INSTITUT PASTEUR

Research in Microbiology 159 (2008) 602–608

www.elsevier.com/locate/resmic

Genome-wide analysis of single nucleotide polymorphisms in *Bordetella pertussis* using comparative genomic sequencing

Ram P. Maharjan [a], Chong Gu [a], Peter R. Reeves [b], Vitali Sintchenko [c,d], Gwendolyn L. Gilbert [c,d], Ruiting Lan [a,*]

[a] *School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, NSW 2052, Australia*
[b] *School of Molecular and Microbial Biosciences, University of Sydney, Sydney, NSW 2006, Australia*
[c] *Centre for Infectious Diseases and Microbiology, Institute of Clinical Pathology and Medical Research, Westmead Hospital, Westmead, NSW 2145, Australia*
[d] *School of Health Sciences, University of Sydney, NSW 2006, Australia*

Received 25 May 2008; accepted 8 August 2008
Available online 26 August 2008

Abstract

Bordetella pertussis is known to be a genotypically homogeneous pathogen but the extent of homogeneity at the genomic level is unknown. A currently circulating *B. pertussis* isolate from Australia was compared with the genome-sequenced Tohama I strain isolated in Japan in the 1950s from a distantly related lineage. Microarray-based comparative genome sequencing (CGS) was used to detect single nucleotide polymorphisms (SNPs) in a total of 1.4 Mb of the 4.09 Mb genome, including 1012 coding-regions, 217 pseudogenes and 268 intergenic regions. The CGS analysis, followed by validation using real-time PCR and DNA sequencing, identified 70 SNPs and five 1–3 bp indels, giving an overall frequency of base changes of 1 per 20 kb. Thirty-two of the 56 SNPs in coding regions were non-synonymous, including five located in virulence-associated genes. The data also allowed us to compare genomic diversity with other ''clonal'' human pathogens such as *Mycobacterium tuberculosis* and *Yersinia pestis*, showing that *B. pertussis* may be one of the least variable pathogenic bacterial species.

Keywords: *Bordetella pertussis*; Single nucleotide polymorphisms; DNA microarray; Comparative genomic sequencing

1. Introduction

Despite extensive vaccination programs, pertussis remains endemic in many countries with ~50 million infections and 300,000 deaths worldwide each year, predominantly in developing countries [8]. Since the early 1990s, steady increases in infection rates have been observed in countries with high vaccination coverage, especially among adolescents and young adults [19]. A number of factors have been suggested to contribute to the increasing incidence of pertussis including waning immunity due to a limited duration of protection from the vaccine, cyclic variation in disease patterns, and emergence of new *Bordetella pertussis* variants evading vaccine induced immunity [16,30,48].

B. pertussis is known to be genetically homogenous and to have evolved only recently from a human-adapted lineage of *B. bronchiseptica* through reductive evolution [11,37]. However the extent of homogeneity at the genomic level is unknown. Multilocus enzyme electrophoresis [32,46] and multilocus sequence typing [11,36,47] have shown little sequence variation in housekeeping genes. More sequence variation has been observed in genes encoding antigens or virulence factors, including vaccine components – pertussis toxin (*ptx*), pertactin (*prn*), fimbriae (*fim2*) and surface proteins [31,36,47] – presumably because they are under greater selection pressure due to widespread vaccine-induced immunity. Pulsed field gel electrophoresis (PFGE) and multilocus variable number tandem repeat (VNTR) analysis (MLVA), have also demonstrated variation among strains of *B. pertussis* [2,4,12,20,26,42]. MLVA of six VNTR loci revealed an average of seven alleles per locus and 46 MLVA

* Corresponding author. Tel.: +61 2 9385 2095; fax: +61 2 938 51591.
E-mail address: r.lan@unsw.edu.au (R. Lan).

0923-2508/$ - see front matter © 2008 Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.resmic.2008.08.004

types in 198 Dutch isolates studied [42]. Microarray analysis of the presence/absence of genes from the Tohama I genome showed gene content variation in *B. pertussis* [6,9,23]. Twenty-six and 10 genomic regions were found to be variably present in the studies of Cummings et al. [9] and Caro et al. [6] respectively. However such data do not provide a quantitative measure of genome-wide nucleotide divergence.

Single nucleotide polymorphisms (SNPs) are the most common form of genetic variation and can be quantified precisely. It has been shown that SNPs best reflect phylogenetic relationships among closely related bacterial strains as they are largely due to mutation. Examples include variation within *Escherichia coli* O157:H7 [50], *Mycobacterium tuberculosis* [13,18] *Mycobacterium leprae* [29], *Salmonella enterica* serovar Typhi [35,41], *Yersinia pestis* [1], and *Bacillus anthracis* [38]. Genome-wide SNPs can be identified by comparing multiple whole genome sequences of the same species (e.g. *M. tuberculosis* [13,14,18], *Y. pestis* [1], and *B. anthracis* [38, 39]). However, this approach is not applicable to *B. pertussis*, as currently only one genome sequence is available.

In this study, we investigated the SNPs in 1497 chromosomal regions of a typical *B. pertussis* isolate from Australia, using a DNA microarray-based comparative genomic sequencing (CGS) method. The genome sequenced strain Tohama I, isolated in Japan in the 1950s, was used as a reference strain.

2. Materials and methods

2.1. Bacterial strains and genomic DNA preparation

We used a recent Australian *B. pertussis* isolate (strain L517) for the CGS analysis. It was isolated in December 2006 from an infected child and belonged to a major clone currently circulating in Australia, as demonstrated by MLVA (unpublished data). The genome sequenced strain Tohama I (NCTC13521), purchased from the National Culture Type Collection (UK), was used as a reference strain. Bacterial isolates were grown on Regan–Lowe medium containing charcoal agar (Oxoid, UK), 10% horse blood and *Bordetella* selective supplement cephalexin (Oxoid, UK) at 37 °C for 3–5 days in a humidified incubator. Genomic DNA was extracted and purified from plate cultures using the phenol–chloroform method as described by Octavia and Lan [35] and stored at −20 °C.

2.2. Comparative genomic analysis

Pair-wise gene-by-gene comparisons between *B. pertussis* (GenBank accession no: NC_002929), *B. parapertussis* 12822 (GenBank accession no: NC_002928) and *B. bronchiseptica* RB50 (GenBank accession no: NC_002927) were conducted using MegaBlast (http://www.ncbi.nlm.nih.gov/blast/megablast.html) available through the Australian National Genetic Information Service to identify orthologs and variation between genomes.

2.3. SNP discovery

All genes and other regions used for SNP discovery are listed in Table S1. We used the DNA tiling microarray-based comparative genomic sequencing (CGS) method available from NimbleGen System Inc, (Madison, WI, USA) to discover SNPs in *B. pertussis*. The NimbleGen system uses two sets of microarrays: an initial mutation mapping array and a resequencing array. The former is used to locate potential SNPs in the test strain by tiling 29-mer oligonucleotide probes across each strand of targeted regions of a genome with spacing of seven nucleotides and overlaps of 22 nucleotides. Resequencing microarrays were employed to identify the exact position and nature of SNPs detected in the mutation mapping array. A total of 48 kb were resequenced using 386,000 probes with four probes per base position in each strand. Details of mutation mapping and resequencing procedures can be found elsewhere [3].

2.4. Validation of SNPs

All SNPs reported by NimbleGen were validated by real-time PCR using hairpin-primers (HP-RT-PCR) as described by Hazbon et al. [21]. Those that could not be confirmed by the HP-RT-PCR method were further investigated by sequencing (automated DNA sequence analyzer ABI3730, Applied Biosystem) after conventional PCR amplification of the targeted regions. The false-positive and negative rates of the CGS were calculated based on results from both HP-RT-PCR and sequencing. The primers used for real-time PCR and conventional PCR are listed in Table S2.

2.5. PCR assays

All conventional PCR reactions were performed in a Bio-Rad MJ Mini-personal thermocycler (BioRad, USA). Each 50 µl reaction contained a mixture of 1 µl of template DNA (∼20 ng), 0.5 µl (30 pmol/µl) of each forward and reverse primer, 0.5 µl of dNTPs (10−mM), 5 µl of 10× Taq polymerase PCR buffer (New England Biolabs), 0.5 µl (2.5 U) of Taq polymerase (New England Biolabs) and sterile MilliQ water to adjust the final volume. After initial denaturation at 94 °C for 2 min, the reaction was performed for 35 cycles at 94 °C for 15 s, 55 °C for 30 s and 72 °C for 1 min followed by 10 min final extension at 72 °C. The PCR products were verified on agarose gels.

Real-time PCR was performed in the Corbett Rotor-Gene 6000 (Corbett Research, Australia). Each 10 µl reaction contained a mixture of 0.5 µl (5 pmol/µl) of each forward and reverse primer, 1 µl (20 ng) of DNA, 5 µl of Sensi-Mix Sybr plus (Bioline, Australia) and MilliQ water to adjust the final volume. The reaction profiles consisted of holding at 95 °C for 10 min to activate Taq polymerase followed by two thermal cycling profiles. The first profile was performed for 10 cycles at 95 °C for 15 s, 69 °C for 20 s and 72 °C for 20 s with a decrease of 1 °C for the annealing step in each cycle, while the second profile was performed for 40 cycles at 95 °C for