

## Global spatial dynamics and vaccine-induced fitness changes of *Bordetella pertussis*

Lefrancq, Noémie; Bouchez, Valérie; Fernandes, Nadia; Barkoff, Alex-Mikael; Bosch, Thijs; Dalby, Tine; Åkerlund, Thomas; Darenberg, Jessica; Fabianova, Katerina; Vestrheim, Didrik F.; Fry, Norman K; González-López, Juan José; Gullsby, Karolina; Habington, Adele; He, Qiushui; Litt, David; Martini, Helena; Piérard, Denis; Stefanelli, Paola; Stegger, Marc; Zavadilova, Jana; Armatys, Nathalie; Landier, Annie; Guillot, Sophie; Hong, Samuel L; Lemey, Philippe; Parkhill, Julian; Toubiana, Julie; Cauchemez, Simon; Salje, Henrik; Brisse, Sylvain

*Published in:*  
Science Translational Medicine

*DOI:*  
[10.1126/scitranslmed.abn3253](https://doi.org/10.1126/scitranslmed.abn3253)

*Publication date:*  
2022

*License:*  
Unspecified

*Document Version:*  
Accepted author manuscript

[Link to publication](#)

### *Citation for published version (APA):*

Lefrancq, N., Bouchez, V., Fernandes, N., Barkoff, A.-M., Bosch, T., Dalby, T., Åkerlund, T., Darenberg, J., Fabianova, K., Vestrheim, D. F., Fry, N. K., González-López, J. J., Gullsby, K., Habington, A., He, Q., Litt, D., Martini, H., Piérard, D., Stefanelli, P., ... Brisse, S. (2022). Global spatial dynamics and vaccine-induced fitness changes of *Bordetella pertussis*. *Science Translational Medicine*, 14(642), [eabn3253].  
<https://doi.org/10.1126/scitranslmed.abn3253>

### Copyright

No part of this publication may be reproduced or transmitted in any form, without the prior written permission of the author(s) or other rights holders to whom publication rights have been transferred, unless permitted by a license attached to the publication (a Creative Commons license or other), or unless exceptions to copyright law apply.

# Global spatial dynamics and vaccine-induced fitness changes of *Bordetella pertussis*

Noémie Lefrancq<sup>1,2\*</sup> & Valérie Bouchez<sup>3,4\*</sup>, Nadia Fernandes<sup>3</sup>, Alex-Mikael Barkoff<sup>5</sup>, Thijs Bosch<sup>6</sup>, Tine Dalby<sup>7</sup>, Thomas Åkerlund<sup>8</sup>, Jessica Darenberg<sup>8</sup>, Katerina Fabianova<sup>9</sup>, Didrik F. Vestrheim<sup>10</sup>, Norman K. Fry<sup>11,12</sup>, Juan José González-López<sup>13,14</sup>, Karolina Gullsby<sup>15</sup>, Adele Habington<sup>16</sup>, Qiushui He<sup>5,17</sup>, David Litt<sup>11</sup>, Helena Martini<sup>18</sup>, Denis Piérard<sup>18</sup>, Paola Stefanelli<sup>19</sup>, Marc Stegger<sup>7</sup>, Jana Zavadilova<sup>20</sup>, Nathalie Armatys<sup>3,4</sup>, Annie Landier<sup>3,4</sup>, Sophie Guillot<sup>3,4</sup>, Samuel L. Hong<sup>21</sup>, Philippe Lemey<sup>21</sup>, Julian Parkhill<sup>22</sup>, Julie Toubiana<sup>3,4,23</sup>, Simon Cauchemez<sup>1</sup>, Henrik Salje<sup>1,2,#</sup> & Sylvain Brisse<sup>3,4,#</sup>

## Affiliations:

1. Institut Pasteur, Mathematical Modelling of Infectious Diseases Unit, UMR2000, CNRS, 75015 Paris, France
2. Department of Genetics, University of Cambridge, Cambridge CB2 3EH, UK
3. Institut Pasteur, Université de Paris, Biodiversity and Epidemiology of Bacterial Pathogens, 75724 Paris, France
4. National Reference Center for Whooping Cough and Other Bordetella Infections, 75724 Paris, France
5. University of Turku UTU, Institute of Biomedicine, Research Center for Infections and Immunity, FI-20520 Turku, Finland
6. Centre for Infectious Disease Control, National Institute for Public Health and the Environment (RIVM), 3720 BA Bilthoven, The Netherlands
7. Statens Serum Institut, Bacteria, Parasites and Fungi / Infectious Disease Preparedness, 2300 Copenhagen, Denmark
8. The Public Health Agency of Sweden, Unit for laboratory surveillance of bacterial pathogens, SE-171 82 Solna, Sweden
9. National Institute of Public Health, Department of Infectious Diseases Epidemiology, CZ-10000 Prague, Czech Republic
10. Norwegian Institute of Public Health, Department of Infectious Disease Control and Vaccine, N-0213 Oslo, Norway
11. Respiratory and Vaccine Preventable Bacteria Reference Unit, Public Health England – National Infection Service, London NW9 5EQ, UK
12. Immunisation and Countermeasures Division, Public Health England – National Infection Service, London NW9 5EQ, UK
13. University Hospital Vall d'Hebron, Microbiology Department, 08035 Barcelona, Spain
14. Universitat Autònoma de Barcelona, Department of Genetics and Microbiology, 08193 Barcelona, Spain
15. Centre for Research and Development, Uppsala University/Region Gävleborg, 80187 Gävle, Sweden
16. Molecular Microbiology Laboratory, Children's Health Ireland, Crumlin, D12 N512 Dublin, Ireland
17. InFLAMES Research Flagship Center, University of Turku, FI-20520 Turku, Finland

18. Department of Microbiology, National Reference Centre for *Bordetella pertussis*, Universitair Ziekenhuis Brussel, Vrije Universiteit Brussel (VUB), B-1090 Brussels, Belgium
19. Department of Infectious Diseases, Istituto Superiore di Sanità, IT-00161 Rome, Italy
20. National Institute of Public Health, National Reference Laboratory for Pertussis and Diphtheria, 100 00 Prague, Czech Republic
21. Department of Microbiology, Immunology and Transplantation, Rega Institute, KU Leuven, 3000 Leuven, Belgium
22. Department of Veterinary Medicine, University of Cambridge, Cambridge CB3 0ES, UK
23. Université de Paris, Department of general paediatrics and paediatric infectious diseases, Necker - Enfants malades Hospital, APHP, 75015 Paris, France

\* These authors contributed equally to this work

# These authors co-supervised this work

Correspondence to:

Dr. Sylvain Brisse: email: [sylvain.brisse@pasteur.fr](mailto:sylvain.brisse@pasteur.fr)

Dr. Henrik Salje: email: [hs743@cam.ac.uk](mailto:hs743@cam.ac.uk)

**Overline:**

*Bordetella pertussis* fitness and spread.

**One Sentence Summary:**

Genomic analyses of *Bordetella pertussis* reveal spatial structure, rate of geographic spread, and fitness shifts in response to vaccine changes.

## **Abstract**

As with other pathogens, competitive interactions between *Bordetella pertussis* strains drive infection risk. Vaccines are thought to perturb strain diversity through shifts in immune pressures, however, this has rarely been measured due to inadequate data and analytical tools. We used 3344 sequences from 23 countries to show that on average there are 28.1 transmission chains circulating within a subnational region, with the number of chains strongly associated with host population size. We estimated it takes 5-10 years for *B. pertussis* to be homogeneously distributed throughout Europe, with the same time-frame required for the United States. Increased fitness of pertactin-deficient strains following implementation of acellular vaccines, but reduced fitness otherwise, can explain long-term genotype dynamics. These findings highlight the role of national vaccine policy in shifting local diversity of a pathogen that is responsible for 160,000 deaths annually.

## Introduction

The role of local population immunity in driving ecological interactions between strains from the same disease system remains a foundational question of infectious disease dynamics and ecology. However, immunity's role in determining genetic diversity has rarely been explored as doing so requires long-term genetic data as well as perturbations in population immunity. Appropriate analytical tools that can quantify changes in lineage fitness are also necessary. *Bordetella pertussis* provides a rare natural experiment in this regard as many countries have switched from whole cell vaccines (WCV) to acellular vaccines (ACV), which have different immunogenic properties(1, 2).

*B. pertussis* is the causative agent of whooping cough, responsible for an estimated 160,700 deaths each year(3). Despite long-standing immunization programs and high vaccine coverage, *B. pertussis* continues to circulate endemically, including with increased transmission in recent decades, the reasons for which remain unclear(4). As with many bacterial systems, the study of *B. pertussis* is complicated by frequent asymptomatic carriage and long-term endemic co-circulation of multiple lineages(5–10). National surveillance systems have been used to reconstruct the evolutionary and spatial history of *B. pertussis*(11–16). There have been major changes in the global patterns of circulating genotypes over the last 50 years, particularly after the implementation of vaccination, including in antigen-coding genes coding for pertussis toxin, fimbriae, and pertactin(17). The changes in circulating genotypes appeared in different continents and appear consistent with selective sweeps. Understanding the drivers of strain dynamics has important public health implications as different genotypes have been linked to differences in virulence, and with the duration and efficacy of vaccine protection(18–20).

Using 3344 whole-genome sequences from 23 countries, we characterized the diversity of *B. pertussis* across different spatial scales (within-district, within-country, intra-continental and global), estimated its rate of expansion, and assessed the importance of the local population size in driving the diversity of circulating lineages. We separately developed an analytical framework that estimated the relative fitness of different lineages and characterized fitness changes following switches in the local vaccine being used. Our analytical approach allowed us to investigate *B. pertussis* diversity within and between locations, long-term genotype dynamics, and a potential role for national vaccine policy in driving genotype changes.

## Results

### Temporal dynamics of *B. pertussis* genetic diversity

We established a consortium of national reference laboratories to sequence the whole genome of 1,331 isolates from 12 European countries. Along with publicly available genomes, this generated a dataset of 3,344 genomes from 23 countries and 5 continents covering an 85-year period (Figure 1A, Table S1-2). Subnational location was recorded for 97% of genomes. To characterize genetic diversity, we built a time-resolved phylogeny (Figure 1B-E). In addition, we identified the genotype of each isolate based on the allele of the promoter region of pertussis toxin (PT, with promoter forms *ptxP1* or *ptxP3*) and type 3 fimbrial protein gene (*fim3-1* or *fim3-2*)(18, 21). PT is the major toxin produced by *B. pertussis* and is responsible for most of the systemic symptoms associated with pertussis disease(22). Isolates with a *ptxP3*-type promoter have been suggested to produce more PT than those with a *ptxP1*-type promoter(18). Fim3 is one of the two fimbriae produced by *B. pertussis* and is involved in adhesion to host cells(23). A polymorphism of the *fim3* gene has been reported to have occurred mainly for isolates with a *ptxP3*-type promoter(17). In addition to these genotypes, we also identified whether each isolate was capable of producing the immunogenic surface protein pertactin (PRN), a key target of vaccine-induced immunity(24). We found that in nearly all countries, the proportion of *ptxP3* isolates increased from 5-20% in the early 1990s to >80% by mid-2000s (Figure 1F). By contrast, *fim3* allele distribution has been more steady over the years (Figure 1G). Last, PRN-deficient strains have displaced PRN-positive strains in most countries (Figure 1H). These findings are consistent with previous efforts that identified changing patterns of genotype diversity(17).

### *B. pertussis* spread across spatial scales

To track the spatial spread of *B. pertussis*, we compared the evolutionary distance of pairs of sequences with their spatial distance. We found a strong log-linear relationship for up to five years, with consistent patterns observed in different countries (Figure 1I). Adjusting for uneven sampling by location and year, we estimated that pairs of isolates separated by a year of evolutionary time from within the same country were separated by an average of 1.9km (95%CI: 0.9-3.2), rising to 41.1km (95%CI: 27.9-53.5) when separated by 5 years (Figure S1), with consistent patterns obtained with alternative approaches for the definition of population centroids (Figure S2).

Recent findings from household transmission studies have identified frequent subclinical transmission of *B. pertussis*(10, 25). Consistent with such widespread subclinical transmission maintaining strain diversity, we found that despite strong overall patterns of spatial structure in sequences, the majority of sequence pairs within any location at any time were only distantly

related. At both national and subnational levels, fewer than one in twenty sequence pairs coming from the same year had a Most Recent Common Ancestor (MRCA) within the prior two years (Figure 2A, Figure S3), with the vast majority of pairs separated by more than 20 years of evolutionary time.

To further explore lineage diversity at subnational levels, we considered pairs of individuals isolated within the same year and having a MRCA of under 2 years to be part of the same transmission chain. On average 3.7% of pairs (95%CI: 3.0-4.4) within each region (defined here as the smallest subnational administrative unit we have for each country; average population size: 6.5 million, average area: 110,000 km<sup>2</sup>) were from the same transmission chain (Figure 2B). There were an average of 28.1 discrete chains (95%CI: 22.5-32.9) circulating within a region. However, the number of chains depended strongly on population size (Figure 2C); each increase in log-population size in the region was associated with a 0.55 (95%CI: 0.38-0.74) increase in the number of circulating chains, with consistent patterns across the different continents (Figure S4). Increases in population density were also associated with small increases in the number of circulating chains but the area of the region was not (Figure S5). These findings suggest a role for density-dependent transmission in *B. pertussis* spread, with increased transmission in areas with high population density (26).

To characterize the longer-term spread of lineages, we calculated the probability that pairs from the same location were separated by an MRCA of specific time intervals, relative to pairs that come from different locations. In France, where data had the greatest spatial resolution, sequences from the same district were 2.2 times (95%CI: 1.4-3.2) as likely to have an MRCA within the two prior years as sequence pairs from different districts (Figure 2D). We found similar spatial clustering in other countries (Figure S6). Further, pairs of sequences coming from the same European country were 12.9 times (95%CI: 5.8-28.0) as likely to have an MRCA within the prior year than sequence-pairs coming from different European countries. We observed a similar pattern of spatial dependence between different US states as between European countries (Figure 2E). Overall, it took approximately 3-4 years for a lineage to be well mixed within a European country (where there is no more spatial dependence between cases). Similarly, it took *B. pertussis* 5-9 years to be well mixed across Europe, 4-7 years to be well mixed across the US, and 8-11 years to be well mixed across different continents (Figure 2D-F, Figure S7). As we considered the relative probability of MRCA within and between locations for isolates sampled from the same year, our results are not affected by biased sampling between locations or in time(26).

### Characterization of the fitness of different genotypes by vaccination era

To explain the observed changes in genotype in each location, we developed a time-series model that estimates the annual proportion of isolates that were of each genotype in each country. We used a single framework to fit the data from all countries in parallel. We assumed that the relative fitness of the different genotypes was the same across countries but allowed for a change in fitness following the switch from WCV to ACV, which occurred at different times in different countries (Figure S8). This simple model was able to recover the time-series (Figure 3A-D, Figure S9), and the observed number (slope = 0.98,  $R^2$  = 0.99, p-value < 2.2e-16, linear regression), overall proportion (slope = 0.76,  $R^2$  = 0.77, p-value = < 2.2e-16), and annual relative change (slope = 0.61,  $R^2$  = 0.59, p-value < 2.2e-16) in genotypes in each country (Figure 3E-G).

We found that the underlying fitness of a genotype was associated with whether WCV or ACV was being used in that country at that time. *ptxP3*/PRN-positive types, irrespective of the *fim3* allele, were the most fit genotypes during the WCV era (relative fitness of 1.11 [95%CI: 1.06-1.18] per year compared to other genotypes circulating at that time, equivalent to a relative fitness of 1.01 [95%CI: 1.00-1.01] per transmission generation). The *ptxP3/fim3-1*/PRN-deficient genotype was the most fit genotype after ACV implementation (relative fitness of 1.30 per year [95%CI: 1.16-1.52], equivalent to a relative fitness of 1.02 per transmission generation [95%CI: 1.01-1.03]) (Figure 4A). We found that the switch from WCV to ACV was associated with an increased fitness of PRN-deficient genotypes, but only in isolates with a *ptxP3* background, with the effect further increased for *fim3-1* genotypes (Figure 4B). On average, PRN-deficient isolates were 1.26 (95%CI: 1.21-1.32) times as fit as PRN-positive isolates after the introduction of ACV (Figure 4C). Where only WCV was used, PRN-deficient isolates were on average 0.94 times (95%CI 0.92-0.95) as fit as their PRN-positive counterparts (Figure 4C). Alternative models that did not allow for changes in fitness following vaccine switch or use a common date across countries for changes in fitness had lower adequacy as measured by Watanabe–Akaike information criterion (Table S2)(27). Estimates for those models were highly consistent (Figure S10). Models that allowed for a delay between ACV implementation and fitness change also did not improve model fit (Figure S11, Table S3), suggesting a rapid impact of the vaccine switch on genotype fitness at the population level. Our model was able to recover true fitness parameters when using simulated data with known parameters, even under conditions of biased sampling (Figure S12). Additionally, to test the predictive power of our model, a held-out analysis was performed. We held out 10% of the country-year data from the model fitting process and compared the prediction with the actual observed values (Figure S13). We repeated this procedure holding out individual countries that provided the greatest number of sequences (Japan, France, and the US) and found that the fitness estimates were largely unchanged



(Figure S14). Further, our population-level findings are consistent with those from experimental studies. In particular, PRN-deficient isolates have been shown to have increased ability to colonize the respiratory tract of ACV-vaccinated mice than PRN-positive isolates, whereas in unvaccinated control mice, PRN-positive isolates outcompeted PRN-deficient isolates(28). Mouse models have also shown increased respiratory tract colonization by *ptxP3* than by *ptxP1* isolates(29, 30).

## Discussion

We have used a large geo-referenced dataset to uncover the spatial spread of an important bacterial threat to human health. We have also identified how vaccine policy decisions can have immediate and far-reaching implications for the fitness of circulating strains, and how small fitness differences at each transmission generation can lead to major changes in lineage composition. Our findings also support and quantify the driving role of acellular vaccines in changing genotype distribution and are consistent with changes in the mutation rate in vaccine antigen genes after ACV implementation(12). A similar role of vaccine-induced changes in lineage composition has been found for pneumococcus(34–36).

Our modeling approach recovered the substantially different genotype dynamics that have been observed across countries, including in Japan where the proportion of isolates that were PRN-deficient initially increased following ACV implementation, but unlike most other countries subsequently decreased after 2005 (Figure 1H)(2, 15, 31). To date, the main hypothesis to explain this decrease has been a change in ACV formulations to ones that do not contain PRN, which became widely available in 2012(2, 15, 31). We provide an alternative explanation for the decrease in PRN-deficient strains. *ptxP1* remained in circulation in Japan after it had largely disappeared elsewhere, and unlike PRN-deficient *ptxP3* strains, PRN-deficient *ptxP1* strains only developed a small fitness advantage post ACV implementation (Figure 4B). Our model suggests that the reduction in PRN-deficient strains may simply have been driven by the rise of PRN-positive *ptxP3* strains, which are marginally more fit than *ptxP1* PRN-deficient strains in the ACV context. We expect that this phenomenon can only delay the evolution towards PRN deficiency, as switches to fitter PRN-deficient forms of *ptxP3 B. pertussis* appear inevitable. It remains unclear whether the growing use of ACV formulations without PRN in Japan can further delay this predictable trend. Our fitness model was unable to capture the genotype dynamics in China, where *ptxP1*-positive strains remain dominant. This could be explained by the uniquely high prevalence of erythromycin resistant strains(14, 32), which might have a large impact on fitness(33) but which our model does not take into account.

This study has some limitations. As with virtually all phylogenetic studies, the availability of sequences differs substantially by country and year. In particular, our inferences are largely based on sequences from high income countries. WCV remains broadly used in low- and middle-income countries. As we also quantified genotype fitness when WCV was the dominant vaccine, our findings are relevant to these undersampled locations, and suggest that, given their increased fitness in WCV environments, *ptxP3*/PRN-positive genotypes will dominate. This is indeed the case for the two countries in our dataset, which have only ever used WCV (Tunisia and Iran). In the event of future switches from WCV to ACV, we could expect PRN-deficient strains to prevail, either through the successful introduction of strains circulating elsewhere or through the loss of PRN expressions in local strains. Despite clear differences in the fitness of strains, some strains with apparent reduced fitness remained in circulation for long time periods. It remains unclear if this persistence is due to spatial heterogeneity in population immunity profiles (for example, from heterogeneous vaccine coverage) that can facilitate ongoing circulation of globally less fit strains, the external introduction of strains from environments where they have greater fitness, or unmeasured markers of fitness not captured by our approach. For example, whereas this study assesses the impact of vaccine type (WCV or ACV) on strain fitness, increased, sustained sequencing efforts would allow us to tackle more precise questions about whether different WCV or ACV vaccines(37–39) or vaccination policies and schedules(40, 41) exert distinct pressures. Last, there may also be other mutations that are linked to changes in fitness that we have not considered (17).

This work provides a detailed view of the strength of spatial structure and rate of geographic spread for a pathogen that is responsible for tens of thousands of deaths each year, highlighting how *B. pertussis* spread is a globally interconnected issue. It further provides a quantitative description of strains fitness and highlights the key role of vaccine policy decision-making in driving ecological change. Multiple co-circulating lineages are a common feature across bacterial disease systems. Our approaches provide an avenue to obtain a much-needed mechanistic understanding of how perturbations in population immunity can drive local genetic diversity.

## Materials and Methods

### Study design

We compiled a dataset of 3,344 *B. pertussis* whole genome sequences (Table S1 and S2), sampled from a 85-year period (1935-2019) that included 1,011 sequenced French isolates from the National Reference Center (NRC) for Whooping Cough and Other Bordetella Infections in France (Institut Pasteur, Paris, France); 320 newly sequenced European isolates, randomly selected among the 2014-2016 isolates from the *B. pertussis* collections of 12 European countries; and 2,013 isolates with publicly available genomes from 13 countries. These publicly available isolates were selected from all high-quality raw sequences deposited on NCBI and ENA at the time of accession (December 2019,  $n = 2,327$ ). The majority of publicly available isolates were generated by the CDC (Atlanta, USA), who shared metadata. Critically, a number of isolates ( $n=314$ ) deposited in NCBI were duplicates from the same isolates and were removed from the analysis.

### Metadata

We collected metadata for each isolate (date of sampling, continent, country) from linked publications, or NCBI when no publication was listed. The vaccine status of patients was not known. For 3,256 sequences coming from 19 countries, the home district or region of the genome was also recorded. Geographic coordinates were extracted for each location. When the geographic precision was region or district, the coordinates of the most populated city were extracted. We also considered centroids of the whole region as a sensitivity analysis. For each country, we extracted information on vaccine coverage as a function of time (1980-2019, when available) from the Global Health Observatory data repository(42). We also collected years of implementation of each type of vaccine in the literature. For the acellular vaccine (ACV), we distinguished the year of the first implementation of any ACV (as a booster or a primary vaccination), and the date of ACV introduction for primary vaccination series (Table S1, Figure S8). To maximize statistical power, we focused the analysis on the type of vaccine (WCV, ACV), rather than the specific details of each vaccine (for example vaccine provider, vaccine strain used, number of antigens present, presence of a pertactin component). We also chose to focus on the use (or not) of each vaccine in each country, rather than on the country-specific vaccine coverage, assuming that the vaccination coverage was sufficient to exert selective pressure on strains.

## **Ethical considerations**

The study was coordinated by the French national reference center for pertussis, whose activities are approved by the French supervisory ethics authority (CNIL, n°1474593). Isolates sequenced in this study come from the *B. pertussis* governmental surveillance laboratories of Belgium, Czech Republic, Denmark, Spain, Finland, France, Ireland, Italy, Netherlands, Norway, Sweden, and the UK. These isolates were all collected as part of existing public health surveillance approved protocols in each country. No personally identifiable information was used as part of this study.

## **Newly sequenced isolates**

The study Accession Numbers on the European Read Archive are PRJEB21744, PRJEB42353 and PRJEB45681. Details and accession numbers of the raw sequence data are listed in data file S1. Details on the sequencing protocol can be found in supplementary materials.

## **Genomic analysis of PRN-deficient isolates**

*De novo* assembly was performed, as previously described(16). Briefly, paired-end reads were clipped and trimmed (AlienTrimmer(43)), corrected (Musket(44)), merged if needed (FLASH(45)), and subjected to a digital normalization procedure with khmer(46). For each sample, remaining processed reads were assembled and scaffolded (SPAdes(47)).

We defined the PRN allele of each isolate with BLASTn(48) using as query a fasta file containing all known PRN alleles.

Next, we defined the pertactin expression (PRN status) of all isolates. As the PRN-deficiency has been shown to be caused by mutations in the promoter or coding regions of the PRN gene(49), we compiled all the events that cause the PRN deficiency based on the literature and on the analysis of French isolates and gathered all the genomic events identified in PRN-negative isolate in a single fasta file. Then, we assessed the PRN status of all isolates with BLASTn(48) using as query the fasta file (Data Files S2 and S3). In addition, we used IS\_mapper/0.1.5.1 from fastq files looking for IS481 from Tohama (BP0080), which is the main insertion element in *B. pertussis*(50) that can insert within the prn gene or promoter. We made use of 839 French isolates for which we have Western blots available(1), and checked the correspondence between genomics and experimental results. In our dataset, we identified 1722 (51.5%) PRN-positive isolates, 1472 (44.0%) PRN-deficient and 150 (4.5%) isolates with unknown status.

### Single Nucleotide Polymorphism (SNP) detection

SNP detection was conducted using an in-house pipeline available at 10.5281/zenodo.6335548. Briefly, adapters and barcodes were stripped from the fastq data and the reads were quality filtered and trimmed using a Phred quality threshold score of 30 using Cutadapt(51). We checked the quality of each fastq file using FastQC(52). After cleaning, the average read depth across genomes was 147X (range: 30-958). Reads were mapped against the complete Tohama I reference genome (Accession number: NC\_002929), using BWA-MEM algorithm(53). On average 98% (range: 88-99.9) of positions were covered with more than 5 reads. Extraction of SNP was achieved with the GATK HaplotypeCaller, with ERC GVCF settings(54). We then built an in-house filtering script using R(55). We kept variants for which the Phred quality score was higher than 30, with a minimum read depth of 5, with at least 2 reads in both the forward and reverse directions. We called a position as 'N' when <40% of the reads were different from the reference, we used the IUPAC code for positions with between 40% and 80% of reads not matching the reference, and we called variants only for positions with >80% reads not matching the reference. Further, we filtered out repeated regions (IS481, IS1002 and IS1663(50)), and phage regions using Phaster(56). We also checked for recombination in our alignment using Gubbins(57). As a result, we obtained an alignment of 8,496 SNPs, consistent with other publications(17).

### Genotyping *ptxA*, *ptxP*, *fim2*, and *fim3*

We genotyped all isolates for the genes pertussis toxin A (*ptxA*), pertussis toxin promoter (*ptxP*), type 2 fimbrial protein (*fim2*) and type 3 fimbrial protein (*fim3*) using BLAST(48). The sequences used as references for the alleles *ptxA*, *ptxP*, *fim2*, and *fim3* are available at 10.5281/zenodo.6335548.

### Phylogenetic analysis

*B. pertussis* strain genotyping has been used to reconstruct the evolutionary and spatial history of observed infections, however, previous efforts have relied on pre-genomics markers or individual country data(11–16, 58, 59). In this project we used whole genome sequences and combined data across multiple countries.

We used the SNP-based alignment to reconstruct the phylogenetic relationships of the isolates. We built maximum-likelihood trees using IQ-tree(60). TempEst was used to check for temporal signal in the data (Figure S15)(61). Maximum Clade Credibility (MCC) trees were inferred using a Markov Chain Monte Carlo (MCMC) Bayesian approach implemented on the program BEAST 1.10.4(62), under a GTR substitution model(63) accounting for the number of constant sites, a relaxed lognormal clock model(64) and a skygrid population size model(65). Three independent

Markov chains were run for 300,000,000 generations each, with parameter values sampled every 10,000 generations. Runs were optimized using the GPU BEAGLE library(66). Chains were manually checked for convergence (ESS values > 200) using the Tracer software(67). We manually removed a 20% burn-in.

Furthermore, we used the discrete trait analysis implemented in BEAST to attribute state characters representing the isolation country of each strain with the Bayesian Stochastic Search Variable Selection (BSSVS)(68). This method estimates the most probable state at each node in the MCC trees, allowing us to reconstruct plausible ancestral states (here, country) on these nodes. Because of the large nature of our dataset (3344 genomes from 23 countries, with 8,496 variable positions) running this model would have been computationally challenging. Thus, we ran the model on a subset of trees (n=10,000) extracted from the posterior distribution of trees generated by the initial BEAST run. We ran three independent Markov chains for 4,000,000 generations each, with parameter values sampled every 1,000 generations. We manually checked for convergence using the Tracer software(67) and removed a 10% burn-in. We used treeannotator to summarize the posterior trees into a Maximum Clade Credibility (MCC) tree. The full MCC tree is presented in Figure S16, and also available the MicroReact platform (<https://microreact.org/project/qWJRauq26T5JWXjNRCux4M-global-phylogeography-b-pertussis>).

### **Expansion rate of *B. pertussis***

We assume that the sequences in our dataset are representative of what was circulating in that location and at that time. In spatially structured transmissions, as pathogens spread away from each other, we would expect there to be increased spatial distance between cases as they become separated by more and more transmission events(26). To explore whether this occurs, we compared evolutionary time between *B. pertussis* bacteria with their spatial separation. For each location considered, we identified all pairs of bacteria that were separated by between 0 and t durations of evolutionary time and calculated the mean spatial distance between the identified pairs. We varied t between 0 and 40 years. We reconstructed 95% confidence intervals using a bootstrap: we resampled all bacteria with replacement, allowing for even sampling by location where the probability of each isolate being selected was equal to the inverse of the number of isolates available from that location. We repeated the resampling 200 times and recalculated the mean distance separating each group of bacteria pairs each time. The 95% confidence intervals are the 2.5% and 97.5% quantiles from the resultant distribution. We separately considered pairs coming from the same country (France, Japan, US). This list of countries was selected as they represent the locations with most sequences available. We separately repeated this analysis using pairs of sequences from across Europe.

### Estimating the probability that a pair of cases are from the same transmission chain

We consider a pair of cases that came within the same year to be part of the same transmission chain if their MRCA was within the past two years of the earlier of the two cases. We choose this cutoff as two years represents a recent introduction into a population, within which pairs are likely to remain transmission-related. We then explored the sensitivity of this cutoff (Figure S17).

We can derive an expression for the probability  $P_x$  that a pair of *B. pertussis* cases are from the same transmission chain within a location  $loc$  (26):

$$P_x(loc) = \frac{\sum_{i=1}^n \sum_{j \neq i}^n I(loc_i = loc_j \cap t_{ij} \leq 1 \text{ year} \cap G_{ij} \leq 2 \text{ years})}{\sum_{i=1}^n \sum_{j \neq i}^n I(loc_i = loc_j \cap t_{ij} \leq 1 \text{ year})}$$

[Eq. 1]

where  $n$  represents the number of bacteria for which sequence data are available,  $t_{ij}$  is the time between the cases, and  $G_{ij}$  is the time to the MRCA from the earlier of  $i$  and  $j$ .  $I$  is an indicator function.

### Estimation of effective number of transmission chains

The reciprocal of  $P_x(loc)$  is an estimate of the size of the pool of discrete transmission chains  $\eta(loc)$  that infect pairs of individuals separated within a location  $loc$ . It also represents the lower limit of the number of chains circulating within a location  $loc$ , as previously shown (26).

We reconstructed 95% confidence intervals using a bootstrap: we resampled all bacteria with replacement, allowing for even sampling by location, over 200 resampling events and recalculate the statistic each time. The 95% confidence intervals are the 2.5% and 97.5% quantiles from the resultant distribution.

### Effective number of transmission chains for different population sizes in regions

We explored whether the effective number of chains circulating within regions depended on the size of the population of this region using Equation 1. We estimated the mean probability  $P_x(loc, pop_1, pop_2)$  that a pair of *B. pertussis* cases within the same region, with a population size between  $pop_1$  and  $pop_2$ , are from the same transmission chain:

$$P_x(loc, pop_1, pop_2) = \frac{\sum_{i=1}^n \sum_{j \neq i}^n I(loc_i = loc_j \cap t_{ij} \leq 1 \text{ year} \cap G_{ij} \leq 2 \text{ years} \cap P_{loc} > pop_1 \cap P_{loc} < pop_2)}{\sum_{i=1}^n \sum_{j \neq i}^n I(loc_i = loc_j \cap t_{ij} \leq 1 \text{ year})}$$

[Eq. 2]

The effective number of transmission chains  $\eta(loc, pop_1, pop_2)$  is then given by the reciprocal of  $P_x(loc, pop_1, pop_2)$ .

We reconstructed 95% confidence intervals using a bootstrap as detailed above.

In a sensitivity analysis, we considered different MRCA cutoffs (1 and 3 years) to define transmission chains, and obtained similar results (Figure S17).

### **Relative risk that a pair of bacteria have a MRCA within a defined period, when coming from the same versus different locations**

To better understand the spread of *B. pertussis* within districts in France, countries in Europe, states in the US and between continents, we characterize the similarity in bacteria within these spatial scales relative to that observed between spatial scales. In each case, we estimate  $RR_{loc}(g_1, g_2)$ , the probability that a pair of bacteria within a location  $loc$  that were isolated within the same year of each other had an MRCA within range having an MRCA within  $g_1$ - $g_2$ , relative to the probability that a pair of bacteria from different locations that were isolated within the same year of each other had an MRCA within that particular range(26). For the range  $g_1$ - $g_2$ , we used sliding windows of time going from 0 years to 20 years.

$$RR_{loc}(g_1, g_2) = \frac{\frac{\sum_{i=1}^n \sum_{j \neq i}^n I(loc_i = loc_j \cap t_{ij} \leq 1 \text{ year} \cap G_{ij} \geq g_1 \cap G_{ij} < g_2)}{\sum_{i=1}^n \sum_{j \neq i}^n I(loc_i = loc_j \cap t_{ij} \leq 1 \text{ year})}}{\frac{\sum_{i=1}^n \sum_{j \neq i}^n I(loc_i \neq loc_j \cap t_{ij} \leq 1 \text{ year} \cap G_{ij} \geq g_1 \cap G_{ij} < g_2)}{\sum_{i=1}^n \sum_{j \neq i}^n I(loc_i \neq loc_j \cap t_{ij} \leq 1 \text{ year})}}$$

[Eq. 3]

By conditioning on spatial and temporal location of sequences, this approach minimizes the impact of underlying sampling biases in which cases were sequenced. We reconstructed 95% confidence intervals using bootstrap, as detailed above.

### **Estimating the fitness of *B. pertussis* strains**

To quantify the fitness of *B. pertussis* strains, we developed a model that makes use of the isolates from 20 countries with more than 10 sequences in our dataset (fewer sequences than



this threshold did not allow for sufficient data points to provide a robust estimate of genotype distributions over time). First, we assigned all tips and nodes in the tree a strain type from one of six possible types based on PRN, *ptxP*, and *fim3* status. We considered only isolates for which the *ptxP3* allele was 1 or 3, the *fim3* allele was 1 or 2, and the PRN expression was known. Although there exist other genotypes, these are rare and most of the other types, for example the *ptxA* alleles, almost completely overlap with the *ptxP* alleles. Assigning the genotypes *ptxP1* versus *ptxP3*, and *fim3-1* versus *fim3-2* was straightforward as they are all monophyletic(17). However, PRN-deficiency is a highly homoplastic trait, so assigning PRN expression to each node was more complicated. Briefly, we know ancestral *B. pertussis* sequences all produce PRN, and that the deficiency appears by mutations in the promoter or coding regions of the PRN gene(49). Moreover, because of the complexity of the mutations, reversions are highly unlikely, and have to our knowledge never been reported. Thus, the PRN expression of each node can be easily reconstructed by finding all the monophyletic clades for which isolates are PRN-deficient. To do this, we used an in-house algorithm (script available at 10.5281/zenodo.6335548).

We separately assigned a country to each node and tip. Country information for the nodes was extracted from the BEAST discrete reconstruction presented previously. To maximize the correct country assignment for nodes, we only consider nodes for which the country's posterior probability was >0.9. In addition, we excluded nodes that were distant by more than 5 years from any tip.

Next, we computed  $f_{i,ref}$ , the relative abundance of each strain  $i$  with respect to a chosen  $ref$  strain. We chose the *ptxP3/fim3-1*/PRN-positive genotype as the reference strain. In a sensitivity analysis, we found that the choice of the reference strain did not affect the results, providing that the chosen genotype was present in all countries, for most years. We then used a simple logistic model to capture the evolution of this abundance, at each time  $t$ :

$$\frac{d f_{i,ref}}{dt} = r_{i,ref} \cdot f_{i,ref} (1 - f_{i,ref}),$$

*with  $i \in \{strains\} \setminus ref$ , and  $f_{i,ref}(t = 0) = f_{i,ref,0}$*

[Eq. 4]

where  $r_{i,ref}$  is the growth rate of that abundance, shared across all countries and  $f_{i,ref,0}$  is the initial relative abundance of the strain  $i$  with respect to a chosen  $ref$ . To control for the varying presence of all circulating strains through time, we present fitness as the *average relative growth rate*  $\underline{r}_i$  for each strain  $i$  with respect to randomly selected strain in the population:

$$\bar{r}_i = \sum_{j \neq i}^n \bar{f}_j (r_{i,ref} - r_{j,ref})$$

[Eq. 5]

where  $n$  is the number of strains, and  $\bar{f}_j$  is the average frequency of strain  $j$  on the period of time considered.

This average relative growth rate  $\bar{r}_i$  can be identified as the selection rate coefficient  $s$  of the strain  $i$  in the population considered (69–71).

We can further multiply the selection rate by the mean generation time  $T$  to obtain the dimensionless selection coefficient  $s_T$ , the relative fitness advantage per transmission generation (70):

$$s_T = s \cdot T = \bar{r}_i \cdot T$$

[Eq. 6]

We used a mean generation interval of 22.8 days (95%CI: 22.1–23.5) to compute  $s_T$  (72).

The selection coefficients  $s$  and  $s_T$  represent one of the most possible direct measures of the fitness advantage of any new variant, and are the best possible predictors of whether or not it is expected to increase in frequency during an outbreak (71, 73).

We fit this model to all countries-specific time series in our dataset with the Rstan package (74), using a Negative Binomial likelihood with an overdispersion parameter  $\delta$ . We fit the frequencies as simplex vectors for each country, with the oldest strain frequency,  $f_{1,0}$ , having a *normal*(1, 0.1) prior, based on a previous global study (17). We used *cauchy*(0, 0.15) for the relative growth rate parameters  $r_{i,ref}$ . We ran this model on 3 independent chains with 2,000 iterations and 50% burn-in. We used 2.5 and 97.5 quantiles from the resulting posterior distributions for 95% credible intervals of the parameters.

### Investigating changes in *B. pertussis* fitness across vaccine eras

To investigate whether *B. pertussis* strain fitness changed across vaccine eras, we compiled dates of vaccination changes (Table S1) and used our fitness model and estimate  $r_i$  on:

- WCV era, defined as the period from the WCV implementation in each country, to the first ACV implementation (as a booster or primary vaccine);
- the ACV era, defined as the period from the first ACV implementation to now.

In a sensitivity analysis, we considered a range of different definitions for these eras, including the mean implementation year of the WCV and ACV across countries, and started at different

vaccine coverages (Table S2). We also considered models without a vaccine switch. Model comparison was done using the Watanabe–Akaike information criterion (WAIC) implemented in the *loo* package (27, 75). Estimates from this sensitivity analysis are presented in Figure S10.

Additionally, to test the predictive power of our model, a held-out analysis was performed. We held out 10% of the country-year data from the model fitting process and compared the prediction with the actual observed values (Figure S13). We repeated this procedure holding out individual countries that provided the greatest number of sequences (Japan, France, and US). We found that the fitness estimates were largely unchanged (Figure S14).

We developed a simulation framework where the true growth rates parameters were known, and assessed the performance of our model to estimate the fitness of different strains in a population (see Supplementary materials).

## **List of supplementary materials**

### **Materials and methods**

Figure S1-S17

Table S1- S3

Data File S1-S3

MDAR Reproducibility Checklist

## References and notes

1. V. Bouchez, D. Brun, T. Cantinelli, G. Dore, E. Njamkepo, N. Guiso, First report and detailed characterization of *B. pertussis* isolates not expressing Pertussis Toxin or Pertactin. *Vaccine* **27**, 6034–6041 (2009).
2. N. Otsuka, H.-J. Han, H. Toyoizumi-Ajisaka, Y. Nakamura, Y. Arakawa, K. Shibayama, K. Kamachi, Prevalence and genetic characterization of pertactin-deficient *Bordetella pertussis* in Japan. *PLoS One* **7**, e31985 (2012).
3. K. H. T. Yeung, P. Duclos, E. A. S. Nelson, R. C. W. Hutubessy, An update of the global burden of pertussis in children younger than 5 years: a modelling study. *Lancet Infect. Dis.* **17**, 974–980 (2017).
4. D. W. Jackson, P. Rohani, Perplexities of pertussis: recent global epidemiological trends and their potential causes. *Epidemiol. Infect.* **142**, 672–684 (2014).
5. D. A. Caugant, M. C. J. Maiden, Meningococcal carriage and disease--population biology and evolution. *Vaccine* **27 Suppl 2**, B64–70 (2009).
6. G. P. DeMuri, E. R. Wald, The Group A Streptococcal Carrier State Reviewed: Still an Enigma. *J Pediatric Infect Dis Soc* **3**, 336–342 (2014).
7. A.-L. Bañuls, A. Sanou, N. T. Van Anh, S. Godreuil, Mycobacterium tuberculosis: ecology and evolution of a human bacterium. *J. Med. Microbiol.* **64**, 1261–1269 (2015).
8. A. van Belkum, Hidden Staphylococcus aureus Carriage: Overrated or Underappreciated? *MBio* **7**, e00079–16 (2016).
9. S. T. Almeida, S. Nunes, A. C. Santos Paulo, I. Valadares, S. Martins, F. Breia, A. Brito-Avô, A. Morais, H. de Lencastre, R. Sá-Leão, Low prevalence of pneumococcal carriage and high serotype and genotype diversity among adults over 60 years of age living in Portugal. *PLoS One* **9**, e90974 (2014).
10. C. J. Gill, C. E. Gunning, W. B. MacLeod, L. Mwananyanda, D. M. Thea, R. C. Pieciak, G. Kwenda, Z. Mupila, P. Rohani, Asymptomatic *Bordetella pertussis* infections in a longitudinal cohort of young African infants and their mothers. *Elife* **10** (2021), doi:10.7554/eLife.65663.
11. A. J. Schmidtke, K. O. Boney, S. W. Martin, T. H. Skoff, M. L. Tondella, K. M. Tatti, Population diversity among *Bordetella pertussis* isolates, United States, 1935-2009. *Emerg. Infect. Dis.* **18**, 1248–1255 (2012).
12. K. L. Sealey, S. R. Harris, N. K. Fry, L. D. Hurst, A. R. Gorringe, J. Parkhill, A. Preston, Genomic analysis of isolates from the United Kingdom 2012 pertussis outbreak reveals that vaccine antigen genes are unusually fast evolving. *J. Infect. Dis.* **212**, 294–301 (2015).
13. S. Octavia, S. Z. Wu, S. Kaur, L. Valinsky, E. Marva, J. Moran-Gilad, R. Lan, Whole-genome sequencing and comparative genomic analysis of *Bordetella pertussis* isolates from the 2007-2008 epidemic in Israel. *J. Infect.* **74**, 204–207 (2017).
14. Z. Xu, Z. Wang, Y. Luan, Y. Li, X. Liu, X. Peng, S. Octavia, M. Payne, R. Lan, Genomic epidemiology of erythromycin-resistant *Bordetella pertussis* in China. *Emerg. Microbes Infect.* **8**, 461–470 (2019).
15. A. Zomer, N. Otsuka, Y. Hiramatsu, K. Kamachi, N. Nishimura, T. Ozaki, J. Poolman, J. Geurtsen, *Bordetella pertussis* population dynamics and phylogeny in Japan after adoption of

acellular pertussis vaccines. *Microb Genom* **4** (2018), doi:10.1099/mgen.0.000180.

16. V. Bouchez, J. Guglielmini, M. Dazas, A. Landier, J. Toubiana, S. Guillot, A. Criscuolo, S. Brisse, Genomic Sequencing of *Bordetella pertussis* for Epidemiology and Global Surveillance of Whooping Cough. *Emerg. Infect. Dis.* **24**, 988–994 (2018).

17. M. J. Bart, S. R. Harris, A. Advani, Y. Arakawa, D. Bottero, V. Bouchez, P. K. Cassiday, C.-S. Chiang, T. Dalby, N. K. Fry, M. E. Gaillard, M. van Gent, N. Guiso, H. O. Hallander, E. T. Harvill, Q. He, H. G. J. van der Heide, K. Heuvelman, D. F. Hozbor, K. Kamachi, G. I. Karataev, R. Lan, A. Lutyńska, R. P. Maharjan, J. Mertsola, T. Miyamura, S. Octavia, A. Preston, M. A. Quail, V. Sintchenko, P. Stefanelli, M. L. Tondella, R. S. W. Tsang, Y. Xu, S.-M. Yao, S. Zhang, J. Parkhill, F. R. Mooi, Global population structure and evolution of *Bordetella pertussis* and their relationship with vaccination. *MBio* **5**, e01074 (2014).

18. F. R. Mooi, I. H. M. van Loo, M. van Gent, Q. He, M. J. Bart, K. J. Heuvelman, S. C. de Greeff, D. Diavatopoulos, P. Teunis, N. Nagelkerke, J. Mertsola, *Bordetella pertussis* strains with increased toxin production associated with pertussis resurgence. *Emerg. Infect. Dis.* **15**, 1206–1213 (2009).

19. H. Hallander, A. Advani, F. Alexander, L. Gustafsson, M. Ljungman, C. Pratt, I. Hall, A. R. Gorringer, Antibody responses to *Bordetella pertussis* Fim2 or Fim3 following immunization with a whole-cell, two-component, or five-component acellular pertussis vaccine and following pertussis disease in children in Sweden in 1997 and 2007. *Clin. Vaccine Immunol.* **21**, 165–173 (2014).

20. F. R. Mooi, N. A. T. Van Der Maas, H. E. De Melker, Pertussis resurgence: waning immunity and pathogen adaptation--two sides of the same coin. *Epidemiology & Infection* **142**, 685–694 (2014).

21. M. van Gent, M. J. Bart, H. G. J. van der Heide, K. J. Heuvelman, F. R. Mooi, Small mutations in *Bordetella pertussis* are associated with selective sweeps. *PLoS One* **7**, e46407 (2012).

22. N. H. Carbonetti, Contribution of pertussis toxin to the pathogenesis of pertussis disease. *Pathog. Dis.* **73**, ftv073 (2015).

23. A. R. Gorringer, T. E. Vaughan, *Bordetella pertussis* fimbriae (Fim): relevance for vaccines. *Expert Rev. Vaccines* **13**, 1205–1214 (2014).

24. L. Ma, A. Caulfield, K. K. Dewan, E. T. Harvill, Pertactin-Deficient *Bordetella pertussis*, Vaccine-Driven Evolution, and Reemergence of Pertussis. *Emerg. Infect. Dis.* **27**, 1561–1566 (2021).

25. N. Ahvenainen, T. Dub, A. Knuutila, A.-M. Barkoff, J. Sane, Q. He, Widespread circulation of pertussis in Finland during 1968-1972 when the whole cell vaccine was in use. *Clin. Microbiol. Infect.* **27**, 1526–1528 (2021).

26. H. Salje, J. Lessler, I. Maljkovic Berry, M. C. Melendrez, T. Endy, S. Kalayanarooj, A. A. Nuegoonpipat, S. Chanama, S. Sangkijporn, C. Klungthong, B. Thaisomboonsuk, A. Nisalak, R. V. Gibbons, S. Iamsirithaworn, L. R. Macareo, I.-K. Yoon, A. Sangarsang, R. G. Jarman, D. A. T. Cummings, Dengue diversity across spatial and temporal scales: Local structure and the effect of host population size. *Science* **355**, 1302–1306 (2017).

27. S. Watanabe, Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. *J. Mach. Learn. Res.* **11**, 3571–3594 (2010).

28. A. Safarchi, S. Octavia, L. D. W. Luu, C. Y. Tay, V. Sintchenko, N. Wood, H. Marshall, P.

McIntyre, R. Lan, Pertactin negative *Bordetella pertussis* demonstrates higher fitness under vaccine selection pressure in a mixed infection model. *Vaccine* **33**, 6277–6281 (2015).

29. A. J. King, S. van der Lee, A. Mohangoo, M. van Gent, A. van der Ark, B. van de Waterbeemd, Genome-wide gene expression analysis of *Bordetella pertussis* isolates associated with a resurgence in pertussis: elucidation of factors involved in the increased fitness of epidemic strains. *PLoS One* **8**, e66150 (2013).

30. A. Safarchi, S. Octavia, L. D. W. Luu, C. Y. Tay, V. Sintchenko, N. Wood, H. Marshall, P. McIntyre, R. Lan, Better colonisation of newly emerged *Bordetella pertussis* in the co-infection mouse model study. *Vaccine* **34**, 3967–3971 (2016).

31. Y. Hiramatsu, Y. Miyaji, N. Otsuka, Y. Arakawa, K. Shibayama, K. Kamachi, Significant Decrease in Pertactin-Deficient *Bordetella pertussis* Isolates, Japan. *Emerg. Infect. Dis.* **23**, 699–701 (2017).

32. Z. Wang, Z. Cui, Y. Li, T. Hou, X. Liu, Y. Xi, Y. Liu, H. Li, Q. He, High prevalence of erythromycin-resistant *Bordetella pertussis* in Xi'an, China. *Clin. Microbiol. Infect.* **20**, O825–30 (2014).

33. Y. Feng, C.-H. Chiu, U. Heininger, D. F. Hozbor, T. Q. Tan, C.-H. W. von König, Emerging macrolide resistance in *Bordetella pertussis* in mainland China: Findings and warning from the global pertussis initiative. *Lancet Reg Health West Pac* **8**, 100098 (2021).

34. T. Azarian, P. P. Martinez, B. J. Arnold, X. Qiu, L. R. Grant, J. Corander, C. Fraser, N. J. Croucher, L. L. Hammitt, R. Reid, M. Santosham, R. C. Weatherholtz, S. D. Bentley, K. L. O'Brien, M. Lipsitch, W. P. Hanage, Frequency-dependent selection can forecast evolution in *Streptococcus pneumoniae*. *PLoS Biol.* **18**, e3000878 (2020).

35. S. W. Lo, R. A. Gladstone, A. J. van Tonder, J. A. Lees, M. du Plessis, R. Benisty, N. Givon-Lavi, P. A. Hawkins, J. E. Cornick, B. Kwambana-Adams, P. Y. Law, P. L. Ho, M. Antonio, D. B. Everett, R. Dagan, A. von Gottberg, K. P. Klugman, L. McGee, R. F. Breiman, S. D. Bentley, Global Pneumococcal Sequencing Consortium, Pneumococcal lineages associated with serotype replacement and antibiotic resistance in childhood invasive pneumococcal disease in the post-PCV13 era: an international whole-genome sequencing study. *Lancet Infect. Dis.* **19**, 759–769 (2019).

36. J. Corander, C. Fraser, M. U. Gutmann, B. Arnold, W. P. Hanage, S. D. Bentley, M. Lipsitch, N. J. Croucher, Frequency-dependent selection in vaccine-associated pneumococcal population dynamics. *Nat Ecol Evol* **1**, 1950–1960 (2017).

37. P. E. M. Fine, J. A. Clarkson, Reflections on the Efficacy of Pertussis Vaccines *Clinical Infectious Diseases* **9**, 866–883 (1987).

38. G. A. M. Berbers, S. C. de Greeff, F. R. Mooi, Improving pertussis vaccination. *Hum. Vaccin.* **5**, 497–503 (2009).

39. A. Wong, A. Opinel, S. J.-B. Combes, J. Toubiana, S. Brisse, Determining Factors for Pertussis Vaccination Policy: A Study in Five EU Countries. *Vaccines (Basel)* **8** (2020), doi:10.3390/vaccines8010046.

40. A.-M. Barkoff, J. Mertsola, D. Pierard, T. Dalby, S. V. Hoegh, S. Guillot, P. Stefanelli, M. van Gent, G. Berbers, D. Vestreheim, M. Greve-Isdahl, L. Wehlin, M. Ljungman, N. K. Fry, K. Markey, Q. He, Pertactin-deficient *Bordetella pertussis* isolates: evidence of increased circulation in Europe, 1998 to 2015. *Euro Surveill.* **24** (2019), doi:10.2807/1560-7917.ES.2019.24.7.1700832.

41. J. Paireau, S. Guillot, F. Aït El Belghiti, S. Matczak, S. Trombert-Paolantoni, V. Jacomo, M.-K. Taha, H. Salje, S. Brisse, D. Lévy-Bruhl, S. Cauchemez, J. Toubiana, Impact of Vaccine Schedule Change on Pertussis Epidemiology in France: A Modelling and Serological Study (2021), doi:10.2139/ssrn.3786698.
42. WHO, Diphtheria tetanus toxoid and pertussis (DTP3) - Immunization coverage estimates by country *Global Health Observatory data repository* (available at <https://apps.who.int/gho/data/node.main.A827?lang=en>).
43. A. Criscuolo, S. Brisse, AlienTrimmer: a tool to quickly and accurately trim off multiple short contaminant sequences from high-throughput sequencing reads. *Genomics* **102**, 500–506 (2013).
44. Y. Liu, J. Schröder, B. Schmidt, Musket: a multistage k-mer spectrum-based error corrector for Illumina sequence data. *Bioinformatics* **29**, 308–315 (2013).
45. T. Magoč, S. L. Salzberg, FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957–2963 (2011).
46. M. R. Crusoe, H. F. Alameldin, S. Awad, E. Boucher, A. Caldwell, R. Cartwright, A. Charbonneau, B. Constantinides, G. Edverson, S. Fay, J. Fenton, T. Fenzl, J. Fish, L. Garcia-Gutierrez, P. Garland, J. Gluck, I. González, S. Guermond, J. Guo, A. Gupta, J. R. Herr, A. Howe, A. Hyer, A. Härpfer, L. Irber, R. Kidd, D. Lin, J. Lippi, T. Mansour, P. McA’Nulty, E. McDonald, J. Mizzi, K. D. Murray, J. R. Nahum, K. Nanlohy, A. J. Nederbragt, H. Ortiz-Zuazaga, J. Ory, J. Pell, C. Pepe-Ranne, Z. N. Russ, E. Schwarz, C. Scott, J. Seaman, S. Sievert, J. Simpson, C. T. Skennerton, J. Spencer, R. Srinivasan, D. Standage, J. A. Stapleton, S. R. Steinman, J. Stein, B. Taylor, W. Trimble, H. L. Wiencko, M. Wright, B. Wyss, Q. Zhang, E. Zyme, C. T. Brown, The khmer software package: enabling efficient nucleotide sequence analysis. *F1000Res*. **4**, 900 (2015).
47. A. Bankevich, S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, S. I. Nikolenko, S. Pham, A. D. Prjibelski, A. V. Pyshkin, A. V. Sirotkin, N. Vyahhi, G. Tesler, M. A. Alekseyev, P. A. Pevzner, SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
48. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
49. A. Zeddeman, M. van Gent, C. J. Heuvelman, H. G. van der Heide, M. J. Bart, A. Advani, H. O. Hallander, C. H. Wirsing von König, M. Riffelman, J. Storsaeter, D. F. Vestreim, T. Dalby, K. A. Krogfelt, N. K. Fry, A. M. Barkoff, J. Mertsola, Q. He, F. Mooi, Investigations into the emergence of pertactin-deficient *Bordetella pertussis* isolates in six European countries, 1996 to 2012. *Euro Surveill.* **19** (2014), doi:10.2807/1560-7917.es2014.19.33.20881.
50. J. Parkhill, M. Sebaihia, A. Preston, L. D. Murphy, N. Thomson, D. E. Harris, M. T. G. Holden, C. M. Churcher, S. D. Bentley, K. L. Mungall, A. M. Cerdeño-Tárraga, L. Temple, K. James, B. Harris, M. A. Quail, M. Achtman, R. Atkin, S. Baker, D. Basham, N. Bason, I. Cherevach, T. Chillingworth, M. Collins, A. Cronin, P. Davis, J. Doggett, T. Feltwell, A. Goble, N. Hamlin, H. Hauser, S. Holroyd, K. Jagels, S. Leather, S. Moule, H. Norberczak, S. O’Neil, D. Ormond, C. Price, E. Rabinowitsch, S. Rutter, M. Sanders, D. Saunders, K. Seeger, S. Sharp, M. Simmonds, J. Skelton, R. Squares, S. Squares, K. Stevens, L. Unwin, S. Whitehead, B. G. Barrell, D. J. Maskell, Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica*. *Nat. Genet.* **35**, 32–40 (2003).
51. M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).



52. S. Andrews, Others, FastQC: a quality control tool for high throughput sequence data. 2010 (2017).
53. H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEMarXiv [*q-bio.GN*] (2013) (available at <http://arxiv.org/abs/1303.3997>).
54. A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, M. A. DePristo, The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
55. R Core Team, R: A Language and Environment for Statistical Computing (2021) (available at <https://www.R-project.org/>).
56. D. Arndt, J. R. Grant, A. Marcu, T. Sajed, A. Pon, Y. Liang, D. S. Wishart, PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res.* **44**, W16–21 (2016).
57. N. J. Croucher, A. J. Page, T. R. Connor, A. J. Delaney, J. A. Keane, S. D. Bentley, J. Parkhill, S. R. Harris, Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res.* **43**, e15 (2015).
58. A. Elomaa, A. Advani, D. Donnelly, M. Antila, J. Mertsola, Q. He, H. Hallander, Population dynamics of *Bordetella pertussis* in Finland and Sweden, neighbouring countries with different vaccination histories. *Vaccine* **25**, 918–926 (2007).
59. M. R. Weigand, M. M. Williams, Y. Peng, D. Kania, L. C. Pawloski, M. L. Tondella, CDC Pertussis Working Group, Genomic Survey of *Bordetella pertussis* Diversity, United States, 2000-2013. *Emerg. Infect. Dis.* **25**, 780–783 (2019).
60. L.-T. Nguyen, H. A. Schmidt, A. von Haeseler, B. Q. Minh, IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
61. A. Rambaut, T. T. Lam, L. Max Carvalho, O. G. Pybus, Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol.* **2**, vew007 (2016).
62. M. A. Suchard, P. Lemey, G. Baele, D. L. Ayres, A. J. Drummond, A. Rambaut, Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* **4**, vey016 (2018).
63. S. Tavaré, Others, Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on mathematics in the life sciences* **17**, 57–86 (1986).
64. A. J. Drummond, S. Y. W. Ho, M. J. Phillips, A. Rambaut, Relaxed phylogenetics and dating with confidence. *PLoS Biol.* **4**, e88 (2006).
65. V. N. Minin, E. W. Bloomquist, M. A. Suchard, Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Mol. Biol. Evol.* **25**, 1459–1471 (2008).
66. D. L. Ayres, M. P. Cummings, G. Baele, A. E. Darling, P. O. Lewis, D. L. Swofford, J. P. Huelsenbeck, P. Lemey, A. Rambaut, M. A. Suchard, BEAGLE 3: Improved Performance, Scaling, and Usability for a High-Performance Computing Library for Statistical Phylogenetics*Systematic Biology* **68**, 1052–1061 (2019).
67. A. Rambaut, A. J. Drummond, D. Xie, G. Baele, M. A. Suchard, Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Syst. Biol.* **67**, 901–904 (2018).

68. P. Lemey, A. Rambaut, A. J. Drummond, M. A. Suchard, Bayesian phylogeography finds its roots. *PLoS Comput. Biol.* **5**, e1000520 (2009).
69. M. Travisano, R. E. Lenski, Long-term experimental evolution in *Escherichia coli*. IV. Targets of selection and the specificity of adaptation. *Genetics* **143**, 15–26 (1996).
70. L.-M. Chevin, On measuring selection in experimental evolution. *Biol. Lett.* **7**, 210–213 (2011).
71. N. G. Davies, S. Abbott, R. C. Barnard, C. I. Jarvis, A. J. Kucharski, J. D. Munday, C. A. B. Pearson, T. W. Russell, D. C. Tully, A. D. Washburne, T. Wenseleers, A. Gimma, W. Waites, K. L. M. Wong, K. van Zandvoort, J. D. Silverman, C. C.-19 W. Group<sup>1</sup>‡, COVID-19 Genomics UK (COG-UK) Consortium‡, K. Diaz-Ordaz, R. Keogh, R. M. Eggo, S. Funk, M. Jit, K. E. Atkins, W. John Edmunds, Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science* **372** (2021), doi:10.1126/science.abg3055.
72. M. A. Vink, M. C. J. Bootsma, J. Wallinga, Serial intervals of respiratory infectious diseases: a systematic review and analysis. *Am. J. Epidemiol.* **180**, 865–875 (2014).
73. T. Day, S. Gandon, Applying population-genetic models in theoretical evolutionary epidemiology. *Ecol. Lett.* **10**, 876–888 (2007).
74. Stan Development Team, RStan: the R interface to Stan (2020) (available at <http://mc-stan.org/>).
75. A. Vehtari, A. Gelman, J. Gabry, Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC *Statistics and Computing* **27**, 1413–1432 (2017).

**Acknowledgments.** We thank Dr. C. Schot for her contribution to isolates and genomic data collection. We thank Gabriele Buttinelli and Luigina Ambrosio (Dept. of Infectious Diseases, Istituto Superiore di Sanità, Rome, Italy) and Carlo Concato (Virology Unit, Bambino Gesù Children's Hospital, IRCCS, Rome, Italy) for their contribution to Italian isolates collection and preparation. We thank Dr. M. Weigand and his team at CDC (Atlanta, USA) for making sequences publicly available and sharing metadata.

**Funding.** The study was supported financially by the French Government Investissement d'Avenir grant ANR-16-CONV-0005 (INCEPTION project, to SB and HS). The National Reference Center for Whooping Cough and Other Bordetella Infections receives support from Institut Pasteur and Public Health France (Santé publique France, Saint Maurice, France). This work was also supported financially by the French Government's Investissement d'Avenir program Laboratoire d'Excellence "Integrative Biology of Emerging Infectious Diseases" (ANR-10-LABX-62-IBEID) to SB and a European Research Council (No. 804744 to HS). PL acknowledges funding from the European Research Council under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 725422-ReservoirDOCS).

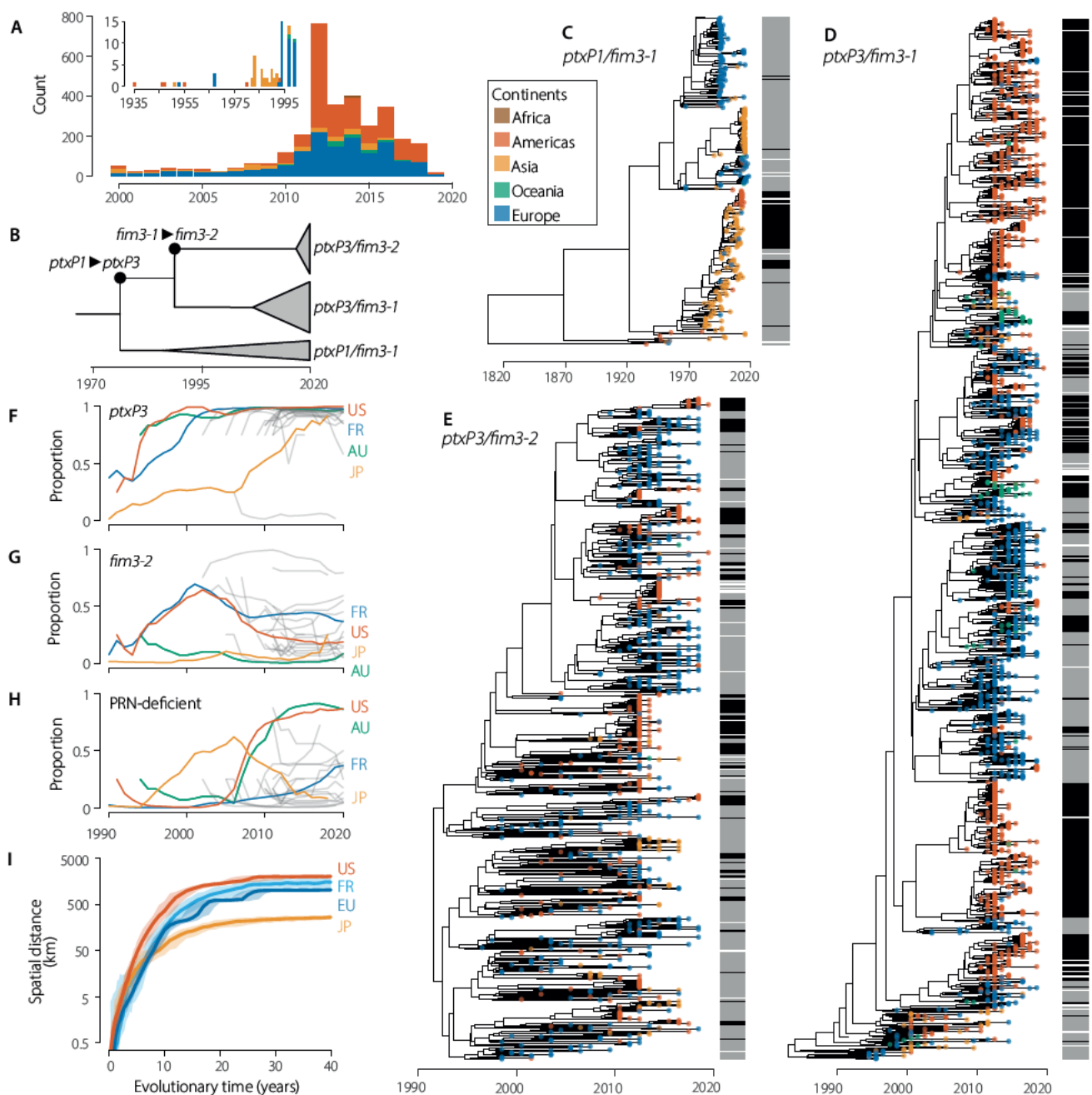
**Author contributions.**

JT, SC, HS and SB conceived the study. NL and HS developed the methods and performed the modelling analysis, supported by SLH, PL, SB and JP. VB, NL and NF performed the genomic analyses. NL, VB, NF, A-MB, TB, TA, KF, DFV, NKF, JGL, KG, AH, QH, DL, HM, DP, PS, JZ, SG, JT and SB participated in isolate and genomic data collection. VB, NA, AL, SG, TB, PS, HM and MS sequenced the isolates. NL, VB, HS and SB accessed and verified all the data. NL, HS and SB wrote the first draft of the manuscript. All authors provided input to the manuscript and reviewed the final version.

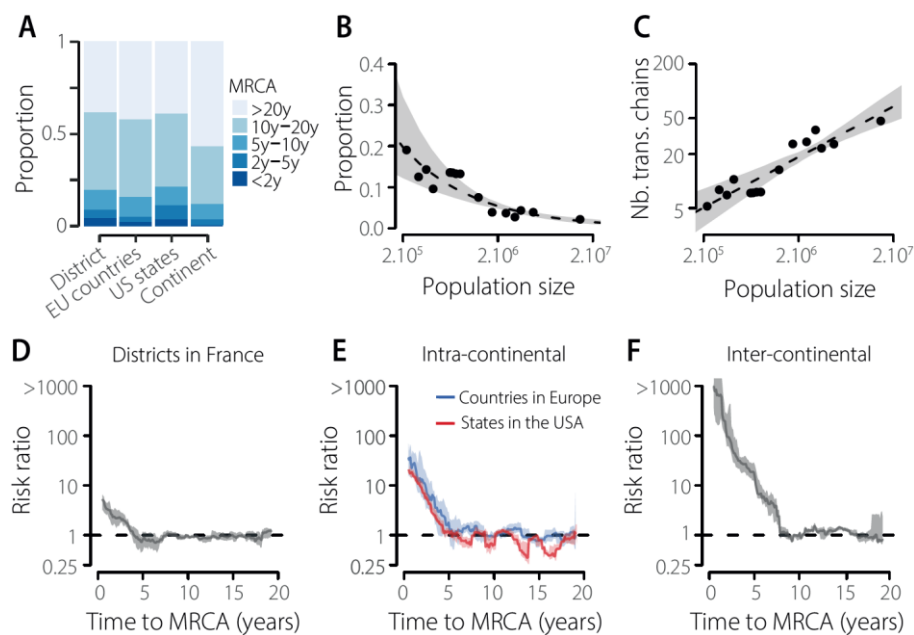
**Competing interests.** The authors declare that they have no competing interests.

**Data and materials availability:** Code is available at [10.5281/zenodo.6335548](https://zenodo.org/record/6335548). All sequences generated within this study were deposited in ENA, with accession numbers and metadata attached to each sequence available in data file S1.

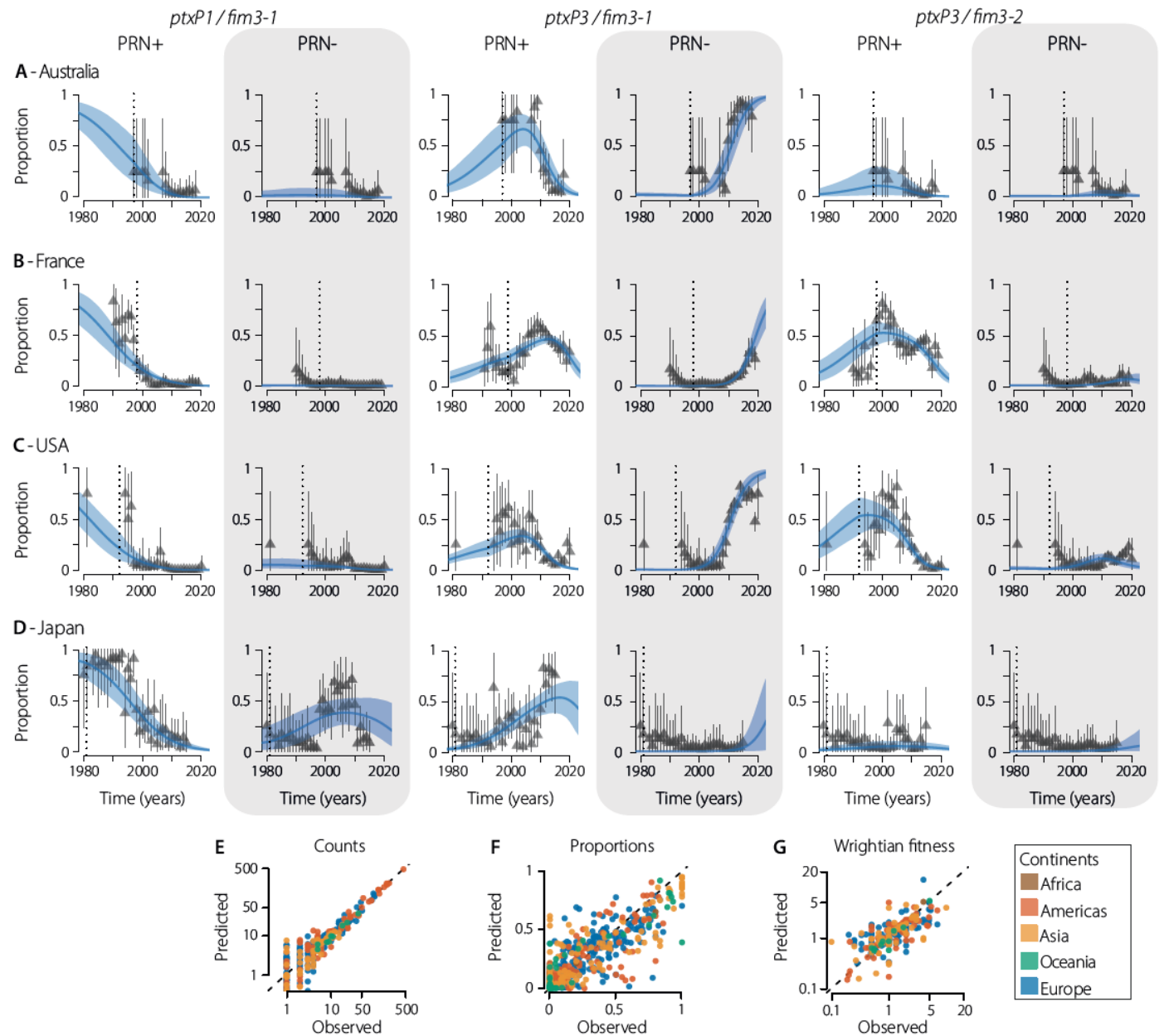
**Figure 1: Origin of sequences and genetic diversity of *Bordetella pertussis*.** (A) Number of sequences by continent as a function of time (see Table S1 for further details). (B) Schematic of the evolutionary relationship of the *ptxP* and *fim3* alleles. (C-E) Maximum clade credibility trees for the different genotypes: *ptxP1/fim3-1* (C), *ptxP3/fim3-1* (D) and *ptxP3/fim3-2* (E). Branch tips are colored by the continent of collection. Horizontal bars denote pertactin (PRN) expression (black: deficient expression, gray: wild-type expression, white: not known). The full tree is available on the MicroReact platform (<https://microreact.org/project/gWJRauq26T5JWXjNRCux4M-global-phylogeography-b-pertussis>). (F-H) Temporal trends in strain frequencies, computed on rolling 7-year windows, for the *ptxP3* (F) and *fim3-2* (G) alleles, and PRN deficiency (H). (I) Median spatial distance between *B. pertussis* pairs from different locations (EU: Europe; FR: France; JP: Japan) separated by different evolutionary times. The global median distance between pairs over the first 5 years is presented in Figure S1. The shaded area represents 95% CIs.



**Figure 2: *Bordetella pertussis* diversity within and between locations. (A)** Breakdown of the MRCA separating pairs of *B. pertussis* isolates from within the same year for different locations: within districts (within 12 European countries, Iran, and Japan), within European countries, within US states, and within continents. **(B)** Proportion of pairs within a region that belong to the same transmission chain (defined as MRCA<2y) as a function of population size (average from 19 countries). Proportions are computed for rolling windows of population sizes. Dots represent data and the dashed line represents model fit assuming an exponential relationship between the proportion of pairs from the same transmission chain and population size. The gray shaded region 95% confidence intervals. **(C)** Number of transmission chains within regions, as a function of population size. Numbers were computed for rolling windows of population sizes. Dots represent the estimates from the data. The dashed line represents model fit and the gray shaded region 95% confidence intervals. **(D-F)** Relative risk that a pair of bacteria have a MRCA within a defined period when coming from the same versus different district in France (D), same versus different country in Europe (E, blue), the same versus different state in the US (E, red), or the same versus different continent (F). Shaded regions in (D-F) represent 95% confidence intervals.



**Figure 3: Changing genotype distributions of *Bordetella pertussis* by country.** (A-D) Fits of the proportion of each genotype for 4 example countries: Australia (A), France (B), USA (C) and Japan (D) (other countries' fits can be found in Figure S9). Gray triangles represent data, gray bars denote binomial 95% confidence intervals. Blue lines and shaded areas represent the median and 95% credible interval of the posterior. Vertical dotted line denotes the year of the first ACV introduction, for each country. (E-G) Predicted *versus* observed counts (E), proportions (F), and Wrightian fitness (G), respectively. The Wrightian fitness of a strain is defined as the ratio of its proportions at time  $t+1$  and time  $t$ . Colors represent the continents. The dotted line denotes the identity line.



**Figure 4: Estimates of fitness of each genotype.** (A) Estimated fitness of each genotype as a function of the vaccine era. The dots and lines represent 2.5, 50, and 97.5 percentiles of the posterior distributions. PRN-deficient strains (PRN-) are shown in red, PRN-positive strains (PRN+) are shown in blue. (B) Effect of vaccine switch (whole-cell vaccine [WCV] to acellular vaccine [ACV]) for the different genotypes and PRN expressions. (C) Effect of PRN-deficiency on the different underlying genotypes. Horizontal lines and shaded areas represent the overall mean across all genotypes. The boxplots represent the 2.5, 25, 50, 75, and 97.5 percentiles of the posterior distributions.

