Shortomics

Comparative genomics of Czech vaccine strains of *Bordetella pertussis*

Ana Dienstbier[1], Derek Pouchnik[2], Mark Wildung[2], Fabian Amman[3], Ivo L. Hofacker[3,4], Julian Parkhill[5], Jana Holubova[6], Peter Sebo[6] and Branislav Vecerek[1*]

[1]Institute of Microbiology v.v.i., Laboratory of post-transcriptional control of gene expression, 14220 Prague, Czech Republic, [2]Laboratory for Biotechnology and Bioanalysis, Center for Reproductive Biology, Washington State University, Pullman, Washington 99164-7520, [3]University of Vienna, Institute for Theoretical Chemistry, Währinger Straße 17, A-1090 Vienna, Austria, [4]University of Vienna, Research group BCB, Faculty of Computer Science, Währinger Straße 24, 1090 Vienna, Austria, [5]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK, [6]Institute of Microbiology v.v.i, Laboratory of molecular biology of bacterial pathogens, 14220 Prague, Czech Republic.

*Corresponding author, Laboratory of post-transcriptional control of gene expression, Institute of Microbiology of the CAS v.v.i., Tel: +420241062507; E-mail: vecerek@biomed.cas.cz

One sentence summary: This is the first report on the genomics of Czech pertussis vaccine strains showing their uniqueness in terms of SNP-based phylogeny and genome organization

Keywords: Bordetella, pertussis, genomics, region of difference, genome rearrangement, vaccine pressure

ABSTRACT

Bordetella pertussis is a strictly human pathogen causing the respiratory infectious disease called whooping cough or pertussis. *B. pertussis* adaptation to acellular pertussis vaccine pressure ~~was~~ has been ~~recently~~ repeatedly highlighted, but recent data indicate that adaptation of circulating strains started already in the era of the whole cell pertussis vaccine (wP) use. We sequenced the genomes of five *B. pertussis* wP vaccine strains isolated in the former Czechoslovakia in the pre-wP (1954 - 1957) and early wP ~~use era~~ (1958 - 1965) eras, when only limited population travel into and out of the country was possible. Four isolates exhibit a similar genome organization and form a distinct phylogenetic cluster with a geographic signature. The fifth strain is rather distinct, both in genome organization and SNP-based phylogeny. Surprisingly, despite isolation of this strain before 1966, its closest sequenced relative appears to be a recent isolate from the US. On the genome content level, the five vaccine strains contained both new and already described regions of difference. One of the new regions contains duplicated genes potentially associated with transport across the membrane. The prevalence of this region in recent isolates indicates that its spread might be associated with selective advantage leading to increased strain fitness.

INTRODUCTION

Bordetella pertussis, the etiological agent of whooping cough (pertussis), is a strictly human Gram-negative bacterium infecting the respiratory tract (Cherry, 2010). Despite massive world-wide vaccination programs, pertussis remains the least-controlled vaccine-preventable infectious disease and it is a major cause of infant morbidity and mortality globally (WHO, 2006). As in many other countries, prior to introduction of the whole cell-based pertussis vaccine (wP), pertussis was the major cause of infant mortality in the former Czechoslovakia. The morbidity due to pertussis steeply declined after the compulsory vaccination was introduced in 1958 (Vysoka-Burianova *et al.*, 1976). As in other countries with high vaccination coverage, the incidence of pertussis in the Czechoslovakia and later on in the Czech Republic started to rise progressively since the 1990s (Raguckas *et al.*, 2007, Fabianova *et al.*, 2010, Sealey *et al.*, 2016) and this trend accelerated strongly upon switch to acellular pertussis subunit vaccine use in 2007 (Chlibek *et al.*, 2017).

While several factors are contributing to pertussis resurgence in the most developed countries, the two prominent ones are the incomplete and short-lived immunity induced by current aP vaccines and the genetic changes in circulating *B. pertussis* strains that lead to escape from immunity by antigenic variation (Mooi *et al.*, 2014, Burdin *et al.*, 2017). However, adaptive mutations occurred already in the wP era, suggesting that the major driving force of *B. pertussis* adaptation is vaccination as such (Bart *et al.*, 2014). *B. pertussis* is generally considered to be a genetically monomorphic pathogen (King *et al.*, 2010, Mooi, 2010) with rather limited extent of sequence variation within the global population (Bart *et al.*, 2014). Nevertheless, *B. pertussis* possesses an efficient mechanism of genome structure diversification due to the presence of almost 250 copies of the insertion sequence IS481 in its genome. These mobile elements allow for intragenomic recombination and excision and/or insertion of the flanked genome regions, leading to genome decay (Parkhill *et al.*, 2003, Cummings *et al.*, 2004), genome rearrangements (Bowden *et al.*, 2016, Weigand *et al.*, 2017) and gene expression alterations (Brinig *et al.*, 2006). Recently, we showed that insertion elements significantly affect the global gene expression profile in *B. pertussis* (Amman *et al.*, 2018). Apparently, *B. pertussis* adaptation goes beyond the changes in the genes coding for antigens that are present in the acellular vaccine and involves also other virulence-associated genes and the genes coding for surface-exposed proteins (Bart *et al.*, 2014). To understand how these vaccination-induced adaptation changes contributed to the current re-emergence of whooping cough, it is important to analyze genomes of strains from the pre-vaccine era. At present, there are over 1000 *B. pertussis* genome sequences deposited in Genbank and JGI GOLD databases (Mukherjee *et al.*, 2017). ~~c~~Close to

350 of them are completely assembled, of which ~~B. pertussis genomic sequences deposited at the GenBank database, of which~~ 330 are the genome sequences of the isolates from the aP vaccine era (since 1990s). In this study we thus sequenced and *de novo* assembled genomes of five historical *B. pertussis* strains that were collected in the former Czechoslovakia between 1954 and 1965. The very same strains were used for the formulation of the DTwP vaccine Alditepera, produced by the Institute of Sera and Vaccines in Prague (Pekarek & Rezabek, 1959, Pekarek & Rezabek, 1959). These strains were not extensively passaged under laboratory conditions and represent a unique set of isolates from the pre-wP vaccine (1954-1957) and early wP vaccine era (1958-1965). The content and organization of the genome of these strains was compared to that of other vaccine strains and recent clinical isolates.

GENOME SEQUENCING AND ANNOTATION

Genomes were sequenced using Illumina MiSeq (paired-end sequencing protocol) and PacBio RSII platforms. Illumina data is deposited under the project PRJEB4543 in Genbank. PacBio reads were assembled using HGAP SMRT Portal protocol and Illumina data was used to further polish the assemblies with Pilon software (Walker *et al.*, 2014). All genomes were *de novo* assembled into single contigs (Supplementary table 1), deposited in GenBank under accession numbers ERS2367611-ERS2367615 and annotated using Prokka software (Seemann, 2014).

PHYLOGENETIC ANALYSIS

Genomic analysis revealed that all Czech vaccine strains belong to *ptxP1* lineage (carry pertussis toxin promoter type 1). *B. pertussis* strains from *ptxP1* lineage form a phylogenetic cluster separate from *ptxP3* strains which emerged in the last 25-30 years (Bart *et al.*, 2014, Weigand *et al.*, 2017). Therefore, to put genomic sequences of the Czech vaccine strains into broader context with previously completely sequenced *ptxP1* and *ptxP2* strains of *B. pertussis*, we performed SNP-based phylogenetic analysis using the kSNP3.0 program with *k* of 23 and maximum parsimony method (Gardner *et al.*, 2015). In total we analyzed 19 *ptxP1* and 4 *ptxP2* strains, which were isolated from various geographic locations (USA, China, Japan, UK, Netherlands) from 1935 to 2012 (Supplementary Table 1). Phylogenetic analysis based on the 851 detected SNPs divided the strains into six major clusters (Figure 1A). Most of the recent isolates (isolated in 2000-2012) containing *ptxP1* allele clustered separately from the old *ptxP1* isolates (isolated in 1935-1965). Surprisingly, one of the old Czech strains, VS67, clustered together with the strain E945, which was isolated in the USA in 2005. The other four Czech strains formed a distinct cluster (cluster 4), separated from the other strains by six synonymous, six non-synonymous and three intergenic SNPs (Supplementary Table 2).

GENOME ORGANIZATION

The sequenced genomes were aligned by progressiveMauve algorithm with default parameters (Darling *et al.*, 2010). Alignment revealed that when compared to the reference strain Tohama I, all genomes contain large-scale structural rearrangements (Figure 1B). According to the genome organization, the strains could be classified into three groups. One group contains three strains: VS377, VS401 and VS366. Strain VS393 differs from this group by a single large inversion, which, among other genes, contains *fha/fim* and type III secretion system loci. VS67 differs from the other four strains by two additional large-scale inversions. In order to determine whether genome organization observed in the Czech strains can be also found among other already characterized strains, we have extracted and compared the permutation matrices from the Mauve alignment utilizing scripts published previously (Weigand *et al.*, 2017). This analysis revealed that Czech strains have a unique genome organization that has not been found so far in the other *ptxP1* and *ptxP2* strains (data not shown).

REGIONS OF DIFFERENCE

Mauve output was used to extract genome regions differentially present among the strains. In total eight such regions of difference were identified among the studied Czech strains and Tohama I (Table 1, Supplementary Figure 1). All RDs are either directly or in close proximity flanked by IS481 elements which indicates the mobile nature of the RDs.

Majority of the identified GRs have been described previously and for them we kept the previously established designation (Brinig *et al.*, 2006, Bart *et al.*, 2010). Two of the newly reported RDs were consecutively named RD30 and RD31 (Table 1). RD30, which is duplicated in VS67, contains genes, which code for the MFS transporter and a putative membrane protein (BP2451 and BP2452). It is possible that the duplicated region allows for enhanced transport of the cargo across the membrane. RD31 consists of 5 genes, but the only gene with assigned function is BP0894 coding for mannose-6-phosphate isomerase.

PREVALENCE OF RDs

Analysis of the association of the identified RDs with other *B. pertussis* strains showed that in many cases the distinct distribution of RDs among the strains correlated with their phylogenetic assignment (Supplementary Table 3). For instance, RD3 and RD5 are associated with pre-aP era *B. pertussis* strains from phylogenetic clusters 5 and 6. RD3 is also present in the Czech strains comprising cluster 4. In contrast, duplication of RD30 is prevalent in aP-era strains from the phylogenetic cluster 3. This suggests that the duplication might provide the currently circulating strains with a selective fitness advantage. The distribution of RD22-24 and RD26 among the strains is very similar, suggesting that the functions encoded within these loci might be linked. These GRs are found in almost all strains except for the cluster 6. RD29 and RD31 are missing in the Czech vaccine strains VS366 and VS393, respectively, indicating that the loss of these regions might have not conferred any advantage to these strains which possibly prevented their further spread.

DISCUSSION

In this study we conducted a comprehensive analysis of the genomes of five *B. pertussis* strains that were collected from 1954 to 1965 in the former Czechoslovakia at times when population travel into and out of the country was very limited. These representative isolates were later used for the development of the local wP vaccine. In contrast to the Japanese vaccine strain Tohama I, the Czech vaccine strains did not undergo massive passaging under laboratory conditions. This is the first study on Czech *B. pertussis* vaccine strains and one of the very few providing complete genome assemblies for the strains from the pre-wP vaccine or early wP use era. *De novo* sequencing of the genomes revealed that all Czech strains contain large-scale genome structural rearrangements compared to the reference strain Tohama I.

SNP-based phylogeny revealed that four of the strains form a separate cluster distinct from other so far analyzed strains, suggesting that at the time of their isolation the geographic factors played a significant role. It is tempting to speculate that following massive immunization by the wP vaccine, these strains disappeared from the population and did not spread globally. On the other hand, the fifth Czech strain V67 clusters together with a recent US isolate suggesting that it belongs to a lineage the descendants of which may still circulate within immunized population.

In agreement with SNP-based phylogenetic analysis, Czech *B. pertussis* strains exhibit a genome organization pattern that distinguishes them from other *ptxP1* and *ptxP2* strains, and contain some new and some previously reported regions of difference (Brinig *et al.*, 2006, Bart *et al.*, 2010). Loss of RD3 and/or RD5 is characteristic of *B. pertussis* strains isolated from other countries during the early wP use period (Kallonen *et al.*, 2011). Accordingly, RD5 was absent from all Czech strains and RD3 was lost from VS67 strain. Four of the RDs (RD22-24 and RD26) absent in Tohama I, are present in the Czech strains thereby supporting earlier reports which demonstrate the presence of these loci in the majority of older European *B. pertussis* isolates and vaccine strains (Kallonen *et al.*, 2011).

To conclude, our study suggests that the analysis of strains from pre-vaccine era is of high importance for our understanding of *B. pertussis* evolution in the light of pertussis resurgence. The impact of the various SNPs, RDs and genome re-arrangements on the physiological fitness and pathogenicity of each particular *B. pertussis* isolate, however, remains to be determined.

FUNDING

This work was supported by the Czech Health Research Council (www.azvcr.cz/) (grant 16-30782A) to BV and by funding from RVO61388971. This work was also supported by Mobility grant from Czech Academy of Sciences (MSM200201702) to AD, as well as by the Austrian FWF project I 1988-B22 to ILH.

ACKNOWLEDGEMENTS

We thank Dr. Bostik and Dr. Vavrova (Sevapharma, Prague) for the provision of the vaccine strains and Dr. Fabianova (NIPH, Prague) for helpful discussions.

REFERENCES

Amman F, D'Halluin A, Antoine R*, et al.* (2018) Primary transcriptome analysis reveals importance of IS elements for the shaping of the transcriptional landscape of Bordetella pertussis. *RNA Biol, 101080/1547628620181462655*.

Bart MJ, van Gent M, van der Heide HG, Boekhorst J, Hermans P, Parkhill J & Mooi FR (2010) Comparative genomics of prevaccination and modern Bordetella pertussis strains. *BMC genomics* **11**: 627.

Bart MJ, Harris SR, Advani A*, et al.* (2014) Global population structure and evolution of Bordetella pertussis and their relationship with vaccination. *mBio* **5**: e01074.

Bowden KE, Weigand MR, Peng Y*, et al.* (2016) Genome Structural Diversity among 31 Bordetella pertussis Isolates from Two Recent U.S. Whooping Cough Statewide Epidemics. *mSphere* **1**.

Brinig MM, Cummings CA, Sanden GN, Stefanelli P, Lawrence A & Relman DA (2006) Significant gene order and expression differences in Bordetella pertussis despite limited gene content variation. *Journal of bacteriology* **188**: 2375-2382.

Burdin N, Handy LK & Plotkin SA (2017) What Is Wrong with Pertussis Vaccine Immunity? The Problem of Waning Effectiveness of Pertussis Vaccines. *Cold Spring Harb Perspect Biol* **9**.

Cherry JD (2010) The present and future control of pertussis. *Clin Infect Dis* **51**: 663-667.

Chlibek R, Smetana J, Sosovickova R, Fabianova K, Zavadilova J, Dite P, Gal P, Naplava P & Lzicarova D (2017) Seroepidemiology of whooping cough in the Czech Republic: estimates of incidence of infection in adults. *Public Health* **150**: 77-83.

Cummings CA, Brinig MM, Lepp PW, van de Pas S & Relman DA (2004) Bordetella species are distinguished by patterns of substantial gene loss and host adaptation. *Journal of bacteriology* **186**: 1484-1492.

Darling AE, Mau B & Perna NT (2010) progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PloS one* **5**: e11147.

Fabianova K, Benes C & Kriz B (2010) A steady rise in incidence of pertussis since nineties in the Czech Republic. *Epidemiol Mikrobiol Imunol* **59**: 25-33.

Gardner SN, Slezak T & Hall BG (2015) kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome. *Bioinformatics* **31**: 2877-2878.

Kallonen T, Grondahl-Yli-Hannuksela K, Elomaa A, Lutynska A, Fry NK, Mertsola J & He Q (2011) Differences in the genomic content of Bordetella pertussis isolates before and after introduction of pertussis vaccines in four European countries. *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases* **11**: 2034-2042.

King AJ, van Gorkom T, van der Heide HG, Advani A & van der Lee S (2010) Changes in the genomic content of circulating Bordetella pertussis strains isolated from the Netherlands, Sweden, Japan and Australia: adaptive evolution or drift? *BMC genomics* **11**: 64.

Mooi FR (2010) Bordetella pertussis and vaccination: the persistence of a genetically monomorphic pathogen. *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases* **10**: 36-49.

Mooi FR, Van Der Maas NA & De Melker HE (2014) Pertussis resurgence: waning immunity and pathogen adaptation - two sides of the same coin. *Epidemiol Infect* **142**: 685-694.

Mukherjee S, Stamatis D, Bertsch J*, et al.* (2017) Genomes OnLine Database (GOLD) v.6: data updates and feature enhancements. *Nucleic acids research* **45**: D446-D456.

Parkhill J, Sebaihia M, Preston A*, et al.* (2003) Comparative analysis of the genome sequences of Bordetella pertussis, Bordetella parapertussis and Bordetella bronchiseptica. *Nature genetics* **35**: 32-40.

Pekarek J & Rezabek K (1959) An endocrinological test for innocuity of the pertussis vaccine. *J Hyg Epidemiol Microbiol Immunol* **3**: 79-84.

Pekarek J & Rezabek K (1959) The investigation of different components of pertussis vaccine obtained by centrifugation. *J Hyg Epidemiol Microbiol Immunol* **3**: 67-78.

Raguckas SE, VandenBussche HL, Jacobs C & Klepser ME (2007) Pertussis resurgence: diagnosis, treatment, prevention, and beyond. *Pharmacotherapy* **27**: 41-52.

Sealey KL, Belcher T & Preston A (2016) Bordetella pertussis epidemiology and evolution in the light of pertussis resurgence. *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases* **40**: 136-143.

Seemann T (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**: 2068-2069.

Vysoka-Burianova B, Burian V, Maixnerova M*, et al.* (1976) Surveillance of pertussis in the CSSR. IV. Immunological surveys of antibodies to pertussis and parapertussis in the Bohemian regions and in Slovakia in 1958 - 1971. *J Hyg Epidemiol Microbiol Immunol* **21**: 229-247.

Walker BJ, Abeel T, Shea T*, et al.* (2014) Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS one* **9**: e112963.

Weigand MR, Peng Y, Loparev V*, et al.* (2017) The History of Bordetella pertussis Genome Evolution Includes Structural Rearrangement. *Journal of bacteriology* **199**.

WHO (2006) Vaccine preventable deaths and the Global Immunization Vision and Strategy, 2006-2015. *MMWR Morb Mortal Wkly Rep* **55**: 511-515.

Dienstbier et al. Figure 1

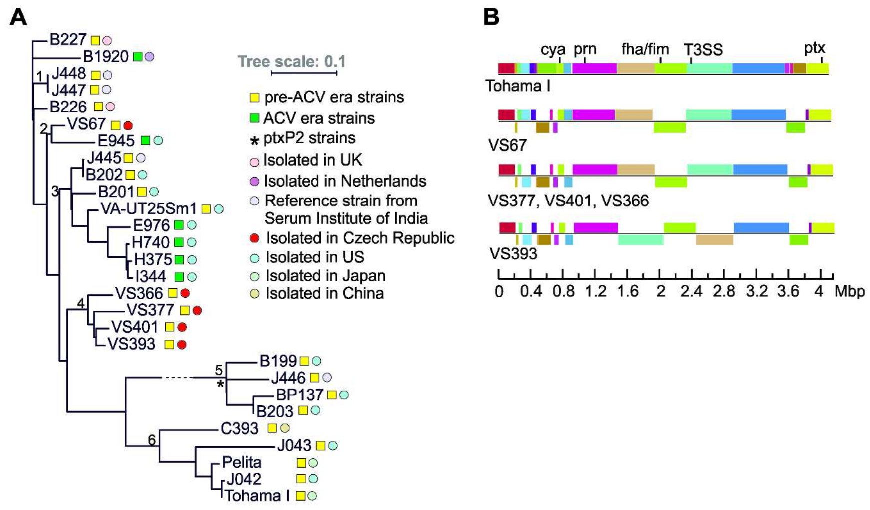


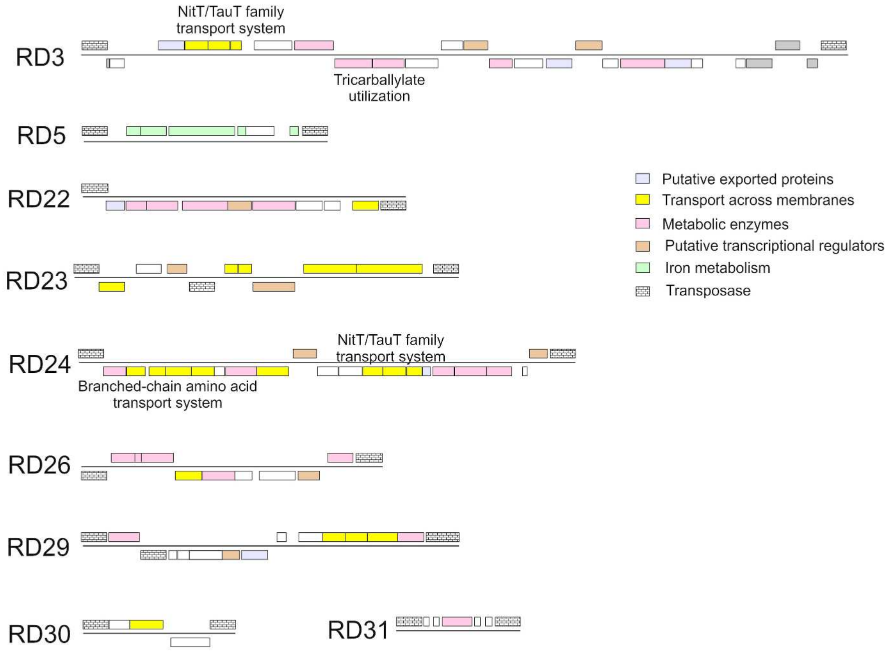
Figure 1. Phylogenetic tree and genome alignments of the analysed *B. pertussis* isolates. (**A**) SNP-based maximum parsimony phylogenetic reconstruction of *ptxP1* and *ptxP2* lineages of *B. pertussis* strains. All uncollapsed nodes have bootstrap value support of >75%. The dashed line indicates the branch length shortened for better visualization. (**B**) Genome alignment of the studied *B. pertussis* Czech isolates. Homologous gene blocks are denoted by the same colour. Genome loci coding for the key virulence factors are marked above the schematic Tohama I genome representation: *cya* – adenylate cyclase; *prn* – pertactin; *fha/fim* filamentous hemagglutinin and fimbriae; T3SS – type III secretion system; *ptx* – pertussis toxin.

355x270mm (300 x 300 DPI)

Table 1. Distribution of regions of difference among Czech vaccine strains

RD	Presence/Absence in strain	Start, RD	Stop, RD	Reference
RD3	Absent in VS67	BP0911	BP0937	Brinig et al., 2006
RD5	Present only in Tohama I	BP1136	BP1141	Brinig et al., 2006
RD22	Absent only in Tohama I	BB0541	BB0534	Brinig et al., 2006
RD23	Absent only in Tohama I	BB0917	BB0921	Bart et al., 2010
RD24	Absent only in Tohama I	BB1140	BB1158	Brinig et al., 2006
RD26	Absent only in Tohama I	BB4888	BB4880	Brinig et al., 2006
RD29	Absent only in VS366	BP2820	BP2832	King et al., 2008
RD30	Repeated 2x in VS67	BP2452	BP2451	This study
RD31	Absent only in VS393	BP0892	BP0896	This study

Dienstbier et al. Supplementary Figure 1



Graphical representation of the genetic content within the regions of difference identified in Czech vaccine strains . Regions are arranged in the VS67 strain gene order. Designation of the regions of difference is shown on the left, open reading frames are depicted as blocks and color coded according to the predicted function, their position indicates the strandedness.

287x264mm (300 x 300 DPI)

Supplementary table 1. Strains used in this study

Strain	Genbank Acces. No.	Country	Year	ptxP	Reference	Additional remarks
B199	CP022361	USA: PA	1935	ptxp2	Unpublished	
B203	CP012128	USA: MI	1939	ptxP2	Weigand *et al.*, 2016)	Sanofi-Pasteur MSD, strain 10536
B202	CP016338	USA: PA	1946	ptxP1	Weigand *et al.*, 2016	Lederle Laboratories, strain 134
J042	CP019869	USA	1947	ptxP1	Unpublished	
J043	CP016887	USA	1947	ptxP1	Unpublished	
C393	CP010963	China	1951	ptxP1	Bowden *et al.*, 2016	CS, Chinese vaccine reference
E476	CP010964	Japan	1954	ptxP1	Bowden *et al.*, 2016	Tohama I, GlaxoSmithKline vaccine reference
B201	CP013075	USA: IN	1955	ptxP1	Weigand *et al.*, 2017	
B227	CP013076	UK	1967	ptxP1	Weigand *et al.*, 2017	
B226	CP016957	UK	1967	ptxP1	Unpublished	
B1920	CP009752	Netherlands	2000	ptxP1	Bart *et al.*, 2014	
VA-UT25Sm1	CP015771	USA: ~~VA~~TX	~~2001~~1977	ptxP1	Unpublished	
E976	CP011175	USA: NY	2005	ptxP1	Weigand *et al.*, 2017	
E945	CP016956	USA: CA	2005	ptxP1	Unpublished	
H375	CP010961	USA: CA	2010	ptxP1	Bowden *et al.*, 2016	
H740	CP011190	USA: GA	2011	ptxP1	Weigand *et al.*, 2017	
I344	CP011255	USA: MN	2012	ptxP1	Weigand *et al.*, 2017	
Bp137	CP010323	USA	ND	ptxP2	Akamatsu *et al.*, 2015	Vaccine strain in Latin America
J448	CP017405	ND	ND	ptxP1	Weigand *et al.*, 2016	Reference strain from Serum Institute of India
J445	CP017402	ND	ND	ptxP1	Weigand *et al.*, 2016	Reference strain from Serum Institute of India
J447	CP017404	ND	ND	ptxP1	Weigand *et al.*, 2016	Reference strain from Serum Institute of India
J446	CP017403	ND	ND	ptxP2	Weigand *et al.*, 2016	Reference strain from Serum Institute of India
Pelita III	CP019957	Japan	ND	ptxP1	Unpublished	Indonesian reference strain, P.T. Bio Farma Indonesia
VS67	ERZ500380	Czech Republic	before 1966	ptxP1	This study	
VS393	ERZ500382	Czech Republic	before 1966	ptxP1	This study	
VS401	ERZ500384	Czech Republic	1954	ptxP1	This study	
VS377	ERZ500383	Czech Republic	before 1966	ptxP1	This study	
VS366	ERZ500381	Czech Republic	1957	ptxP1	This study	

REFERENCES

Akamatsu MA, Nishiyama MY, Jr., Morone M*, et al.* (2015) Whole-Genome Sequence of a Bordetella pertussis Brazilian Vaccine Strain. *Genome Announc* **3**.

Bart MJ, Harris SR, Advani A*, et al.* (2014) Global population structure and evolution of Bordetella pertussis and their relationship with vaccination. *MBio* **5**: e01074.

Bowden KE, Weigand MR, Peng Y*, et al.* (2016) Genome Structural Diversity among 31 Bordetella pertussis Isolates from Two Recent U.S. Whooping Cough Statewide Epidemics. *mSphere* **1**.

Weigand MR, Peng Y, Loparev V, Batra D, Burroughs M, Johnson T, Juieng P, Rowe L, Tondella ML & Williams MM (2016) Complete Genome Sequences of Bordetella pertussis Vaccine Reference Strains 134 and 10536. *Genome Announc* **4**.

Weigand MR, Peng Y, Loparev V*, et al.* (2016) Complete Genome Sequences of Four Bordetella pertussis Vaccine Reference Strains from Serum Institute of India. *Genome Announc* **4**.

Weigand MR, Peng Y, Loparev V*, et al.* (2017) The History of Bordetella pertussis Genome Evolution Includes Structural Rearrangement. *J Bacteriol* **199**.

Supplementary table 2. List of SNPs distinguishing the cluster 4 strains

SNP No.	Position	Allele	Intergenic/Synonymous/Non-synonymous	Annotation	Gene in E476 (Tohama I)
1	3967296	G/A	Synonymous	cytochrome oxidase subunit I	RD16_18905
2	3310990	G/A	Synonymous	molybdate ABC transporter substrate-binding protein	RD16_15605
3	1684734	G/A	Intergenic	histidine kinase	upstream of RD16_07985
4	1482030	C/T	Synonymous	ATP synthase	RD16_06985
5	1645729	G/A	Non-synonymous (A/T)	aminopeptidase	RD16_07800
6	2176522	C/T	Intergenic	MarR family transcriptional regulator	upstream of RD16_10235
7	1921439	G/A	Synonymous	hypothetical protein	RD16_09100
8	760046	C/T	Non-synonymous (G/D)	DNA methylase	RD16_03690
9	3477892	C/T	Non-synonymous (A/T)	ABC transporter substrate-binding protein	RD16_16365
10	33773	G/A	Non-synonymous (A/V)	2-methyl citrate dehydratase	RD16_00170
11	371766	A/G	Intergenic	glutamate dehydrogenase	Upstream of RD16_01860
12	1336320	C/T	Non-synonymous (E/K)	hypothetical protein	pseudogene
13	2157160	G/A	Synonymous	sodium transporter	pseudogene
14	4018189	C/T	Non-synonymous (G/D)	AraC family transcriptional regulator	RD16_19175
15	2401358	C/T	Synonymous	chemotaxis protein	RD16_11360

Supplementary table 3. Prevalence of identified genomic regions in strains used in this study

Strain	Genbank Acces. No.	RD3	RD30	RD29	RD31	RD24	RD22/ RD26	RD23	RD5
B199	CP022361	+		+	+	+	+	+	+
B203	CP012128	+		+	+	+	+	+	+
B202	CP016338		+	+	+	+	+	+	
J042	CP019869	+		+	+				+
J043	CP016887	+		+	+				+
C393 (CS)	CP010963	+		+	+	+	+		+
E476 (Tohama I)	CP010964	+		+	+				+
B201	CP013075		+	+	+	+	+	+	
B227	CP013076		+	+	+	+	+	+	
B226	CP016957		+	+	+	+	+	+	
B1920	CP009752		+	+	+		+	+	
VA-UT25Sm1	CP015771	+	+	+	+	+	+	+	
E976	CP011175		+	+	+	+	+	+	
E945	CP016956		+	+	+	+	+	+	
H375	CP010961		+	+	+	+	+	+	
H740	CP011190		+	+	+	+	+	+	
I344	CP011255		+	+	+	+	+	+	
Bp137	CP010323	+		+	+	+	+	+	+
J448	CP017405			+	+	+	+	+	
J445	CP017402		+	+	+	+	+	+	
J447	CP017404			+	+	+	+	+	
J446	CP017403	+		+	+	+	+	+	+
Pelita III	CP019957	+		+	+				+
VS67	ERS2367611		+	+	+	+	+	+	
VS393	ERS2367613	+		+		+	+	+	
VS401	ERS2367615	+		+	+	+	+	+	
VS377	ERS2367614	+		+	+	+	+	+	
VS366	ERS2367612	+			+	+	+	+	