech Republic (n = 1), Finland (n = 4), Germany (n = 5), Hungary (n = 30), Iceland (n = 9), Italy (n = 89), Moldova (n = 2), Russia (n = 3), Serbia (n = 1), Slovakia (n = 1), Spain (n = 32), Sweden (n = 13), UK (n = 12).

the trajectories of 25% of the best fitting simulated deleterious variants ( $s \sim \mathcal{U}[0-1]$  and  $T_{\text{onset}} \sim \mathcal{U}[500-10,000]$ ; [supplemental material and methods](#)), relative to only 1% of the best fitting simulated neutral variants (OR = 33, 95% CI = [5–240],  $p < 10^{-10}$ ; [Figure S5B](#); [Table S4](#)). These observations collectively support a history of negative selection driving the evolution of the TB-risk P1104A variant after the Bronze Age.

To quantify the degree of deleteriousness of *TYK2* P1104A during European history, we verified that allele frequency trajectories were informative to assess negative selection, and,

encouragingly, we observed a strong positive correlation between drops in allele frequencies and  $s$  values ([Figure S6A](#)). We first hypothesized that negative selection started with the arrival of agriculture in Europe,<sup>20</sup> a period that includes the upper bound estimation for the most recent common ancestor of the *M. tuberculosis* complex  $\sim 6,000$  ya.<sup>18,19</sup> However, such an early onset of selection ( $T_{\text{onset}} = 10,000$ ) was clearly rejected by our simulations (Hotelling's T-squared test  $p = 5.4 \times 10^{-4}$ ; [Figure S6B](#); [supplemental material and methods](#); [Table S4](#)), as no simulated variants were able to reproduce the frequency increase of P1104A until the Bronze



**Figure 2. Present-day geographic distribution and age estimation of the *TYK2* P1104A mutation**

(A) Frequencies from present-day Europeans (EUR;  $f = 0.03$ ;  $n = 503$ ), sub-Saharan Africans (AFR;  $f = 0$ ;  $n = 504$ ), Americans (AMR;  $f = 0.01$ ;  $n = 347$ ), Middle Easterners (MDE;  $f = 0.02$ ;  $n = 163$ ), and Central Asians (CAS;  $f = 0.01$ ;  $n = 363$ ) are shown (supplemental material and methods). Figure was built with 1000 Genomes Project data<sup>24</sup> and modified to include Middle Easterners and Central Asians. The presence of *TYK2* P1104A among American populations from the 1000 Genomes Project reflects recent admixture with Europeans,<sup>24</sup> with all populations sharing a unique 6 kb-long haplotype around *TYK2* P1104A, although the allele is absent from Native Americans.

(B) Left panel: posterior distribution for the age (in thousands of years) of the *TYK2* P1104A mutation, according to the best fitting simulations with variable onset of selection, using 10,000,000 simulations and all available summary statistics. CI boundaries are shown with dashed black lines. Right panel: the proposed demographic model, showing the point estimate for  $T_{\text{age}}$  (mode = 30 kya, purple red circle) and the 95% confidence interval (black vertical segment across the purple circle) for the age of the mutation.

Age. Conversely, when allowing the onset of selection to vary across the last 10,000 years, using the former large priors ( $T_{\text{onset}} \sim \mathcal{U}[500-10,000]$  ya and  $s \sim \mathcal{U}[0-1]$ ), our best simulations did not significantly differ from P1104A (i.e., the simulation set was not rejected; Hotelling's T-squared test  $p = 0.09$ ) and revealed that scenarios with recent onsets of negative selection were those best fitting the data (Figure S6B).

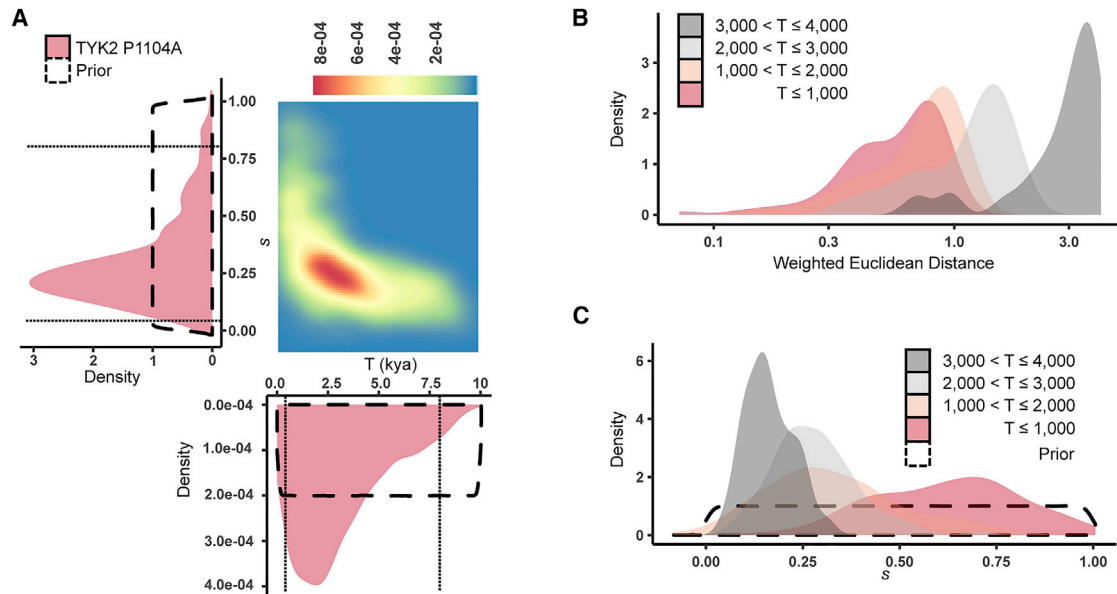
To explain the strongest frequency increase and decrease for P1104A, we modeled allele frequencies of  $K = 5$  ancient populations (Late Neolithic, Bronze Age, Iron Age, and Middle Ages) and present-day Europeans, and assumed large priors for model parameters (supplemental material and methods, Table S3). We found that negative selection on P1104A homozygous carriers started 1,937 ya (95% CI [500–7,912]), with a selection coefficient of 0.21 (95% CI [0.06–0.82]) (Figures 3A–3C). This onset of selection is consistent with a neutral evolution for the allele until the Bronze Age, suggesting that drift and admixture are sufficient to explain the increase of P1104A frequency until this epoch. These estimations should not be biased owing to read mapping bias of the reference allele in the ancient genome dataset,<sup>32</sup> given that 1104A is the alternative allele (supplemental material and methods). Furthermore, parameter estimation was found to be robust to the choice of the summary statistics used, with the 95% CIs of the estimates including the true simulated value 93% of the time (Figures S6C and S6D). Although our analysis showed that the more recent the onset of selection was the closer the frequency trajectory estimation was to the empirical data (Figure S6A), the fit was found to be similar within the last ~2,000 years (Figure 3B), consistent with our estimation. With respect to the selection coefficient, the posterior distributions of  $s$  were shifted to 1 as  $T_{\text{onset}}$  became closer to 0, and the general posterior distribution

for the strength of negative selection was similar to that of onsets of selection occurring between 1,000 and 3,000 ya (Figures 3A and 3C). Importantly, consistent ABC estimates of the strength and the onset of selection were found when either excluding the Iron Age, i.e., the epoch with smallest sample size ( $s = 0.19$ ; 95% = [0.03–0.83];  $T_{\text{onset}} = 1,670$  ya; 95% CI = [500–8,388] ya) or when using the whole European frequency trajectory, i.e., from the Paleolithic to the present ( $s = 0.21$ ; 95% = [0.04–0.84];  $T_{\text{onset}} = 1,567$  ya; 95% CI = [500–8,367]).

Using the same approach, we estimated the selection coefficient of another mutation, *TYK2* I684S, a missense variant that is neither in linkage disequilibrium with P1104A nor associated with TB risk,<sup>11</sup> and found values that were compatible with neutrality ( $s = 0.02$ ; 95% CI [0–0.19]; Figures S7A and S7B). Thus, our analyses support the notion that, despite the reported protective effects of P1104A against some immune-related disorders,<sup>22,23</sup> TB has exerted pressure on the *TYK2* P1104A variant over the last ~2,000 years, with a 20% relative fitness reduction for homozygotes at each generation since.

Finally, we sought to apply the same approach to reported pathogenic variants, by cross-matching the ClinVar database<sup>33</sup> with aDNA variants present in our cohort that fall into the uncertainty range of P1104A in the Bronze Age ([0.04–0.10], Figure 1B). Among the resulting three variants with a “pathogenic” clinical significance annotation, only one (*HFE* C282Y [MIM: 613609]) presents a frequency decrease across the last four epochs. *HFE* C282Y is a known disease-causing variant underlying hemochromatosis, an autosomal-recessive autoimmune disease (*HFE1* [MIM: 235200]) that impairs mineral metabolism, which can affect the growth and clearance of intra- and extra-cellular pathogens.<sup>34</sup> *HFE* C282Y reached its maximum frequency, of





**Figure 3. Estimation of the onset and strength of negative selection driving the evolution of *TYK2* P1104A**

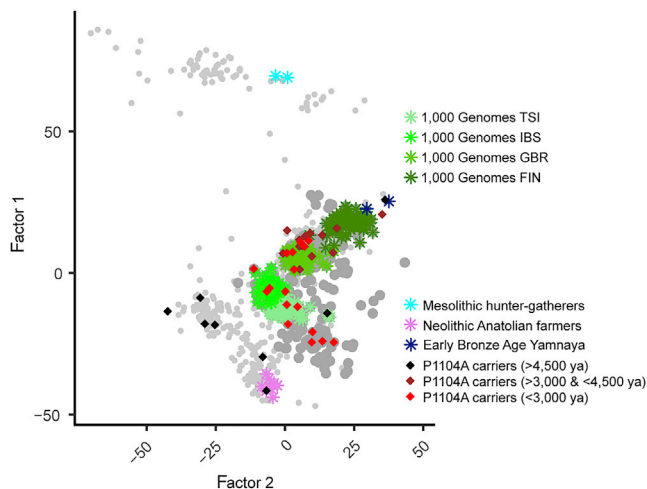
(A) Joint (as a 2D density plot) and individual (as purple densities) posterior distributions for the onset (in thousands of years) and the strength of negative selection ( $s$ ) for *TYK2* P1104A, based on the best fitting simulations with variable onset of selection, using European summary statistics from the Late Neolithic epoch onward (10,000,000 simulations). CI boundaries are shown with dashed black lines. (B) Distribution of the weighted Euclidean distances between the best fitting simulations and the observed data, under the proposed demographic model, for (from right to left)  $3,000 < T_{\text{onset}} \leq 4,000$ ,  $2,000 < T_{\text{onset}} \leq 3,000$ ,  $1,000 < T_{\text{onset}} \leq 2,000$ , or  $500 < T_{\text{onset}} \leq 1,000$ . (C) Posterior distributions for the *TYK2* P1104A's negative selection coefficient, based on the best fitting simulations with variable onset of selection, for the same groups of onsets of selection as in (B) using the same color code.

nearly 10%, during the Middle Ages and then decreased to its present-day frequency of 4%. Consistent with our expectations, we found a similarly strong selection coefficient of 0.20 (mode = 0.22; 95% CI [0.03–0.76]; Figure S7B), and an onset of negative selection during the Middle Ages (mode = 724 ya; 95% CI [500–7,508]).

A potential limitation of our approach, which is inherent to most aDNA studies, is genetic discontinuity due to large population replacements or to sampling bias for geographical locations.<sup>35</sup> For example, different sampling proportions from northern and southern Europeans across epochs may result in genetic discontinuity in our dataset, given that the former present higher Eastern steppe ancestry than the latter after the Bronze Age.<sup>36</sup> We thus repeated our ABC setup for northern and southern Europeans using a geographical division,<sup>37</sup> designed to distinguish high and low levels of Steppe ancestry (Figure S8). Despite much lower sample sizes, we found evidence for negative selection in both northern ( $s = 0.24$ ; 95% CI: [0.02–0.87]) and southern ( $s = 0.13$ ; 95% CI: [ $1.6 \times 10^{-4}$ –0.81]) European homozygotes, with a slightly left-shifted posterior distribution in southern Europe, where the sample size is more limited (Figure S9). We also found, using factor analysis,<sup>38</sup> that P1104A carriers scattered throughout European sub-structured populations, across all epochs after its introduction to Europe (Figure 4).

In addition, ancestry proportions were similar between P1104A carriers and the rest of the dataset at each epoch (Table S1). Notably, the observed ancestry shift between Bronze Age and present-day samples (from 0.29 to 0.36 for the whole

dataset [Table S3], representing a 24% relative increase, and from 0.23 to 0.39 for P1104A carriers [Table S1]) does not, on its own, explain the frequency decline of the allele after the Bronze Age (from 0.074 to 0.029, representing a 61% relative decrease). Yet, we performed an ABC estimation accounting for ancestry variation across epochs (supplemental material and methods). Using the estimated Anatolian ancestry of our dataset at each epoch from the Late Neolithic onward, we estimated very similar values for the strength and onset of negative selection for *TYK2* P1104A at the pan-European level ( $s = 0.27$ ; 95% CI: [0.08–0.93];  $T_{\text{onset}} = 2,045$  ya; 95% CI [500–8,690]; Figures S10A and S10B). Similarly, we found comparable estimations for northern and southern Europeans ( $s = 0.26$ ; 95% CI [0.06–0.83];  $T_{\text{onset}} = 1,046$  ya; 95% CI [500–6,934]; Figures S10C and S10D; and  $s = 0.24$ ; 95% CI [0.02–0.85];  $T_{\text{onset}} = 3,229$  ya; 95% CI [500–8,963]; Figures S10E and S10F, respectively). Conversely, we found no evidence of selection for *TYK2* I684S ( $s = 0.02$ ; 95% CI: [0–0.69]), as expected, and a weaker signal of negative selection for *HFE* C282Y ( $s = 0.12$ ; 95% CI: [0–0.76]). Collectively, these findings suggest that the observed frequency drop of P1104A after the Bronze Age is not due to major geographical and/or temporal differences in ancestry components in our aDNA dataset, but instead to the action of natural selection. Moreover, when re-estimating the age of P1104A without modern data from Middle Easterners and Central Asians, as they are not entirely representative of ancestral Anatolian farmers and steppe herders, respectively,<sup>39,40</sup> we obtained almost identical results (mode = 30,303 ya; 95% CI [23,113–60,273]).



**Figure 4. TYK2 P1104A carriers scatter throughout the genetic diversity of the dataset**

Factor analysis (Factor 2 versus Factor 1) for 701 high-quality ancient genomes from the full set of 1,013 used in this work, and 363 pseudo-haploid present-day genomes (TSI, IBS, GBR, and FIN, green colors from lighter to darker, respectively) from various European populations from the 1000 Genomes Project (supplemental material and methods). Two Mesolithic hunter-gatherers (cyan), eight Neolithic Anatolian farmers (violet), and two Early Bronze Age individuals associated with the Yamnaya (>80% steppe ancestry) culture (blue) account for the three major ancestries existing in present-day Europeans, which are, in turn, correlated with their respective epochs. P1104A carriers are shown with black (>4,500 ya), brown (>3,000 ya and <4,500 ya), or red (<3,000 ya) diamonds. Other individuals, older (light) or younger (dark) than 3,000 ya, are represented by gray dots.

In this attempt to define the mode and tempo of TB pressure on Europeans, we found that the only common variant known to underlie monogenic predisposition to TB has evolved under strong negative selection in Europe after the Iron Age. In doing so, we provide population genetic evidence for the high burden of TB in Europeans over the last two millennia, in line with the dating of *M. tuberculosis* lineage 4 at 1,943 ya<sup>41</sup> and of strains found in 18<sup>th</sup> century Hungarian mummies at 1,604 ya, or in mummified remains of the 17<sup>th</sup> century Bishop Peder Winstrup of Lund between 929 and 2,084 ya.<sup>19,31</sup> Notably, the TB-associated mutation ranks among the top 2.7% of variants, present in the studied capture array, with similar frequencies in the Bronze Age (0.04–0.10) that have decreased the most since this period (Table S3; supplemental material and methods). Such variants might also include targets of negative selection (Table S5). A selection coefficient of 0.20 would entail >2,500,000 cumulative deaths over the last 2,000 years due to P1104A homozygosity, representing 1%–2% of all TB-related deaths in the 19<sup>th</sup> century Europe (Figure S11). This figure is consistent with a previous estimation of 1% of TB cases due to the at-risk genotype among present-day Europeans.<sup>7</sup> We anticipate that the same population genetics framework could be used to delineate other human genetic variants, of yet unknown function, that have drastically decreased or increased in frequency across time due to microbial pressure. Thus, adopting an evolutionary

approach represents a promising alternative to investigate the genetic sources of present-day disparities, between individuals and populations, in susceptibility to infection.

## Data and code availability

Pseudo-haploid ancient and modern genome data are available at <https://reich.hms.harvard.edu/allen-ancient-dna-resource-aadr-downloadable-genotypes-present-day-and-ancient-dna-data> (V42.4: March 1, 2020 release). Code to perform ABC estimations from simulated frequency data are available at [https://github.com/h-e-g/SLiM\\_aDNA\\_selection](https://github.com/h-e-g/SLiM_aDNA_selection).

## Supplemental Information

Supplemental Data can be found online at <https://doi.org/10.1016/j.ajhg.2021.02.009>.

## Acknowledgments

We thank David Reich for helpful comments on an earlier version of the manuscript. We also thank Guillaume Achaz, Jérémy Choin, Evelyne Heyer, Nina Marchi, Anna-Lena Neehus, Olivier Neyrolles, and Mary O'Neill for data sharing and discussions. The Laboratory of Human Genetics of Infectious Diseases is supported by the Howard Hughes Medical Institute, the Rockefeller University, the St. Giles Foundation, the National Institutes of Health (NIH) (R01AI088364), the Meyer Foundation, the Agence Nationale de la Recherche (ANR) under the “Investments for the Future” program (ANR-10-IAHU-01), the Fondation pour la Recherche Médicale (FRM) (EQU201903007798), *Institut National de la Santé et de la Recherche Médicale* (INSERM), and the University of Paris. The laboratory of Human Evolutionary Genetics is supported by the Institut Pasteur, the Collège de France, the Centre Nationale de la Recherche Scientifique (CNRS), the Agence Nationale de la Recherche (ANR) grants LIFECHANGE (ANR 17 CE12 0018 02) and CNSVIRGEN (ANR-19-CE15-0009-02), the French Government’s Investissement d’Avenir program, Laboratoires d’Excellence “Integrative Biology of Emerging Infectious Diseases” (ANR-10-LABX-62-IBEID) and “Milieu Intérieur” (ANR-10-LABX-69-01), the Fondation pour la Recherche Médicale (Equipe FRM DEQ20180339214), the Fondation Allianz-Institut de France, and the Fondation de France (n°00106080). G.K. was supported by the Imagine Institute with the grant “Imagine Thesis Award.”

## Declaration of Interests

The authors declare no competing interests.

Received: October 12, 2020

Accepted: February 5, 2021

Published: March 4, 2021

## Web resources

OMIM, <https://www.omim.org/>

## References

1. Anderson, R.M., May, R.M., and Anderson, B. (1992). *Infectious Diseases of Humans: Dynamics and Control* (Oxford: Oxford University Press).
2. Cairns, J., and Singer, A.L. (1997). *Matters Of Life And Death: Perspectives On Public Health, Molecular Biology, Cancer, And The Prospects For The Human Race* (Diane Pub Co).
3. Casanova, J.-L., and Abel, L. (2018). Human genetics of infectious diseases: Unique insights into immunological redundancy. *Semin. Immunol.* *36*, 1–12.
4. Karlsson, E.K., Kwiatkowski, D.P., and Sabeti, P.C. (2014). Natural selection and infectious disease in human populations. *Nat. Rev. Genet.* *15*, 379–393.
5. Quintana-Murci, L. (2019). Human Immunology through the Lens of Evolutionary Genetics. *Cell* *177*, 184–199.
6. Paulson, T. (2013). Epidemiology: A mortal foe. *Nature* *502*, S2–S3.
7. Kerner, G., Ramirez-Alejo, N., Seeleuthner, Y., Yang, R., Ogishi, M., Cobat, A., Patin, E., Quintana-Murci, L., Boisson-Dupuis, S., Casanova, J.-L., and Abel, L. (2019). Homozygosity for *TYK2* P1104A underlies tuberculosis in about 1% of patients in a cohort of European ancestry. *Proc. Natl. Acad. Sci. USA* *116*, 10430–10434.
8. Furin, J., Cox, H., and Pai, M. (2019). Tuberculosis. *Lancet* *393*, 1642–1656.
9. Boisson-Dupuis, S., Bustamante, J., El-Baghdadi, J., Camcioglu, Y., Parvaneh, N., El Azbaoui, S., Agader, A., Hassani, A., El Hafidi, N., Mrani, N.A., et al. (2015). Inherited and acquired immunodeficiencies underlying tuberculosis in childhood. *Immunol. Rev.* *264*, 103–120.
10. Abel, L., Fellay, J., Haas, D.W., Schurr, E., Srikrishna, G., Urbanowski, M., Chaturvedi, N., Srinivasan, S., Johnson, D.H., and Bishai, W.R. (2018). Genetics of human susceptibility to active and latent tuberculosis: present knowledge and future perspectives. *Lancet Infect. Dis.* *18*, e64–e75.
11. Boisson-Dupuis, S., Ramirez-Alejo, N., Li, Z., Patin, E., Rao, G., Kerner, G., Lim, C.K., Krementsov, D.N., Hernandez, N., Ma, C.S., et al. (2018). Tuberculosis and impaired IL-23-dependent IFN- $\gamma$  immunity in humans homozygous for a common *TYK2* missense variant. *Sci. Immunol.* *3*, 3.
12. Quintana-Murci, L., and Clark, A.G. (2013). Population genetic tools for dissecting innate immunity in humans. *Nat. Rev. Immunol.* *13*, 280–293.
13. Kwiatkowski, D.P. (2005). How malaria has affected the human genome and what human genetics can teach us about malaria. *Am. J. Hum. Genet.* *77*, 171–192.
14. Laval, G., Peyr egne, S., Zidane, N., Harmant, C., Renaud, F., Patin, E., Prugnolle, F., and Quintana-Murci, L. (2019). Recent Adaptive Acquisition by African Rainforest Hunter-Gatherers of the Late Pleistocene Sick-Cell Mutation Suggests Past Differences in Malaria Exposure. *Am. J. Hum. Genet.* *104*, 553–561.
15. Louicharoen, C., Patin, E., Paul, R., Nuchprayoon, I., Witoonpanich, B., Peerapittayamongkol, C., Casademont, I., Sura, T., Laird, N.M., Singhasivanon, P., et al. (2009). Positively selected G6PD-Mahidol mutation reduces *Plasmodium vivax* density in Southeast Asians. *Science* *326*, 1546–1549.
16. Band, G., Rockett, K.A., Spencer, C.C., Kwiatkowski, D.P.; and Malaria Genomic Epidemiology Network (2015). A novel locus of resistance to severe malaria in a region of ancient balancing selection. *Nature* *526*, 253–257.
17. Shriner, D., and Rotimi, C.N. (2018). Whole-Genome-Sequence-Based Haplotypes Reveal Single Origin of the Sick-Cell Allele during the Holocene Wet Phase. *Am. J. Hum. Genet.* *102*, 547–556.
18. Bos, K.I., Harkins, K.M., Herbig, A., Coscolla, M., Weber, N., Comas, I., Forrest, S.A., Bryant, J.M., Harris, S.R., Schuene-mann, V.J., et al. (2014). Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. *Nature* *514*, 494–497.
19. Sabin, S., Herbig, A., V agene, A.J., Ahlstr om, T., Bozovic, G., Arcini, C., K uhnert, D., and Bos, K.I. (2020). A seventeenth-century *Mycobacterium tuberculosis* genome supports a Neolithic emergence of the *Mycobacterium tuberculosis* complex. *Genome Biol.* *21*, 201.
20. Wolfe, N.D., Dunavan, C.P., and Diamond, J. (2007). Origins of major human infectious diseases. *Nature* *447*, 279–283.
21. Mathieson, I. (2020). Human adaptation over the past 40,000 years. *Curr. Opin. Genet. Dev.* *62*, 97–104.
22. Jostins, L., Ripke, S., Weersma, R.K., Duerr, R.H., McGovern, D.P., Hui, K.Y., Lee, J.C., Schumm, L.P., Sharma, Y., Anderson, C.A., et al.; International IBD Genetics Consortium (IBDGC) (2012). Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* *491*, 119–124.
23. Beecham, A.H., Patsopoulos, N.A., Xifara, D.K., Davis, M.F., Kempainen, A., Cotsapas, C., Shah, T.S., Spencer, C., Booth, D., Goris, A., et al.; International Multiple Sclerosis Genetics Consortium (IMSGC); Wellcome Trust Case Control Consortium 2 (WTCCC2); and International IBD Genetics Consortium (IBDGC) (2013). Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nat. Genet.* *45*, 1353–1360.
24. Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., Abecasis, G.R.; and 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature* *526*, 68–74.
25. Beaumont, M.A., Zhang, W., and Balding, D.J. (2002). Approximate Bayesian computation in population genetics. *Genetics* *162*, 2025–2035.
26. Gravel, S., Henn, B.M., Gutenkunst, R.N., Indap, A.R., Marth, G.T., Clark, A.G., Yu, F., Gibbs, R.A., Bustamante, C.D.; and 1000 Genomes Project (2011). Demographic history and rare allele sharing among human populations. *Proc. Natl. Acad. Sci. USA* *108*, 11983–11988.
27. Skoglund, P., and Mathieson, I. (2018). Ancient Genomics of Modern Humans: The First Decade. *Annu. Rev. Genomics Hum. Genet.* *19*, 381–404.
28. Lazaridis, I., Patterson, N., Mittnik, A., Renaud, G., Mallick, S., Kirsanow, K., Sudmant, P.H., Schraiber, J.G., Castellano, S., Lipson, M., et al. (2014). Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* *513*, 409–413.
29. Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., Brandt, G., Nordenfelt, S., Harney, E., Stewardson, K., et al. (2015). Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* *522*, 207–211.
30. Albers, P.K., and McVean, G. (2020). Dating genomic variants and shared ancestry in population-scale sequencing data. *PLoS Biol.* *18*, e3000586.
31. Kay, G.L., Sergeant, M.J., Zhou, Z., Chan, J.Z.M., Millard, A., Quick, J., Szikossy, I., Pap, I., Spigelman, M., Loman, N.J., et al. (2015). Eighteenth-century genomes show that mixed

- infections were common at time of peak tuberculosis in Europe. *Nat. Commun.* *6*, 6717.
32. Günther, T., and Nettelblad, C. (2019). The presence and impact of reference bias on population genomic studies of pre-historic human populations. *PLoS Genet.* *15*, e1008302.
  33. Landrum, M.J., Lee, J.M., Benson, M., Brown, G.R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., et al. (2018). ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* *46* (D1), D1062–D1067.
  34. Miller, H.K., Schwiesow, L., Au-Yeung, W., and Auerbuch, V. (2016). Hereditary Hemochromatosis Predisposes Mice to *Yersinia pseudotuberculosis* Infection Even in the Absence of the Type III Secretion System. *Front. Cell. Infect. Microbiol.* *6*, 69.
  35. Silva, N.M., Rio, J., and Currat, M. (2017). Investigating population continuity with ancient DNA under a spatially explicit simulation framework. *BMC Genet.* *18*, 114.
  36. Olalde, I., Brace, S., Allentoft, M.E., Armit, I., Kristiansen, K., Booth, T., Rohland, N., Mallick, S., Szécsényi-Nagy, A., Mittnik, A., et al. (2018). The Beaker phenomenon and the genomic transformation of northwest Europe. *Nature* *555*, 190–196.
  37. Mathieson, S., and Mathieson, I. (2018). FADS1 and the Timing of Human Adaptation to Agriculture. *Mol. Biol. Evol.* *35*, 2957–2970.
  38. François, O., and Jay, F. (2020). Factor analysis of ancient population genomic samples. *Nat. Commun.* *11*, 4661.
  39. Narasimhan, V.M., Patterson, N., Moorjani, P., Rohland, N., Bernardos, R., Mallick, S., Lazaridis, I., Nakatsuka, N., Olalde, I., Lipson, M., et al. (2019). The formation of human populations in South and Central Asia. *Science* *365*, 365.
  40. Skourtanioti, E., Erdal, Y.S., Frangipane, M., Balossi Restelli, F., Yener, K.A., Pinnock, F., Matthiae, P., Özbal, R., Schoop, U.D., Guliyev, F., et al. (2020). Genomic History of Neolithic to Bronze Age Anatolia, Northern Levant, and Southern Caucasus. *Cell* *181*, 1158–1175.e28.
  41. O’Neill, M.B., Shockey, A., Zarley, A., Aylward, W., Eldholm, V., Kitchen, A., and Pepperell, C.S. (2019). Lineage specific histories of *Mycobacterium tuberculosis* dispersal in Africa and Eurasia. *Mol. Ecol.* *28*, 3241–3256.