PHD

Investigating the genome of Bordetella pertussis using long-read sequencing

Ring, Natalie

Award date:
2020

Awarding institution:
University of Bath

Link to publication

Investigating the genome of *Bordetella pertussis* using long-read sequencing

submitted by

Natalie Ring

for the degree of Doctor of Philosophy

of the

University of Bath

Department of Biology and Biochemistry

July 2020

ORCID ID: 0000-0002-8971-8507

Acknowledgements

Firstly, many thanks to Dr Stefan Bagby and Dr Andrew Preston for their supervision, guidance, and patience throughout the last three (and then some more) years, as well as for coming up with the project idea and finding funding for it in the first place.

On that note, I am indebted to the University of Bath and Oxford Nanopore Technologies for funding this project, and to ONT for letting me present (and paying my transport and accommodation fees!) at several of their international meetings. Thanks also to the Microbiology Society and Genetics Society for funding travel to two conferences in 2018, and to Bath's Department of Biology and Biochemistry for funding my trip to the International Bordetella Symposium in 2019.

My first first-author research paper (Chapter 2) would not have been possible without the bucket load of data I produced at ONT HQ in summer 2017. I am very grateful to ONT for the provision of many flow cells and reagents during that time, and to Dan Fordham and Phill James for sharing their lab and expertise with me. Additional ONT thanks go to Dan Turner, my industrial supervisor.

Thanks to the Nanopore Group in Santa Cruz for helping Stefan get started with the R7 MinION sequencing in 2015, and for one final R9.4 run which helped me at the beginning of my PhD in 2016. Likewise, acknowledgement is due to Robert Chapman and Joshua Spear, Bagby Lab final year project students in 2015 and 2016, who performed some of the original nanopore sequencing of *Bordetella pertussis*, laying the groundwork for this PhD project.

Josh Quick and Nick Loman at the University of Birmingham hosted me for several days over the summer of 2018, helping me to extract and sequence some (fairly) ultra-long *B. pertussis* DNA from various strains for Chapter 4, even if my assembly masterplan didn't work in the end. Thanks to them for their time, as well as for being nanopore gurus in general.

Many thanks go to Audrey and colleagues at New Zealand's Institute of Environmental Science and Research for selecting, preparing and shipping their *B. pertussis* samples around the world, after being contacted out of the blue by us in May 2018. Without them, Chapter 3 wouldn't exist!

Thanks also to Ayla, Mohsina, Mareike and everyone else who has come and gone from 4S 0.30 for their support and comradery, for watching Crufts and various sports with me during quiet moments, and for listening to me moan when things weren't working more often than was probably interesting!

Massive thanks to Iain MacArthur and Stacy Ramkissoon for their endless help and moral support during my ventures into "proper" microbiology lab work, and particularly to Stacy for her friendship and positivity since Day One (including so many lunches, WhatsApp gossips, crazed phone calls and trips to Disneyland!).

To my parents, sister and brother-in-law, friends from back home, and parents-in-law, thanks for providing much needed distraction through trips to Auckland, Mayrhofen, Cornwall and beyond, as well as for all your support and, again, listening to more moaning than was probably strictly necessary. Thanks also to my nephew Arthur for being an endless source of entertainment and pride since November 2017, and acknowledgment in advance to KulaRing number 2, due September 2020.

I will be forever grateful to my husband and the most patient person I know, Jamie McBrien. Thanks for letting me drag you back to Bath (and now to Edinburgh), for listening to roughly a thousand different iterations of essentially the same presentation, and for your help with the VBA code used in Chapter 5 – finally, a use for your least useful skill! Lastly, thanks to McBRing junior for unknowingly keeping me company whilst I finished this thesis - we look forward to meeting you in November 2020!

Abstract

Whooping cough, the respiratory disease caused by the bacterium *Bordetella pertussis*, has been resurgent for the last thirty years. Several reasons have been suggested for this resurgence, including increased awareness and improved diagnosis techniques, waning immunity conferred by the whooping cough vaccine, and genetic shifts of circulating bacteria away from the vaccine strains. These genetic changes may have been accelerated by the switch in many countries from a whole cell vaccine to an acellular vaccine containing just one to five *B. pertussis* antigens. Aside from certain key genes, however, variation between *B. pertussis* strains appears to be very limited on the level of single genes. Instead, in recent years, a picture of genome level inter-strain variation has been emerging, beginning with the revelation that genomic rearrangements, mediated by the numerous insertion sequence elements in the *B. pertussis* genome, are common. Investigations of whole genome variation, alongside classic molecular epidemiological studies, may therefore be important to our understanding of how genetic changes in *B. pertussis* are contributing to whooping cough resurgence.

Many genome level changes may be observable only in closed genome sequences. The highly repetitive *B. pertussis* genome, which contains up to 300 identical copies of a 1,000 bp insertion sequence, has traditionally been difficult to resolve to the single-contig level, with most genomes assembled using Illumina sequencing data consisting of at least as many contigs as there are insertion sequence copies. Here, I first define a sequencing and data processing pipeline, utilising nanopore long-read and Illumina short-read sequencing to enable the assembly of accurate, closed *B. pertussis* genome sequences. Using this hybrid sequencing pipeline, I then investigate the genomes of 66 *B. pertussis* strains isolated in New Zealand between 1982 and 2008. New Zealand commonly sees a higher rate of incidence of whooping cough than most other countries, and no isolates from the country had previously been sequenced. Several of the genomic features of the New Zealand isolates match those observed in many other countries, including a selective sweep from strains carrying the *ptxP1* allele to the *ptxP3* allele and a recent rapid increase in the number of strains which are unable to produce pertactin, one of the antigens usually included in the acellular vaccine. Nonetheless, the data also indicate that the strains circulating in New Zealand might be more genetically similar than those circulating in other countries, particularly in recent years, and particularly during whooping cough outbreaks. This strain screen is the first of its kind to use nanopore sequencing, and to include traditional analysis of genotypes with analysis of genome level variation, such as rearrangements and copy number variations. Next, I attempt to investigate an ultra-long genomic duplication identified whilst testing the hybrid assembly pipeline on five UK *B. pertussis* strains. This work ultimately shows that the complexity of the *B. pertussis* genome can make *in vitro* studies into the links between genotype and phenotype difficult. Finally, I use the closed genome sequences of every *B. pertussis* strain sequenced with long read technologies to investigate any recent changes in filamentous haemagglutinin, another of the antigens included in the acellular vaccine. Studies of the genes coding for this antigen have typically been limited by its length and repetitive nature, which have hindered attempts to assemble its whole sequence. This work reveals a homopolymeric locus which may be prone to slippage and which, under selective pressure, could therefore lead to an increase in the numbers of strains which are deficient in this vaccine antigen.

Overall, the work in this thesis demonstrates how long-read sequencing can reveal previously unstudied or intractable aspects of *B. pertussis* biology, along with defining an affordable method for using nanopore long-read sequencing to assemble and study closed *B. pertussis* genomes.

Conflict of Interest Statement

This PhD was 50% funded by Oxford Nanopore Technologies. In addition, the sequencing runs described in Chapter 2 were conducted at ONT HQ, using reagents and flow cells provided by ONT. Care has been taken in this thesis to maintain objectivity regarding the advantages and disadvantages of nanopore sequencing as far as possible.

List of Abbreviations

ACV	Acellular vaccine
BUSCO	Benchmarking Universal Single-Copy Orthologs
CDC	Centers for Disease Control
CLIMB	Cloud Infrastructure for Microbial Bioinformatics
CNV	Copy number variant
DTaP	Diphtheria-Tetanus-Pertussis vaccine containing ACV
DTwP	Diphtheria-Tetanus-Pertussis vaccine containing WCV
ESR	New Zealand's the Institute of Environmental Science and Research, funded by the New Zealand Ministry of Health
FFPE	Formalin-fixed, paraffin-embedded
FHA	Filamentous haemagglutinin
Fim2/3	Fimbrial 2/ Fimbrial 3
gDNA	Genomic DNA
HGP	Human Genome Project
HMM	Hidden Markov Models
HPC	High performance computing
Indel	Insertion/deletion mutation
IS	Insertion Sequence
MLEE	Multilocus enzyme electrophoresis
MLST	Multilocus sequence type
MLVA	Multiple-locus tandem repeat analysis
MRC	Medial Research Council
NCBI	National Center for Biotechnology Information
ONT	Oxford Nanopore Technologies
ORF	Open reading frame
PacBio	Pacific Biosciences
PBS	Phosphate-buffered saline
PCR	Polymerase chain reaction
PFGE	Pulsed field gel electrophoresis
PRN	Pertactin (protein)

PT	Pertussis toxin (protein)
qPCR	Quantitative polymerase chain reaction
RAPD	Randomly amplified polymorphic DNA
RFLP	Restriction fragment length polymorphism
rRNA	Ribosomal RNA
RT-qPCR	Reverse transcription quantitative polymerase chain reaction
SBL	Sequencing-by-ligation
SBS	Sequencing-by-synthesis
SNP	Single nucleotide polymorphism
SPRI	Solid phase reversible immobilization
SRA	Sequence read archive
ssDNA	Single-stranded DNA
Tdap	Tetanus-diptheria-acellular pertussis vaccine (booster)
*vag*s	Virulence-activated genes
vrg*s*	Virulence-repressed genes
WCV	Whole cell vaccine
WGS	Whole genome sequencing
WHO	World Health Organisation

Contents

Chapter 1: Introduction

Parts of this chapter have been adapted from:

"How genomics is changing what we know about the evolution and genome of *Bordetella pertussis*".

Natalie Ring, Jonathan Abrahams, Stefan Bagby, Andrew Preston and Iain MacArthur

Advances in Microbiology, Infectious Diseases and Public Health, doi: 10.1007/5584_2019_401

July 2019

"Sometimes, when it's going badly, she wonders if what she believes to be a love of the written word is really just a fetish for stationary"

- David Nicholls, One Day

1.1 Bordetella pertussis *and whooping cough*

The *Bordetellae*

The *Bordetella* genus contains a growing number of gram-negative coccobacilli. There are currently 16 named species in the genus, with many additional "*Bordetella sp.*" and "*Bordetella genomosp.*" members which have yet to be characterised and named (see **Table 1.1** for further details and references). *Bordetella* species occupy diverse ecological niches, including soil, water, plants and animals, having been isolated from a wide variety of settings, from human wounds to mural paintings in an ancient tomb. Phylogenies based on 16S rRNA for all species found so far suggest that the origins of the genus were likely environmental (Hamidou Soumana, Linz and Harvill, 2017). **Figure 1.1** shows a simple phylogeny based on representative 16S rRNA sequences for the 16 named species and, indeed, the environmental species are found throughout the tree, whereas the animal-adapted species are restricted to one major branch. Recently, metagenomic studies have resulted in a surge in the number of environmental *Bordetella* identified, with new findings reported in the literature regularly (for example: Calon et al., 2019; Qi et al., 2019; Zheng, X. et al., 2019). The species which are currently best characterised, however, are the mammal-adapted classical *Bordetellae*: *Bordetella bronchiseptica*, *Bordetella pertussis* and *Bordetella parapertussis*.

Table 1.1 Named *Bordetella* species according to NCBI's Taxonomy Browser, as of June 2019

***Bordetella* Species**	**Taxon ID**	**Originally isolated from**	**Primary Host(s)**	**Key publication**
ansorpii	288768	Human epidermal cyst	Humans	Ko et al. (2005)
avium	521	Turkeys with respiratory disease	Poultry	Kersters et al. (1984)
bronchialis	463025	Respiratory samples from cystic fibrosis patients	Humans	Vandamme et al. (2015)
bronchiseptica	518	Dogs with respiratory disease	Various animals, including humans	Ferry (1912)
flabilis	463014	Respiratory samples from cystic fibrosis patients	Humans	Vandamme et al. (2015)
hinzii	103855	Immunocompromised humans and poultry with respiratory disease	Humans, poultry, rabbits	Vandamme et al. (1995)
holmesii	35814	Humans with blood infection	Human	Weyant et al. (1995)
muralis	1649130	Environmental (tomb mural painting)	Environmental	Tazato et al. (2015)
parapertussis	519	Humans with respiratory disease	Humans, sheep	Eldering and Kendrick (1938)
pertussis	520	Humans with respiratory disease	Humans	Bordet and Gengou (1906)
petrii	94624	Environmental (anaerobic bioreactor)	Environmental, humans	von Wintzingerode et al. (2001)
pseudohinzii	1331258	Mice with respiratory disease,	Mice	Ivanov et al. (2016)
sputigena	1416810	Respiratory samples from cystic fibrosis patients	Humans	Vandamme et al. (2015)
trematum	123899	Humans with wounds or otitis media	Humans	Vandamme et al. (1996)
tumbae	1649139	Environmental (tomb mural painting)	Environmental	Tazato et al. (2015)
tumulicola	1649133	Environmental (tomb mural painting)	Environmental	Tazato et al. (2015)

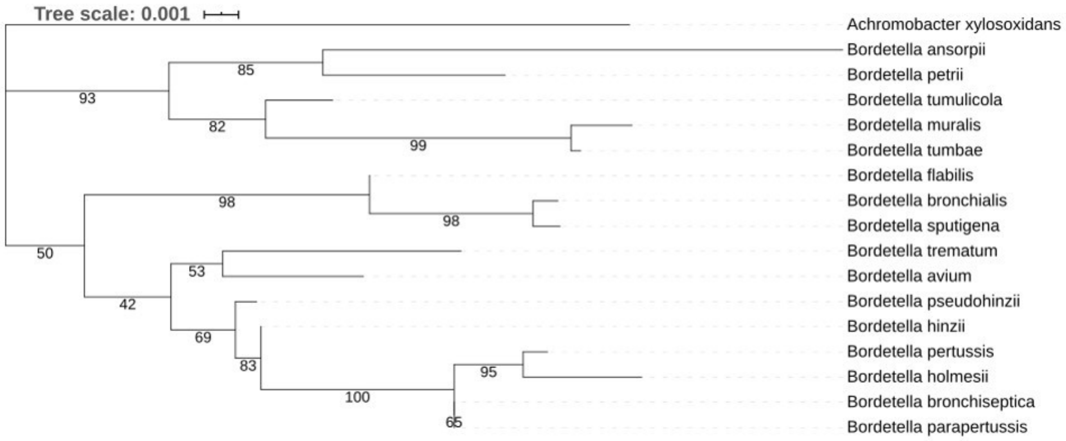


Figure 1.1 The phylogenetic relationships between the 16 named *Bordetella* species according to NCBI's Taxonomy Browser (as of July 2020), with *Achromobacter xylosoxidans* (also from the family *Alcaligenaceae*) as an outgroup. The classical *Bordetellae* cluster together, along with *B. holmesii*. Clustal W was used to align the 16S rRNA sequence for each species, a tree was constructed with IQ-Tree 2, and the tree was visualised with iTOL, with up to 2000x bootstrapping (Thompson, Higgins and Gibson, 1994; Letunic and Bork, 2019; Minh et al., 2020). A phylogeny constructed from whole genome sequences may show slightly different relationships, particularly between species which are found on the same branches; however, for some of the lesser-studied *Bordetellae*, only 16S rRNA sequences are currently available.

All three classical *Bordetella* species can cause respiratory disease. *B. bronchiseptica* was first isolated in 1911, when it was identified as a cause of distemper (specifically, kennel cough) in dogs (Ferry, 1912). *B. bronchiseptica* has subsequently been found to be capable of infecting numerous different mammals, including cats, rabbits, pigs, horses, seals, koalas and, rarely, humans (McGowan, 1911; Gallagher, 1965; Fisk and Soave, 1973; McKenzie, Wood and Blackall, 1979; Goodnow, 1980; Baker and Ross, 1992; Bjornstad and Harvill, 2005). In contrast, both *B. pertussis* and *B. parapertussis* are highly niche-restricted. Two distinct lineages of *B. parapertussis* have been identified, one which infects sheep and one which causes a mild form of whooping cough in humans (Heininger et al., 1994; van der Zee, Groenendijk, et al., 1996). *B. pertussis* is found only in the human nasopharynx, where it is responsible for causing the vast majority of cases of whooping cough.

Whooping cough infection has three distinctive stages

Whooping cough, also known as pertussis, is a respiratory disease which most severely affects infants younger than 12 months, but which can affect anyone of any age. Recent estimates suggest that 24.1 million children are infected with whooping cough annually, resulting in almost 161,000 deaths (Yeung et al., 2017). A typical case of whooping cough follows a distinctive pattern, comprising three clinically-defined stages. The catarrhal stage, which includes symptoms much like the common cold such as a sore throat, runny nose and a non-productive cough, lasts around two weeks. The catarrhal stage is followed by the spasmodic stage, which is characterised by regular lengthy episodes of paroxysmal coughing which can cause cyanosis, and which typically end with a deep inspiration accompanied by the classic "whooping" sound and, often, vomiting. After around two weeks of this acute stage, the convalescent stage gradually develops. During the convalescent stage, coughing episodes slowly decrease in frequency as the body recovers, although complications such as secondary infections leading to pneumonia or, more rarely, seizures and encephalopathy, can still occur (Lauria and Zabbo, 2020). Convalescence can last from two weeks up to several months (Cherry, 1999; Kent and Heath, 2014).

Whooping cough is rarely diagnosed during its earliest stage, because the symptoms are so reminiscent of less serious illnesses. Diagnosis therefore tends to occur during the spasmodic or convalescent stages, by which time culture testing is unlikely to be positive because most bacteria have already been cleared by the host. Diagnosis by polymerase chain reaction (PCR) or serology is therefore recognised as more sensitive, although the specificity of serology can be affected by the age of the patient: infants under three months may not develop detectable antibodies, resulting in false negatives, whilst testing within one year of immunisation can produce false positives (Kent and Heath, 2014). Antibiotics can be prescribed for whooping cough infection, but often do not alter the clinical course of the infection, likely due to the common delay in diagnosis (Altunaiji et al., 2007).

Whooping cough in the pre-vaccine era

Whooping cough appears to be a relatively young disease. Writings which describe any illness with the distinctive symptoms of whooping cough are largely absent from historical records until the 1400s. Reports from Persia seem to describe the first outbreaks of a disease which sounds very similar to modern day whooping cough in the late 1400s, whilst a paper by Nils Rosen von Rosenstein from the 18th century claims the first cases occurred in France in 1414 (Mattoo and Cherry, 2005; Aslanabadi et al., 2015). Less than a century later, Guillaume de Baillou produced what has long been considered to be the first description of a whooping cough epidemic, in Paris in 1578 (Hardy, 1993). Further outbreaks were reported regularly in Europe from then onwards, and the disease was named "pertussis", meaning violent cough, in 1679 (Mattoo and Cherry, 2005). The etiological agent behind pertussis, however, remained a mystery until the beginning of the 20th century.

The bacterium *B. pertussis* was first identified by Dr Jules Bordet at the Pasteur Institute in Paris in 1900, when he observed a "small ovoid Gram-negative bacterium" in expectorate sampled from Bordet's own five month-old daughter, who was suffering from pertussis. Six years later in Brussels, Bordet and Octave Gengou used a medium they had developed themselves, Bordet-Gengou (BG) medium, to isolate bacteria from expectorate from Bordet's son, Paul, who had also developed the disease (Cavaillon, Sansonetti and Goldman, 2019). Bordet and Gengou's medium was prepared from potatoes, serum, agar and blood, and is still widely used to grow *Bordetella* species. The bacterium was originally named *Haemophilus pertussis* due to the observation that freshly isolated bacteria were haemolytic, although it was also noted that after *in vitro* culture, the bacteria were no longer haemolytic. The genus was eventually renamed *Bordetella* in honour of Bordet later in the twentieth century, after more species, *B. bronchiseptica* and *B. parapertussis*, were isolated, and the species were collectively found to differ sufficiently from other *Haemophilus* to be reclassified (Bordet and Gengou, 1906; Howson, Howe and Fineberg, 1991; Guiso, 2009).

Whole cell vaccination reduced global whooping cough incidence, but was associated with reactogenicity

Once the bacterium responsible for whooping cough had been identified, efforts to develop an effective whooping cough vaccine began. Thorsvald Madsen described the first large-scale use of a whooping cough vaccine containing whole *B. pertussis* cells in 1925. Although somewhat successful in controlling a whooping cough outbreak in Europe, the vaccine had potential serious adverse effects, with two infant deaths occurring within 48 hours of immunisation (Madsen, 1933; Howson, Howe and Fineberg, 1991). Throughout the 1930s, further experimental vaccines were used to both prevent whooping cough and to treat ongoing infections, with varying levels of efficaciousness (Cherry, 1996;

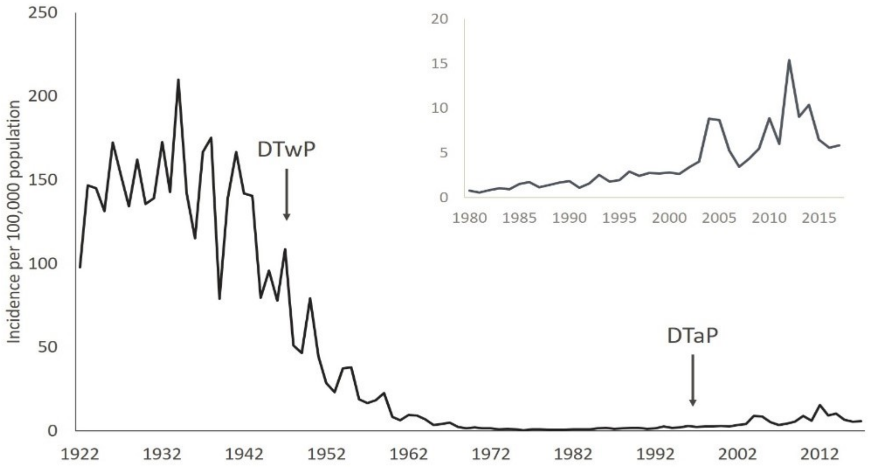


Figure 1.2 Whooping cough incidence in the USA, 1922-2017. The whole cell vaccine, labelled DTwP here, was made widely available from 1948 and was replaced by the acellular vaccine, labelled DTaP, in 1996. Annual notification data from Centers for Disease Control (CDC) Surveillance and Reporting website, annual population data from US Census Bureau. Inset shows years of resurgence only.

Mattoo and Cherry, 2005). By the 1940s and 1950s, however, more fully developed whole-cell vaccines (WCVs), containing chemically killed bacteria, were deployed and widely used in many countries, usually combined with vaccines for diphtheria and tetanus (collectively called the DTwP vaccine).

As seen in **Figure 1.2**, in the pre-vaccine era the average yearly incidence of reported whooping cough was 157 per 100,000 people in the United States, and 230 per 100,000 people in the United Kingdom, although it was suspected that fewer than 20% of cases were diagnosed, suggesting the true incidence was much higher. Pertussis infection was endemic; in an average year, up to 150,000 people were affected and around 300 deaths were linked with whooping cough infection in the United Kingdom. In addition, epidemics occurred on average every three years (Cherry, 1984; Mattoo and Cherry, 2005; Amirthalingam, Gupta and Campbell, 2013). The introduction of the DTwP appeared to be effective: for example, in England and Wales the incidence in infants under 1 year fell from 12.1 per 1,000 population in 1954-1957 to only 1.6 per 1,000 population in 1970-1973 (Amirthalingam, Gupta and Campbell, 2013). However, the three-yearly epidemic cycle continued, albeit with greatly reduced infection rates (Fine and Clarkson, 1982).

A perception grew over several decades that the pertussis component of the DTwP was responsible for side effects ranging from vaccine-site soreness and swelling to encephalopathy or even death. This perception peaked in the 1970s, leading to a dramatic decrease in DTwP uptake (Byers and Moll, 1948; Cody et al., 1981). In the United Kingdom, it was estimated that the rate of uptake fell from around 80% to as low as 9% in some areas of the country (British Medical Journal, 1981). Contemporary and subsequent research has shown that the most severe adverse effects were likely not caused by the vaccine itself; rather, the vaccine was given at around the same time that such illnesses would have occurred in the affected infants anyway (Cherry, 1990, 1992; Blumberg et al., 1993; Cherry et al., 1993; Cherry, 1996; Moore et al., 2004). Nonetheless, public confidence in the vaccine had been damaged.

Acellular vaccination replaced WCV in many countries throughout the 1990s and 2000s

Although uptake of the DTwP stabilised somewhat after the scare in the 1970s, throughout the 1980s and 1990s a new type of whooping cough vaccine was developed. The new vaccine does not contain whole bacterial cells; instead it contains one to five purified *B. pertussis* antigenic proteins. As such, the new vaccine is termed the acellular pertussis vaccine (ACV). The antigens included in the ACV are summarised in **Table 1.2**, whilst **Table 1.3** shows which versions of the pertussis vaccine are in use in the countries from which data is considered in this thesis. Compared to the WCV, the ACV results in fewer adverse effects such as fever, site soreness and inflammation, although the frequency of each of these events still increases with each subsequent vaccine dose (Mattoo and Cherry, 2005).

Like the WCV before it, the ACV is usually given in combination with diphtheria and tetanus toxoids (DTaP) and sometimes also with polio, *Haemophilus influenzae* type b and hepatitis B (the "6-in-1", DTaP-IPV-Hib-HepB vaccine) (Chen, Z. and He, 2017). In the UK, the three- or five-pertussis antigen DTaP vaccine is administered to infants at 8, 12, and 16 weeks, followed by a pre-school booster between 3 and 5 years. A further booster vaccine, the Tdap, is available for adolescents and adults; the Tdap contains lower doses of the *B. pertussis* antigens, to mitigate the adverse effects observed from repeated doses of pertussis vaccines (Amirthalingam, Gupta and Campbell, 2013). Since the 2012 whooping cough outbreak, a Tdap booster has been recommended for pregnant women in the UK (PHE, 2018). **Figure 1.3** shows the distribution of vaccine uptake across the globe: in general, most developed countries have coverage higher than 90%, whereas developing countries tend to be slightly lower. Similarly, most developed countries now use the ACV, whilst many developing countries have continued to use the WCV, likely due to the increased expense associated with producing the purified antigens for the ACV compared to killed whole bacteria for the WCV (Chen, Z. and He, 2017).

One to five *B. pertussis* virulence factors are included in the acellular pertussis vaccine

Most formulations of the ACV contain pertactin (PRN). Pertactin is a 69 kDa autotransporter protein. Due to an Arg-Gly-Asp-binding motif in the N-terminal domain, PRN is thought to be an adhesin (Mattoo and Cherry, 2005). The N-terminal domain is transported across the *B. pertussis* cell membrane by the C-terminal domain, which takes the form of a transmembrane β-barrel (Henderson, I.R. and Nataro, 2001). Outside of the membrane, the N-terminal domain is thought to play a part in adherence to host cells; this has been shown to be likely in *Salmonella* and *Escherichia coli* (Charles et al., 1989; Emsley et al., 1996; Everest et al., 1996). Although Δ*prn* deletions have not been shown to have any effect on adherence *in vitro* or *in vivo*, it was noted during early trials of the ACV that anti-PRN antibodies are important during infection, hence PRN is included in many ACV formulations (see **Table 1.3**) (Lipscombe et al., 1991; Roberts et al., 1991; Cherry et al., 1998; Storsaeter et al., 1998).

Table 1.2 Antigens included in the acellular pertussis vaccine, and their characteristics

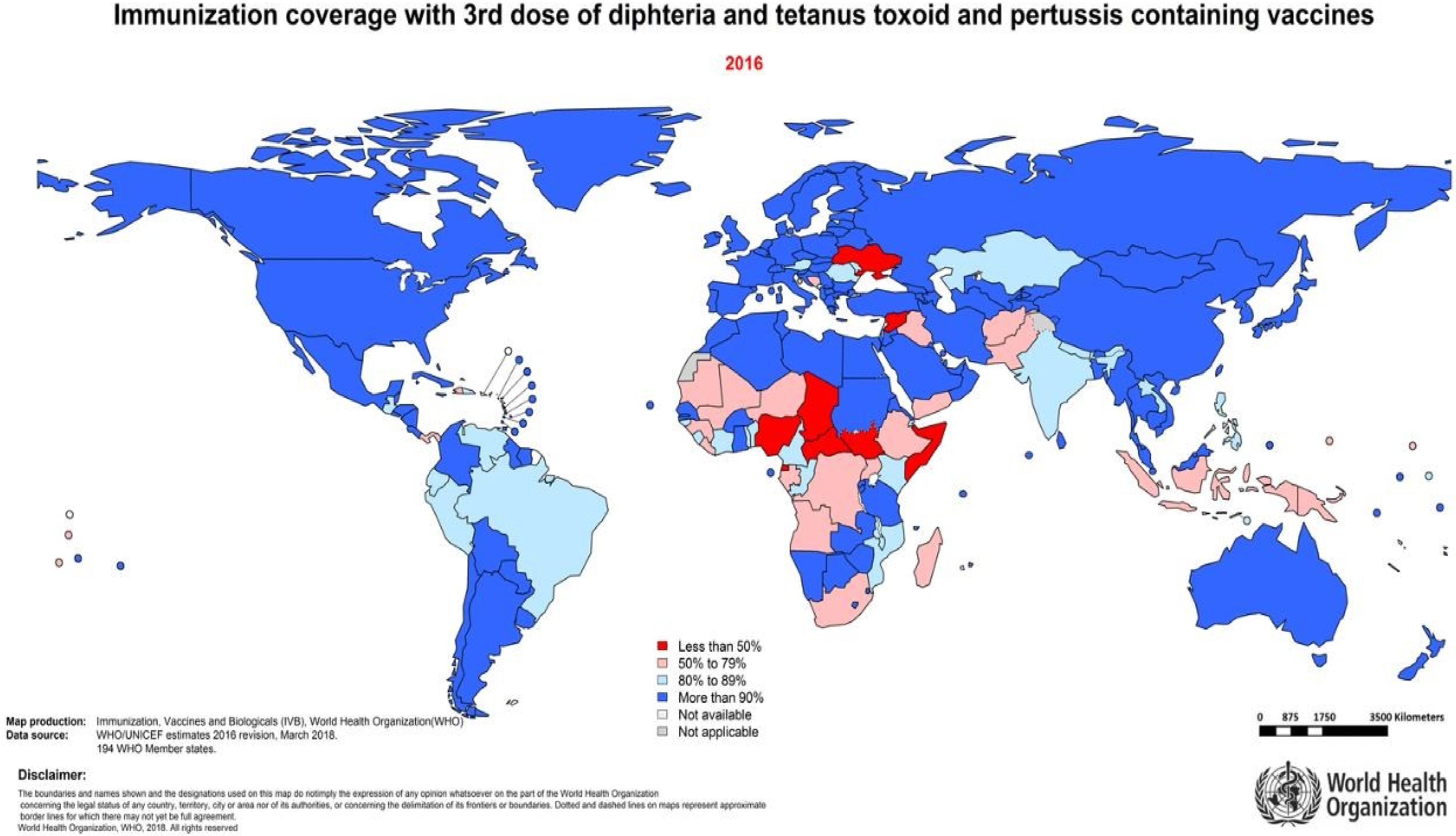
Antigen	Short name	Type of virulence factor	Protein size / kDa	Important genes	Gene IDs	Frequency of inclusion in ACV
Pertactin	PRN	Adhesin	69	*prn*	BP1054	Most
Pertussis toxin	PT	Toxin	105	*ptxA-E*	BP3783-3787	All
Filamentous haemagglutinin	FHA	Adhesin	220	*fhaB, fhaC, ctpA, sphB1*	BP1879, BP1884, BP0609, BP0216	Most
Fimbrial 2	Fim2	Adhesin	22 (monomer)	*fim2, fimB-D*	BP1119, BP1881-3	Some
Fimbrial 3	Fim3	Adhesin	22 (monomer)	*fim3, fimB-D*	BP1568, BP1881-3	Some

Table 1.3 Current pertussis vaccination strategies in all countries considered in this thesis

Country	Year WCV introduced	Year ACV introduced	PT	PRN	FHA	Fim2 and Fim3	Vaccine coverage (three dose)* /%	Source
Belgium	1950s	1999-2002	✓	✓	?	X	98	Barkoff et al. (2019)
Denmark	1961	1997	✓	X	X	X	97	Barkoff et al. (2019)
Finland	1952	2005	✓	✓	✓	X	91	Barkoff et al. (2019)
France	1959	2004	✓	?	✓	X	96	Barkoff et al. (2019)
Italy	1961	1995	✓	✓	✓	X	95	Barkoff et al. (2019)
Netherlands	1953	2005	✓	✓	✓	?	93	Barkoff et al. (2019)
Norway	1952	1998	✓	✓	✓	X	94	Barkoff et al. (2019)
Sweden†	1953	1996	✓	?	✓	X	98	Barkoff et al. (2019)
UK	1957	2004	✓	✓	✓	?	94	Barkoff et al. (2019)
USA	1948	1996	✓	✓	✓	?	94	Klein (2014)
Canada	1943	1997	✓	✓	✓	?	91	Canadian Public Health Association (2019)
Australia	1953	1999	✓	✓	✓	?	95	Pillsbury, Quinn and McIntyre (2014)
Poland	1960	NA	-	-	-	-	95	Gzyl et al. (2004)
Japan	1947	1981	✓	✓	✓	✓	99	Noble et al. (1987)
China	1973	2009	✓	X	✓	X	99	Zheng, Y. et al. (2018)
New Zealand	1945	2000	✓	✓	✓	X	93	Reid (2006)
Israel	1957	2002	✓	✓	✓	X	98	Stein-Zamir et al. (2010)
Brazil	1950	NA	-	-	-	-	83	Domingues, Teixeira and Carvalho (2012); Guimarães, Carneiro and Carvalho-Costa (2015)
Mexico	1960	2007	✓	X	✓	X	88	Aquino-Andrade et al. (2017); Carrillo (2017)

* Coverage statistics from WHO (WHO, 2019b)
✓ Antigen included in ACV
? Antigen included in some available ACV formulations but not others

† Sweden ceased Pertussis vaccination between 1979 and 1996
X Antigen not included in ACV
- ACV not used



Immunization coverage with 3rd dose of diphtheria and tetanus toxoid and pertussis containing vaccines

2016

Map production: Immunization, Vaccines and Biologicals (IVB), World Health Organization(WHO)
Data source: WHO/UNICEF estimates 2016 revision, March 2018.
194 WHO Member states.

Disclaimer:

The boundaries and names shown and the designations used on this map do notimply the expression of any opinion whatsoever on the part of the World Health Organization concerning the legal status of any country, territory, city or area nor of its authorities, or concerning the delimitation of its frontiers or boundaries. Dotted and dashed lines on maps represent approximate border lines for which there may not yet be full agreement.

World Health Organization, WHO, 2018. All rights reserved

Figure 1.3 Estimated global coverage of three DTP (DTwP/DTaP) doses as of 2016. Many countries now have greater than 90% coverage (WHO, 2019a).

Every formulation of the ACV contains pertussis toxin (PT), which is a 105 kDa ADP-ribosylating toxin. The mature protein is comprised of five proteins subunits, S1 to S5, which are coded for by the *ptxA* to *ptxE* genes. PT is an A-B toxin, in which S1 (the *ptxA* protein) forms the A subunit and the remaining proteins form a pentameric ring-shaped B subunit including one copy of S2, S3 and S5, and two copies of S4 (Tamura et al., 1983; Locht and Keith, 1986; Nicosia et al., 1986). The B subunit binds to, and delivers the A subunit into, eukaryotic host cells (Kaslow and Burns, 1992; Xu, Y. and Barbieri, 1995, 1996). Once delivered, the A subunit transfers an ADP-ribose to GTP-binding proteins, thereby disrupting the normal function of G proteins in G protein-coupled cell signalling and transport (Katada, Tamura and Ui, 1983; Hoshino et al., 1990; Shinoda, Katada and Ui, 1990). This disruption of G protein activity results in a number of deleterious effects within the host cell, including upset electrolyte and fluid balance, and a reduced ability to recruit neutrophils and macrophages (Meade et al., 1984; Carbonetti, 2007; Andreasen and Carbonetti, 2008). These effects, particularly immune supression, are thought to play a vital role during infection establishment, and also appear to be the cause of many of the symptoms of whooping cough, such as leucocytosis (Carbonetti, 2010). Some questions over the functions of PT during whooping cough infection remain. For example, PT is only expressed in *B. pertussis*; although the gene is present in the genomes of both *B. bronchiseptica* and *B. parapertussis*, mutations in the promoter region (*ptxP*) are thought to prevent expression (Arico and Rappuoli, 1987). Yet, even without PT, *B. parapertussis* can also cause whooping cough, albeit less severely than *B. pertussis* (Heininger et al., 1994). Nevertheless, PT can be considered the major virulence factor of *B. pertussis* and, indeed, some formulations of the ACV contain only PT and are still effective at preventing whooping cough.

Like pertactin, filamentous haemagglutinin (FHA) is included in most formulations of the ACV around the world. Most of what we know about FHA has been learned by studying the FHA protein of *B. bronchiseptica* in mammals other than humans; however, the protein and its mechanisms are thought to behave the same way in *B. pertussis* as in *B. bronchiseptica* (Nash and Cotter, 2019). FHA is a secreted protein which acts as a membrane-bound adhesin via the same Arg-Gly-Asp motif as PRN, as well as being released. As an adhesin it has been shown that FHA alone is sufficient for the adherence of bacterial cells to eukaryotic host epithelial cells in culture, whilst as a released protein it contributes to the evasion of the host's immune response (Urisu, Cowell and Manclark, 1986; Relman et al., 1989; Cotter et al., 1998; Melvin et al., 2015). Knockout studies in a variety of mammals have shown that FHA is required, but not sufficient, to establish *Bordetella* infection in mammalian respiratory tracts (Inatsuka, Julio and Cotter, 2005; Julio et al., 2009; Henderson, M.W. et al., 2012; Melvin et al., 2015). FHA is known as a prototypical member of the two-partner-secretion (TPS) pathway family of proteins, the two partners being FHA and FhaC (Nash and Cotter, 2019). FHA, coded for by the 10 kbp *fhaB* gene, is initially synthesised as a 370 kDa precursor protein, FhaB. The transmembrane, β-barrel, FhaC protein facilitates the translocation of FhaB across the bacterial membrane, and along the way the C-terminal prodomain of FhaB is proteolytically cleaved to produce the final 220 kDa FHA protein. Nash and Cotter (2019) proposed the most comprehensive mechanism yet for the processing of FhaB into FHA. The proposed mechanism involves sequential, multi-step cleavage and degradation of the prodomain by an unknown periplasmic protease, periplasmic CtpA, and extracellular SphB1. The process culminates in the release of the mature, highly immunogenic, 220 kDa FHA protein from the membrane, triggered by a currently unknown signal. As with pertussis toxin, questions about the biology of FHA remain. For example, Nash and Cotter (2019) also speculate that another, as yet largely undefined, role for the full length FhaB protein exists, potentially related to immune suppression and the binding of adenylate cyclase toxin (ACT), another *Bordetella* virulence factor.

The final two components of the ACV, Fim2 and Fim3, are included less frequently than the others. The fimbriae are members of the type 1 pili family, which are long, hair-like structures on the cell surface, composed of thousands of pilin subunits (Scheller and Cotter, 2015). *B. pertussis* strains can express Fim2 fimbriae alone (serotype 1-2), Fim3 fimbriae alone (serotype 1-3), or a combination of both fimbriae (serotype 1-2-3) (Ashworth, Irons and Dowsett, 1982). The most heavily studied type 1 pili mechanism is that of uropathogenic *E. coli*, and *Bordetella* pili biogenesis is thought to proceed by the same mechanism (Busch, Phan and Waksman, 2015). A variety of proteins are involved in the biogenesis of these structures, including FimB, FimC and FimD, which are believed to act as a periplasmic chaperone, outer membrane usher and tip adhesin, respectively (Willems, van der Heide and Mooi, 1992). The pilin subunits are delivered sequentially to the transmembrane β-barrel usher protein (FimC) by the chaperone (FimB) and, as the chain of pilin subunits grows, it is translocated across the outer membrane into the extracellular space. Each chain has a tip adhesin protein, FimD, at the extracellular end, which is thought to be involved in the adherence of *Bordetella* cells to eukaryotic epithelial cells and monocytes. Genes for other pilin subunits, *FimA, FimN* and *FimX*, are also found in the *B. bronchiseptica* and *B. pertussis* genomes, but their proteins are not believed to be produced in either species (Scheller and Cotter, 2015). Interestingly, in *B. pertussis* and *B. bronchiseptica*, many of the *Fim* genes are found on an operon between the FHA genes *FhaB* and *FhaC*, but *fim2* and *fim3* are both found elsewhere, separately, in the genome. As a result of their co-translation with the *Fha* genes, experimental investigation of the fimbriae in *Bordetella* has not been an easy task. However, although experimental evidence for the role of the fimbriae in *Bordetella* adherence to eukaryotic cells is limited, and despite their relatively rare inclusion in the ACV, the presence of Fim2 and Fim3 has been shown to significantly enhance vaccine-mediated protection (Olin et al., 1997; Geuijen et al., 1998; Scheller and Cotter, 2015).

B. pertussis virulence depends on the BvgAS two-component system

Lacey (1960) described three distinct phenotypic phases in which the classical *Bordetella* species could exist. These distinct phases were designated the X, I and C modes, and the bacteria could seemingly modulate between different phases according to their environmental signals (Mattoo and Cherry, 2005). It is now known that, like many other bacteria, *Bordetella* species sense and respond to their surrounding environment using a two-component system (Gross, Arico and Rappuoli, 1989). In *Bordetella*, this system is the Bvg (*Bordetella* virulence gene) system, and it is almost identical in *B. bronchiseptica*, *B. parapertussis*, and *B. pertussis* (Mattoo and Cherry, 2005). As shown in **Figure 1.4**, the BvgAS system comprises two proteins: a histidine sensor kinase, BvgS, which spans the inner membrane to sense environmental stimuli and activate a response regulator, BvgA, via a phosphorylation cascade (Cotter and Jones, 2003; Melvin et al., 2014).

BvgAS is active at 37^{o}C in the absence of chemical modulators; these conditions correspond to Lacey's X mode, which is now known as the Bvg(+) phase. Bvg(+) conditions initiate the autophosphorylation of the BvgS histidine kinase (HK) domain, triggering a phosphorylation cascade which culminates in the phosphorylation and activation of BvgA. Activated BvgA promotes the expression of many genes in *B. pertussis*, including key virulence factors such as adhesins and toxins like filamentous hemagglutinin (FHA) and pertussis toxin (PT) (Moon et al., 2017). Together, the BvgA-activated genes are known as the "virulence-activated genes" (*vag*s). One of the genes activated by BvgA is that of the

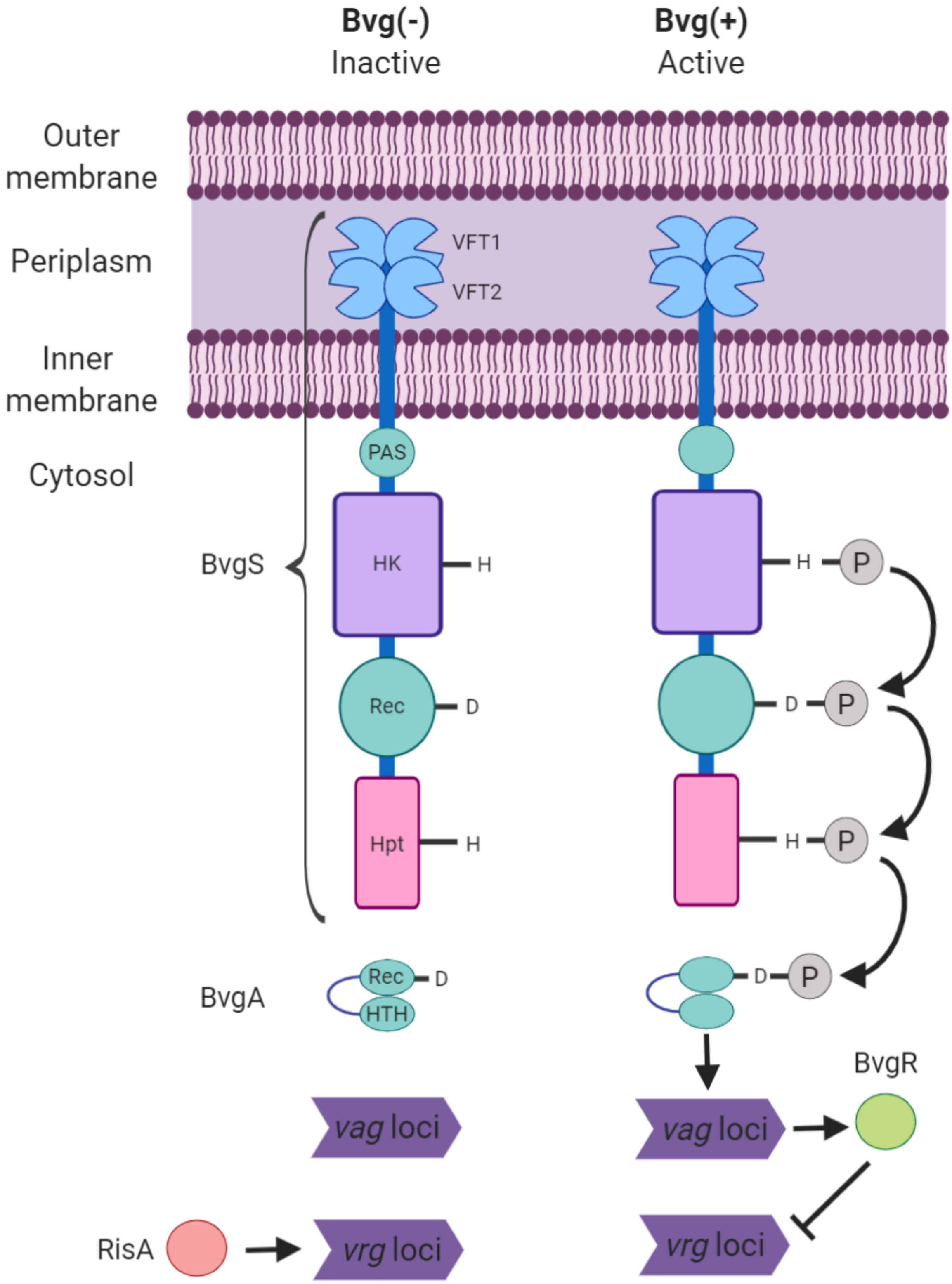


Figure 1.4 The BvgAS two-component system of *B. pertussis*. Under Bvg(+) conditions, the venus flytrap (VFT) domains of BvgS undergo a conformational change, leading to autophosphorylation of the histidine kinase (HK) domain at a conserved histidine (H) residue. The phosphoryl group is then transferred in turn to the receiver (Rec) domain and histidine phosphotransferase (Hpt) domains of BvgS, and finally to conserved aspartate (D) residue in the BvgA Rec domain. The phosphorylated BvgA dimerises and acts as a transcription factor for the virulence-activated genes (*vags*), including the Bvg repressor protein, BvgR, which represses the virulence-repressed genes (*vrgs*). Under Bvg(-) conditions, no phosphorylation of BvgS or BvgA occurs, meaning expression of the *vags* is not promoted, and expression of the *vrgs* is not repressed. The expression of the *vrgs* can therefore be promoted by RisA. Bvg(i) exists as an intermediate between these two phases.

Figure inspired by Cotter and Jones (2003); Dupre et al. (2015); Hoffman (2017).

Bvg repressor protein, BvgR. Whilst the *vag*s are activated by BvgA, the transcription of another set of genes is repressed by BvgR (Merkel and Stibitz, 1995; Merkel, Barros and Stibitz, 1998; Merkel et al., 2003). These "virulence-repressed genes" (*vrg*s) include those involved in the outer membrane, those required for biofilm formation and motility, and those involved in a large number of metabolic pathways (Cotter and Jones, 2003; Hoffman, 2017; Chen, Q. and Stibitz, 2019; Hoffman et al., 2019). Bvg(+) is considered to be the default "on-state" for *B. pertussis*, and mutant bacteria which are unable to phosphorylate BvgA and therefore enter the Bvg(+) phase cannot colonise respiratory tracts in mice (Martinez de Tejada et al., 1998; Moon et al., 2017).

Lacey's C mode, now known as the Bvg(-) phase, is less well-characterised than Bvg(+). Bvg(-) can be induced *in vitro* by reduced temperature (lower than around 27°C) or the presence of chemical modulators such as ≥ 40 mM $MgSO_4$ or ≥ 10 mM nicotinic acid, but *in vivo* triggers for the Bvg(-) in nature are as yet undetermined (Melton and Weiss, 1993; Chen, Q. and Stibitz, 2019). In the Bvg(-) phase, the expression of the *vag*s, including the major *B. pertussis* virulence factors and BvgR, is not promoted, as BvgA does not become phosphorylated. Consequently, the expression of the *vrg*s is not repressed; in the Bvg(-) phase expression of the *vrg*s is promoted by RisA (Croinin, Grippe and Merkel, 2005; Stenson et al., 2005). Expression of the *vrg*s has been shown to be detrimental to virulence in animal models; Δ*bvgR* knockout bacteria are unable to cause disease in mice. Instead, this class of genes is currently believed to play a role during transmission and survival outside of the optimal environment. Bvg(-) bacteria can survive in non-optimal conditions and locations for longer than Bvg(+) bacteria, as well as growing more quickly and not expressing the antigens commonly observed during infection (Akerley and Miller, 1993; Akerley, Cotter and Miller, 1995; Martinez de Tejada et al., 1998; Merkel et al., 1998).

Very little is known about the third *Bordetella* phase, Lacey's I mode, which is now known as the Bvg-intermediate, or Bvg(i) phase. Bvg(i) occurs during the transition between Bvg(+) and Bvg(-), or vice-versa; the BvgAS system functions not simply as a biphasic switch, but as a gradient of expression and repression of *vag*s and *vrg*s (Cotter and Miller, 1997; Deora et al., 2001; Stockbauer et al., 2001). Like Bvg(+) and Bvg(-), there is a set of genes which are characteristically expressed under Bvg(i) conditions, typically temperatures lower than 37°C but higher than 27°C, concentrations of $MgSO_4$ lower than 40 mM but higher than 15 mM, or 0.4 to 2 mM nicotinic acid *in vitro* (Mattoo and Cherry, 2005). Under these conditions, BvgS is less active, hence levels of phosphorylated BvgA are lower. Consequently, expression of the *vag*s and repression of the *vrg*s is limited (Williams, C.L. et al., 2005). In addition, a set of Bvg(i)-specific genes is thought to exist. The first of these to be described codes for an adhesin, Bvg-intermediate phase protein A, or BipA (Stockbauer et al., 2001). The *bipA* promoter is activated by low levels of phosphorylated BvgA, but repressed by higher levels of BvgA. Thus, BipA is expressed maximally as bacteria transition between fully active and fully inactive BvgS, when it is involved in biofilm formation (Nishikawa et al., 2016). Bvg(i) is thought to play a role very early in infection, potentially during colonisation, but the majority of the genes involved, and their specific functions, remain largely undiscovered.

Whooping cough has been resurgent since the late 1980s and early 1990s

The World Health Organisation estimates that, globally, 90% of the target population receive at least one dose of pertussis-containing vaccine, and 85% receive the full three recommended doses; these levels of coverage have been stable since 2015 (WHO, 2018). Despite these relatively high levels of vaccination, the incidence of whooping cough has been increasing in many countries since the 1980s

(De Serres et al., 1995; Cherry, 1996; de Melker et al., 1997). Throughout this period, notable epidemics have occurred in the USA in 2005, 2010 and 2012, the UK in 2012, Australia between 2008 and 2012, Japan between 2008 and 2012, and, most recently, New Zealand between 2017 and 2019 (Octavia et al., 2012; Winter et al., 2012; Miyaji et al., 2013; Bowden et al., 2014; Clark, 2014; Lam et al., 2014; Sealey et al., 2015; Stuff NZ, 2019).

A key feature of the whooping cough resurgence is a slight shift in the epidemiological profile of the disease, as seen in **Figure 1.5**. Throughout the pre-vaccine era, whooping cough was a disease of childhood: around 95% of cases were seen in children younger than 10 years old. During the whole-cell vaccine era, there was a slight shift towards the younger end of this range, with infants under one year old representing 53.5% of all cases diagnosed (Mattoo and Cherry, 2005). Since the late 1990s, however, a growing proportion of cases have been seen in older children and adolescents. In a 2004 US epidemic, adolescents comprised 36% of cases, ultimately leading to the introduction of a Tdap booster for 11- or 12-year olds. Despite wide uptake of this booster, epidemics in 2010 and 2012 saw elevated infection rates in fully vaccinated children and adolescents aged 7-10 and 13-14 years old (Clark, 2014).

To a certain extent, the increase in overall recorded incidence can be explained by a heightened awareness of the disease and the introduction of new, more sensitive, diagnosis methods such as PCR and serology. It is believed that when diagnoses were primarily made by culture, fewer than 20% of whooping cough infections were diagnosed, mainly because by the time diagnosis was attempted, the patient's bacterial load was usually too small to detect via culture (Mattoo and Cherry, 2005). The ability of PCR and serology to detect smaller bacterial loads will therefore have resulted in an increase in the number of cases diagnosed, with notifications approaching the true number of infections for the first time. James D. Cherry, a paediatrician and researcher who has been a global expert in whooping cough for many decades, believes that the perceived "resurgence" of the disease is a myth: that, although more people are being diagnosed with the disease nowadays, the number of people

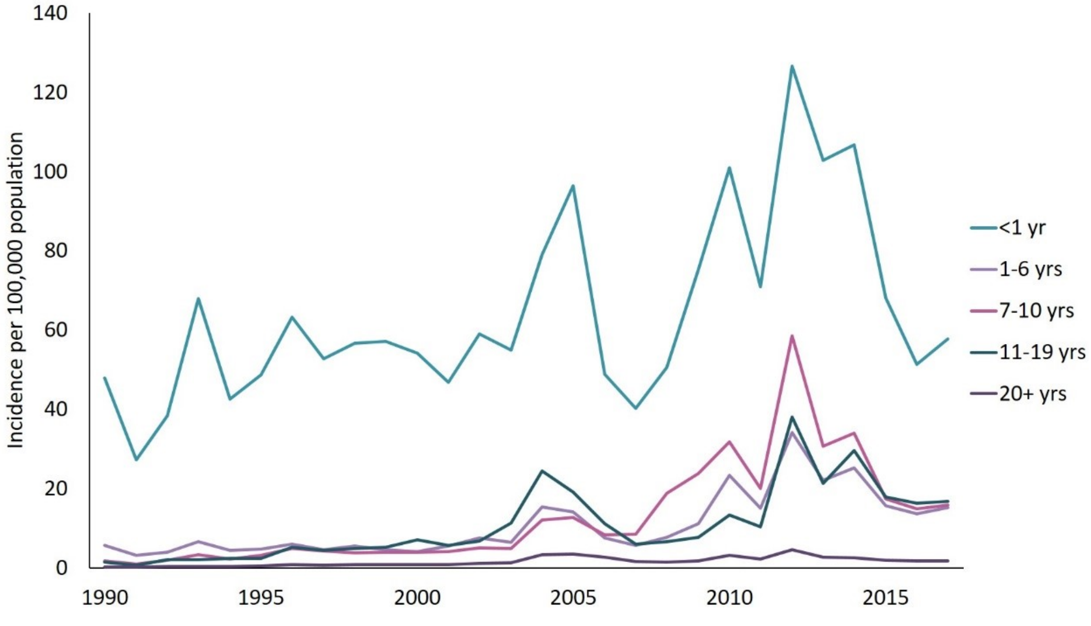


Figure 1.5 Whooping cough incidence across age groups in the USA, 1990-2017. Incidence has increased in every age group and, although <1 year remains the group mostly highly affected, there has been a notable increase in the relative proportion of 1-19 year olds.

Data from Centers for Disease Control (2019b).

infected has not truly increased (Cherry, 2015). However, if this was correct, we might expect to have seen an initial increase in diagnoses when the new diagnosis techniques were introduced, followed by a plateau at whatever the true rate of incidence was. Instead, the measured incidence of the disease has continued to increase on an almost yearly basis, as seen in **Figure 1.5**. In addition, a 2014 survey of 19 countries by the World Health Organisation indicated that at least five countries (Australia, Chile, Portugal, USA and the UK) demonstrated a clear resurgence of whooping cough, separate from any increase in notifications resulting from detection bias (World Health Organisation, 2014).Therefore, although some of the increased incidence is due to people now being diagnosed who previously would not have been, other factors must also be at play.

One additional factor contributing to the recent resurgence of whooping cough is waning immunity. Even natural immunity to pertussis, gained through prior infection, is known to wane substantially in 7-20 years. The immunity conveyed by the whole-cell vaccine was also thought to wane within a few years of vaccination; children vaccinated several years previously were commonly found to be infected with whooping cough, although their disease usually seemed to be less severe (Wendelboe et al., 2005). The persistence of the three-to-five yearly epidemic cycle also suggests that the WCV was successful in preventing severe manifestations of whooping cough, but not in preventing its transmission (Clark, 2014). As indicated by the epidemiological shift in age groups infected, however, immunity conveyed by the acellular vaccine appears to wane faster than either naturally- or WCV-conveyed immunity. Case-controlled studies from a 2010 California epidemic suggest that the protection conveyed by the ACV falls to 71% within five years of vaccination, and that the odds of pertussis infection increase 42% with every passing year (Klein et al., 2012; Misegades et al., 2012). It is nonetheless important to note that the whooping cough incidence has increased across all age groups even in countries which still use the WCV, like Brazil and China (World Health Organisation, 2014; Guimarães, Carneiro and Carvalho-Costa, 2015). In addition, particularly severe whooping cough epidemics were observed over the decade prior to the switch from WCV to ACV, including the Netherlands, Australia, Canada and the US (Bass and Wittler, 1994; De Serres et al., 1995; Andrews, Herceg and Roberts, 1997; de Melker et al., 1997). Thus, like increased awareness and improved diagnosis techniques, the more rapidly waning immunity conveyed by the ACV compared to the WCV does not fully explain the resurgence of whooping cough.

A final factor which may be contributing to the resurgence of whooping cough, and the focus of this thesis, is changes to *B. pertussis* at the genetic level, such that the strains used to prepare vaccines no longer fully represent the strains which are circulating. Such changes may have occurred by chance, as a result of natural genetic drift, or, more likely, they may have happened as a result of vaccination applying a selection pressure against the vaccine strains, thus selecting for strains which have diverged. Shifts in the circulating *B. pertussis* population were observed soon after the first whole-cell vaccines were introduced. The original WCVs contained cells with the fimbrial serotype 1,2 and, as early as the 1950s and 1960s, a rapid shift towards serotype 1,3 was noticed in vaccinated populations but not unvaccinated populations. This shift was thought to have contributed to a reduction in vaccine efficacy; the vaccine was subsequently changed to include both serotypes 1,2 and 1,3, which seemingly resulted in efficacy improvements (Preston, N.W., 1976). Thus, the concept of whooping cough vaccination causing circulating *B. pertussis* populations to shift, resulting in lowered efficacy, has been previously established at the protein level. Since the advent of whole genome sequencing (WGS) in the 1990s and 2000s, attention has turned to also observing such shifts at the genetic level, investigating their influence on *B. pertussis* evolution, and understanding their potential contribution to reduced vaccine efficacy.

B. pertussis evolved from *Bordetella bronchiseptica* by insertion sequence-mediated genomic streamlining

Multilocus sequence typing (MLST) studies showed that the majority of *B. bronchiseptica* isolates can be assigned into one of two distinct complexes, Complex I and Complex IV (Diavatopoulos et al., 2005; Park et al., 2012). The higher resolution of WGS has allowed a deeper investigation of the relationships between *B. bronchiseptica*, *B. parapertussis* and *B. pertussis*. It appears that *B. pertussis* and *B. parapertussis* are most closely related to different *B. bronchiseptica* complexes, *B. pertussis* to Complex IV and *B. parapertussis* to Complex I (Linz et al., 2016). Nonetheless, the two species cause very similar pathologies in humans and they both appear to have evolved from their *B. bronchiseptica*-like ancestor primarily through insertion sequence (IS)-mediated genome reduction.

The genome of *B. bronchiseptica* contains very few, if any, IS elements. The original reference strain to be whole genome sequenced by Parkhill et al. (2003) did not contain any IS elements, although some strains sequenced subsequently contain a handful of IS *481*, IS *1001*, IS *1002* or IS *1663*, as seen in **Table 1.4**. The average *B. parapertussis* genome contains around 30 IS elements, mainly IS *1001* with a few copies of IS *1002*, and no copies of IS *481* or IS *1663*. The *B. pertussis* genome contains the most IS elements by far, with over 250 copies of IS *481* on average, fewer than 20 copies of both IS *1002* and IS *1663*, and no copies of IS *1001*. The appearance and subsequent expansion of these IS elements via homologous recombination throughout the genomes of their *B. bronchiseptica*-like ancestors is thought to have led to the (independent) speciation of *B. pertussis* and *B. parapertussis* (Parkhill et al., 2003; Preston, A., Parkhill and Maskell, 2004; Diavatopoulos et al., 2005).

IS-mediated reductions in genome size and the number of functional genes was key in the speciation process of *B. pertussis*. As shown in **Table 1.4**, the genome size and number of genes in an average *B. bronchiseptica* are 5.21 Mbp and around 5,000, respectively. By contrast, the average genome size and number of genes in a *B. pertussis* isolate are only 4.11 Mbp and around 4,000, respectively. The *B. pertussis* genome has lost many genes present in *B. bronchiseptica*, including those involved in metabolism and membrane transport (Parkhill et al., 2003). The insertion of IS elements into genes has also produced over 350 pseudogenes in *B. pertussis*, compared to only around 20 in *B. bronchiseptica*. The extensive expansion and recombination between IS elements in *B. pertussis* has

Table 1.4 Characteristics of all classical *Bordetella* closed genomes available in NCBI's RefSeq database, March 2019 [2]

Characteristic	***B. pertussis***	***B. parapertussis***	***B. bronchiseptica***
Number of closed genomes	421	18	19
Genome size / Mb [1]	4.11 (4.04 – 4.39)	4.78 (4.77 – 4.90)	5.21 (5.08 – 5.34)
Number of predicted genes[1]	3,979 (3,856 – 4,239)	4,501 (4,490 – 4,574)	4,911 (4,774 – 5,090)
Number of proteins [1]	3,615 (3,425 – 3,866)	4,166 (4,157 – 4,184)	4,804 (4,663 – 4,993)
G+C% [1]	67.7 (67.7 – 67.8)	68.1 (67.8 – 68.1)	68.2 (68.1-68.4)
IS *481* [1]	256 (234 - 273)	-	1 (0-3)
IS *1001* [1]	-	22 (22-28)	1 (0-1)
IS *1002* [1]	8 (5 – 10)	9 (0-9)	0 (0-5)
IS *1663* [1]	17 (16 – 24)	-	2 (0-14)

[1] Figures shown are mean and (range)

[2] Table adapted from Ring et al. (2019)

sculpted a streamlined and specialised bacterium, and likely explains why the pathogen is entirely niche-restricted to the human nasopharynx.

At the base level, *B. pertussis* strains exhibit limited inter-strain variation, although certain genes have been evolving more rapidly since the introduction of vaccination

Prior to the wide availability of WGS (the "pre-genomics" era), a variety of alternative molecular techniques were used to investigate diversity in *B. pertussis* genomes. These techniques included multilocus enzyme electrophoresis (MLEE), multiple-locus variable number tandem repeat analysis (MVLA), restriction fragment length polymorphism (RFLP) analysis, ribotyping, randomly amplified polymorphic DNA (RAPD) typing and pulsed field gel electrophoresis (PFGE) (Musser et al., 1986; Moissenet et al., 1996; van der Zee, Vernooij, et al., 1996; van der Zee et al., 1997; Vandamme et al., 1997; Mooi et al., 2000; Kurniawan et al., 2010). Using these pre-genomics techniques to investigate strains isolated in the pre-vaccine and WCV eras, a picture began to emerge of a *B. pertussis* population which had continued to shift since the introduction of vaccination but which, on the whole, exhibited very little inter-strain variation compared to other bacteria (van der Zee, Vernooij, et al., 1996; Mooi et al., 1998; van Loo et al., 1999; Kurniawan et al., 2010)). However, pre-genomics molecular techniques provided limited resolution, relying on small numbers of already well characterised genomic features (e.g. MLVA, ribotyping, serotyping) or on visualisation of restriction enzyme reactions which could have been influenced by mutations within enzyme recognition sites (e.g. MLEE, RFLP, PFGE).

The higher resolution of WGS means that many investigations into diversity within bacterial populations are now centred around the identification of single nucleotide polymorphisms (SNPs) between strains. SNPs can be used to estimate relatively accurately the relationships between strains, the time since certain lineages diverged from others, and how quickly genomes are mutating and evolving. Investigations of SNPs in *B. pertussis*, however, are limited by the same limited variation which was observed using pre-genomics techniques. For example, by comparing six Dutch strains, Bart et al. (2010) estimated *B. pertussis* to have a SNP rate between strains of 1 per 8,675 bases, in contrast to 1 per 3,000 bases in *Mycobacterium tuberculosis* or 1 per 6,700 bases in *E. coli* (Fleischmann et al., 2002; Gutacker et al., 2002; Zhang, W. et al., 2006).

Another benefit of WGS has been the ability to conduct higher throughput strain screens. Bart, Harris, et al. (2014) used WGS to screen 343 strains which had been isolated from 19 different countries across a 90-year period (1920-2010). This landmark longitudinal study, the first to consider genetic information from across the entire genome of so many strains, had two key findings. Firstly, phylogenetic analysis of the 343 strains revealed two deep lineages, one of which (branch I) contained only 1.7% of the strains screened, and which has ostensibly disappeared since the year 2000. Branch I and branch II seemingly diverged around 2,000 years ago, which the authors speculate may represent two independent introductions of *B. pertussis* into humans. Most recent *B. pertussis* isolates clustered to a sub-branch of branch II (IIb), and there was seemingly no geographical structure in the phylogenetic tree, reflecting the limited inter-strain diversity previously observed. Secondly, allele typing showed that the genes coding for the antigens included in the ACV have undergone a number of allelic profile shifts, away from the alleles represented by the vaccine strains (**Figure 1.6**). This finding was supported by Sealey et al. (2015) and, more recently, Etskovitz et al. (2019), who showed that the antigens included in the ACV have begun to evolve more rapidly since the switch from WCV to ACV.

One of the most striking allelic shifts observed in longitudinal studies like Bart, Harris, et al. (2014) is the appearance and rapid spread of a new pertussis toxin promoter allele, *ptxP3,* since the late 1980s. *ptxP3* is characterised by a single base pair (bp) G-to-A mutation in the recognition sequence to which BvgA, *ptx*'s transcriptional regulator, binds. This SNP appears to increase binding affinity in BvgA, increasing the transcription and, thus, production of pertussis toxin in *ptxP3*-carrying strains. The increased expression of this important virulence factor is accompanied by altered expression of proteins involved in complement resistance and, together, these changes result in the transmissibility and severity of whooping cough infections caused by *ptxP3* strains (Mooi et al., 2009; Bart et al., 2010; King et al., 2013; de Gouw et al., 2014).

The rapid selective sweep of *ptxP3* may have been related to the use of the ACV. Advani et al. (2011) found a higher and earlier prevalence of *ptxP3*-carrying strains in Gothenburg compared to the rest of

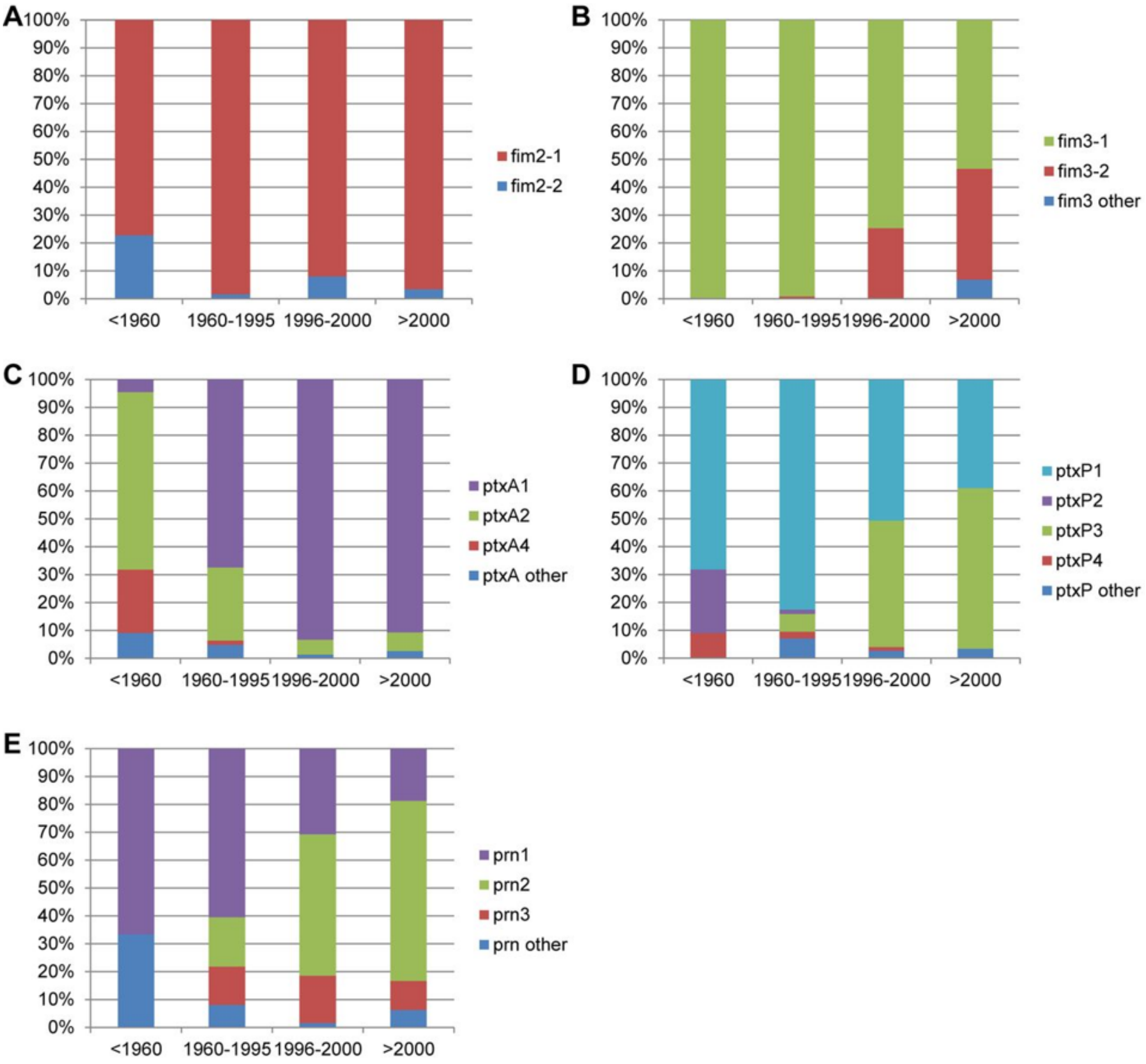


Figure 1.6 Allelic changes in the genes which code for the ACV antigens since the introduction of whooping cough vaccination, away from the alleles represented in the vaccine strains. Notable selective sweeps have been seen in *ptxP* (from *ptxP1* towards *ptxP3*), *prn* (from *prn1* to *prn2*), *ptxA* (from *ptxA2* to *ptxA1*) and *fim3* (from *fim3-1* towards *fim3-2*). Most recent strains now have the allelic profile *ptxA1-ptxP3, prnA2, fim3-2, fim2-1*.

Figure from Bart, Harris, et al. (2014).

Sweden. Gothenburg used first a single-component PT-only ACV, followed by a two-component PT+FHA ACV, whilst the rest of Sweden was using a three-component PT+FHA+PRN ACV. Advani et al. suggest that the accelerated appearance of *ptxP3* could have been because of the greater pressure against PT conveyed by PT-only or PT+FHA vaccines; although the one- and two- component vaccines did not cause the SNP which produces *ptxP3*, they caused the allele to spread more rapidly because of the pressure they apply against the vaccine allele, *ptxP1*. However, the appearance and rapid spread of strains carrying *ptxP3* may have occurred even if no whooping cough vaccines were in use; a lone SNP can arise by chance, and the fitness benefit conveyed by the enhanced virulence of strains carrying this mutation means that the allele would have spread rapidly with or without vaccine pressure.

A similar selective sweep, which was almost definitely due to pressure applied by the ACV, is the appearance and proliferation of strains which are deficient in one or more of the ACV antigens. In the 1990s, prior to the introduction of the ACV, only a handful of strains were isolated which had mutations resulting in their non-ability to express PRN, in a variety of countries around the world (Mastrantonio et al., 1999; Miyaji et al., 2013; Weigand et al., 2018). Bart et al.'s longitudinal study, which included 323 strains isolated prior to 2007, did not find any PRN-deficient strains, although one has since been resequenced and found to be deficient after all (Bart, Harris, et al., 2014; Zomer et al., 2018). From the mid-2000s onwards, however, an increasing number of PRN-deficient strains have been isolated. For example, in Australia the percentage of strains found to be PRN-deficient increased from 5% to 78% in the four year period between 2008 and 2012, whilst 640 of 753 strains isolated in the United States between 2011 and 2013 were PRN-deficient (Lam et al., 2014; Martin et al., 2015). Strong evidence suggests that PRN-deficiency has increased so rapidly in response to the introduction of the ACV. Martin et al.'s study of 753 US strains found a significant association between ACV status and likelihood of being infected by PRN-deficient *B. pertussis* instead of PRN-positive *B. pertussis* (Martin et al., 2015). In addition, Barkoff et al. studied PRN-deficiency in European *B. pertussis* strains isolated between 1998 and 2015, and found that the longer a country had been using the ACV, the higher the chance that strains from the country would be PRN-deficient, as shown in **Figure 1.7** (Barkoff et al., 2019). A variety of different PRN deficiency-causing mechanisms have been observed, including insertion of IS *481* into the gene, large deletions and deleterious SNPs. No single mechanism appears to be predominant (Hegerle et al., 2012; Queenan, Cassiday and Evangelista, 2013; Barkoff et al., 2019).

A slightly different PRN deficiency trend has been seen in Japan, however. Japan introduced the ACV in 1981, many years before most countries. The number of PRN-deficient strains appears to have increased initially, reaching 41% of all strains isolated between 2005 and 2007. At every time point since then, the proportion of PRN-deficient strains appears to have decreased: to 35% between 2008 and 2010, and to 25% between 2011 and 2013. Furthermore, PRN was removed from Japan's ACV in 2012; a significant further decrease was observed between 2014 and 2016, with only 8% of strains now found to be PRN-deficient (Hiramatsu et al., 2017). Limited recent data from Ontario suggests the same trend may also now be occurring in Canada. The first PRN-deficient strains were isolated in 2011, when 16.7% were deficient. The proportion of deficient strains increased rapidly, reaching 70.8% in 2016. As seen in Japan, after reaching a peak, the number of PRN-deficient strains now appears to be decreasing: a sharp fall was seen from 2016 to 2017, when only 46.2% of strains were deficient (Tsang et al., 2019). No more recent data is available as yet to confirm whether this trend of decreasing

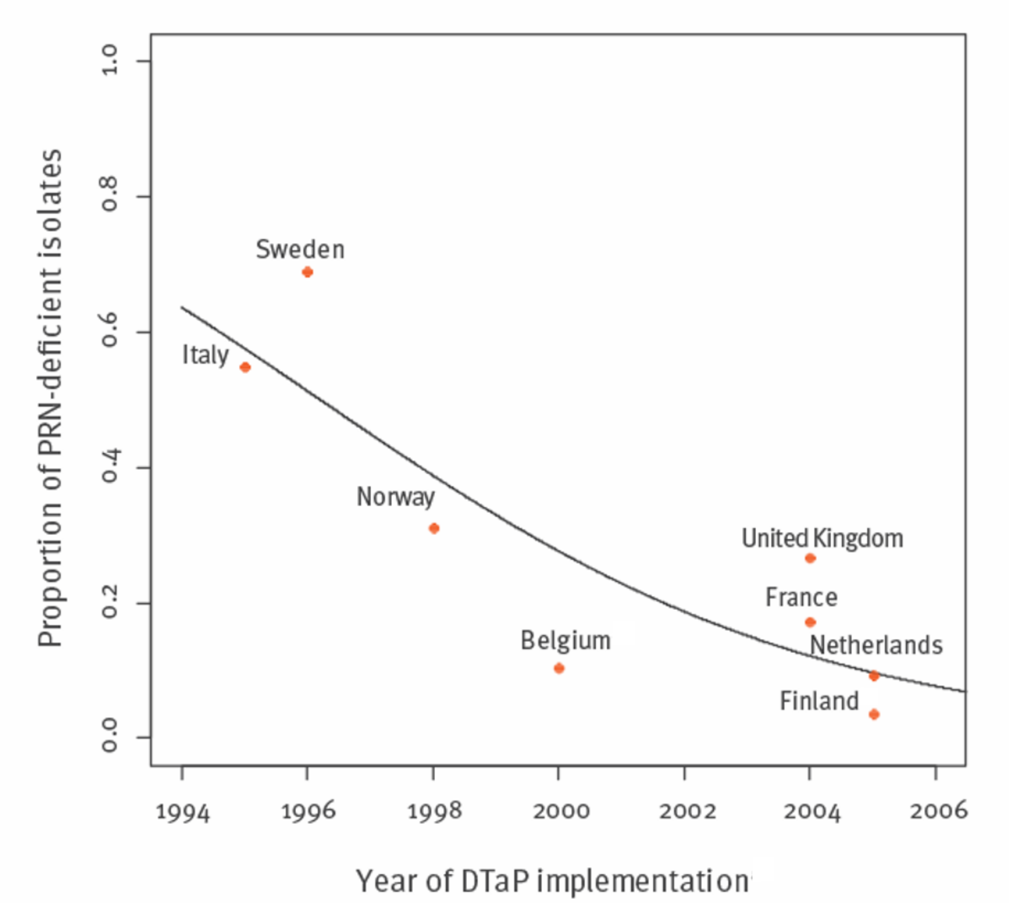


Figure 1.7 Barkoff et al. (2019) found a significant correlation between the year of introduction of the ACV and the proportion of *B. pertussis* isolates found to be Prn-deficient.

Figure reproduced with permission from Barkoff et al (2019) and Ring et al. (2019).

deficiency has continued. Nonetheless, the observations in these two countries indicate that there may yet be further twists in the ACV-mediated PRN deficiency story.

A smaller number of strains have also been isolated since the late 2000s which are deficient in other ACV antigens. A few strains have been found which were unable to produce PT. Interestingly, these strains were always also PRN-deficient (Bouchez et al., 2009; Williams, M.M. et al., 2016; Weigand et al., 2018). Additionally, and slightly more frequently, FHA-deficient strains have been isolated in recent years. Like PT deficiency, FHA deficiency often appears to occur in tandem with PRN deficiency (Bart et al., 2015; Weigand et al., 2018). Like PRN deficiency, PT and FHA deficiency can be caused by the insertion of IS481 into the gene, deletions, or deleterious SNPs in the protein coding genes themselves or in related regulatory genes.

Pertactin is thought to play an important role in the ACV: a clinical study showed that a three-component ACV containing PT, FHA and PRN was substantially more effective than a two-component ACV containing only PT and FHA, for example (Hewlett, 1997). Thus, PRN-deficient strains may have a selective advantage over PRN-producing strains in countries where a PRN-containing ACV is in use. Additionally, as Δ*prn* deletion strains do not seem to have a negative effect on virulence *in vivo* or *in vitro*, the benefits of PRN deficiency likely outweigh any costs to the deficient bacteria (Lipscombe et al., 1991; Roberts et al., 1991; Cherry et al., 1998; Storsaeter et al., 1998). Both PT and FHA, however, are thought to play more vital roles than PRN during *B. pertussis* infection, hence being deficient in either may result in a less virulent strain. Therefore, despite the selective pressure applied by their inclusion in the ACV, we may not see a rapid proliferation in strains which are deficient in these antigens; although deficiency-causing mutations may be occurring more frequently, the deficient strains are less likely to cause infection, therefore they will not be isolated and observed *in vitro*. However, as the FHA gene is long (10 kbp) and contains repetitive regions, it has often not been included in screens such as Bart, Harris, et al. (2014).

B. pertussis has a highly repetitive genome

Whilst the wide availability of WGS since the early 2000s has allowed numerous high-throughput studies of *B. pertussis*, these studies have been limited by the apparent monomorphic nature of the bacterium: very few inter-strain differences are detectable at the base level, aside from the changing allelic profile and increasing deficiency in a few key genes. On the other hand, the presence of so many IS elements in the *B. pertussis* genome, and the tendency of such regions to move within the genome, means that genome-level differences between strains may exist (Bentley et al., 2008; Siguier et al., 2014). In other species, for example *Pseudomonas aeruginosa* and *E. coli*, genome-level variations such as rearrangements have been shown to result in altered gene expression and phenotypic diversity (Sousa, de Lorenzo and Cebolla, 1997; Darch et al., 2014). However, detailed investigation of genome-level variations would require the assembly of closed genome sequences, which was not possible with any of the early-2000s WGS technologies. These technologies, such as Illumina's fluorophore-based sequencing, Roche 454's pyrosequencing and Ion Torrent's pH-based sequencing, produce sequencing reads of lengths in the order of hundreds of base pairs. *B. pertussis*'s repetitive IS elements are over a thousand base pairs long, hence have confounded attempts to assemble closed genomes using short sequencing reads. The *B. pertussis* genomes sequenced throughout the 2000s and 2010s using short-read sequencing have therefore tended to consist of at least several hundred pieces (contigs), ostensibly one contig per IS copy in the genome.

During the speciation of *B. pertussis* from its *B. bronchiseptica*-like ancestor, extensive genome arrangements occurred, largely due to IS-mediated recombination. It is likely that IS-mediated rearrangements are an ongoing phenomenon in the *B. pertussis* genome, as it has been demonstrated that reductions in genome size, which occur via the same mechanism, are also ongoing (King et al., 2010). IS-mediated rearrangements could result in phenotypic variation between strains, either through altering the distance of genes from the origin of replication, or through the recently discovered inward- and outward-facing promoters within the IS *481* sequence itself, either of which could cause changes in gene expression or regulation (Amman et al., 2018). Supporting this, Brinig et al. (2006) used microarrays to reveal altered gene expression in isolates with different genomic arrangements. Furthermore, PFGE studies have tended to suggest that each *B. pertussis* isolate can be assigned a "pulsed-field type", representing different genome arrangements in isolates which are otherwise similar, for example in terms of SNPs (Bisgard et al., 2001; Advani, Donnelly and Hallander, 2004; Advani et al., 2013; van Gent et al., 2015).

There has therefore long existed strong evidence that the assembly of closed *B. pertussis* genome sequences could reveal previously unresolved genome-level inter-strain variations, almost certainly including genome rearrangements and deletions, and potentially other additional IS-mediated changes, such as duplications. Thus, the arrival of long-read sequencing technologies on the scene in the early 2010s presented a significant opportunity in the field of *B. pertussis* genomics.

1.2 *The evolution of genome sequencing*

First generation sequencing facilitated the Human Genome Project

Genome sequencing has undergone several periods of rapid progress. The first full protein-coding gene to be fully sequenced, in the early 1970s, coded for the 129 amino acid (387 bp) coat protein of the RNA bacteriophage MS2. This first gene was sequenced by Min Jou et al. (1972), using a very lengthy, manual, ribonuclease and gel electrophoresis method. The entire bacteriophage MS2 RNA genome contains only three genes, and it was not long before the entire 3,569 bp genome had been sequenced: the first genome, RNA or DNA, ever fully sequenced (Fiers et al., 1976). However, larger whole genomes were thought to be too large and difficult to sequence in full using these manual methods and, for the next decade, sequencing projects continued to focus on a few select genes or the smaller phage genomes, such as the 5,375 bacteriophage ϕX174 (PhiX), which was the first full DNA genome ever sequenced, in the late 1970s (Sanger et al., 1977; Sanger et al., 1978; Green, Watson and Collins, 2015; Wetterstrand, 2016).

The Human Genome Project (HGP), begun in 1990, was a game changer for genome sequencing, requiring the collaboration of hundreds of researchers around the world, costing over $500 million, and taking over a decade to complete. This enormous undertaking paved the way for the sequencing of whole genomes of all kinds of organisms, from every kingdom of life. The sequencing of whole genomes was facilitated by a sequencing technique developed by Sanger, Nicklen and Coulson (1977). Originally known as the "chain termination method", this sequencing technique is still considered to be the most reliable, albeit most expensive and manual, sequencing method in existence. In the years since its inception, the chain termination method has come to be known, simply, as "Sanger sequencing".

Sanger sequencing makes use of dideoxynucleotides (ddNTPs) which, when incorporated in a DNA strand, terminate chain elongation by DNA polymerase. The DNA fragment to be sequenced is amplified, a short primer sequence is ligated to the template strand, and DNA polymerase is used to synthesise new complimentary strands. Synthesis requires the addition of nucleotides, A, C, G and T. In nature, deoxynucleotides (dNTPs) are incorporated into the new strand; in Sanger sequencing, both dNTPs and ddNTPs are present in the reaction mix. ddNTPs lack a hydroxyl group which is present in dNTPs and which is necessary for DNA polymerase to extend a DNA fragment. Numerous new copies of the original DNA fragment are synthesised by the DNA polymerase adding dNTPs one-by-one, until a ddNTP is incorporated instead and the synthesis halts. Because the ddNTPs are incorporated at random, if enough copies of the original fragment have been synthesised, fragments of all possible lengths, from 1 bp up to the total length of the original, will be present in the reaction mix. The original chain termination method utilised radiolabelled nucleotides to characterise the identity of each base in the sequence. In modern Sanger sequencing, however, the ddNTPs are tagged with fluorophores. Using capillary electrophoresis, which can resolve fragments to as little as 1 bp length difference, combined with laser excitation and detection, the identity of each base in the sequence can be deduced according to its fluorescence, as shown in **Figure 1.8** (Sanger and Coulson, 1975; Sanger, Nicklen and Coulson, 1977; Branton and Deamer, 2018).

Due to the single-base resolution of capillary electrophoresis, Sanger sequencing is highly accurate (Karger and Guttman, 2009). Technological developments during the 1980s allowed automation of the process, greatly increasing the speed at which DNA could be sequenced. This, in turn, made the

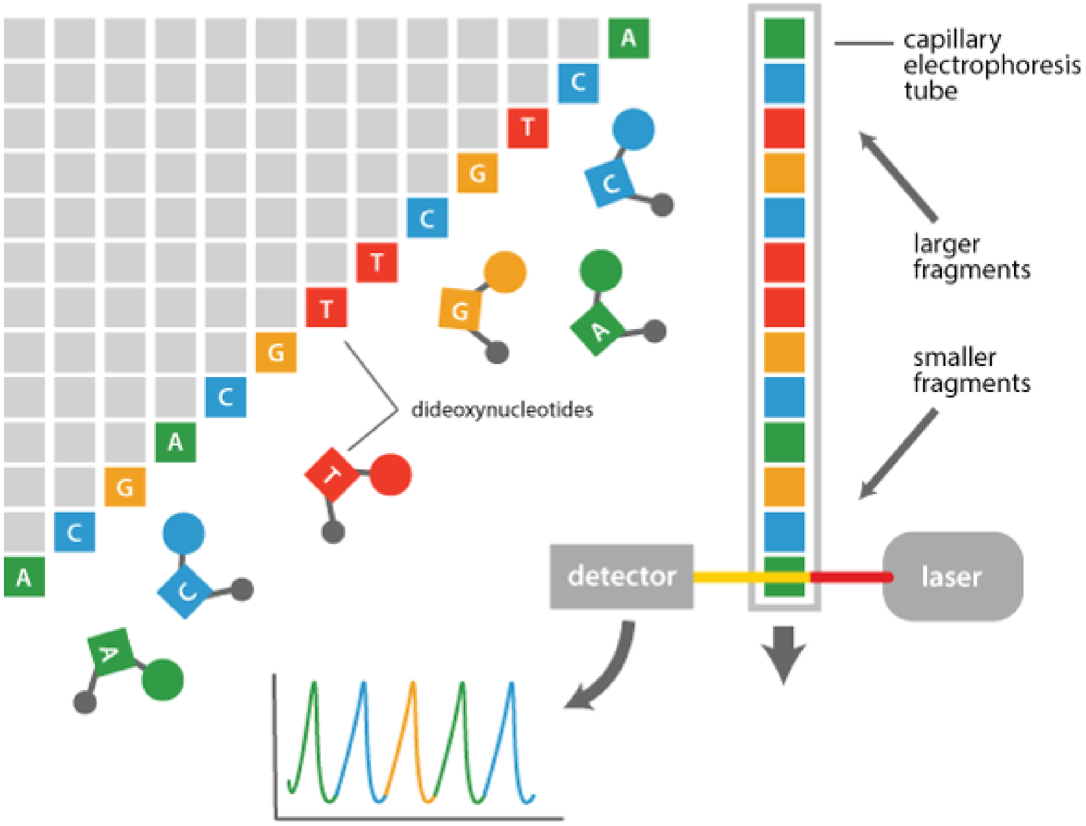


Figure 1.8 The "chain termination" technique, now known as Sanger sequencing. DNA fragments of all possible lengths are produced when labelled ddNTPs are incorporated at random during DNA polymerase elongation, as the ddNTP prevents further DNA polymerase activity. The fragments are separated according to size using capillary electrophoresis, which can resolve fragments to as little as 1 bp length difference. As the fragments pass through the capillary in size order, a laser excites the ddNTP's fluorophore label, and the identity of the fluorophore is detected and recorded. Thus, original sequence of the DNA can be deduced.

Image from ABM Knowledge Base (2017).

prospect of sequencing the full 3 billion base pairs of the human genome far more realistic. However, the length of reads produced by Sanger sequencing is limited by the ability of the capillary electrophoresis process to resolve fragments to single base pairs; typically, reads produced by Sanger sequencing have a maximum length of 500-800 bases (Heather and Chain, 2016; Branton and Deamer, 2018). In addition, despite automation, the process is relatively slow, as evidenced by the length of time taken to sequence the original human genome during the HGP. In order to speed up the HGP, many hundreds of researchers, in many different research centres around the world, were involved in the sequencing which, in turn, significantly increased the cost of the project. Nonetheless, the HGP proved that sequencing whole genomes was feasible, essentially giving rise to the field of genomics. However, faster and cheaper technologies would be required to conduct more high-throughput whole genome sequencing.

Technological advances enabled massively parallel Next Generation Sequencing

The early 2000s saw the rapid development of several new competing sequencing technologies, collectively known as the Next Generation Sequencing (NGS) technologies. The main aim of each of these was to decrease the time and cost involved in sequencing genomes. This aim was achieved through the development of massively parallel sequencing; a high number of short DNA fragments being sequenced at the same time. Genomes could now be sequenced in days or weeks instead of years, and the average cost of sequencing a genome decreased to thousands of dollars instead of millions by the early 2010s (Wetterstrand, 2016).

One of the first NGS methods to be developed was pyrosequencing, which was pioneered by Pål Nyrén and his colleagues in the late 1990s, and released commercially in the 2000s by 454 Life Sciences, which was later taken over by Roche (Nyren, 1987; Ronaghi et al., 1996; Ronaghi, Uhlen and Nyren, 1998). The pyrosequencing method makes use of pyrophosphate ions, which are released during DNA synthesis when a complementary base is incorporated into the growing strand, to fuel a chain reaction which culminates in the release of light, as illustrated in **Figure 1.9**. ATP sulfurylase catalyses the synthesis of ATP from pyrophosphate and adenosine pyrosulfate (APS). Luciferase can then use the ATP produced from the pyrophosphate to catalyse the conversion of luciferin to oxyluciferin, producing light as a by-product. In pyrosequencing, the template strand of interest is bathed in DNA polymerase, APS, ATP sulfurylase, luciferin and luciferase. dNTPs are introduced to this mixture in turn; when the next complementary dNTP is introduced, DNA polymerase will incorporate a base into the growing complementary strand, releasing a pyrophosphate ion which subsequently results in the release of light. The light can be detected, hence the identity of each base in the DNA can be deduced according to which dNTP was present in the mix when the light was released. The amount of light released corresponds to the number of bases which have been incorporated; if the next several bases in the template strand are all the same (a homopolymeric tract), more light will be detected. However, homopolymeric tracts are a source of error in pyrosequencing, because the amount of light detected can be misinterpreted, and resolution between single base pair differences is difficult (Balzer, Malde and Jonassen, 2011). Because dATP would catalyse the luciferase reaction without the need for

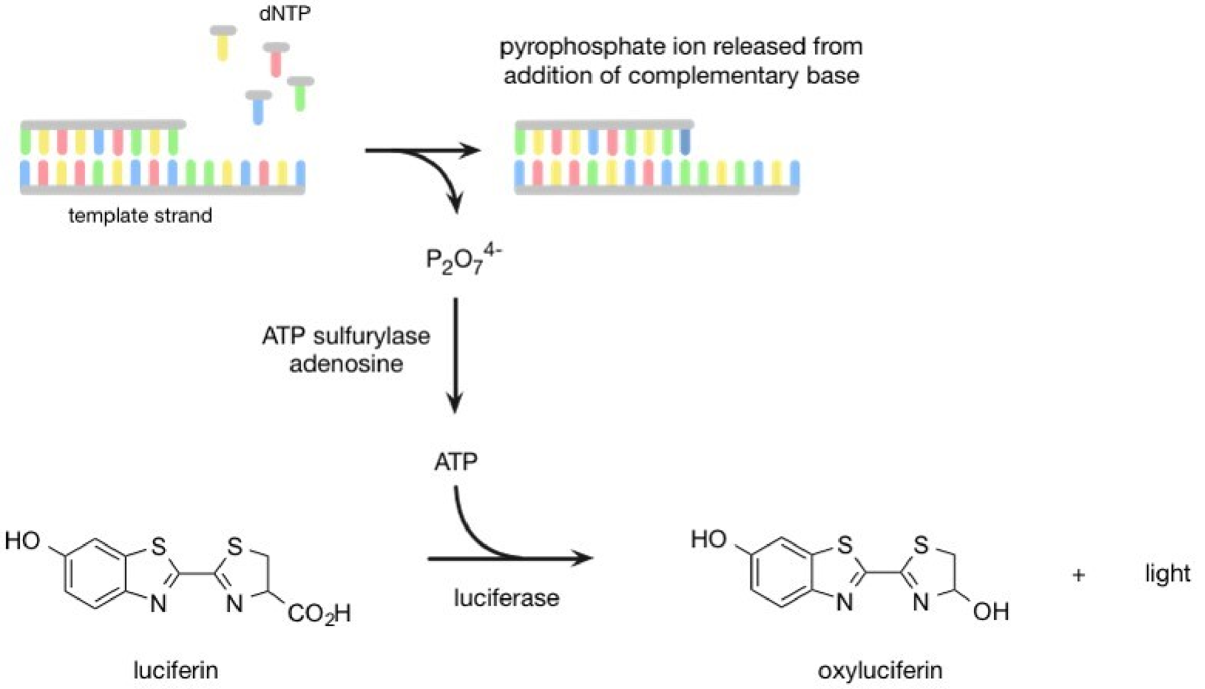


Figure 1.9 The pyrosequencing method developed by Pål Nyrén and colleagues and commercialised by 454 Life Sciences and Roche.

Figure adapted from atdbio (2019).

pyrophosphate release, dATPαS is used instead. dATPαS can be incorporated into the complementary strand by DNA polymerase, but cannot be used by luciferase.

Massively parallel pyrosequencing was achieved through the use of microbeads. The DNA of interest is sheared into fragments less than a thousand base pairs long, and denatured to form single-stranded DNA (ssDNA). The ssDNA fragments are adhered to individual microbeads, and PCR is used to amplify the DNA on the beads, so that each microbead is coated with millions of copies of the same fragment. The microbeads are placed into microwells on a single plate, each of which can contain hundreds of thousands of microwells; thus, by detecting the output of each microwell individually, hundreds of thousands of DNA fragments can be sequenced at the same time (Margulies et al., 2005; Heather and Chain, 2016). This ability to sequence large numbers of DNA fragments at the same time, facilitated by technological advances - in microfabrication to produce the microbeads and in high-resolution imaging to detect signals produced during the sequencing process - is the defining feature of NGS, and is key in many other DNA sequencing methods developed during the 2000s (Shendure and Ji, 2008; Heather and Chain, 2016).

Another such NGS technique which employs a massively parallel, microbead-based, method is Ion Torrent (Rothberg et al., 2011). The Ion Torrent was the first sequencer to use a non-imaging method; instead, the Ion Torrent method deduces the sequence of DNA strands by monitoring fluctuations in pH as a strand complementary to the fragment of interest is synthesised. The binding of a nucleotide to a growing DNA strand releases protons, hence the incorporation of a nucleotide can be detected as a pH change. During Ion Torrent sequencing, as with pyrosequencing, dNTPs are delivered in an iterative fashion to the reaction mix, and a pH change is only detected when the dNTP matching the next base in the DNA sequence is delivered. Despite the elegant theory behind this method, Ion Torrent sequencing shares one of pyrosequencing's major flaws: it is prone to errors when sequencing homopolymeric sequences, because more than one dNTP may be incorporated at the same time, and the extra signal can be missed or difficult to interpret accurately.

Most NGS technologies are similar to Sanger sequencing, in that they use DNA polymerase-based methods which are collectively known as "sequencing-by-synthesis" (SBS). SBS methods work by using DNA polymerases to synthesise new DNA strands complementary to the template strand of interest, whilst monitoring the order in which bases are added to the new strand. Each SBS technology uses a slightly different method to monitor which bases are being added, each with its own accompanying strengths and weaknesses. Not all NGS sequencers use SBS methods, however. Applied Biosystems' "sequencing by oligonucleotide ligation and detection" (SOLiD) sequencing system uses, for example, a sequencing-by-ligation (SBL) method (McKernan et al., 2009). The SOLiD method is complicated and requires several sequencing cycles per strand, but the basic principle involves the use of DNA ligase, instead of DNA polymerase, to hybridize short fragments instead of individual bases to the template strand of interest, releasing differently coloured fluorophores to help identify each base along the way. Unlike Ion Torrent and pyrosequencing, the SOLiD method is not prone to homopolymer errors. Conversely, it has been reported to have a high error rate when sequencing palindromic sequences (Huang et al., 2012).

The most commonly-used NGS method also monitors the release of fluorophores to decipher DNA sequences. Originally called Solexa, Illumina sequencing is an SBS technology which has become the gold standard for WGS since its inception in the mid-2000s (Bennett, 2004; Bentley et al., 2008; Greenleaf and Sidow, 2014). As shown in **Figure 1.10**, the Illumina method binds short fragments of ssDNA to the surface of a flow cell, where they are amplified by PCR to produce clonal clusters. Sequencing then commences using cycles of DNA polymerase and DNA-blocking enzymes to sequentially add single fluorescently-tagged dNTPs, building a new strand complementary to that being sequenced. As each dNTP is incorporated into the nascent strand, its fluorophore is released and can be detected by the sequencer. Each of the nucleotides (dATP, dTTP, dGTP, dCTP) is tagged with a differently-coloured fluorescent molecule, thus the sequence of the template strand can be deduced according to the colour of the fluorescence released. The use of clonal clusters ensures that this fluorophore signal is strong enough to be distinguished from background noise, as the release of a single fluorophore would not be. The accuracy of Illumina sequencing is generally high, with error rates 0.1% or below (Fox et al., 2014; Pfeiffer et al., 2018). However, the inclusion of a PCR amplification step can introduce bias, mainly in genomes with a particularly high or low GC content (Kanagawa, 2003; Benjamini and Speed, 2012; Chen, Y.C. et al., 2013). This kind of bias results in the over- or under-representation in certain DNA fragments in sequencing libraries. GC-bias is a problem which has affected every sequencing method discussed thus far, even those which do not use PCR, like Sanger sequencing (Aird et al., 2011).

Most NGS sequencers are large and expensive. Illumina, for example, produce a wide range of sequencing machines, with outputs ranging from 0.14 Gb of data in a 9.5-hour run using the Illumina iSeq, to the ultra-high-throughput Illumina NovaSeq which has an output of 3 Tb of data in a 44-hour run. Even the iSeq requires a bench-top and weighs 16 kg (Illumina, 2019a, b). The running cost per sample for Illumina sequencers is relatively low, but the start-up cost of Illumina sequencing remains a barrier for many laboratories, with the iSeq costing around $20,000, the NovaSeq costing almost $1

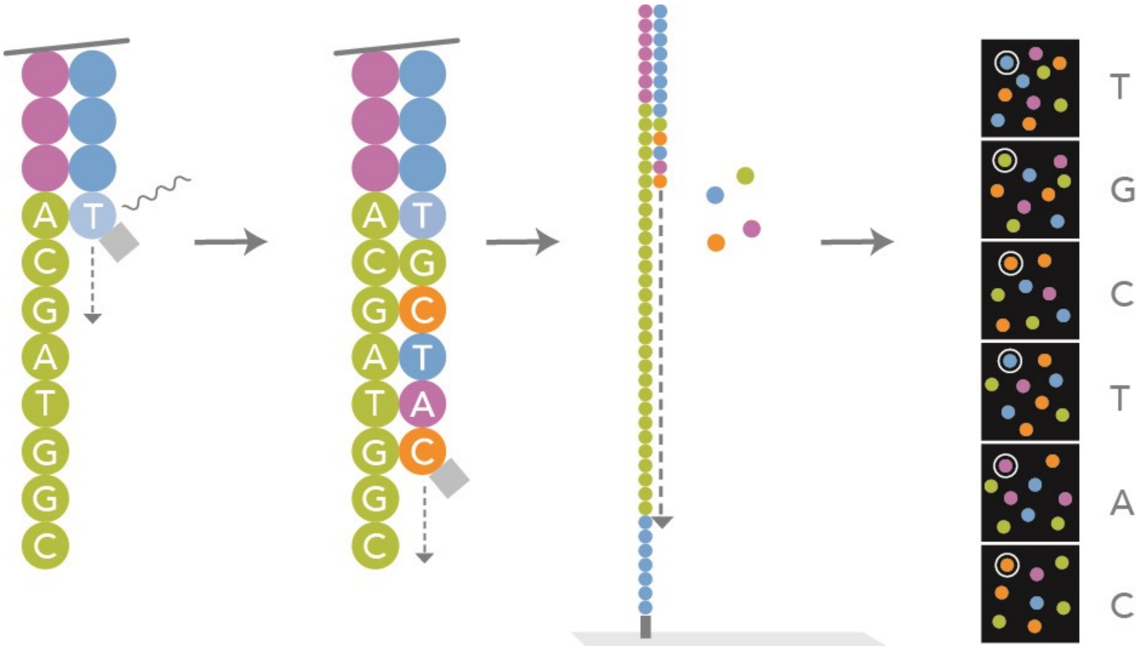


Figure 1.10 Illumina has been the gold-standard sequencing method during the 2010s. The strand of interest is bound to one of the thousands of adaptor sequence tethered to the flow cell. Numerous cycles of PCR produce a clonal cluster, in which all strands are identical to the template strand. Sequencing then progresses in a cyclic fashion: DNA polymerase is attached to the template strand, and sequentially adds fluorophore-tagged dNTPs to build the complementary strand. After the addition of each dNTP, synthesis is temporarily blocked enzymatically, and the fluorophore is released. The fluorophore is laser-detected, sequencing is unblocked, and the next dNTP can be incorporated. Each type of dNTP is tagged with a different colour, thus the identity of each base in the strand can be deduced according to the colour sequence detected from each clonal cluster.

Figure from Illumina (2017).

million, and the other Illumina sequencers costing anywhere in between (Herper, 2017; Proffitt, 2018). Nonetheless, the establishment of large sequencing service centres, like MicrobesNG at the University of Birmingham, means that WGS has become increasingly widely available throughout the late 2000s and 2010s (MicrobesNG, 2019).

In 1995, the 1.83 Mb genome sequence of *Haemophilus influenzae Rd* was published, the first bacterial whole genome ever to be sequenced, and a major leap forward in the genomics field (Fleischmann et al., 1995). This was duly followed by the first draft human genome sequence in 2001, and a slightly more complete draft in 2004 (Lander et al., 2001; International Human Genome Sequencing Consortium, 2004). So rapid has been the development and widespread uptake of whole genome sequencing since then that the number of curated genome assemblies across all kingdoms of life available from the NCBI database is now (as of September 2019) over 270,000, representing almost 50,000 different organisms and including 213 different human genome assemblies. The vast majority of the genomes sequenced thus far have been prokaryotic: over 210,000 of the genome assemblies stored in the NCBI database are prokaryotic, and over 206,000 of those are bacterial (almost 18,000 of these assemblies belong to the model gram-negative bacterium, *E. coli*) (NCBI, 2019). Our understanding of bacterial genomics has increased at an unprecedented rate, to the extent that we now know enough to be able to design new bacterial species, like the synthetic bacterial genome of *Caulobacter ethensis*-2.0 published in April 2019 (Venetz et al., 2019). However, most of the genome assemblies available from the NCBI database are not closed; they consist of multiple fragments, known as contigs. Until 2011, for example, only one closed genome sequence for *B. pertussis* existed: that of the reference strain, Tohama I, which was sequenced in 2003 (Parkhill et al., 2003). The reason for the lack of closed genomes is the length of sequencing read produced by the NGS technologies used to sequence most of the available genome assemblies.

As mentioned at the end of section 1.1, sequencing read length is limited in NGS technologies. In Illumina sequencing, reads are limited to the length of the tethered strands in the clonal clusters, usually 150-300 bp. The sequencing of the identical strands in clonal clusters, whilst initially ensuring higher accuracy, is prone to phasing issues. The base within the strand being called at any given moment tends to get out of sync between the different strands within each cluster; the longer the sequencing continues, the more out of sync the sequencing will be, rapidly deviating beyond the level of phasing which can be corrected informatically. This means sequencing longer strands would result in lower accuracy (Nakamura et al., 2011; Loman, 2013). In addition, the read lengths for all sequencing methods which use PCR to produce clonal clusters or fragment-covered microbeads will be limited by the efficiency and accuracy of PCR reactions, which decreases with increasing fragment length (Knierim et al., 2011; Jia et al., 2014).

As shown in **Figure 1.11**, repetitive regions of DNA sequence are difficult or impossible to assemble into single fragments using short reads. Instead, reads which are longer than the longest repetitive region in a genome are required to produce a closed sequence for that genome. It has been suggested that the “golden threshold” of read lengths for the assembly of microbial genomes is 7 kbp, because 7 kbp should span even the longest repetitive region in any microbial genome (Koren and Phillippy, 2015). During the 2010s, therefore, the focus of the newest DNA sequencers has been the production of read lengths in the thousands of bases, as opposed to the hundreds of bases of NGS.

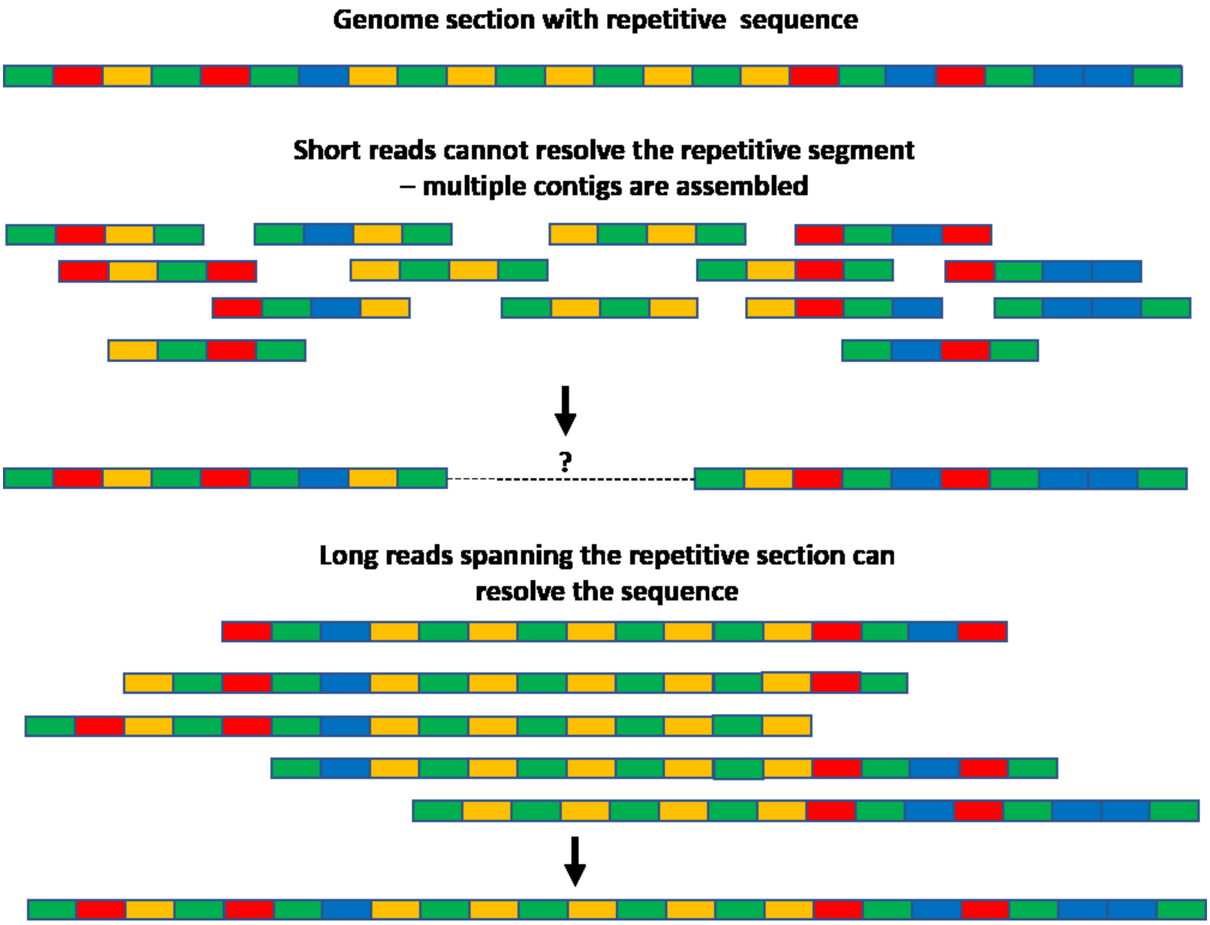

Genome section with repetitive sequence
Short reads cannot resolve the repetitive segment – multiple contigs are assembled
?
Long reads spanning the repetitive section can resolve the sequence

Figure 1.11 Repetitive regions of sequence can be resolved by reads longer than the repetitive region

Long-read sequencing of *B. pertussis* with Pacific Biosciences instruments succeeded in producing closed genome sequences

The first long-read sequencing technology to be commercially released was Pacific Biosciences (PacBio) sequencing. Theoretically to be acquired by Illumina by the end of 2019 for $1.2 billion, PacBio sequencing represents another step-change in genomics, facilitating widespread assembly of closed microbial genome sequences for the first time (Eisenstein, 2019). The PacBio sequencing method is a variation of classic SBS, sharing much in common with Illumina sequencing. In Illumina sequencing, the DNA to be sequenced is bound to the flow cell, and DNA polymerase and fluorescently tagged dNTPs are in solution. By contrast, in PacBio sequencing, the DNA polymerase is bound to wells in the flow cell, and the DNA to be sequenced is in solution with the dNTPs. As shown in **Figure 1.12**, DNA strands are captured by the DNA polymerase and the complementary strand is synthesised, releasing fluorescence each time a base in incorporated (Eid et al., 2009). The location of each DNA polymerase well is monitored precisely, meaning the fluorescence from single DNA strands can be distinguished from background noise. This means that PCR is not required, thus some of the bias intrinsic to NGS techniques is avoided (Kanagawa, 2003). The PacBio method is able to produce read lengths much longer than NGS technologies; the average read produced by the PacBio RSII or Sequel sequencers is 10-16 kbp, and reads of up to 60 kbp have been regularly reported (Rhoads and Au, 2015; Weirather et al., 2017; Ardui et al., 2018). PacBio sequencing read length does have an upper limit, however, as the lifespan of the DNA polymerases is limited.

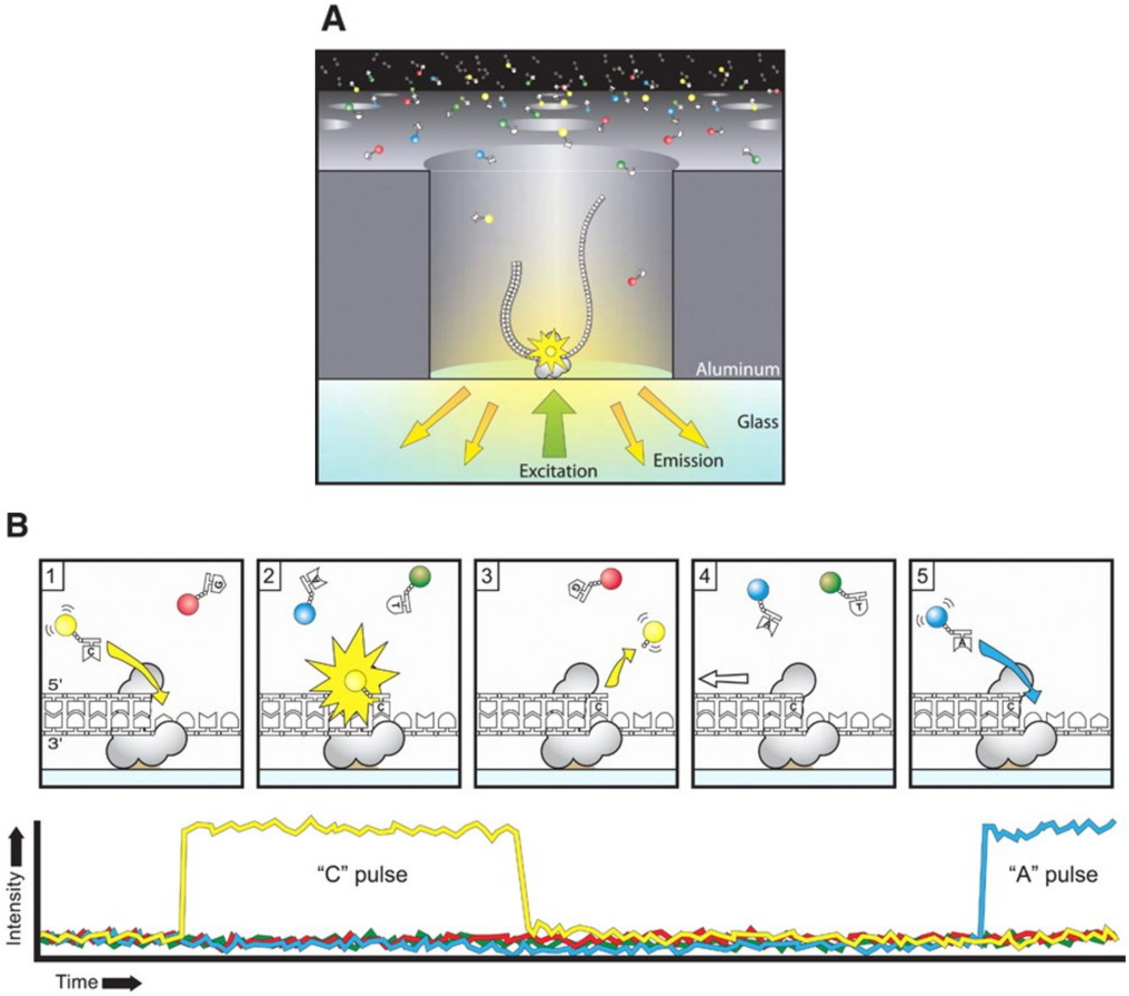


Figure 1.12 PacBio's SMRT sequencing method. DNA polymerases are attached to the bottom of wells across the flow cell. DNA strands to be sequenced are introduced in solution, and can be captured by the polymerases. Sequencing then progresses in a manner largely similar to the Illumina SBS method: DNA polymerase catalyses the synthesis of the complementary strand, using fluorescently tagged dNTPs which are also supplied in solution. The location of each polymerase-containing well is monitored precisely for fluorescence, thus the sequence of the strand is deduced.

Figure adapted from Eid et al. (2009)

Sequencing individual DNA strands instead of clonal clusters or clonal microbeads results in a slightly lower raw accuracy. Consequently, PacBio sequencing includes an internal consensus step: template strands are circularised and processed multiple times by the same DNA polymerase, giving multiple calls for each base in the strand, which can be averaged. Early PacBio sequenced had a raw error rate of up to 18% (Nagarajan and Pop, 2013). However, developments to the sequencing chemistry and data analysis tools have led to rapid raw accuracy improvements. In addition, unlike most NGS technologies which make systematic errors (such as Ion Torrent's homopolymer issues, or SOLiD's palindromic sequence issues), PacBio sequencing has a random error profile. This means that errors can be corrected informatically, with high enough sequence coverage. Together, these factors have led to reported PacBio accuracies of over 99.9% (Eisenstein, 2019).

The first study to take advantage of PacBio for *B. pertussis* was Bart, Zeddeman, et al. (2014). Using PacBio's long reads, Bart et al. were able to produce closed, fully annotated, genomes for two B. pertussis strains: B1917 and B1920. These were the first closed *B. pertussis* genomes to be produced since the original sequencing of the Tohama I reference strain using Sanger sequencing (Parkhill et al., 2003). The arrangement of the B1917 and B1920 genomes differed significantly, with three large

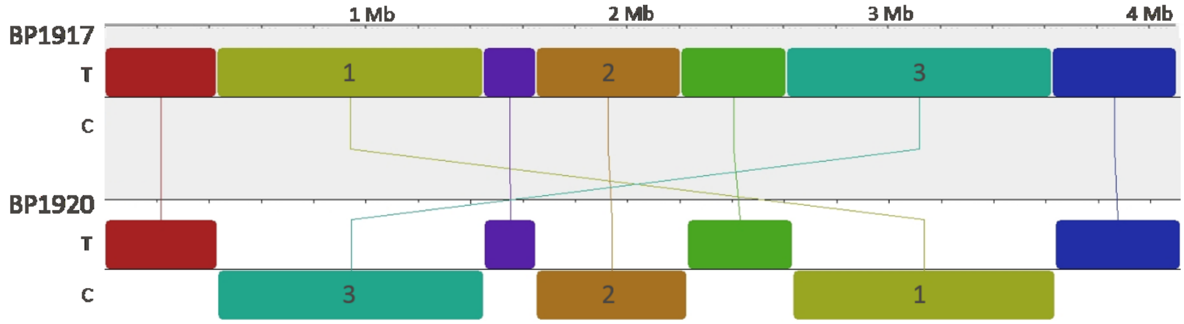


Figure 1.13 Different arrangements of the B. pertussis genome seen in BP1917 and BP1920. Three large sections of the genome (1, 2, and 3) are inverted in BP1920, as shown by their appearance on the complementary strand when the two whole genome sequences are aligned using progressiveMauve (Darling, Mau and Perna, 2010). The Genbank accession numbers for the genome sequences produced by Bart, Zeddeman, et al. (2014) for BP1917 and BP1920 are CP009751 and CP009752, respectively.

Figure and legend from Ring et al. (2019).

inversions and a variety of deletion and/or insertion events between the pair (**Figure 1.13**). Having proven the ability of PacBio to produce closed *B. pertussis* genomes, Bart et al. (2015) next sequenced 11 *B. pertussis* strains which represented the pandemic *ptxP3* lineage, again using PacBio sequencing to produce 10 kbp-long reads. These additional strains were predictably similar in terms of SNPs and indels, but again showed extensive differences in the arrangement of their genomes.

Since its commercial rollout in the early 2010s, the cost of PacBio sequencing has continued to rapidly decrease. Thus, high-throughput strain screens are becoming increasingly feasible. **Figure 1.14** shows the dramatic increase in closed genome sequences for the classical Bordetella species available from the NCBI's RefSeq database since 2014. In 2011, only one closed sequence was available each for *B. bronchiseptica*, *B. parapertussis* and *B. pertussis*. As of September 2019, 20 *B. bronchiseptica*, 21 *B. parapertussis* and 550 *B. pertussis* closed genomes were available (not including those produced in this thesis). A large number of the closed *B. pertussis* genomes have been generated by the USA's Centers for Disease Control (CDC), mostly using PacBio long reads in hybrid with Illumina short reads. The hybrid assembly approach improves the accuracy of assemblies produced using long-read sequencing, as long-read sequencing still has an intrinsically higher error rate than short-read sequencing (Au et al., 2012; Koren et al., 2012).

Bowden et al. (2016) at the CDC conducted the first whooping cough outbreak screen to use long and short reads in hybrid, sequencing 31 strains isolated in the USA during 2010 and 2012 whooping cough outbreaks. 21 different arrangement profiles were seen in the 31 genomes, most consisting of inversions around the origin of replication, as appears to be common in bacterial genomic rearrangement (Boccard, Esnault and Valens, 2005). They also validated the arrangements using whole genome optical mapping. This showed that, as predicted, the boundaries between rearranged sections were composed of a repeated element: an insertion sequence, or the rRNA operon. 89% of the boundaries consisted of the most abundant insertion sequence, IS 481.

The most thorough investigations of *B. pertussis* genomic rearrangement thus far were also conducted by the CDC, and used a hybrid assembly strategy. Weigand et al. (2017) combined PacBio long reads with Illumina short reads to close the genomes of 257 strains, dating from 1939 to 2014. When clustered based on their arrangement profiles, most isolates clustered according to allelic profile. This suggests that most structures are relatively stable, as supported by a clinical isolate which showed the

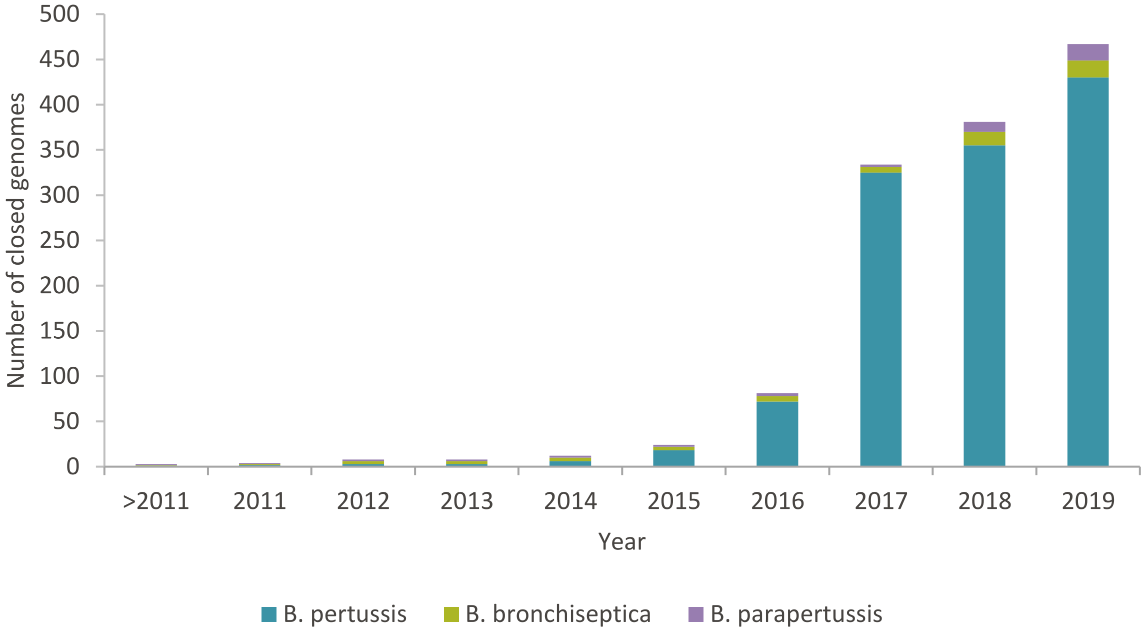


Figure 1.14 The numbers of closed classical *Bordetella* genomes available on RefSeq has increased rapidly since long-read sequencing technologies became widely available.

Figure from Ring et al. (2019).

same structure before and after 11 serial passages. Furthermore, these findings suggest that lineages are conserved not just in terms of SNPs, but also in genomic arrangement. Interestingly, Weigand et al. note that, on average, only half of their predicted IS 481 target sites are occupied in any given genome, suggesting a potential for further IS-mediated structural changes in future generations, assuming these sites are not non-permissive. This study was followed by Weigand et al. (2019), in which closed genome sequences were produced for 469 *B. pertussis* isolates and 167 isolates from other *Bordetella* species with multiple IS elements. This showed that, across the genus, rearrangements tended to take the form of large, symmetrical inversions. In addition, whilst the scope for different genome arrangements is very large, only a small subset of arrangements tends to be observed, suggesting that certain gene orders are more favourable than others.

Like NGS technologies, PacBio sequencers are large and expensive: the most recent sequencer, the Sequel II, is reported to have a basic price of almost $500,000 (Genome Web, 2019). This means that only larger sequencing centres, like the CDC, can afford to run on-site PacBio sequencing, although a number of sequencing service providers have started to add PacBio sequencing as an option. In 2014, however, a competing long-read sequencer was released to early access users, one which promised to reduce start-up costs to less than $1,000, and to make DNA sequencing portable: Oxford Nanopore Technology (ONT)'s MinION nanopore sequencer.

Nanopore sequencing has developed over three decades

Nanopore sequencing is the brainchild of David Deamer and Daniel Branton, who began developing the concept in 1989 (Deamer, Akeson and Branton, 2016). The theory behind nanopore sequencing is simple: single strands of DNA can be drawn through nano-scale pores in a membrane by electrophoresis, disrupting the flow of current across the membrane in a manner distinctive of which bases are translocating at any given moment. However, the journey from conception to commercialisation was long and required much optimisation, trial and engineering of numerous proteins. From conception to the release of the prototype nanopore sequencers to a group of “early access” users in 2014 took 25 years, and even since its release, the technology has continued to develop rapidly.

The earliest research into nanopore sequencing all focussed on the *Staphylococcus aureus* transmembrane protein, α-hemolysin. Work in the 1980s and early 1990s showed that the 1.5 nm pore formed by α-hemolysin was big enough to accommodate ssDNA (~1.2 nm diameter), but not double-stranded DNA (dsDNA, >2 nm diameter) (Menestrina, 1986; Song et al., 1996). John Kasianowicz worked extensively with α-hemolysin throughout the 1990s, establishing the conditions under which polynucleotides could be drawn through the pore by inserting the protein into a lipid bilayer between two compartments filled with KCl (Kasianowicz et al., 1996). By applying a positive voltage difference of greater than 80 mV across the membrane, RNA homopolymers were seen to move from the “cis” compartment into the “trans” compartment. The flow of current was disrupted as the RNA moved, for lengths of time proportional to the length of the strand. Having shown that translocation of RNA was possible, the next trials included homopolymeric double-stranded and single-stranded DNA. qPCR was used to quantify the strands in the trans compartment, showing that only ssDNA had been able to translocate and thus proving that the width of the α-hemolysin pore did limit translocation to single strands alone (Kasianowicz et al., 1996).

The next important developments, from the late-1990s to late-2000s, focussed on proving whether the *S. aureus* α-hemolysin pore could discriminate between different nucleotide bases according to their level of current disruption. Along the way, Akeson et al. (1999) used synthesized RNA strands consisting of 70 Cs followed by 30 As to show that two distinct levels of current disruption could be recorded, although it was found that these different levels were due to the different secondary structures formed by the polyC segment compared to the polyA, rather than differences between individual bases. Soon afterwards, Meller et al. (2000) showed that ssDNA polynucleotides of identical length but different base sequences produced different current blockage signals. Five years later, Ashkenasy et al. (2005) synthesised a variety of polyC strands which each had a single adenosine substitution at a different position. Each type of strand produced a different flow of current through the α-hemolysin pore, which could be used to distinguish at which base the adenosine had been substituted, essentially proving that it was possible to resolve sequences at the level of single bases. This was followed by several years of work by David Stoddart, showing that individual bases could be resolved even in non-homopolymeric sequences, and further characterising how the structure of the α-hemolysin pore interacted with the ssDNA strands to disrupt the flow of current (Stoddart et al., 2009; Stoddart, Heron, et al., 2010; Stoddart, Maglia, et al., 2010).

By 2010, therefore, it had been well established that the idea of using nanopores to sequence DNA or RNA with single base resolution was feasible. However, the studies to that point had almost all used terminal hairpin complexes to suspend the strand of interest within the pore, which greatly improved the signal-to-noise ratio, but would not be suitable for a high-throughput sequencing method (Deamer, Akeson and Branton, 2016). For high-throughput sequencing, strands would need to flow freely from the cis chamber to the trans chamber, leaving the pore unblocked for subsequent strands. Therefore, a mechanism for slowing down the translocation of strands through the pore was needed; unhindered, the voltage bias required to drive strands through a pore would result in a translocation speed of less than 10 µs per nucleotide. To ensure that current fluctuations lasted long enough to be detected, a processive enzyme would be required, to ratchet strands through the pore at a detectable rate, calculated to be at least 100 µs per nucleotide (Deamer, Akeson and Branton, 2016; Byrd and Raney, 2018).

Previous work had already demonstrated the utility of processive enzymes in slowing the rate of DNA translocation through a protein pore (Church et al., 1998; Benner et al., 2007; Hornblower et al., 2007). Cockroft et al. (2008) then successfully showed than an A family DNA polymerase could allow single-nucleotide resolution. This method included the attachment of a hairpin structure to the end of the DNA strand of interest. The DNA strand could translocate through the pore, driven by the positive voltage difference, up to the hairpin structure. The DNA polymerase could then theoretically ratchet the DNA strand back through the pore one nucleotide at a time. However, this type of polymerase required a voltage alternating between -30 and 30 mV applied to the membrane, to force the DNA polymerase to bind to and then dissociate from the DNA strand being sequenced. The voltage difference needed to drive the DNA strand through the pore initially would therefore force the A family DNA polymerase to remain dissociated from the DNA, hence at most only two nucleotide additions could be observed in sequence using this type of polymerase. It was soon discovered that phi29 DNA polymerase was more resistant to the necessary voltage, staying tightly associated with the DNA long enough to sequence strands up to tens of nucleotides long, using the α-hemolysin pore (Lieberman et al., 2010; Cherf et al., 2012; Byrd and Raney, 2018). The mechanism is illustrated in **Figure 1.15**.

Early prototype nanopore sequencers were therefore produced in 2011 and 2012, which used phi29 DNA polymerase to ratchet DNA strands through a pore, this time an engineered MspA porin protein. Although single base resolution was possible using α-hemolysin, it had become apparent that the physical characteristics of the pore were sub-optimal. Chiefly, its length (~5nm) allowed 12 nucleotides to traverse the membrane at the same time, meaning that the current signal was dependent on 12-mers. This made the processing of the current trace into single bases very computationally intensive and difficult (Meller, Nivon and Branton, 2001; Deamer, Akeson and Branton, 2016). Consequently, a pore protein with a much shorter aperture, but same diameter, was trialled and developed: the MspA porin from *Mycobacterium smegmatis* (Niederweis et al., 1999; Faller, Niederweis and Schulz, 2004). The aperture of the MspA pore was almost ten times shorter (~0.6 nm) than that of α-hemolysin and, with some targeted mutations, was engineered to produce a protein which enabled far easier discrimination between single nucleotides than the original nanopores tested (Butler et al., 2008; Derrington et al., 2010). The trials of the phi29 DNA polymerase-MspA prototype sequencers were successful, and spurred on the development of more widely available commercial nanopore sequencing (Cherf et al., 2012; Manrao et al., 2012).

The first commercially available nanopore sequencer, the MinION, was announced by Oxford Nanopore Technologies in 2012, before being released to early access users two years later. The early MinION used the same MspA pore as the prototypes, but now used a helicase enzyme to ratchet the DNA through the pore one nucleotide at a time, as shown in **Figure 1.15**. Like α-hemolysin before it, phi29 DNA polymerase turned out to be sub-optimal for large scale nanopore sequencing due to a number of inhibitive errors, such as periodically skipping forward or backwards by multiple bases (Heron, 2018). Thus, a helicase enzyme, Dda, from bacteriophage T4 was engineered instead (Byrd and Raney, 2018). Because no shearing of DNA is required by the helicase-nanopore method, DNA strands of theoretically any length could be sequenced by the MinION.

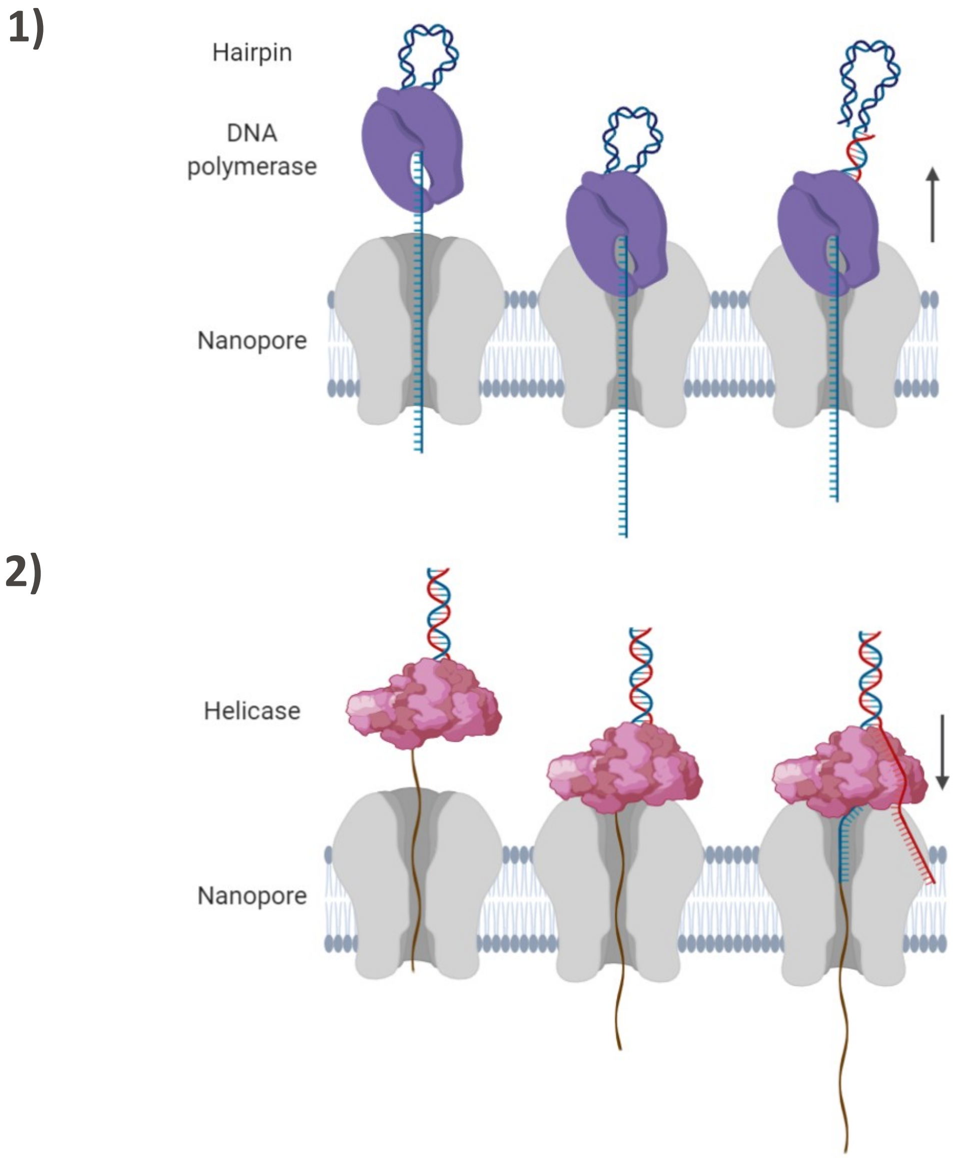


Figure 1.15 DNA polymerase (1) vs DNA helicase (2) as a DNA processing enzyme.

1) DNA polymerase is bound to a short hairpin polymer, which is then ligated to the 3’ end of the ssDNA fragment of interest. The voltage difference across the membrane draws the DNA strand through the nanopore to the trans chamber. The DNA polymerase cannot pass through the pore, and is activated when the two come into contact. The activated polymerase then ratchets the DNA back through to the cis chamber, one nucleotide at a time, resynthesising the complementary strand in the process.

2) A synthetic polymer strand is ligated to the 5’ end of a dsDNA fragment to be sequenced, both of which are captured through a helicase enzyme. The anionic polymer is drawn through the nanopore by the voltage difference across the membrane. The helicase cannot pass through the pore, and is activated when the two make contact. The activated helicase then steps along the dsDNA strand, unzipping it and feeding nucleotides through the pore to the trans chamber one at a time. The complementary strand is released into the cis chamber.

Figure inspired by Byrd and Raney (2018).

The MinION uses a "two-electrode self-referencing" electrochemistry system, which allows for a steady flow of ionic current from the cis chamber to the trans. This system uses platinum electrodes in both chambers, and a bright yellow potassium ferricyanide/ferrocyanide solution to donate and accept electrons respectively, maintaining a stable flow of current throughout a sequencing run:

$$\text{Trans electrode: } [\mathrm{Fe(CN)}_6]^{3-} + e^{-} \rightleftharpoons [\mathrm{Fe(CN)}_6]^{4-}$$

$$\text{Cis electrode: } [\mathrm{Fe(CN)}_6]^{4-} \rightleftharpoons [\mathrm{Fe(CN)}_6]^{3-} + e^{-}$$

A synthetic membrane separates the cis and trans chambers. The cis chamber is a single compartment, with a single electrode, whereas the trans compartment consists of 2,048 wells, each of which contains its own electrode. Each well can contain a single nanopore, meaning a MinION flow cell can contain up to 2,048 individual pores (although a variety of factors during production and shipping mean the actual number of pores on any flow cell is usually significantly lower than this). An application-specific integrated circuit (ASIC) chip sits below the pore wells, and can detect current from 512 of them at any given moment. Each ASIC channel is therefore responsible for sensing current from four wells; the MinION software, MinKNOW, periodically scans all the pores and determines which set of 512 (if 512 are still functional) should be sensed by the ASIC chip (Clarke, 2018). This system is reminiscent of the massively parallel sequencing strategies of NGS and, indeed, early users were optimistic that the MinION could achieve comparable sequencing yields.

A pilot study using early nanopore sequencing to sequence *B. pertussis* was unsuccessful

When the first MinIONs were made available in 2014, preliminary gDNA sequencing was carried out by the Bagby group for two *B. pertussis* strains, UK48 and UK76. Like PacBio sequencing before it, the long reads that the MinION was capable of generating should have been well able to produce fully resolved, closed *B. pertussis* genome sequences. The preliminary sequencing used the Mk1 MinION, the first publicly available sequencing chemistry, R7.3, and cloud-based basecalling with Metrichor. gDNA was extracted from UK48 and UK76 by Tom Belcher at the University of Bath and shipped to the Nanopore Group at the University of California, Santa Cruz. Two 48-hour sequencing runs were completed for UK76, and one was completed for UK48. These runs demonstrated the long-read capabilities of MinION-based nanopore sequencing, with mean read lengths of 5.2 kbp for UK48 and 6.5 kbp for UK76. However, the throughput of the R7.3 flow cells was poor: the single 48-hour run for UK48 produced only 7.1x coverage of the *B. pertussis* genome, and even the two more successful 48-hour UK76 runs only produced a total of 47.8x coverage. Such low yields were characteristic of early MinION flow cells; one 2015 study from Nature Methods reports a total yield of only 133.6 Mb from four sequencing runs (Loman, Quick and Simpson, 2015). In addition, the variation in performance observed between the UK48 and UK76 runs is also reflected in word-of-mouth reports from around the same period (Robison, 2017; Stack Exchange, 2017; van der Helm, 2017).

Attempts to *de novo* assemble the UK76 genome using the 47.8x coverage generated in the two R7.3 runs were unsuccessful. Five different assembly tools were tested: ABruijn, Miniasm, Canu, SMARTdenovo and SPAdes (Bankevich et al., 2012; Ruan, 2015; Li, 2016; Lin et al., 2016; Koren et al., 2017). ABruijn (now called Flye), Miniasm and Canu are still three of the most popular *de novo* assembly tools within the nanopore community and, as of 2019, are known to produce highly contiguous and usually accurate genome assemblies. Tasked with assembling the R7.3 UK76 reads, however, all three, plus SPAdes and SMARTdenovo, were unable to produce a single closed genome sequence, as shown in **Figure 1.16**. This inability to produce a closed sequence despite relatively high coverage in reads longer than the repetitive *B. pertussis* IS481 sequence was likely due to R7.3's high raw error rate, which was reportedly as high as 34% (Jain et al., 2015).

Ultimately, the 2015 pilot study showed that nanopore sequencing using the R7.3 flow cell chemistry was not suitable for the assembly of closed *B. pertussis* genome sequences: the low yield, unpredictability of flow cell performance and high raw error meant that multiple flow cells would be required to produce high enough coverage for accurate consensus assembly, which would be cost-prohibitive for most labs, particularly those running low-throughput sequencing.

A test of new nanopore sequencing chemistry and Mk 1b MinION proved *de novo* assembly of *B. pertussis* genomes was possible

In 2016, however, a new type of sequencing chemistry was released by ONT. R9.4 flow cells contain a new pore protein (the *E. coli* lipoprotein, CsgG), which was engineered to provide improved accuracy, thanks to an even shorter, more specific base-sensing region. In addition, improvements to the helicase enzyme allowed for higher sequencing speeds (up to 450 bases per second, compared to 70 bases per second for R7), thus greater yield per run (Oxford Nanopore Technologies, 2016; Clarke, 2018). The Mk1 MinION was retired and the Mk1b, which is adapted for R9.4 flow cells, was introduced. A new basecalling algorithm was also released, which used neural networks instead of Hidden Markov Models (HMM) and was reported to further improve the raw accuracy of reads produced by R9.4 flow cells.

In December 2016, another sample of UK76 gDNA was sent to the Nanopore Group at the University of California Santa Cruz, where it was sequenced using an R9.4 flow cell and Mk1b MinION. In 48

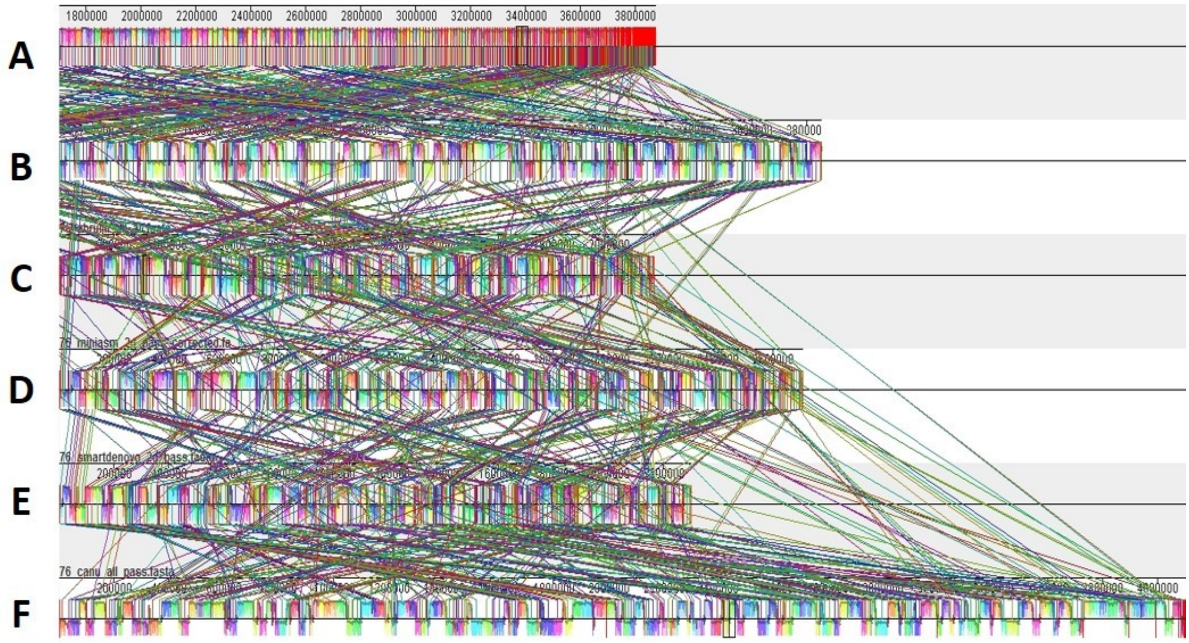


Figure 1.16 Attempts to assemble UK76 reads generated by R7.3 MinION flow cells were unsuccessful. Assembling Illumina short reads usually produces around 300, unresolved, contigs (A). Five assembly tools which had a long-read-only setting were tested with nanopore-generated UK76 long reads. ABruijn (B), Miniasm (C), SMARTdenovo (D), Canu (E) and SPAdes (F) all failed to assemble the nanopore long reads into single contigs. Additionally, each of the assemblers arranged the contigs in a different layout when aligned with progressiveMauve (Darling, Mau and Perna, 2010). Long-read-only assembly of R7.3 reads was therefore unsuited to investigation of *B. pertussis* genome arrangement.

hours, around 9.3 Gb of data was produced, equating to almost 2,500x coverage of the *B. pertussis* genome: nearly 50 times higher than the two R7.3 UK76 runs. Tests with Miniasm (the quickest *de novo* assembly tool) showed that this large R9.4 data set could be assembled into a single-contig, closed genome assembly. However, Miniasm produces low accuracy assemblies, because it does not incorporate any correction or polishing steps. Attempts to assemble the 9.3 Gb of UK76 data using Canu, which includes a thorough read correction step, indicated that assembly of such a large data set would take an impractical length of time, even on a high-performance computing (HPC) cluster. Alternatively, a subset of only one twentieth of the reads could also be assembled into a single contig, suggesting that each R9.4 flow cell produced enough data, of high enough accuracy, to sequence multiple strains together in the same sequencing run.

ONT produce several library preparation kits which allow for multiple samples to be sequenced together, using read barcoding. Each barcode is a short (tens of base pairs) sequence, which is ligated onto every DNA fragment in a DNA library. As shown in **Figure 1.17**, during sequence barcoding, also known as multiplexing, different barcodes are added to different samples, which can then be pooled and sequenced together, followed by computational demultiplexing (Smith et al., 2010). As of 2017, an increasing number of groups had shown the utility of barcoded nanopore sequencing to achieve closed bacterial genome sequences, usually in hybrid with short reads generated by Illumina sequencing (for example: Bayliss et al., 2017; Wick et al., 2017a). Thus, one of the first aims of this project was to test ONT's barcoding protocols with *B. pertussis*, and establish a workflow which allowed for the rapid assembly of closed genome sequences.

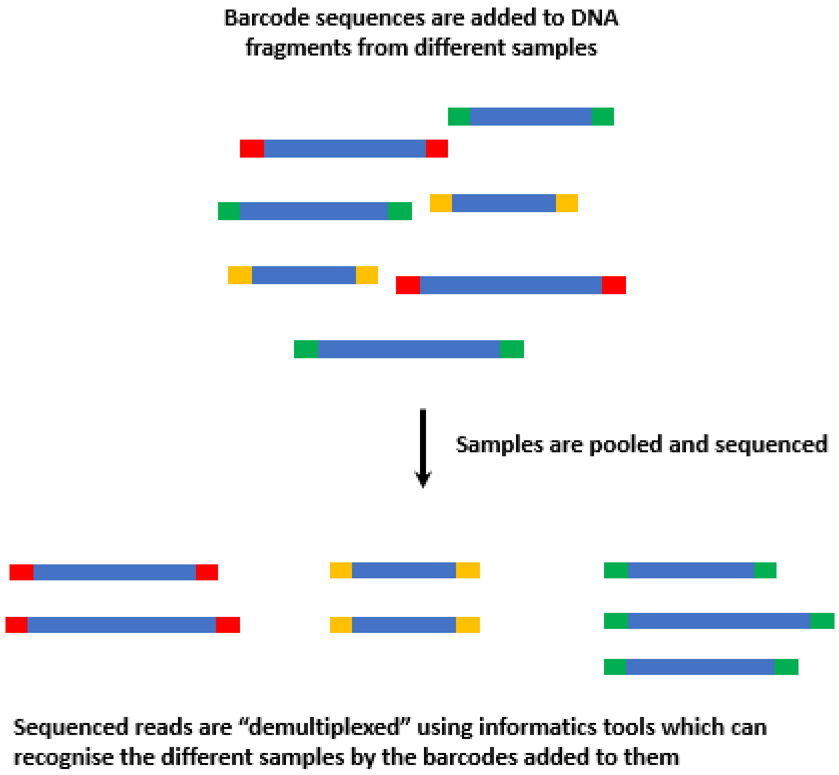

Barcode sequences are added to DNA fragments from different samples
Samples are pooled and sequenced
Sequenced reads are "demultiplexed" using informatics tools which can recognise the different samples by the barcodes added to them

Figure 1.17 Barcodes can be added to DNA strands of interest in order to sequence multiple samples using the same flow cell

1.3 *Aims of this project*

The *B. pertussis* genome requires long-read sequencing for full resolution and assembly of single-contig, closed genome assemblies. Closed genome sequences can reveal differences, such as rearrangements, between strains which otherwise appear to have highly similar genomes. Thus, closed genomes may help in the development of a deeper understanding of *B. pertussis* evolution which, in turn, could play a role in understanding the ongoing resurgence of whooping cough. Nanopore sequencing offers the opportunity to produce closed genomes in a high-throughput manner, at a cost which is affordable for smaller labs, who are not sequencing specialists. The overarching aim of this project was therefore to explore how nanopore sequencing, and long-read sequencing more generally, can be used to study *B. pertussis* genomes and begin to investigate some of the genomic phenomena which may only become apparent through the comparison of closed genome sequences.

The work in this thesis can be split into two separate, but related, project areas:

Chapters 2 and 3 detail the development and use of a workflow for whole genome sequencing and *de novo* genome assembly of *B. pertussis* isolates using multiplexed nanopore sequencing. The workflow is developed and optimised using a subset of five previously characterised isolates in Chapter 2 (Ring et al., 2018), then in Chapter 3 the workflow is adapted and used to investigate a set of 66 uncharacterised samples isolated in New Zealand between 1982 and 2018.

Chapters 4 and 5 explore the potential offered by long-read sequencing to identify and investigate genome-level inter-strain variations. In Chapter 4, an ultra-long genome duplication identified during workflow development (Chapter 2) is explored in more depth, including additional sequencing and investigation of potential phenotypes arising in affected strains (Abrahams et al., In review). In Chapter 5, all currently available *B. pertussis* genomes which have been closed using long-read sequencing are mined to examine any ongoing changes in a gene which codes for one of the antigens used in the acellular whooping cough vaccination but which has previously been excluded from allele typing due to its length and repetitive regions.

Chapter 2: Resolving the complex *Bordetella pertussis* genome using barcoded nanopore sequencing

The work presented in this chapter is published as a peer-reviewed publication:

“Resolving the complex *Bordetella pertussis* genome using barcoded nanopore sequencing”.

Natalie Ring, Jonathan S. Abrahams, Miten Jain, Hugh Olsen, Andrew Preston and Stefan Bagby.

Microbial Genomics, **4,** doi: 10.1099/mgen.0.000234

November 2018

A statement of authorship for this chapter can be found in the appendix: supplementary form **SF2.1**.

“Well, the thing about a black hole, its main distinguishing feature is it’s black. And the thing about space - the colour of space, your basic space colour - is it’s black. So how are you supposed to see them?”

- Holly, Red Dwarf

2.1 *Commentary text – preliminary tests and chapter summary*

The work in this chapter establishes a current optimal workflow for the use of barcoded nanopore sequencing to whole genome sequence *B. pertussis* strains. Prior to the main bulk of the chapter (the paper itself), several preliminary tests were conducted.

PCR barcoding vs native barcoding

Two methods exist for attaching barcode sequences to DNA fragments during nanopore sequencing library preparation. The first, "barcoding-by-PCR", uses specific PCR primers for each barcode, followed by several rounds of PCR. This method has the benefit of amplifying the starting mass of DNA, meaning a low yield during extraction can be overcome. Nonetheless, the use of PCR also comes with disadvantages; for example, certain sequence structures are less likely to be successfully amplified (e.g. highly repetitive or palindromic sequences), which introduces bias into the eventual sequencing library (Kanagawa, 2003). The second barcoding method, "native barcoding", enzymatically ligates barcode sequences onto the gDNA fragments in the same way sequencing adapters are ligated, without the need to introduce PCR bias. However, the obvious disadvantage of not using PCR is the need for a higher starting mass of DNA.

The "low input PCR barcoding" library preparation was trialled with five strains in June 2017 (those sequenced in the main body of the chapter, UK36, UK38, UK39, UK48 and UK76), using the standard protocol as per ONT's instructions (SQK-LWB001) and a starting mass of 50 ng per strain. Solid phase reversible immobilization (SPRI) beads are used several times throughout the library preparation. A common side-effect of SPRI clean-ups is gDNA loss and, during the 2015 pilot study, *B. pertussis* gDNA appeared to be particularly susceptible (Personal Communication, Stefan Bagby, 2017). The same issue was encountered during the PCR barcoding library preparation: after starting with 50 ng gDNA per strain, each strain retained at most 13.3 ng after the first SPRI clean-up step. The PCR amplification step of the preparation could have negated this loss; indeed, after the PCR step, yields of 70 ng and above were achieved for three of the five strains. However, the yields for the remaining two strains were much lower, at ~45 ng and ~15 ng each, meaning that these strains were underrepresented in the final sequencing library. After 48 hours of sequencing on an R9.4 MinION flow cell, the sequencing yield and mean read length were also low: 1.11 Gb in total, with a mean read length of 1,210 bp. In addition, 44% of all reads had no recognisable barcode attached. This meant that the highest coverage yielded for any strain was only 50x which, given nanopore sequencing's high raw error rate, was unlikely to be sufficient for *de novo* assembly of closed genome sequences, as seen in the 2015 pilot study. The native barcoding library preparation, which requires a starting mass of 1.5 µg gDNA, was therefore used for this study instead of PCR barcoding. Although the native barcoding protocol also requires SPRI clean-ups, the much higher starting mass of DNA means that more should be retained in the final sequencing library.

FFPE repair test

The native barcoding library preparation protocol (SQK-LSK108 with EXP-NBD103) includes an optional FFPE end-repair step. FFPE repair can improve sequencing yields, by repairing damaged fragments which would otherwise be unsuitable for adapter ligation. FFPE repair is recommended for samples which have been repeatedly freeze-thawed, or which are very old. The *B. pertussis* gDNA sequenced here was freshly extracted and had not been freeze-thawed; nonetheless, half of each strain's gDNA sample was repaired with NEBNext FFPE reagents (NEB) to test whether sequencing yield or raw accuracy could be improved by including this optional step prior to library preparation. However, predictably, neither the yield (n=5, paired t-test p=0.39) nor the raw accuracy (n=5, paired t-test p=0.937) was significantly improved for the five strains tested. For full statistics, see **Supplementary table S2.1**. The "FFPE-repaired" and "non-FFPE-repaired" reads for each strain were therefore pooled for all subsequent analysis detailed in the main body of this chapter.

Basecaller comparison

The natively barcoded sequencing run was originally basecalled using the basecalling tool built into the sequencing software, MinKNOW (vJune2017). The *de novo* assembly process was therefore first optimised using the MinKNOW-basecalled data. However, in mid-2017, a new standalone basecalling tool, Albacore, was made publicly available by ONT. Word-of-mouth feedback suggested that this new basecalling tool was significantly more accurate than the algorithm included in the MinKNOW software in June 2017. Therefore, the raw sequencing data files (fast5s) were re-basecalled with Albacore, and the MinKNOW and Albacore reads for each of the five sequenced strains were compared to an Illumina-only assembly for each strain, to compare raw accuracy. The mean identity for the MinKNOW reads was 2.46% lower than that of the Albacore reads, meaning Albacore produced significantly more accurate basecalls than MinKNOW for our five strains (n=5, paired t-test p<0.001). For full statistics, see **Supplementary table S2.1**. Consequently, the pipeline optimisation process was repeated using the Albacore-basecalled reads, as thoroughly detailed in the main body of this chapter.

Commentary text – chapter summary and context

Five *B. pertussis* strains from the 2012 UK whooping cough outbreak were sequenced using ONT's native barcoding library preparation and a single R9.4 MinION flow cell, and a wide variety of data analysis tools for basecalling, demultiplexing, pre-assembly read correction, *de novo* assembly and post-assembly polishing were trialled, in all possible combinations. Long-read-only strategies were also compared with hybrid (long nanopore reads plus short Illumina reads). The optimal long-read-only workflow available at the time of these trials (although newer and better tools are now available for almost every step) utilised Albacore basecalling, Porechop demultiplexing, pre-assembly read correction with Canu, *de novo* assembly with Flye and post-assembly polishing with Nanopolish. The optimal hybrid workflow replaced the *de novo* assembly and polishing steps: for hybrid assembly, these steps were both carried out using Unicycler. Four of the five strains sequenced were resolved into closed genomes; we discovered that the genome which remained unclosed contained a duplication of almost 200,000 bp, which resisted resolution.

2.2 *Abstract*

The genome of *Bordetella pertussis* is complex, with high G+C content and many repeats, each longer than 1000 bp. Long-read sequencing offers the opportunity to produce single-contig *B. pertussis* assemblies using sequencing reads which are longer than the repetitive sections, with the potential to reveal genomic features which were previously unobservable in multi-contig assemblies produced by short-read sequencing alone. We used an R9.4 MinION flow cell and barcoding to sequence five *B. pertussis* strains in a single sequencing run. We then trialled combinations of the many nanopore user community-built long-read analysis tools to establish the current optimal assembly pipeline for *B. pertussis* genome sequences. This pipeline produced closed genome sequences for four strains, allowing visualization of inter-strain genomic rearrangement. Read mapping to the Tohama I reference genome suggests that the remaining strain contains an ultra-long duplicated region (almost 200 kbp), which was not resolved by our pipeline; further investigation also revealed that a second strain that was seemingly resolved by our pipeline may contain an even longer duplication, albeit in a small subset of cells. We have therefore demonstrated the ability to resolve the structure of several *B. pertussis* strains per single barcoded nanopore flow cell, but the genomes with highest complexity (e.g. very large duplicated regions) remain only partially resolved using the standard library preparation and will require an alternative library preparation method. For full strain characterization, we recommend hybrid assembly of long and short reads together; for comparison of genome arrangement, assembly using long reads alone is sufficient.

2.3 *Data Summary*

1. Final sequence read files (fastq) for all five strains have been deposited in the Sequence Read Archive, BioProject PRJNA478201, accession numbers SAMN09500966, SAMN09500967, SAMN09500968, SAMN09500969, SAMN09500970.

2. A full list of accession numbers for Illumina sequence reads is available in **Supplementary table S2.1** (https://figshare.com/s/003465e08ba1e8fc8780).

3. Assembly tests, basecalled read sets and reference materials are available from figshare: https://figshare.com/projects/Resolving_the_complex_Bordetella_pertussis_genome_using_barcoded_nanopore_sequencing/31313 .

4. Genome sequences for *B. pertussis* strains UK36, UK38, UK39, UK48 and UK76 have been deposited in GenBank, accession numbers: CP031289, CP031112, CP031113, QRAX00000000, CP031114.

5. Source code and full commands used are available from Github: https://github.com/nataliering/Resolving-the-complex-Bordetella-pertussis-genome-using-barcoded-nanopore-sequencing .

2.4 *Impact Statement*

Over the past two decades, whole genome sequencing has allowed us to understand microbial pathogenicity and evolution to an unprecedented degree. However, repetitive regions, like those found throughout the *Bordetella pertussis* genome, have confounded our ability to resolve complex genomes using short-read sequencing technologies alone. We have used nanopore sequencing, which can generate reads longer than these problematic repetitive regions, to resolve multiple *B. pertussis* genomes with a single flow cell. The resolved genomes can be used to visualize previously predicted genome rearrangements and, in addition, the inability of our long reads to resolve some of our genomes has allowed us to infer the presence of previously unidentified ultra-long duplications in two of our five strains. Thus, our findings point towards unanticipated genome-level genetic variation in strains which appear otherwise monomorphic at the nucleotide level. This work expands the recently emergent theme that even the most complex genomes can be resolved with sufficiently long sequencing reads. Our optimization process, moreover, shows that the analysis tools currently favoured by the sequencing community do not necessarily produce the most accurate assemblies for all organisms; pipeline optimization may therefore be beneficial in studies of unusually complex genomes.

2.5 *Introduction*

Bordetella pertussis is the pathogenic bacterium which causes most cases of whooping cough (pertussis). Pertussis was a major medical burden prior to the international introduction of vaccination in the 1940s and 1950s. Widespread vaccine uptake greatly reduced incidence of the disease in developed countries. Original whole-cell vaccines were replaced by new acellular vaccines throughout the 1990s and early 2000s. The acellular vaccines contain one to five of the *B. pertussis* protein antigens pertactin (PRN), pertussis toxin (PT), filamentous haemagglutinin (FHA), and the fimbrial proteins Fim2 and Fim3. Despite continued high levels of pertussis vaccination coverage, since the early 1990s the number of cases of whooping cough has increased in many countries (Burns, Meade and Messionnier, 2014; Jakinovich and Sood, 2014).

Suggested causes for this resurgence include improved diagnostic tests and awareness, waning immunity as a result of the switch to acellular vaccination, and genetic divergence of circulating B. pertussis from the vaccine strains due to vaccination-induced selection pressure (Ausiello and Cassone, 2014; Clark, 2014; Sealey, 2015). A global survey of strains from the pre-vaccine, whole-cell vaccine and acellular vaccine eras showed that the genome of *B. pertussis*, widely regarded as a monomorphic and slowly evolving organism, has been evolving since the introduction of the whole-cell vaccine (Bart, Harris, et al., 2014). Analysis of strains from several recent epidemics showed the rate of evolution of the genes encoding vaccine antigens has increased since the switch to the acellular vaccine (Octavia et al., 2012; Bowden et al., 2014; Lam et al., 2014; Sealey et al., 2015).

The *B. pertussis* genome contains up to 300 copies of a 1053 bp insertion sequence (IS), IS *481*. A smaller number of copies of IS *1002* (1040 bp) and IS *1663* (1014 bp) contribute further complexity to the genome. These regions of repetition mean that assembly of closed, single-contig *B. pertussis* genomes using short-read sequencing, which produces reads shorter than the IS repeats, has been particularly difficult: most genome sequences available on NCBI comprise several hundred contigs, or at least one contig per IS copy. Over the last decade, many studies have shown that reads longer than the longest repeat are required to resolve regions of high complexity (Chin et al., 2013; Conlan et al., 2014; Koren and Phillippy, 2015; Loman, Quick and Simpson, 2015; Wick et al., 2017a; Jain, Koren, et al., 2018; Jain, Olsen, et al., 2018; Schmid et al., 2018). Assembly of closed genomes may reveal genomic features which were previously unobservable in multi-contig assemblies; this is particularly true for genomes which contain a high number of IS copies, as insertion sequences are known to impact genomic structure via rearrangement, deletion and, more rarely, duplication (Bentley et al., 2008; Siguier et al., 2014).

In 2016, Bowden *et al*. (2016) were the first to use long reads, together with Illumina short reads, to conduct a survey of *B. pertussis* strains which had circulated during two whooping cough epidemics, in the USA, in 2010 and 2012. Assembling closed genomes for these epidemic isolates revealed extensive genomic arrangement differences between isolates which appeared to be otherwise closely related. Bowden *et al*. concluded that further comprehensive whole genome studies are required to fully understand the international resurgence of whooping cough. More recently, Weigand *et al*. showed that the *B. pertussis* genome continues to undergo structural rearrangement on a frequent basis, usually mediated by IS *481* (Weigand et al., 2017). As well as causing structural rearrangement, IS elements have also repeatedly been shown to be responsible for the ongoing reduction of the *B. pertussis* genome via gene deletion (Parkhill et al., 2003; Preston, A., Parkhill and Maskell, 2004; Caro et al., 2006; Heikkinen et al., 2007).

Bowden *et al*. and Weigand *et al*. both used Pacific Biosciences (PacBio) long read sequencing, which has high start-up costs, and lacks the portability needed for on-the-ground epidemic surveillance. In contrast, Oxford Nanopore Technology (ONT)'s MinION nanopore sequencer has relatively low start-up costs. Recent improvements to flow cell yield and the introduction of barcoded library preparation make per-sample MinION costs comparable to those of PacBio or Illumina (Bayliss et al., 2017; Ton et al., 2017; Wick et al., 2017a). In addition, the pocket-sized MinION sequencer is portable, enabling in-the-field sequencing (Edwards et al., 2016; Quick et al., 2016; Pomerantz et al., 2017).

Here we test the ability of barcoded nanopore sequencing, together with a variety of available data analysis tools, to resolve the genomes of five *B. pertussis* strains from a UK epidemic, which were previously unclosed and comprised many contigs assembled using short reads sequenced with Illumina's MiSeq (Sealey et al., 2015). We then briefly investigate the resulting genomes to identify any previously unobserved features, with a particular focus on the genome of one strain which remained unresolved by our hybrid assembly strategy.

2.6 *Methods*

Full method and bioinformatics procedures are described at:

https://github.com/nataliering/Resolving-the-complex-Bordetella-pertussis-genome-using-barcoded-nanopore-sequencing.

All data analysis was carried out using the Medical Research Council's Cloud Infrastructure for Microbial Bioinformatics (CLIMB) (Connor et al., 2016).

Strain isolation and Illumina sequencing

Five strains originally isolated during the UK 2012 whooping cough epidemic were obtained from the National Reference Laboratory, Respiratory and Vaccine Preventable Bacteria Reference Unit, at Public Health England. Short-read sequencing data were generated previously, using genomic DNA (gDNA) extracted using a DNeasy Blood and Tissue kit (Qiagen), multiplex library preparation and Illumina sequencing (Sealey et al., 2015). Full details, including accession numbers, are included in **Supplementary table S2.1**.

DNA extraction

Strains were stored at −80°C in PBS/20 % glycerol at the University of Bath. They were grown on charcoal agar plates (Oxoid) for 72 h at 37°C. All cells were harvested from each plate and resuspended in 3 ml PBS. The optical density of each cell suspension was measured at 600 nm, and volumes of suspension equating to an OD of 1.0 (~2×10^9 B. pertussis cells) in 180 µl were pelleted in a microcentrifuge for 2 min at 12 000 g. gDNA was extracted from each pellet using a GenElute bacterial genomic DNA kit (Sigma Aldrich) according to the manufacturer’s instructions, including the optional RNAase A step and a two-step elution into 200 µl elution buffer (10 mM Tris-HCl, 0.5 mM EDTA, pH 9.0). QuBit fluorometry (dsDNA HS kit; Invitrogen) was used to measure gDNA concentration, and Nanodrop spectrometry (ThermoFisher Scientific) was used to assess gDNA purity.

Nanopore library preparation and MinION sequencing

In total, 1.5 µg of gDNA per strain was concentrated using a 2.5× SPRI clean-up (AMPure XP beads; Beckman Coulter), eluting into 50 µl of nuclease-free (NF) water (Ambion). Then, 48 µl of this was sheared to 20 kbp using g-tubes (Covaris), according to the manufacturer’s instructions.

Sequencing libraries were prepared for all samples using ONT's 1D ligation sequencing kit (SQK-LSK108) with native barcoding (EXP-NBD103), according to the manufacturer’s instructions. Ten barcodes were used, two per strain (see **Supplementary table S2.2** for full details). After library preparation, different volumes of samples were combined to produce an equi-mass pool for eight samples; two samples had much lower concentration after library preparation so were pooled in full. A total mass of 712.5 ng was pooled in 208.9 µl NF water, which was concentrated to 50 µl by 2.5× SPRI clean-up. Full details of mass pooled per sample are given in **Supplementary table S2.2**. This pooled library (712.5 ng in 50 µl) was used for sequencing adapter ligation.

The final sequencing library was loaded onto an R9.4 flow cell and sequenced for 48 h using a MinION MK1b device with MinKNOW sequencing software (protocol NC_48h_Sequencing_Run_FLOMIN106_SQK-LSK108).

Additional basecalling and demultiplexing

The fast5 files were basecalled using ONT’s Albacore (v2.1.3) program, with barcode binning. As suggested by Wick *et al*. (2017a), Porechop was then used to demultiplex the Albacore reads, keeping only those for which Albacore and Porechop agreed on the bin. The Albacore+Porechop fastq files were deposited in the Sequence Read Archive (SRA) with accession codes SAMN09500966 to SAMN0900970. Full details of all read sets (including reads output by MinKNOW’s concurrent basecalling algorithm) are given in **Supplementary table S2.1**.

Assembly of short-read-only drafts

Assuming the available Illumina data to have typically low error, short-read-only genome sequences were assembled for each strain using ABySS (v2.0.3) (Simpson et al., 2009). Prior to assembly, reads were prepared using Trimmomatic (v0.34) (Bolger, Lohse and Usadel, 2014), which trimmed the first

10 bases of each read, and discarded any reads whose four-base sliding-window q-score fell below 32. These assemblies had low contiguity, but theoretically high accuracy.

Comparison of raw reads

A shell script, 'summary_stats', was used to give the total number of reads, mass sequenced and minimum, maximum, mean and median read lengths for each set of raw reads. Summary_stats uses seq_length.py (Gummy-Bear, 2014) and all_stats. All are available from https://github.com/nataliering/Resolving-the-complex-Bordetella-pertussis-genome-using-barcoded-nanopore-sequencing.

Raw percentage identity was estimated by comparing each read set to the B. pertussis reference genome (Tohama I, NC_002929.2). As the UK 2012 strains were not expected to be identical to the Tohama I sequence, read error was also estimated by comparison with the respective Illumina-only assemblies. The comparison was conducted using BWA MEM (Li, 2013) and SAMtools stats (Li et al., 2009), which produces a long output file including 'error rate' [% identity was calculated from this: 100 – (error rate*100)]. Raw_error (https://github.com/nataliering/Resolving-the-complex-Bordetella-pertussis-genome-using-barcoded-nanopore-sequencing/blob/master/raw_error) produces a stats file using this method, given a read set and reference genome. Using the same BWA MEM output, raw read coverage of the Tohama I reference genome was checked using SAMtools depth and visualization with a rolling window in R.

Finally, raw G+C content was calculated using GC_calculator, which outputs the percentage G+C content of a given fasta file (https://github.com/nataliering/Resolving-the-complex-Bordetella-pertussis-genome-using-barcoded-nanopore-sequencing/blob/master/GC_calculator).

Assembly tool testing – nanopore only

The Albacore+Porechop reads for one strain, UK36, were used to test a variety of de novo assembly strategies. Four community-built assembly tools were trialled: ABruijn (now called Flye, v1.0 and v2.3.2 respectively), Canu (v1.5), Miniasm with Minimap/Minimap2 (v0.2-r128, v0.2-r123 and v2.0-r299-dirty, respectively) and Unicycler (v0.4.4) (Li, 2016; Lin et al., 2016; Koren et al., 2017; Wick et al., 2017b; Li, 2018).

Canu has a standalone option to conduct pre-assembly read correction. This was used to correct the 359× coverage UK36 read set to 40× coverage of more accurate reads. Each assembly tool was then trialled with and without pre-assembly read correction. As Canu's read correction step is relatively time-consuming as regards CPU, an alternative was also trialled. Filtlong (https://github.com/rrwick/Filtlong) does not correct reads, but produces read sets comprising the longest and most accurate reads, up to a given level of coverage; 40× and 100× coverage were trialled here.

Finally, Racon (v.1.2.0) was tested to determine whether the draft assemblies could be improved by post-assembly polishing (Vaser et al., 2017). After each Racon polish, the accuracy of the assembly produced was estimated. If an improvement was observed, another round of polishing was conducted, up to a total of five rounds. Once two successive rounds of polishing showed no further improvement, no further Racon polishes were conducted. For Unicycler, no manual Racon polishes were conducted, because Racon polishing is part of the Unicycler assembly process. After Racon polishing, each assembly was further polished with a single round of Nanopolish (v0.9.0) (Loman, Quick and Simpson, 2015).

Testing exhaustive combinations of each of these steps produced 28 draft assemblies for each of the four assembly tools tested (ABruijn/Flye, Canu, Miniasm and Unicycler), a total of 112 draft UK36 assemblies (see **Supplementary table S2.3** for all combinations).

Assembly tool testing – hybrid

As Illumina reads were already available for the strains sequenced here, a variety of hybrid de novo assembly strategies were also tested. Using Pilon (v1.22), the best nanopore-only assembly for each of the assembly tools was polished with the Illumina reads, up to a total of five rounds (Walker et al., 2014). In addition, a hybrid assembly was produced using Unicycler's hybrid mode, which both combines read sets for assembly, and conducts several rounds of Racon and Pilon polishes automatically. Finally, the hybrid assembly mode of SPAdes (v3.12.0) was tested. These hybrid tests produced another 22 draft assemblies (**Supplementary table S2.4**) (Bankevich et al., 2012).

Assessing assembly accuracy

Summary_stats was used to determine the number of contigs, and contig length for each draft assembly. The percentage identity of each draft compared to the Illumina-only draft was estimated using a method developed by Wick *et al.* (2018a). Their chop_up_assembly.py and read_length_identity.py scripts were used to generate percentage identity values for 10 kbp sections along the entirety of each assembly, and a custom shell script, assembly_identity (https://github.com/nataliering/Resolving-the-complex-Bordetella-pertussis-genome-using-barcoded-nanopore-sequencing/blob/master/assembly_identity), was used to calculate the mean percentage identity of the whole.

Quality metrics for each assembly were produced using Quast (v4.5) and BUSCO (v1.22) (Gurevich et al., 2013; Simão et al., 2015). In addition, a method developed by Watson, Ideel (https://github.com/mw55309/ideel), was used to assess the effect of any erroneous indels in the final UK36 hybrid assembly (Watson, 2018).

Comparing genome arrangement

After the best nanopore-only and hybrid assembly pipelines were identified for UK36, the pipelines were used to produce draft assemblies for the remaining four strains. The hybrid assembly for each strain was annotated with Prokka (v1.12) using the proteins from Tohama I as a reference (Seemann, 2014b). The genomes were also submitted to GenBank (accession numbers CP031289, CP031112, CP031113, QRAX00000000 and CP031114).

The arrangement of each nanopore-only assembly was compared to that of each hybrid using progressiveMauve (v20150226 build 10) (Darling, Mau and Perna, 2010). Finally, the nanopore-only assemblies for each strain were compared to each other, also using progressiveMauve. Prior to these alignments, each draft was manually rearranged so that the first gene after the *B. pertussis* origin of replication, *gidA,* was at the beginning of the sequence. gidA_blast (https://github.com/nataliering/Resolving-the-complex-Bordetella-pertussis-genome-using-barcoded-nanopore-sequencing/blob/master/gidA_blast) locates the *gidA* sequence in the draft to enable manual rearrangement. Later, this same process was used to identify IS element copies in the assembled genomes. If a tool assembled the complementary strand instead of the template (as identified by the results of gidA_blast), a reverse complement of the draft sequence was generated using reverse_complement (https://github.com/nataliering/Resolving-the-complex-Bordetella-pertussis-genome-using-barcoded-nanopore-sequencing/blob/master/reverse_complement).

2.7 *Results*

Sequencing yield

During the 48 h MinION sequencing run, 1 803 648 reads were generated, equating to 9.73 Gbp of sequence. In total, 28.78 % of these reads (574 053 reads, 2.8 Gbp) were not assigned to the correct barcode bin during demultiplexing, leaving 6.93 Gbp (1 229 595 reads) of useable sequencing data (**Figure 2.1**). Normalized yield per barcode (taking into account nanograms of gDNA included in the pooled sequencing library) was particularly high for one barcode (NB11, 15.28 Mbp ng^{-1}) but otherwise relatively consistent, ranging from 7.38 to 10.28 Mbp ng^{-1} with a mean yield of 9.06 Mbp ng^{-1} (se=0.37). Mean read length for the full read set was 5689 bp. Read lengths ranged from 4 to 201 977 bp.

The Albacore-demultiplexed reads were re-demultiplexed using Porechop, which keeps only the reads for which both tools agree on the barcode identified. This additional step resulted in a small but

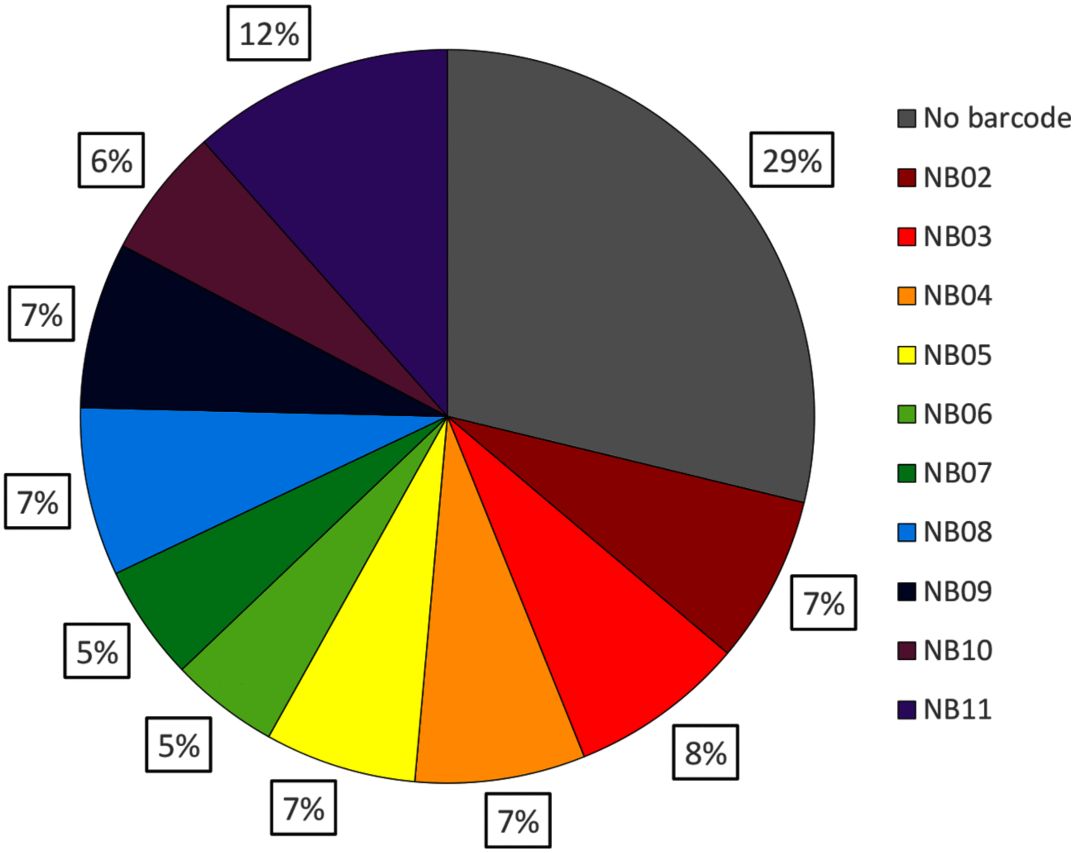


Figure 2.1 The Albacore+Porechop reads were used to assess barcode distribution. This showed that a large portion of the raw reads was placed into the ‘no barcode’ bin, meaning Albacore and Porechop either did not agree on a barcode, or no recognizable barcode was present. Otherwise, the barcodes were largely well distributed.

significant improvement in identity compared to Illumina-only assembly: 82.43 to 82.52 % (n=5, paired t-test P<0.001). Consequently, the reads used for data pipeline testing were those that had been basecalled and demultiplexed by Albacore, followed by Porechop re-demultiplexing.

For full results, including which barcode was assigned to each sample, see **Supplementary table S2.1**.

Assembly tool testing – nanopore-only

Table 2.1 shows the quality measurements for the best nanopore-only assembly per tool trialled. All tools tested were able to resolve the nanopore long reads for UK36 into a complete, closed contig, using default assembly options with no manual intervention. In total, 112 different tool combinations were trialled. Alignment of drafts from different tools using progressiveMauve revealed that each tool also assembled the genome into the same arrangement. However, the length of the draft assemblies showed some variation: 3.984 to 4.134 Mbp, with a mean length of 4.108 Mbp.

Comparing like-for-like assemblies before and after polishing shows that Nanopolish improves identity by 0.216 % on average (n=16, paired t-test P<0.001). Polishing with Racon produced inconsistent results: identity decreased after Racon polishing of ABruijn and Flye drafts, increased by 2.01 % after the optimal number of polishing rounds for pre-corrected non-ABruijn/Flye drafts (n=3), and increased by 15.15 % after optimal rounds for non-ABruijn/Flye drafts with no pre-correction (n=4). The mean number of Racon polishes required to reach optimal percentage identity (after which it began to decrease) was 4.75 (n=7).

The assembly with greatest percentage identity compared to the Illumina-only draft (99.59 %) combined pre-assembly read correction with Canu, assembly with ABruijn and post-assembly polishing with Nanopolish. The assemblies were also assessed using BUSCO, which searches draft assemblies for copies of benchmarking universal single-copy orthologues (BUSCOs). BUSCOs are sets of core genes which are likely to appear universally in related organisms. A set of 40 such core genes from the *Escherichia coli* genome are used as the gram-negative bacterial BUSCOs; if a genome has been assembled accurately, the tool BUSCO is more likely to be able to identify these 40 genes within its sequence. Of the drafts assessed here, the ABruijn assembly contained the highest number of identifiable BUSCOs (37 full and two partial, of the full set of 40; see **Table 2.1** for full results).

Assembly tool testing – hybrid

A number of hybrid assembly strategies were trialled, including polishing a long-read assembly with short reads, scaffolding short-read contigs with long reads, and using both short and long reads together during assembly (**Table 2.2** shows the best draft produced by each tool). Scaffolding short-read contigs with long reads using SPAdes produced one of the highest accuracy assemblies (99.68 %), but did not fully resolve the genome, as six contigs remained. No further polishing was attempted with this SPAdes assembly, as polishing would not close the remaining gaps between the contigs.

The best hybrid assemblies per tool were significantly more accurate than the best nanopore-only assemblies per tool, with a mean identity improvement of 0.11 % (hybrid n=6, nanopore-only n=5, paired t-test P<0.001). In addition, all hybrids contained all 40 identifiable BUSCOs, and all except the SPAdes hybrid were single closed contigs and showed the same arrangement when aligned using progressiveMauve. The best single-contig hybrid assembly, with 99.68 % identity, was produced using Unicycler’s hybrid option, which uses SPAdes, Minimap, Miniasm, Racon and Pilon. **Supplementary table S2.5** shows the results from all nanopore-only and hybrid tests.

Table 2.1 Best *de novo* assembly options and quality measurements for nanopore-only assemblies

Assembler	Pre-assembly read correction	Pre-assembly read filtering / x coverage	Rounds of Racon polishing	Polishing with Nanopolish	Contigs	Assembly length / Mbp	% identity compared to Illumina-only	BUSCOs present/fragment/ missing (of 40)
ABruijn	Yes	No	0	Yes	1	4.105	99.59	37/2/1
Canu	Yes	No	4	Yes	1	4.133	99.54	36/1/3
Flye	Yes	No	0	Yes	1	4.108	99.56	35/3/2
Miniasm + Minimap	Yes	No	5	Yes	1	4.111	99.55	37/0/3
Unicycler	Yes	No	8*	Yes	1	4.107	99.55	35/2/3

*The rounds of Racon listed for Unicycler were carried out as part of the Unicycler protocol; no manual rounds of Racon were conducted

Table 2.2 Best *de novo* assembly options and quality measurements for hybrid assemblies

Assembler	Pre-assembly read correction	Pre-assembly read filtering / x coverage	Assembly includes short reads?	Rounds of Racon polishing	Polishing with Nanopolish	Rounds of Pilon polishing	Contigs	Assembly length / Mbp	% identity compared to Illumina-only	BUSCOs present/fragment/ missing (of 40)
ABruijn	Yes	No	No	0	Yes	3	1	4.109	99.67	40/0/0
Canu	Yes	No	No	4	Yes	3	1	4.130	99.66	40/0/0
Flye	Yes	No	No	0	Yes	3	1	4.108	99.67	40/0/0
Miniasm + Minimap	No	No	No	5	Yes	4	1	4.107	99.66	40/0/0
SPAdes	Yes	No	Yes	n/a	n/a	n/a	6	4.105	99.68	40/0/0
Unicycler	Yes	No	Yes	4*	No	8*	1	4.107	99.68	40/0/0

*The rounds of Racon and Pilon listed for Unicycler were carried out as part of the Unicycler protocol; no manual rounds of polishing were conducted for this assembly

Assembly and annotation of all strains

Using the nanopore-only and hybrid pipelines defined through the tests described here (**Figure 2.2**), draft genomes were assembled for all five UK strains sequenced during our barcoded run. The assemblies were assessed for percentage identity compared to each strain's Illumina-only assembly, G+C content, genome length and number of key IS element features; they were also annotated using Prokka. The full results of this analysis are shown in **Tables 2.3**, **S2.6** and **S2.7**.

The hybrid assembly for one strain, UK76, had slightly lower percentage identity (99.54 %) than the other strains, each compared to their respective Illumina-only ABySS assembly. Discounting UK76, the assemblies had a mean identity of 99.69 % (n=4). The G+C content of the strains varied little: the content for all strains was 67.70 % when rounded to two decimal places. The number of genes predicted by Prokka was also relatively consistent, varying from 3757 to 3804.

The UK36 proteins predicted by Prokka were assessed by Ideel, which searched the Trembl database for similar proteins (Bairoch and Apweiler, 2000). The length of the Prokka-predicted proteins was divided by those of the identified similar Trembl proteins; a perfect match would equal 1.0. This method, therefore, indicates whether indels in a draft sequence cause frameshifts which subsequently lead to truncated (or over-long) protein prediction. After manual curation to remove results which represented genes known to be fully present in other *Bordetella* species but truncated in *B. pertussis*, over 98 % of Prokka-predicted genes had a Prokka:Trembl length ratio of greater than 0.9. This suggests that the residual error in the hybrid assemblies does not cause substantial annotation problems, so the hybrid assemblies for all five strains were submitted to GenBank (accession numbers CP031289, CP031112, CP031113, QRAX00000000 and CP031114).

Comparison of genomic structure of all strains

All strains were assembled into single contigs using the nanopore-only pipeline. These assemblies were aligned using progressiveMauve (**Figure 2.3**), displaying genomic rearrangement between strains; three, UK36, UK38 and UK39, shared exactly the same arrangement, whilst UK48 and UK76 were rearranged.

Of the hybrid assemblies, two strains, UK48 and UK76, had longer genomes than the others (4.112 and 4.113 Mbp, respectively, compared to 4.108 Mbp), which corresponds with them also having more copies of the most abundant IS element, IS *481*. All strains but one were assembled into single contigs. The remaining strain, UK48, was assembled into five contigs (N50=3.934 Mbp). Of these, three were shorter than 500 bp, and were subsequently discarded. The remaining two contigs were 3 934 355 and 178 023 bp. Mapping the raw UK48 reads to the Tohama I reference sequence revealed a section of almost 200 kbp, located between 1.35 and 1.53 Mbp, which had double the read depth of the rest of the reference; the doubled read depth suggests that this section of the genome is duplicated in UK48. No other strain had a similarly duplicated section, although the coverage of UK76 was enriched by around 25 % at the same locus (**Figure 2.4**), potentially indicating a heterogeneous UK76 population, of which a subset (i.e. 25 %) of cells carries a duplication. These abnormalities are also present in the Illumina reads, which were obtained approximately 5 years before our nanopore reads (**Figure 2.4**), but had not been identified previously using the short reads alone.

Table 2.3 Assembly statistics for five UK *B. pertussis* strains, assembled using our hybrid pipeline

Pipeline	Strain	Contigs	Genome length / Mbp	GC content / %	% identity compared to Illumina-only	# genes predicted	IS481 copies	IS1002 copies	IS1663 copies
NANOPORE-ONLY	UK36	1	4.108	67.69	99.47	4698	258	8	17
	UK38	1	4.108	67.69	99.49	4741	258	8	17
	UK39	1	4.109	67.70	99.48	4588	258	8	17
	UK48	1	4.114	67.70	99.47	4610	262	8	17
	UK76	1	4.113	67.70	99.32	4608	262	8	17
HYBRID	UK36	1	4.107	67.70	99.68	3757	258	8	17
	UK38	1	4.108	67.70	99.69	3757	258	8	17
	UK39	1	4.108	67.70	99.69	3804	258	8	17
	UK48	2	4.112	67.70	99.68	3763	262	8	17
	UK76	1	4.113	67.70	99.54	3753	262	8	17

2.8 *Discussion*

Are residual unresolved ultra-long repeats present in some strains?

Our primary aim in this study was to determine whether long reads produced by nanopore sequencing using ONT's MinION can be used to produce closed *B. pertussis* genome sequences, which will enable visualization of large-scale inter-strain genomic differences, and may further reveal previously hidden genomic features. Our nanopore-only assembly pipeline produced closed-contig assemblies for all five strains sequenced here, allowing visualization and validation of previously predicted IS-mediated genomic rearrangements. In addition, the inability of our long reads to produce a closed hybrid assembly for UK48 has revealed a separate, unpredicted, genomic feature in our UK strains.

The region of enriched coverage between 1.35 and 1.53 Mbp in the Tohama I reference genome observed in the UK48 reads (**Figure 2.4**) is likely to indicate a large (almost 200 kbp) duplication of that region which is present in UK48 but not in the reference. A less obvious duplication may also be present in the genome of UK76: a 300 kbp region from 1.38 to 1.68 Mbp shows 125 % coverage. The presence of the same abnormalities in other read sets for both strains suggests that they have not been caused by contamination (**Figure 2.4**). Similar duplicated regions have been observed previously in a very small number of French and Finnish strains (fewer than five) through microarray-based studies in 2006 and 2007 (Caro et al., 2006; Heikkinen et al., 2007). More recently, Weigand *et al.* (2016; 2018) noted complex duplications in two US strains and two Indian vaccine-reference strains; these genomes were long-read sequenced with PacBio, but resolution of the duplications was only possible with optical mapping. The locus found to be duplicated in these previous studies was the same as that we predict is duplicated in UK48 and UK76; however, at 180 and 300 kbp, our predicted duplications are longer than any of those observed previously. The identification of two additional strains carrying a duplication of the same region suggests that IS-mediated duplication is occurring

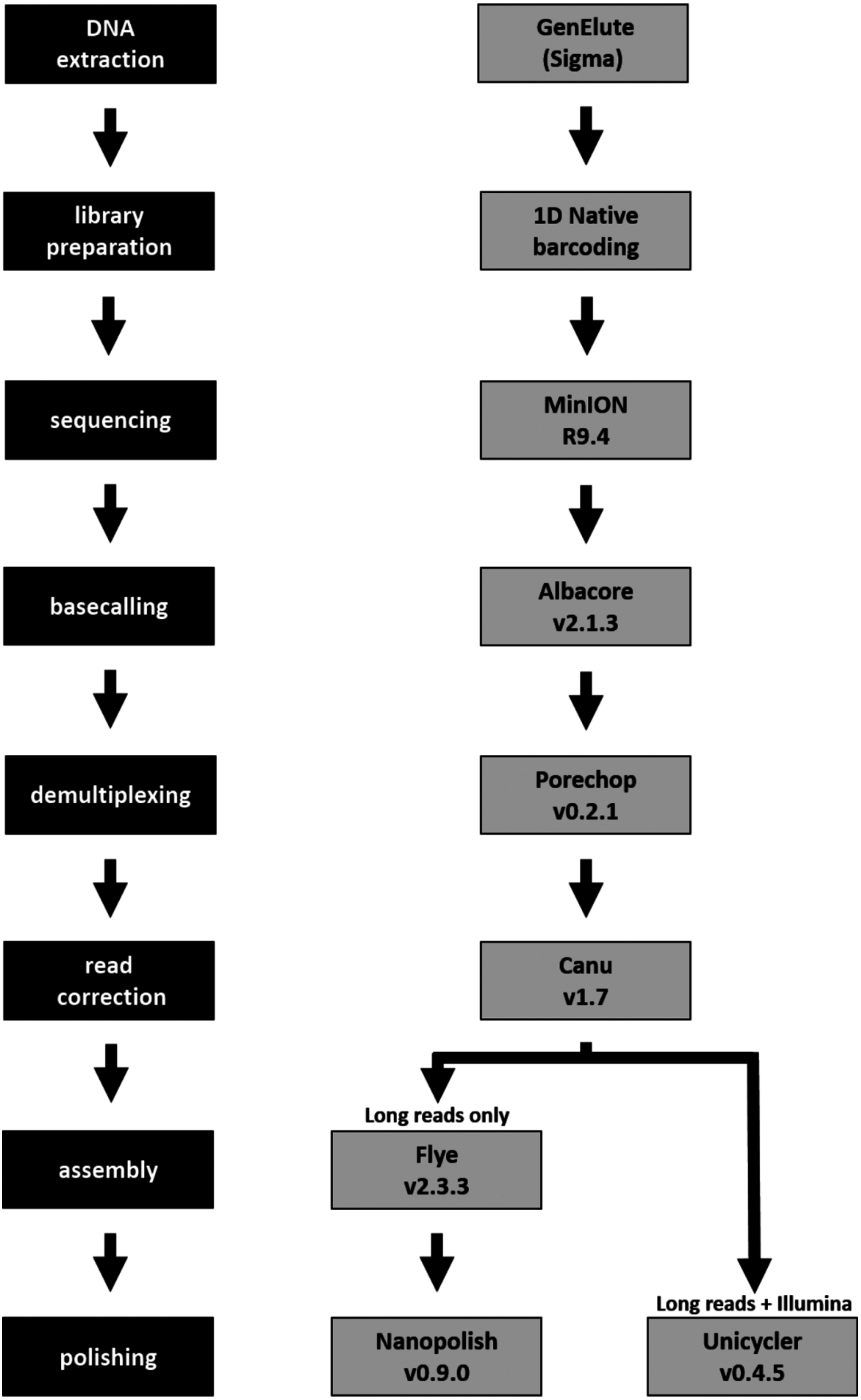


Figure 2.2 Our nanopore-only and hybrid sequencing pipelines, developed through extensive testing of available tools.

more frequently in *B. pertussis* than previously believed. Furthermore, the apparent heterogeneity of our UK76 culture suggests that only a portion of UK76 cells may carry the duplication, a phenomenon previously unobserved in any duplication-carrying B. pertussis isolate. Finally, the locus of the duplication itself, which contains many motility-related genes, may have interesting implications for an organism traditionally described as non-motile.

Neither our nanopore-only pipeline using Flye nor our hybrid pipeline using Unicycler was able to resolve the duplication correctly, however. The nanopore-only pipeline produced closed contigs for all five strains, seemingly missing the duplication completely, whilst the hybrid pipeline produced a multi-contig assembly for UK48 and the same closed contig as the nanopore-only pipeline for UK76. Our UK48 reads ranged from 73 to 108 575 bp with a mean length of 6243 bp, whilst the UK76 reads ranged from 4 to 70 486 bp with a mean length of 5480 bp; if the key to resolving long repeats is to use reads longer than the longest repeat, we will need high coverage of ultra-long reads in the order of hundreds of thousands of bases to resolve these putative duplications via sequencing (Jain, Koren, et al., 2018; Schmid et al., 2018). Nanopore sequencing is currently the only sequencing method theoretically capable of producing such long reads; methods to obtain ultra-long reads are under development by the nanopore community, with reports of reads in the order of millions of bases (Payne et al., 2018a).

Accuracy of long-read sequencing is improving but error estimation is challenging

In addition to our primary aim, we also compared a variety of de novo assembly strategies to determine the current optimal pipeline for producing the most accurate genome sequences for *B. pertussis*.

Without a recent, closely related reference sequence, error estimation in *B. pertussis* assemblies is inexact. Comparison with the Tohama I reference sequence will identify basecalling errors which are false positives, having arisen due to natural variation between different strains (that is, true SNPs will be identified as errors). Moreover, the validity of Tohama I as a representative of all B. pertussis strains is questionable (Caro, Bouchez and Guiso, 2008). The Illumina reads available for four of our sequenced strains (UK36, 38, 39 and 48) showed 98.44 % identity with the Tohama I sequence, suggesting natural genetic variation between Tohama I and these UK strains of around 1.5 %. The false positive rate is thus around 1.5 % when using Tohama I to assess assembly accuracy. On the other hand, comparison with Illumina-only assemblies requires short read data to be available, and assumes the Illumina reads to be close to 100 % accurate, which could be a flawed assumption. The Illumina reads for UK76, for example, had raw identity of only 87.32 % compared to Tohama I. With no distinctive features noted for UK76 in our assembly or in the original comparison of UK epidemic strains, it is unlikely that the UK76 genome is truly 11% less like Tohama I than the other strains sequenced here (Sealey et al., 2015). It seems more likely that the Illumina reads are inaccurate; if this is the case, our assessments of the accuracy of our UK76 assemblies were skewed. This could explain why our UK76 hybrid assembly had a slightly lower estimated accuracy than the other strains. Compared to Tohama I, our hybrid UK76 assembly showed 98.49 % identity, similar to the identity of our other hybrid assemblies (n=5, mean=98.57 %), suggesting that the inaccuracies of the raw Illumina reads do not translate into inaccuracies in the final assembly; only our estimation of accuracy by comparison to the Illumina-only draft is affected. Overall, neither comparison to the Tohama I reference nor comparison to an Illumina-only assembly is ideal for assessing error when working with

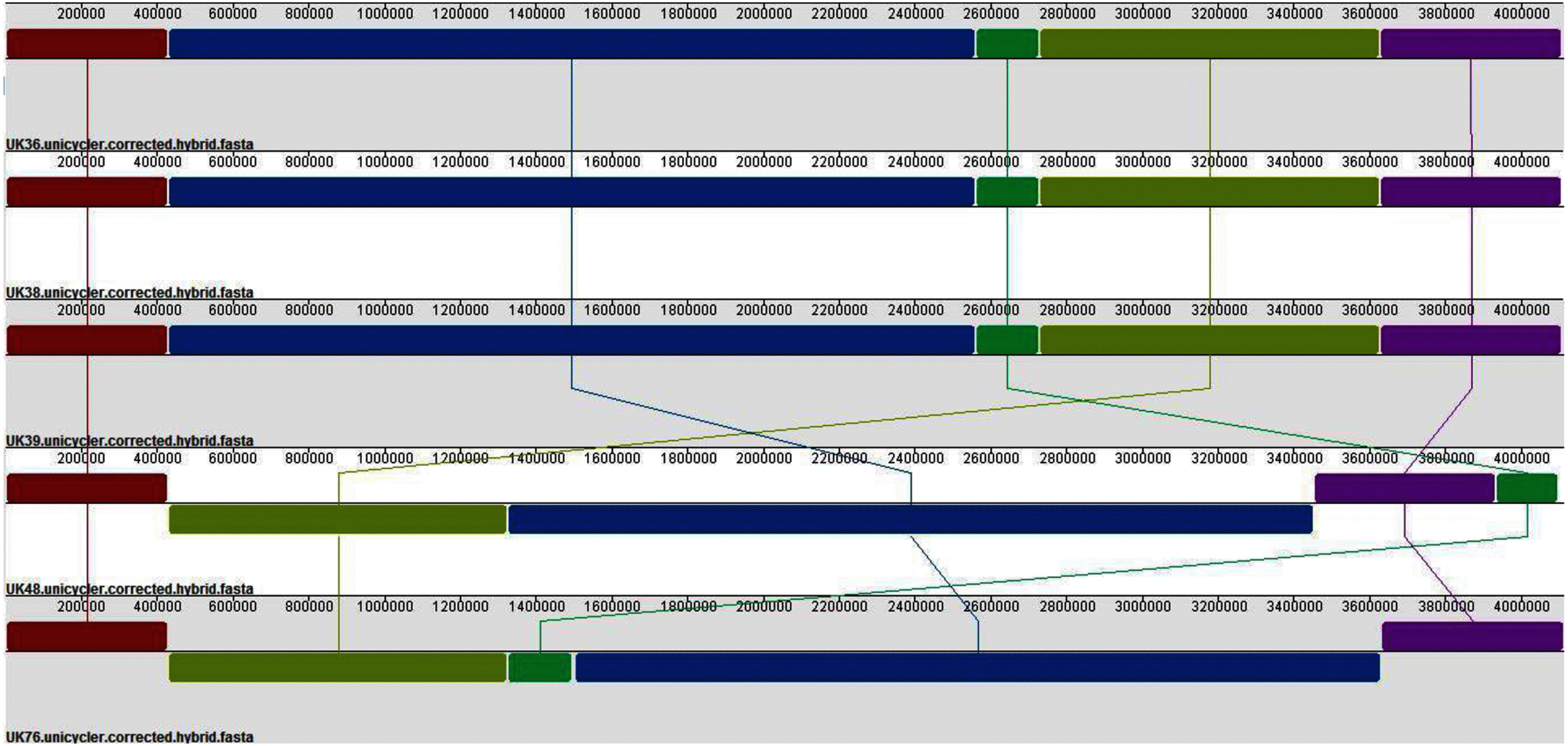

UK36.unicycler.corrected.hybrid.fasta
UK38.unicycler.corrected.hybrid.fasta
UK39.unicycler.corrected.hybrid.fasta
UK48.unicycler.corrected.hybrid.fasta
UK76.unicycler.corrected.hybrid.fasta
200000
400000
600000
800000
1000000
1200000
1400000
1600000
1800000
2000000
2200000
2400000
2600000
2800000
3000000
3200000
3400000
3600000
3800000
4000000

Figure 2.3 Alignment of our five sequenced strains, showing genomic rearrangement. Our five UK B. pertussis strains (UK36, UK38, UK39, UK48 and UK76) were assembled using our nanopore-only pipeline, resulting in single, closed-contig, assemblies. The closed assemblies were aligned with progressiveMauve, which showed that even strains which are closely temporally related can display different genomic arrangements.

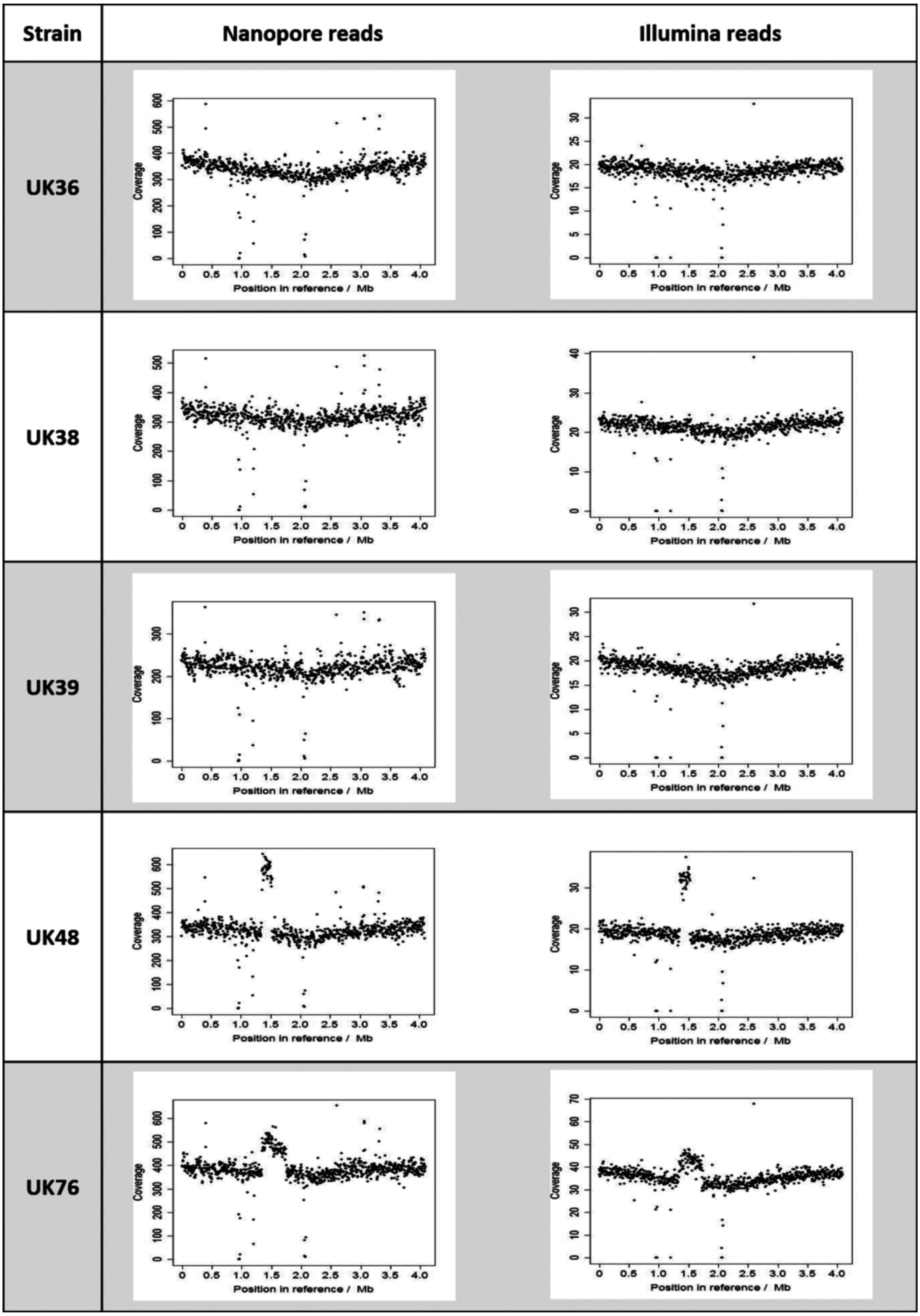


Figure 2.4 Alignment of nanopore reads to the Tohama I reference sequence compared to alignment of Illumina reads to the Tohama I reference sequence. Raw reads from each sequencer were aligned to the reference using BWA MEM, followed by coverage calculation with SAMtools depth. Plots were constructed by calculating a rolling average of coverage, with a window size of 5000 bp. The coverage of three strains (UK36, UK38 and UK39) was consistent across the whole reference genome, whereas UK48 and UK76 coverage was enriched in certain locations. In UK48, a large section from 1.35 to 1.53 Mbp into the reference appears to have exactly twice as much coverage as the rest of the genome. In UK76, a section from 1.38 to 1.68 Mbp is enriched by 25%. The coverage abnormalities seen in UK48 and UK76 are present in both sets of reads, suggesting they are not the result of a quirk in sequencing method, or contamination.

novel strains, and neither strategy gives us a completely accurate estimate, but using a combination of both comparisons allows a good estimate of assembly error.

Having estimated our hybrid assemblies to be, on average, 99.69% accurate, we can conclude that roughly 13 000 bases in each 4.1 Mbp draft genome are incorrect. Whilst these incorrectly called bases will not influence comparisons of genome arrangement, residual base errors in draft genome sequences assembled using long reads remain a concern, with the potential to falsely identify SNPs or prevent accurate protein prediction (Watson and Warr, 2019). Incorrect sequencing of homopolymers is a known weakness of many sequencing methods, including nanopore sequencing, and our assemblies are no exception (Jain, Koren, et al., 2018). Indeed, a base-level manual comparison of one of our hybrid assemblies with a more accurate Illumina-only draft using progressiveMauve revealed that every difference occurred in a homopolymeric tract, with the hybrid sequence having inserted or deleted bases. Two options for correct SNP identification, therefore, are manual correction of known homopolymeric indels, and simply ignoring SNPs which appear to occur in homopolymeric regions. The manual correction option would be time-consuming, whilst the second option could result in false negatives. Nevertheless, until improved pore chemistry or basecalling tools are available which do not produce homopolymeric indels, the use of either option means that SNP identification is still possible, even in assemblies which are less than 100% accurate.

Correct prediction of proteins appears to be of less concern than SNP identification in our hybrid assemblies: all 40 potential bacterial BUSCOs were present in full for all of our strains, and both Quast and Prokka were able to identify the majority of the Tohama I reference proteins in the same assemblies. In addition, assessment of our UK36 hybrid using Watson's Ideel pipeline suggested that, although we know some errors remain, they do not substantially inhibit the correct prediction of full-length proteins during annotation (Watson, 2018). It is here, however, that we can see clearly the benefit of the hybrid assemblies over the nanopore-only assemblies: although the mean accuracy of the nanopore-only assemblies (99.48%) was only 0.2% lower than that of the hybrids, none of the nanopore-only assemblies contained full copies of all 40 BUSCOs.

Does the de Bruijn graph method assemble highly repetitive prokaryotic genomes more accurately than other commonly used methods?

The opinion of the sequencing community has long been that de Bruijn graph assembly is not as effective for error-prone long reads as other *de novo* assembly methods (Pop, 2009; Lu, Giordano and Ning, 2016). This is likely because the de Bruijn graph method requires the reads to be cut into much shorter kmer-length fragments, meaning the benefit of longer reads could be lost. The tool which consistently produced the most accurate nanopore-only *B. pertussis* assemblies was therefore unexpected: the percentage identity and indel rates of our assemblies produced by ABruijn, which is a de Bruijn graph assembler, were better by far than those of the Canu, Miniasm or Unicycler assemblies. The recent version change of ABruijn to Flye seems to have negatively affected these metrics in some of our strains; however, whilst the ABruijn assemblies were better than the Flye assemblies, the Flye assemblies were still better than those produced by other tools. Another recent study, which assembled highly complex and repetitive *Pseudomonas koreensis* genomes using ultra-long nanopore reads, also found Flye to produce the most accurate assemblies (Schmid et al., 2018). This suggests that the de Bruijn method, and particularly ABruijn/Flye, might be optimal for prokaryotic genomes which contain a high number of repeats.

Two possible pipelines for *B. pertussis* genome sequence resolution

We have shown here that resolution of five *B. pertussis* genomes per MinION flow cell is possible, whether using long reads alone or in combination with short reads. Sequencing five strains using one flow cell produced a mean yield of over 300× *B. pertussis* genome coverage per strain, which probably exceeds that required to achieve comparable results. A draft produced from using roughly half of our reads (175× coverage) for UK36, pre-corrected and assembled with Flye, had an identity of 99.467%, whilst the same assembly produced by the full (360× coverage) read set had an identity of 99.474%. This suggests that twice as many strains could be de novo assembled per flow cell without a notable drop in accuracy. Thus, resolution of ten *B. pertussis* genomes per MinION flow cell should be possible.

If short reads are also available, we have shown that hybrid assembly, using pre-correction with Canu followed by Unicycler, remains the most accurate method. Indeed, for now, for full strain characterization (including comparison of genome arrangement, SNP identification and allele-typing), hybrid assemblies are required. For comparison of genome structure and arrangement only (e.g. **Figure 2.3**), however, our nanopore-only pipeline, which uses Canu pre-correction, Flye assembly and post-assembly polishing with Nanopolish, can produce single contig assemblies of adequate accuracy for all but the most unusual *B. pertussis* genomes.

Continued improvement of long-read data processing tools

Although the pipelines we have defined here produce the most accurate *B. pertussis* genome sequences currently possible, the tools available for the analysis of nanopore sequencing data are continually improving. A recent update to Racon added the ability to polish assemblies with Illumina reads; a brief comparison of this with Pilon, however, showed little improvement to our data, so we did not add short-read Racon polishing to our suite of tests. Alternative basecallers such as Chiron or the currently in-development Guppy, which use entirely new basecalling algorithms, may also offer further accuracy improvements and could be trialled with existing and future B. pertussis data sets, particularly if Illumina short reads are not available for hybrid assembly (Teng et al., 2018).

We tested the most commonly used *de novo* assembly tools suitable for long reads and, at the time of writing, are not aware of any newly released tools. However, minor (or sometimes major, in the case of ABruijn to Flye) updates are common. New polishing tools are also being developed: ONT's own Medaka, for example, is claimed to rival Nanopolish in terms of speed and assembly improvement capabilities (Oxford Nanopore Technologies, 2018a). In addition, MaSuRCA was not trialled here due to the low Illumina coverage (the manual suggests 50×+ for hybrid assemblies, whereas we had only 37.5× coverage for UK36) (Zimin et al., 2017). Ultimately, for the foreseeable future, no data pipeline including nanopore reads should be set in stone; we will continue to trial new tools and to update our pipeline where appropriate, and would suggest that similar pipeline optimization may be required for each organism to be sequenced.

2.9 *Commentary text – conclusions*

The work presented in this paper established a workflow which can be used to rapidly sequence and *de novo* assemble closed genomes for larger numbers of *B. pertussis* isolates. The key findings (below) from this body of work lead directly to the research questions addressed in Chapter 4, which investigates the duplications discovered here, and to the development of sequencing methodology for Chapter 3, which screens a larger number of isolates from New Zealand.

Key findings:

- The most recent long-read basecalling, assembly, and other data processing tools tended to produce the most accurate data, likely because they are developing as ONT's sequencing technology does. As the technology develops, the Nanopore Community will be a valuable resource to keep track of which library preparation, flow cells and tools are currently favoured. Periodic re-optimisation (to a lesser extent) of the workflow may also be prudent
- For analyses which include variant calling, hybrid assembly will produce more accurate results than assembly of long reads alone
- A significant portion of sequencing data is lost during the demultiplexing process. This limits the number of strains which could be sequenced on the same flow cell; a yield of, for example, 10 Gb should not be assumed to contain 10 Gb of usable data. The loss of data can be explained in two ways. Firstly, some DNA strands may survive through the library preparation process without having had a barcode ligated, thus will produce reads with no known barcode during demultiplexing. Secondly, the high raw error raw of MinION sequencing may mean that even some reads which do have barcodes may be added to the "unknown barcode" bin during demultiplexing, because the barcode has been misread
- Large duplications, of unknown phenotypic consequence, may be present in the genomes of some *B. pertussis* strains, which would not be noticed during *de novo* assembly with short reads, and would only be apparent during *de novo* assembly with long reads because they mean that the genome cannot be resolved into a single contig. All sequenced strains should therefore be investigated using the coverage analysis method demonstrated here, as this method shows duplications as areas of enhanced coverage.

Chapter 3: Comparative genomics of *Bordetella pertussis* isolates from New Zealand, a country with an uncommonly high incidence of whooping cough

“It's a dangerous business, Frodo, going out of your door. You step into the road, and if you don't keep your feet, there's no knowing where you might be swept off to.”

- J. R. R. Tolkien, The Fellowship of the Ring

3.1 *Abstract*

Whooping cough, the respiratory disease caused by *Bordetella pertussis*, has seen a wide-spread resurgence over the last several decades. Predicted causes for this resurgence include changes to the bacterium at the genomic level, potentially in response to the introduction of an acellular whooping cough vaccine in the late 1990s. Previously, we developed a pipeline to assemble the repetitive *B. pertussis* genome into closed sequences using hybrid nanopore and Illumina sequencing. Here, this sequencing pipeline was used to conduct a more high-throughput, longitudinal screen of 66 strains isolated between 1982 and 2018 in New Zealand. New Zealand suffers from a higher incidence of whooping cough than almost any other vaccine-using country, usually at least twice as many cases per 100,000 people than the USA and UK and often even higher, despite similar rates of vaccine uptake. We believe these strains represent the first New Zealand *B. pertussis* isolates ever sequenced, and the screen is the first to use nanopore sequencing. The analyses here show that, on the whole, genomic trends in New Zealand *B. pertussis* isolates, such as changing allelic profile in vaccine-related genes and increasing pertactin deficiency, have paralleled those seen elsewhere in the world. At the same time, phylogenetic comparisons of the New Zealand isolates with global isolates suggest that a number of strains are circulating in New Zealand which are closely related to each other, and which cluster separately from other global strains. Whilst in apparent continuous circulation, these strains appear to be more highly represented during outbreak periods. The results of this study add to a growing body of knowledge regarding recent changes to the *B. pertussis* genome, and are the first genetic investigation into New Zealand's high incidence of whooping cough.

3.2 *Data Summary*

1. Nanopore and Illumina fastq sequence files for all strains have been deposited in the NCBI's Sequence Read Archive, BioProject PRJNA556977. A full list of accession numbers for all sequence read files is provided in **Supplementary table S3.1**.
2. Genome sequences for all strains have been deposited in the NCBI's GenBank, accession numbers shown in **Supplementary table S3.1**.

3.3 *Introduction*

A (brief) history of whooping cough vaccination in New Zealand

The whooping cough vaccine has been available in New Zealand since 1945, and part of the immunisation schedule since 1960. The original schedule included three doses of the whole cell pertussis vaccine with diphtheria and tetanus (DTwP), at 3, 4 and 5 months old. In 1980, the schedule for DTwP vaccination included two doses, at 3 and 5 months of age. A third, earlier, dose was added at 6 weeks in 1984. The fourth dose was added in 1996, at 2 years, and the fifth was added in 2002, at 4 years (Reid, 2006, 2012).

New Zealand switched from the whole cell vaccine to a three-antigen (PT-FHA-PRN) acellular pertussis vaccine (DTaP) in 2000. After the switch to the acellular vaccine, the timings of some of the doses were changed; the current New Zealand pertussis vaccination schedule consists of doses at 6 weeks, 3

Table 3.1 Changes to the pertussis vaccination schedule in New Zealand

	Year						
	1960	**1980**	**1984**	**1996**	**2000**	**2002**	**2008**
6 weeks			DTwP	DTwPH	DTaP	DTaP-IPV	DTaP-IPV-Hep/Hib
3 months	DTwP	DTwP	DTwP	DTwPH	DTaP	DTaP-IPV	DTaP-IPV-Hep/Hib
4 months	DTwP						
5 months	DTwP	DTwP	DTwP	DTwPH	DTaP	DTaP-IPV	DTaP-IPV-Hep/Hib
15 months				DTwPH	DTaP-Hib	DTaP-Hib	
4 years						DTaP-IPV	DTaP-IPV
11 years							Tdap

DTwP: Diptheria, Tetanus, whole cell Pertussis
DTwPH: Diptheria, Tetanus and whole cell Pertussis and *Haemophilus influenzae* type b
DTaP: Diptheria, Tetanus and acellular Pertussis (under 7s)
DTaP: Diptheria, Tetanus, acellular Pertussis and Polio
DTaP-IPV: Diptheria, Tetanus and acellular Pertussis and *Haemophilus influenzae* type b
DTaP-IPV-Hep/Hib: Diptheria, Tetanus, acellular Pertussis, Polio, Hepatitis B and *Haemophilus influenzae* B
Tdap: Diptheria, Tetanus and acellular Pertussis (booster for over 10s, including adults)

months, 5 months, 4 years and 11-12 years (Reid, 2006, 2012). **Table 3.1** shows the major changes to the vaccine schedule from 1960 onwards.

Historically immunisation coverage in New Zealand has been low

Prior to 2008, vaccination coverage in New Zealand was only surveyed sporadically. The last surveys before regular records were kept were in 1991/2 and 2005. In 1991/2, fewer than 60% of two-year-olds were fully immunised; by 2005, this had increased to 77.4% (NZ Ministry of Health, 2005). In 1993/4, national immunisation coverage targets were created to initiate an effort to increase coverage closer to the levels required for effective herd immunity (often cited at 90-95%) (NZ Ministry of Health, 2005).

Since 2008, comparatively robust immunisation data has been kept by each of New Zealand's District Health Boards, on behalf of the Ministry of Health, and published on a three-monthly cycle (New Zealand Ministry of Health, 2019). These figures show a marked improvement in immunisation coverage over the last decade, and certainly since the 1991/92 coverage survey (**Figure 3.1**). Immunisation coverage in New Zealand at all age points up to 5 years is now comparable to mean coverage in the United Kingdom, although coverage in the UK also varies significantly according to area (approaching 100% in some areas) (NHS, 2017).

Rates of whooping cough in New Zealand are high

Since the whooping cough vaccination was introduced globally, New Zealand has noted a high rate of infections and hospitalisations compared with other developed countries; some sources indicate disease rates 5 to 10 times higher in New Zealand than the UK or USA. For instance, during the 1980s, hospitalisation for whooping cough was required for 0.37 per 100,000 people in the USA, compared with 3.75 per 100,000 in New Zealand (Reid et al., 1994). In addition, epidemic periods have often been more severe in New Zealand than other countries. In a 1993 epidemic, disease notifications in the USA were seven times lower than in a New Zealand 1996 epidemic (Somerville et al., 2007). During concurrent epidemics in the UK and New Zealand in 2012, 122.3 cases were seen per 100,000 people in New Zealand, compared to 20 per 100,000 in the UK (ESR, 2012; Sealey et al., 2015).

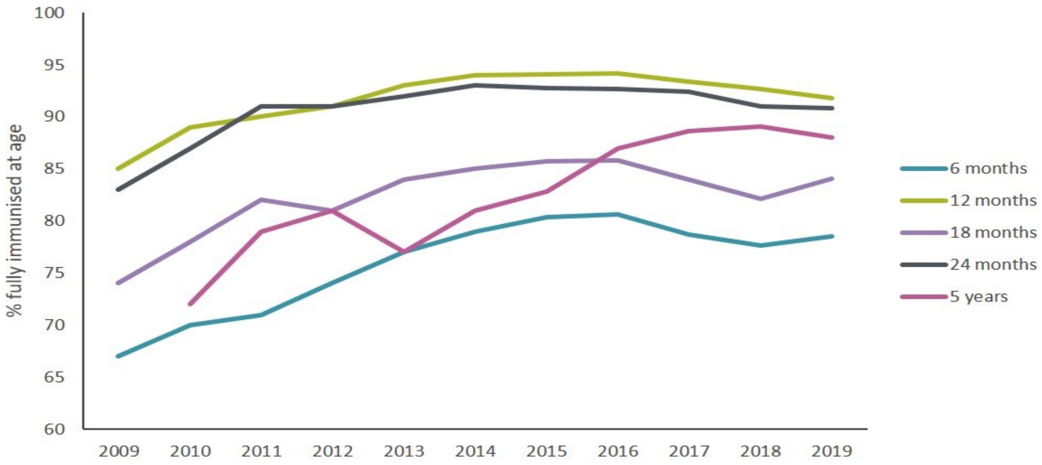


Figure 3.1 Rates of vaccination uptake in New Zealand between 2009 and 2017, across the five main immunisation time points. **Supplementary Table S3.2** shows the raw data from which this figure was constructed. The data shown here represents the whole country; some areas have much lower or higher rates of immunisation coverage. Data was not available for the 2009 5 years time point.

The latest outbreak of whooping cough in New Zealand was announced in December 2017 (New Zealand Ministry of Health, 2017) and, between the identified beginning of the outbreak (16 October 2017) and 4 May 2018, 2,411 cases were reported, representing an infection rate of 50.8 per 100,000 (ESR, 2018). By May 2019, notifications of whooping cough cases had returned to the "expected" level, after a total of 4,697 cases in the 19-month outbreak period (ESR, 2019). The most recent previous outbreak lasted for 16 months, between August 2011 to December 2013, totalling 11,000 cases (247.6 cases per 100,000 across the whole epidemic period) (New Zealand Ministry of Health, 2017).

The higher rates of whooping cough in New Zealand have often been explained by the lower rates of vaccine uptake in the country, particularly in the youngest age categories (Somerville et al., 2007). However, despite an increase in immunisation coverage across all age groups over the last decade, non-epidemic cases of whooping cough in New Zealand have remained at least two times, and often between three and four times higher than in England and Wales or the USA (**Figure 3.2**). Theoretically, this residual higher incidence could be explained by differences in population density, with spread of disease being more likely in more densely populated areas. The portions of the population living in urban settings in New Zealand, the USA and UK are comparable, however, at 87%, 83% and 83% respectively (Worldometer, 2020). Furthermore, during the latest outbreak, the highest rates of incidence were seen outside of the most populated areas, with the West Coast, Nelson Marlborough and Wairarapa district health boards reporting the highest rates, as opposed to those which contain the three largest cities (Auckland, Christchurch and Wellington) (ESR, 2019).

Whooping cough only became a notifiable disease in New Zealand in 1996, hence age incidence data from prior to this is patchy or unavailable. **Figure 3.3a** shows the incidence in each age group in the late 1999s, just before the acellular vaccine was introduced. This incidence agrees with that seen elsewhere in the world during the whole cell vaccine era, with infants representing a majority of cases, and incidence steadily decreasing with age (Mattoo and Cherry, 2005). In 1999-2000, a slight increase in incidence is seen in the 5-9 years age group compared to the 1-4 years age group. In 2002 a vaccine booster dose was added to the schedule at age 4 years old. This booster dose appears to have been successful, as no similar increase is seen in the 5-9 years age group in the 2008-2009 or 2018-2019 data. It is also notable that, in almost every age group, incidence was higher in 1999-2000 than in

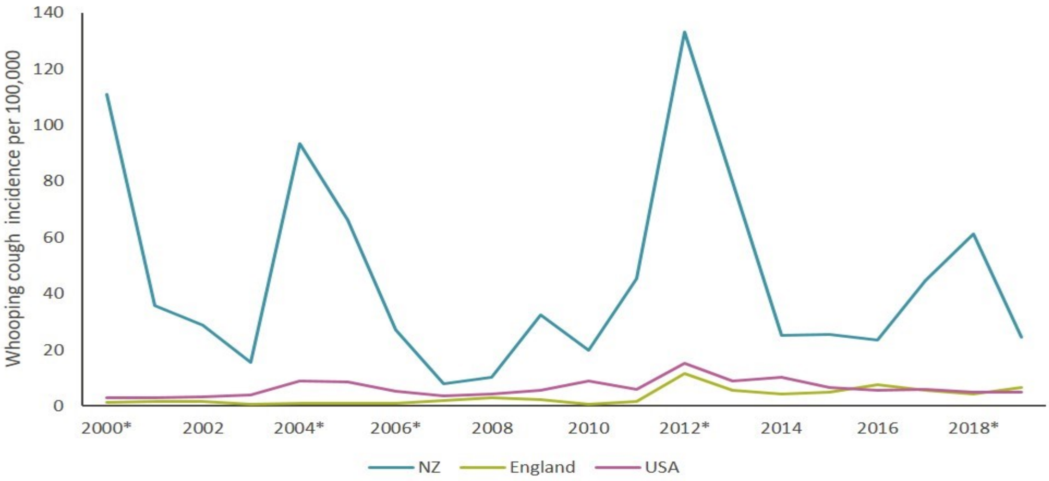


Figure 3.2 Whooping cough cases per 100,000 between 2000 and 2017 in New Zealand, the USA and England & Wales. **Supplementary Table S3.3** shows the raw data used to construct this figure.

*denotes years within a New Zealand epidemic period.

either 2008-2009 or 2018-2019, both of which include outbreak periods. This lower incidence in the more recent years is likely a result of the increasing vaccine coverage in New Zealand since the 1990s. Incidence remained the highest in infants in 2008-2009 and 2018-2019, although a shift towards a slightly higher incidence in adolescents compared with 1999-2000 can be seen; this pattern also matches that seen elsewhere in the world since the introduction of the acellular vaccine (Mattoo and Cherry, 2005). **Figure 3.3b** shows the age distribution and vaccination status for cases notified between 1st January 2017 and 31st May 2018, where vaccination status was known. This period includes six months of the latest outbreak period. At every age group, the number of cases in vaccinated patients was higher than those in unvaccinated patients, suggesting that waning immunity or vaccine inefficacy may have played a part in this outbreak.

Why screen strains from New Zealand using hybrid sequencing?

No *B. pertussis* sequencing reads or submitted genomes on the NCBI database are currently tagged as being from New Zealand (although this does not preclude the possibility of New Zealand genomes being available but lacking location metadata). *B. pertussis* isolates are, however, stored "pending new method development" at the Institute of Environmental Science and Research Ltd (ESR)'s Invasive Pathogen Laboratory (ESR Invasive Pathogen Laboratory, 2019). This unusual lack of genome screening in a developed country therefore makes New Zealand's bank of *B. pertussis* isolates potentially very informative.

Extensive screens of circulating strains during epidemics in the UK, Australia and USA (multiple), have been conducted in recent years (Octavia et al., 2012; Bowden et al., 2014; Lam et al., 2014; Sealey et al., 2015; Bowden et al., 2016) , each contributing to our understanding of how *B. pertussis* is evolving under selection pressure from the whooping cough vaccine, and supporting Mooi, Zeddeman and van Gent's 2015 contention that strain characterization is vital, alongside classical epidemiology, to address the problem of resurging Pertussis (Mooi, Zeddeman and van Gent, 2015). Interestingly, although the strains involved in the USA and UK epidemics were polyclonal and relatively disparate, the strains identified in the prolonged Australian epidemic from 2008 to 2011 (156 cases per 100,000)

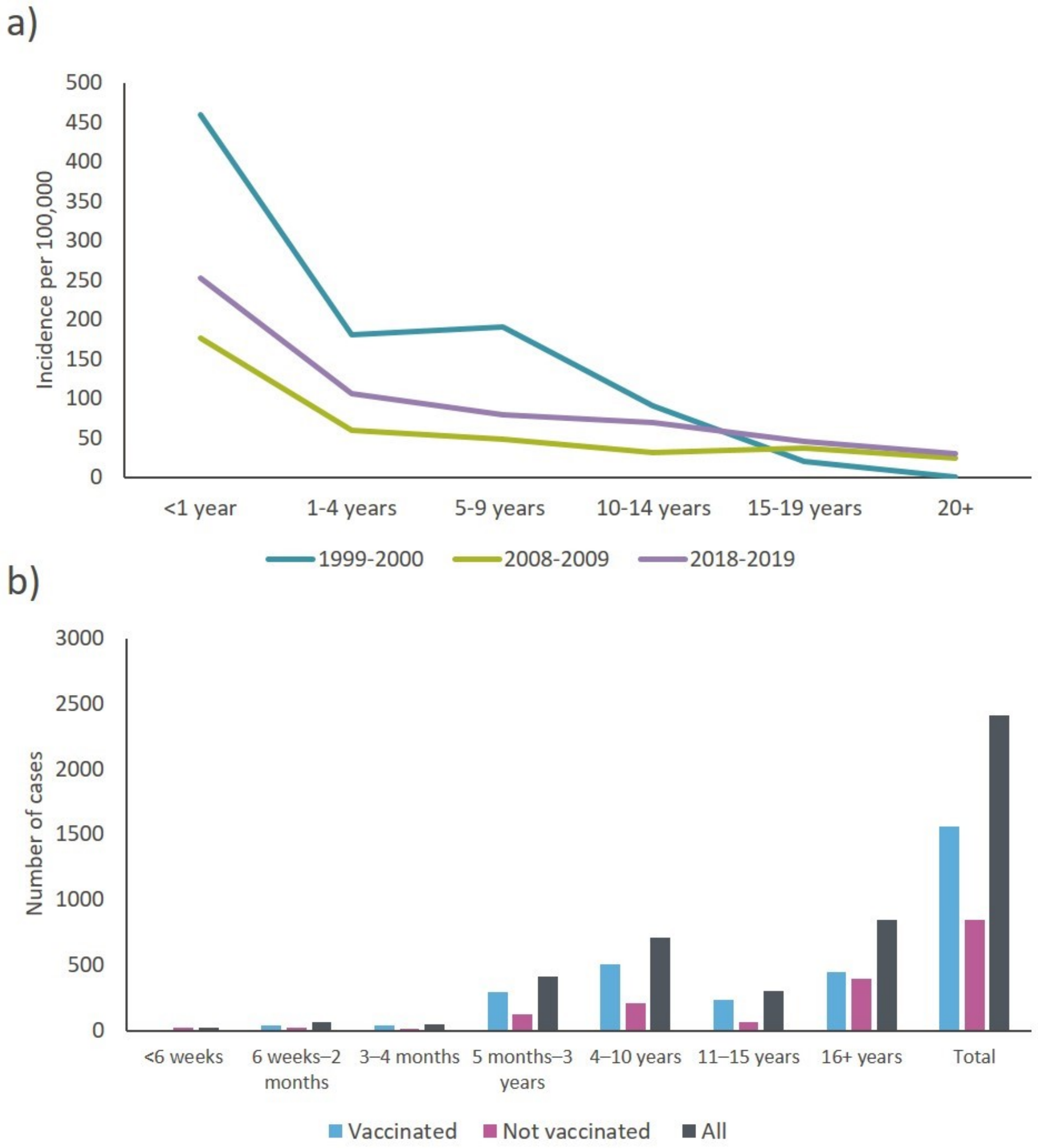


Figure 3.3 Age stratification of whooping cough cases in New Zealand, 1996-2019.

a) Incidence per 100,000 in different age categories in 1999-2000, 2008-2009 and 2018-2019. 1999-2000 incidence from Grant (2004), 2008-2009 incidence from ESR (2009), 2018-2019 incidence from (ESR, 2019).

b) Age distribution and vaccination status of whooping cough cases in New Zealand, 1st January 2017 to 31st May 2018. Not all cases are shown, as for many the vaccination status was unknown. **Supplementary Table S3.4** shows the raw data from which this figure was constructed.

showed higher levels of similarity, with 86% represented by just three “closely-related SNP profiles” (Octavia et al., 2012). This suggests that, especially during epidemics, certain strains may circulate more commonly than others, although this has not been observed in every country.

A landmark screen of global strains by Bart *et al.* in 2014 showed that the genome of *B. pertussis* has been evolving since the introduction of first the whole cell vaccine, and more rapidly since the switch to acellular vaccination (Bart, Harris, et al., 2014). Changes which have occurred include a swift, near global, adoption of a new, more virulent, allele of the pertussis toxin promoter (*ptxP3*) since its first appearance in the 1980s, and the appearance of strains which do not express pertactin; PT and PRN are both antigenic proteins included in the acellular vaccine, therefore it is likely that these changes are related to the introduction of vaccination, and to the switch from WCV to ACV. However, a 2015 comparison of strains from Finland, Sweden and China by Xu *et al.* suggested that in countries where vaccine uptake has traditionally been lower or delayed (such as China or, indeed, New Zealand), the rate of allelic change has also been delayed (Xu, Y. et al., 2015).

Bowden *et al.*'s 2016 comparison of epidemic strains from Vermont and California was the first of its kind to use long-read sequencing for genome surveillance, and resulted in the identification of a variety of structural differences between the strains which would have otherwise gone unvalidated using only short reads, Pulsed Field Gel Electrophoresis or genome mapping (Bowden et al., 2016). However, access to Pacific Biosciences sequencing is restricted by the cost of the sequencers themselves. The availability of previously unsequenced isolates in New Zealand therefore presents an opportunity to demonstrate the utility of an alternative long read sequencer, Oxford Nanopore Technology's MinION, to conduct an outbreak strain screen, based on the hybrid workflow developed in Chapter 2 (Ring et al., 2018).

3.4 *Research questions*

Research question 1: Were the strains circulating during the recent New Zealand outbreak closely related, as in the Australian 2008-2011 outbreak, or polyclonal, as in the UK and USA 2012 outbreaks?

Research question 2: Has New Zealand's historic lower immunisation coverage caused allele frequencies in New Zealand to change more slowly than observed in Bart et al.'s 2014 global survey (and other recent relevant surveys)?

Research question 3: Are rates of whooping cough infection unusually high in New Zealand due to the circulation of a unique, geographically-specific, hypervirulent strain (or strains)?

3.5 *Methods*

Strain isolation

66 strains, collected between 1982 and late 2018, were provided from the New Zealand Institute of Environmental Science and Research (ESR)'s Invasive Pathogens, Special Bacteriology and Culture Collection by Audrey Tiong and Angela Brounts. **Figure 3.4** shows the locations and years in which, and the age of the patient from whom, strains were isolated. Strains were grown and heat-killed at the ESR Kenepuru Science Centre, Porirua, New Zealand, and shipped on ice to the University of Bath, United Kingdom. On arrival, the heat-killed cells were stored at -20°C. Where serotyping data was available (46/66 strains), these results were provided with the strains by the ESR. Full details, including accession numbers, are included in **Supplementary Table S3.9**.

DNA extraction and Illumina sequencing

Heat-killed cells were resuspended in 1 ml PBS and OD_{600} was measured. Volumes of suspension equating to an OD of 2.0 (~4x10^9 *B. pertussis cells*) were pelleted in a microcentrifuge for 2 min at 12,000 g. gDNA was extracted from each pellet using the QIAamp DNA mini kit (Qiagen) according to

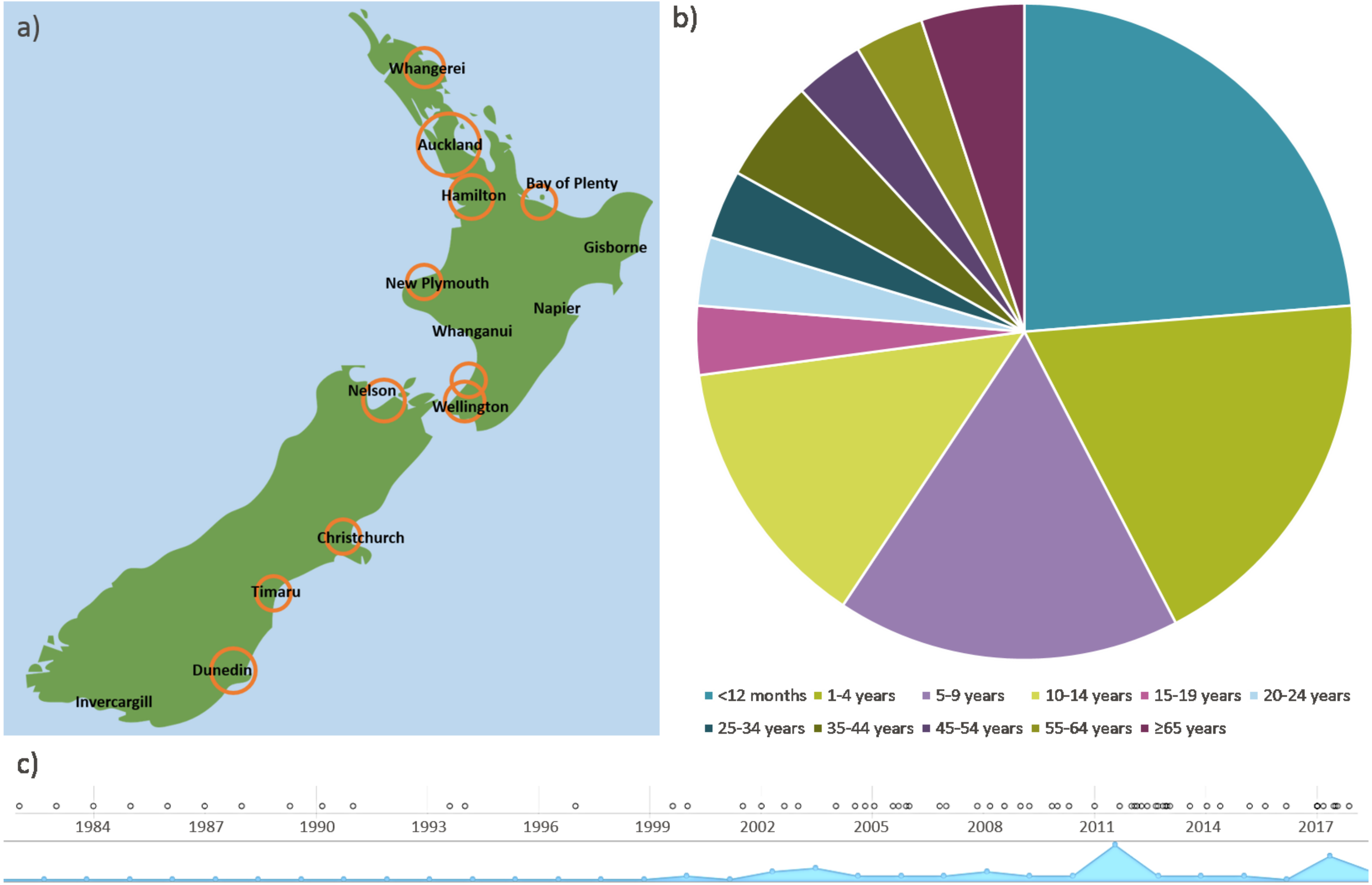


Figure 3.4 (a) Location, (b) patient age and (c) isolation year of the 66 *B. pertussis* samples received from the New Zealand Institute of Environmental Science and Research (ESR)'s Invasive Pathogens, Special Bacteriology and Culture Collection. The sizes of the location markers in a) are proportional to the number of isolates received from that location. Map and timeline were generated by Microreact (Argimon et al., 2016).

the manufacturer's instructions, including a single-step elution into 200 µl of elution buffer (buffer AE).

At least 2.5 µg gDNA from each strain, suspended in 50 µl elution buffer, were sent for Illumina MiSeq sequencing at the Milner Centre, University of Bath in May 2019. In March 2020, 29 strains were sent to Novogene for resequencing with NovaSeq (see **Supplementary table S3.9** for full details).

Nanopore library preparation and sequencing

Sequencing libraries were prepared for all samples using ONT's Rapid Barcoding Kit (SQK-RBK004), according to manufacturer's instructions, including the optional 1x SPRI concentration and clean-up step (using Promega ProNex size selection beads) after library pooling and before addition of RAP adaptors. Between 10 and 12 barcodes were used per MinION flow cell (see **Supplementary Table S3.10** for full details).

Each pooled sequencing library was loaded onto an R9.4 MinION flow cell and sequenced for 48 h using a MinION Mk1b device with MinKNOW sequencing software, including concurrent Guppy fast Flip-flop basecalling to allow monitoring of translocation speed and q-score. If translocation speed dropped below 300 bases per second during the 48-h sequencing run, the flow cell was refuelled, as per manufacturer's instructions for SQK-RBK004. If the translocation speed falls below 300 bases per second, the overall yield of the run can be negatively affected. In addition, the MinKNOW algorithms which monitor the current across the flow cell membrane and produce the raw fast5 "squiggle" plots have been optimised for speeds between 300 and 425 bases per second. Consequently, if the translocation speed slows, the quality scores of the subsequent reads can also begin to decrease. Translocation speeds fall when ATP levels in the running buffer fall beneath optimal levels, which can occur when high masses of DNA are loaded. ATP is used by the adaptor proteins which feed DNA strands through the pore proteins. Hence, lowered ATP can result in reduced protein function and slower DNA translocation. Adding more flow cell running/flush buffer (FLB, ONT), which contains ATP, during the run can recover decreased translocation speeds, as shown in **Figure 3.5**. Using the mid-run refuelling method, high masses of DNA can be loaded to maximise pore throughput without negatively affecting translocation speed and quality scores, thus increasing the overall yield and mean quality score of the run.

Basecalling, demultiplexing and adaptor-trimming

Deepbinner (v0.2.0) (Wick, Judd and Holt, 2018b) was used to demultiplex the raw fast5 files, using the "realtime" setting with "rapid" option. This placed all fast5 files into separate bins, one for each barcode. ONT's Guppy fast Flip-flop basecaller (v3.1.5+781ed57) was then run on each of these barcode bins, with its own rapid barcode demultiplexing option enabled. This placed each of the fastq files basecalled from the fast5s in the Deepbinner barcode bins into further barcode bins, resulting in twelve Deepbinner barcode bin fast5 directories (plus one "unclassified" bin, containing reads with

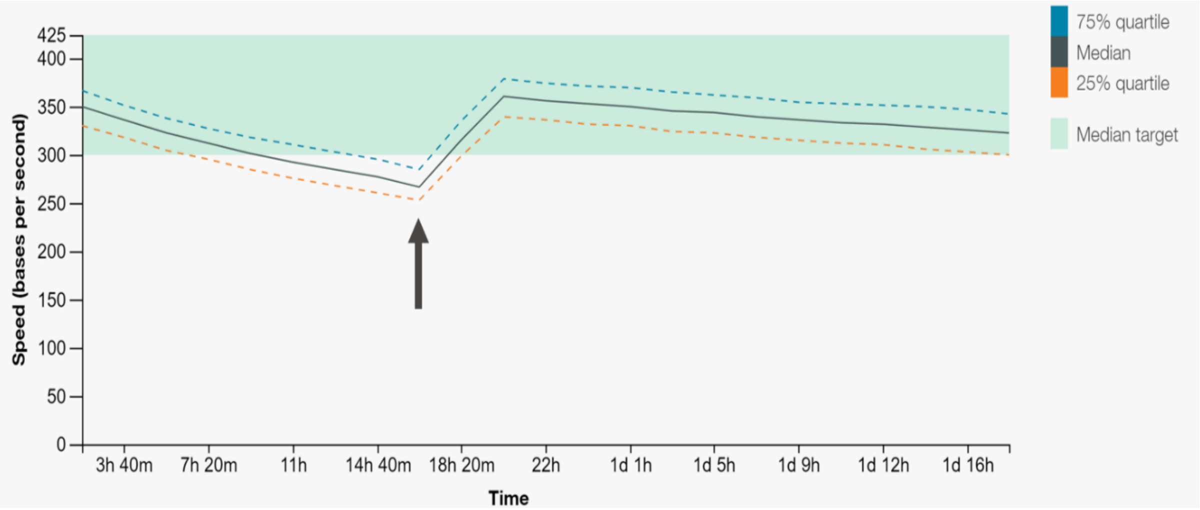


Figure 3.5 Refuelling with FLB during a sequencing run can recover falling translocation speeds and thus increase overall flow cell yield and mean read quality score. Here, the translocation speed of DNA strands through the pore proteins fell below 300 bases per second around 11 hours into the sequencing run. The flow cell was refuelled with 250 µl extra FLB at around 16 hours (see grey arrow), recovering the translocation speed to around 350 bases per second.

unclassifiable barcodes), each containing up to twelve further barcode bin fastq directories (plus "unclassified"). Theoretically, most of the reads from any specific Deepbinner bin should have been placed in a bin corresponding to the same barcode by Guppy, but in some cases the two tools may disagree over the barcode identity; for example, most of the reads in the "barcode01" Deepbinner bin should have also been placed in the "barcode01" bin by Guppy after basecalling, but some may have been placed in other bins. Only reads which were identified as having the same barcode by both tools were retained for further processing - see **Figure 3.6** for schematic. Finally, Porechop (v0.2.4) (Wick, 2017) was used to trim the barcodes and other adaptor sequences from the demultiplexed reads. Occasionally (usually 0-2% of reads), two individual reads can be read as a single read by the sequencing software; Porechop corrects chimeric reads by detecting sequencing adaptors in the middle of single reads and splitting them into multiple reads instead.

Hybrid genome assembly

Illumina MiSeq data was provided pre-trimmed by the Milner Genomics Centre. Illumina NovaSeq data provided by Novogene was trimmed using Trimmomatic (v0.36, Bolger, Lohse and Usadel, 2014) with the options PE, HEADCROP:10, SLIDINGWINDOW:4:25 and MINLEN:100. Genome assembly was attempted for all strains using the hybrid pipeline presented in Chapter 2 (**Figure 2.2**). Nanopore reads were pre-corrected using Canu (v1.8, Koren et al., 2017), followed by hybrid assembly with nanopore and Illumina reads using Unicycler (v0.4.7, Wick et al., 2017b). However, some of the available Illumina data was of poor quality (reads <70 bp, non-uniform length) or low coverage (<40x). Unicycler uses an Illumina-centric hybrid assembly method, using SPAdes to first assemble the Illumina data into contigs, then the nanopore reads to attempt to bridge the contigs. The low quality of some of the Illumina data therefore prevented Unicycler from assembling closed genomes for these strains. After confirming that the nanopore long reads alone were sufficient for the assembly of closed genomes for these strains, the best long-read-centric hybrid assembly strategy identified in Chapter 2 was used to assemble genomes for the strains with low quality Illumina data (full details of which strategy was used for each strain are shown in **Supplementary Table S3.9**).

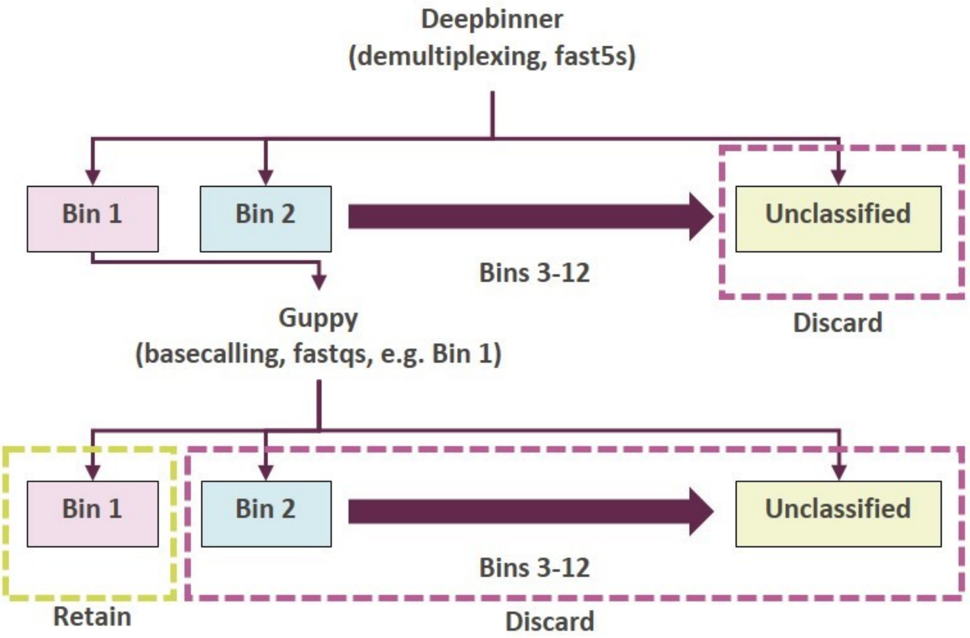


Figure 3.6 Demultiplexing and basecalling by Deepbinner and Guppy. Deepbinner demultiplexes the raw fast5 files, placing them in barcode bins (one for each barcode, plus one for "unclassified" reads). Guppy then basecalls and demultiplexes the fast5s within each of the Deepbinner bins, placing the resulting fastqs in barcode bins (one for each barcode, plus one for "unclassified"). Theoretically, most reads should be placed in corresponding barcode bins by both tools (i.e. a read from the Deepbinner barcode01 bin should be placed in the Guppy barcode01 bin after basecalling). Only reads for which both tools agreed a barcode were retained.

The best long-read-centric hybrid assembly strategy in Chapter 2 used Flye to assemble Canu-corrected nanopore reads, followed by polishing with Nanopolish once, and polishing with Pilon three times using the Illumina short reads. As shown in **Table 2.2**, the estimated accuracy of this strategy was only 0.01% lower than that of Unicycler. Since the work in Chapter 2 was conducted, however, Nanopolish has been succeeded by Medaka as the favoured tool for nanopore-based polishing. Medaka is optimised to polish assemblies which have already been polished four times with Racon. The following assembly strategy was therefore used to assemble the strains which Unicycler was unable to assemble: Canu-corrected nanopore reads were assembled with Flye (v2.7b-b1526) and polished four times with Racon (v1.4.11), then polished with Medaka (v0.11.3, https://github.com/nanoporetech/medaka), both using the nanopore reads alone. Finally, the assembly was polished three times using Pilon (v1.22) with the short Illumina reads.

Illumina reads were not available for three strains (NZ1, NZ5 and NZ29). Genome assemblies were produced for these strains using nanopore reads alone, using the Flye-Racon-Medaka strategy above, without the Pilon polishing steps. These strains were not included in SNP analysis, phylogenies or allele typing, but they were included in analysis of genome arrangement and numbers of IS elements.

Assessing assembly completeness and quality

Closed sequences were checked for misassemblies using BUSCO (v4.0.2, Seppey, Manni and Zdobnov, 2019). Version 4.0.2 of BUSCO is very different to v1.22, used in Chapter 2. Instead of 40 gram-negative representative genes, 124 generic bacterial BUSCO genes were used here (using the command line option "-l bacteria_odb10").

In addition, Illumina reads were assembled with ABySS (v2.0.2, Simpson et al., 2009; Jackman et al., 2017), and the resulting contigs were used to estimate the accuracy of the hybrid assemblies, as described in Chapter 2.

Quantifying IS elements

After assembly, the numbers of four IS elements (IS *481*, IS *1001*, IS *1002*, IS *1663*) in each closed genome were counted by searching for their sequences using blastn (IS element accession numbers shown in **Table 3.2**). IS *1001* was included to screen for any *B. parapertussis* isolates mislabelled as *B. pertussis*. The blastn option "qcov_hsp_perc" was set to 50, to ensure all counted hits represented at least half a copy of the relevant IS element present in the genome.

Comparing genome arrangement

Nanopore-only assemblies were produced for all strains, using the Flye-Racon-Medaka strategy detailed above. These assemblies were used for genome arrangement comparisons, to ensure that assembly strategy did not contribute to any observed arrangement differences. Closed genome sequences were aligned against each other using progressiveMauve (v20150226 build 10, Darling, Mau and Perna, 2010). The results were manually inspected and grouped into types (within each type, no arrangement differences were visible). A representative of each New Zealand arrangement type was aligned against 29 global strains (**Supplementary Table S3.12**), representing each of the CDC arrangement types defined in Weigand et al. (2017)'s landmark *B. pertussis* genomic arrangement study, in order to place the New Zealand arrangements in a wider global context.

Allelic profile typing

Allele type was assigned to the genes coding for the ACV proteins and the promoter for pertussis toxin (*ptxA-E*, *prn*, *fim2*, *fim3*, *fhaB,* and *ptxP*) using a custom-made MLST scheme with Seemann's MLST tool (v2.16.2, Seemann, 2019), using the hybrid assembled genomes. The commands and custom scheme are available from https://github.com/nataliering/FHA_screening. Instructions for using a custom scheme are available from https://github.com/tseemann/mlst. To make the ACV-gene MLST scheme, all known alleles were downloaded for each gene from PubMLST (Jolley, Bray and Maiden, 2018) and processed using tfa_prepper (https://github.com/nataliering/FHA_screening). The exceptions were *prn* and *fim3*, for which the alleles defined in Bart, Harris, et al. (2014) (supplemental text sd6) were used for consistency, as the nomenclature for both on PubMLST appears to be different.

Prediction of duplications

As in Chapter 2, strains were screened for potential large copy number variants by using BWA MEM (v0.7.15-r1140, Li, 2013) to map raw Illumina reads to the Tohama I reference sequence (NC_002929.2) and assessing the read coverage at every position in the genome using SAMtools depth (v1.8-25-g2c7cd7c, Li et al., 2009).

Table 3.2 IS elements counted in NZ genomes

IS element	NCBI accession
IS *481*	M22031.1
IS *1001*	X66858.1
IS *1002*	NP_881859.1
IS *1663*	NP_881791.1

Prediction of pertactin, pertussis toxin and filamentous haemagglutinin deficiency

The final closed genome sequence for each strain was annotated using Prokka (v1.14.6, Seemann, 2014a), with the Tohama I reference proteins for NC_002929.2 from GenBank as a guide. The resulting annotations for *prn*, *ptxA-E* and *fhaB* were screened for presence/absence, and for insertion of IS *481*, IS *1002*, or IS *1663*. Additionally, the assembled *prn* sequences were screened for the presence of other mutations previously identified in pertactin-deficient strains (Pawloski et al., 2014).

Phylogenies

To place the New Zealand strains in a global context, 73 global strains representing different continents, time periods and allelic profiles were included in the analysis. A further 125 global strains were included in more detailed trees for each of New Zealand's four whooping cough outbreaks since 2004. Illumina reads were downloaded from the NCBI's SRA using fasterq-dump (v2.10.5) and trimmed using Trimmomatic (v0.36) using the options HEADCROP:10, SLIDINGWINDOW:4:20 and MINLEN:30 (see **Supplementary Table S3.11** for full details including accession numbers).

Several different trees were constructed: one containing only the New Zealand strains, one containing the New Zealand strains along with the 73 global strains, and one for each of the four New Zealand whooping cough outbreaks since 2004. Snippy (v4.6.0, Seemann, 2014b) and SNP-sites (v2.1.3, Page et al., 2016) were used to perform variant calling and define the core genome, using the paired-end Illumina fastq files for all test strains, with the Tohama I genome as a reference. A maximum-likelihood phylogeny was inferred using IQ-Tree 2 (v2.0.4, Minh et al., 2020). IQ-Tree's built-in ModelFinder (Kalyaanamoorthy et al., 2017) module was used to select the best model for each dataset (the New Zealand and global trees used K3Pu+F+ASC, the 2004-2006 and 2008-2011 outbreak trees used K3P+ASC, the 2012 outbreak tree used TVMe+ASC, and the 2017-2018 outbreak tree used K2P+ASC). Up to 2,000 bootstrapping steps were conducted, using the built-in Ultrafast bootstrap module (Hoang et al., 2018). The best maximum likelihood tree for each dataset was output in Newick format, which was then visualised using Microreact (Argimon et al., 2016) or iTOL (Letunic and Bork, 2019).

3.6 *Results*

12 isolates can be sequenced per R9.4 RevD flow cell using the rapid library preparation

Although 12 barcodes were available, only 10 barcoded strains were included in the first rapid barcoding (SQK-RBK004) run here. Our previous trials (Chapter 2, Ring et al. 2018) had shown that the yield of a single R9.4 flow cell provided enough coverage of 10 strains, but we had not tried to sequence using more than 10 barcodes. In addition, the previous benchmarking was conducted using the 1D ligation library preparation with native barcoding; the rapid barcoding library preparation used during the sequencing of the New Zealand isolates in this chapter does not include the multiple DNA end-repair and clean-up steps, hence was likely to produce a significantly lower yield. However, the raw yield of this first run using the SQK-RBK004 kit with an R9.4 RevD flow cell was over 20 Gb, which provided over 180X coverage for each of the 10 strains, even after some of the data was lost during demultiplexing. Therefore, during subsequent SQK-RBK004 MinION runs, 12 strains were barcoded and sequenced on each flow cell.

In total, 6 MinION R9.4 RevD flow cells were used to sequence the 66 New Zealand isolates. The amount of usable data, after demultiplexing, varied from 3.94 to 12.76 Gb. This corresponded to coverage between 42X and 484X per isolate, with a mean coverage of 207X. The yield of each run tended to correlate with the number of pores available at the beginning of the sequencing, which reflects the quality of the flow cell. The overall mean read length across all sequencing runs was 6,282 bp. The individual run mean read lengths ranged from 4,908 to 7,917 bp. The coverage and read lengths were sufficient to assemble closed genome sequences for all isolates (see next section).

Using nanopore and Illumina reads, closed genomes were produced for all 66 isolates, but some assemblies were not complete

After basecalling and demultiplexing, the reads for each strain were corrected using Canu. 63 strains were then assembled in hybrid with paired Illumina data; 34 strains were assembled using Flye, and 29 were assembled using Unicycler. The closed genome sequences produced by Flye were polished to improve their accuracy. First, the assemblies were polished using the long reads; 4 successive Racon polishes were conducted, followed by a single long-read polish with Medaka. Finally, the assemblies were polished with Illumina short reads, using 3 successive rounds of Pilon. The Unicycler assembly pipeline already includes Racon and Pilon polishing steps. Illumina data was not available for 3 strains (NZ1, NZ5 and NZ29), hence these were assembled with the Flye-Racon-Medaka strategy, without the Pilon polishes. Between the Flye and Unicycler strategies, closed genome sequences were produced for all strains.

The accuracy of the hybrid assemblies was estimated after each step, by comparison with highly-fragmented ABySS assemblies produced using each strain's Illumina data. The mean identity after Flye long-read-only assembly was 99.60% (n=37), increasing to 99.63% after the Racon and Medaka long-read polishing steps (n=37). Polishing with short reads further increased the mean identity to 99.70% (n=37), which is 0.01% higher than the mean identity of the UK genomes assembled in Chapter 2. The individual identities of each hybrid Flye strain compared to their own Illumina data ranged from 99.68 to 99.72%: the maximum percent identity was therefore higher than the maximum of the UK genomes (99.69%). In total, 17 New Zealand genomes assembled with Flye had a percent identity higher than 99.69%.

The mean identity after Unicycler hybrid assembly was 99.65% (n=26). However, the identity of one strain, NZ60, was noticeably lower than every other strain, at 98.11%. Excluding this strain, the mean identity after Unicycler hybrid assembly was 99.71% (n=25), which is 0.01% higher than the mean Flye identity. The individual identities for each strain assembled with Unicycler ranged from 98.11% to 99.73%, 0.01% higher than the maximum hybrid Flye identity. 24 Unicycler hybrid assemblies had identities higher than the best identity achieved in Chapter 2 (99.69%).

The final closed genome sequences were checked for misassemblies using BUSCO. A newer version of BUSCO was used here than in Chapter 2; instead of checking for 40 universal bacterial orthologues, the newer version of BUSCO checked for 124. For reference, the UK genome sequences from Chapter 2 were also checked with the newer version of BUSCO. The 5 hybrid-assembled UK genome sequences contained 40 complete BUSCOs using the older version of the tool, and all 5 also contained 124 complete BUSCOs using the newer version. They were therefore considered to be 100% complete and largely error-free. However, not all of the hybrid-assembled New Zealand genome sequences contained 124 complete BUSCOs. Every sequence assembled using Unicycler (n=26) was 100%

complete according to BUSCO. The sequences assembled with Flye had a mean completeness of 98.03% (n=37); however, the Flye assemblies polished with NovaSeq data (n=3) were all 100% complete, compared to none of the assemblies polished with the MiSeq data (n=34). This suggests that the higher coverage and/or more consistent read lengths of the NovaSeq data enabled Pilon to polish the incomplete Flye assemblies more effectively. As noted by Watson and Warr (2019), residual uncorrected errors in the incomplete genomes likely resulted in some BUSCOs being fragmented or truncated, meaning they could not be predicted by the Prodigal protein prediction algorithm used by BUSCO.

No strains have any ultra-long copy number variations, although two strains share the same 14 kbp duplication

Trimmed nanopore and Illumina reads were aligned with the Tohama I reference genome (NC_002929.2) and processed as described in Chapter 2 to produce coverage plots, to screen for potential copy number variants (CNVs). The resulting plots can be seen in full in **Supplementary Figures S3.1-S3.7**.

Certain previously observed regions of deletion were observed in all strains. A ~29 kbp region from BP0911 to BP0937 (RD3, **Supplementary Table S3.5**), flanked by copies of IS *481* was previously identified as deleted in all French strains since 1980 and all Finnish clinical strains since 1977 in Caro et al. (2006) and Heikkinen et al. (2007), respectively. The RD3 region was deleted in all New Zealand isolates since 1982 studied here. Likewise, the ~7 kbp region from BP1135 to BP1141 (RD2, **Supplementary Table S3.6**) identified as absent in most French strains and every Finnish strain since 1953 in Caro et al. (2006) and Heikkinen et al. (2007) respectively was also absent in all New Zealand strains here.

One region which fluctuated noticeably between different strains was the region from BP1948 to BP1966 (see **Supplementary Table S3.7** for full details), which was deleted in some strains but not others. This 22.7 kbp region was originally noted by Caro et al. (2006), who named it region of difference 4, “RD4”. Caro et al. saw RD4 in only a small number of isolates belonging to their PFGE subgroup IV-β, however Bouchez et al. (2008) later found the deletion to be present in a wider selection of isolates, including every isolate they studied between 1999 and 2007. A similar pattern was seen here, with only 35.7% (5/14) of strains isolated in the 1980s and 1990s lacking the region, increasing to 90.4% (47/52) of strains isolated in the 2000s and 2010s. Interestingly, four of the five strains lacking the deletion in this second time period were all isolated during the same whooping cough outbreak, between 2004 and 2006. Full details of which genomes contained the RD4 deletion are given in **Supplementary Table S3.9**. **Figure 3.7** shows one strain from the 2004-2006 outbreak, NZ32, which does not show the RD4 deletion, compared with another strain from the strain outbreak, NZ30, which does show the deletion.

The other notable CNV seen in this New Zealand strain cohort became apparent in the same way as the ultra-long duplication in the UK48 genome in Chapter 2: Unicycler was unable to assemble the genomes of NZ35 and NZ36 into single contigs. Instead, a nearly-complete contig of around 4.1 Mb was assembled, which Unicycler labelled as “incomplete”, along with a much smaller contig of around 14 kb, which Unicycler labelled as both “complete” and “circular”. The “circular” label suggested that this 14 kb segment contained the same sequence at the beginning and end, most likely an IS element. The coverage analysis (**Figure 3.8**) for both strains revealed the duplication of a 13.2 kbp region flanked

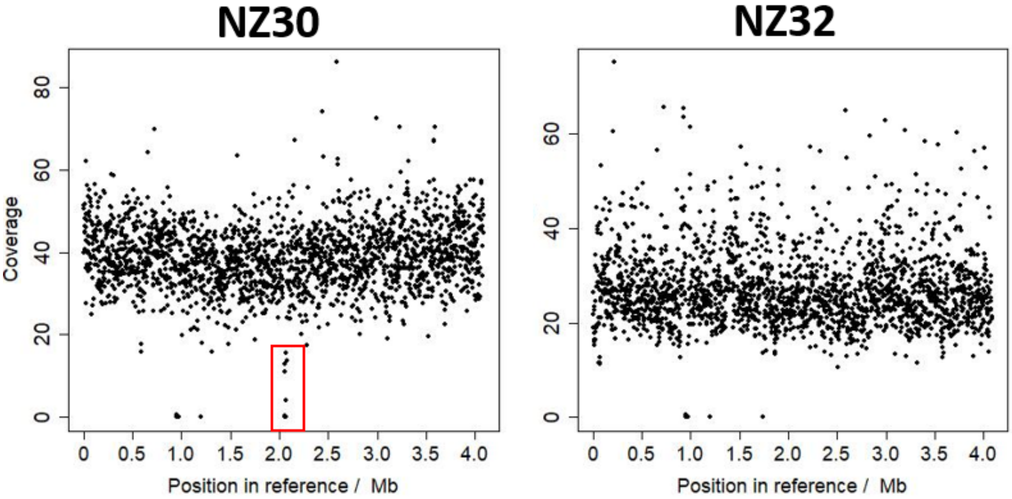


Figure 3.7 Coverage plots for two strains isolated during the 2004-2006 whooping cough outbreak in New Zealand. The RD4 deletion can be seen in the genome of NZ30 (the red box), but not in the genome of NZ32. Plots were constructed by calculating a rolling average of coverage, with a window size of 2000 bp.

by IS *481* (14.3 kbp long in total, including the flanking IS *481*), around 1.8 Mb into the reference genome sequence. This region contains 11 open reading frames (ORFs), including two copies of *cphA* (cyanophycin synthetase), one pseudogene, and a variety of hypothetical exported or membrane proteins. Full details of the region are shown in **Supplementary Table S3.8**. This CNV was previously identified in extremely high copy numbers in UK54, and explored more thoroughly in Abrahams et al. (In review).

Flye was able to assemble the genomes of both strains into single contigs, due to its long-read-centric rather than short-read-centric assembly strategy, so the Flye-Racon-Medaka-Pilon pipeline detailed in the Methods section was used for these two strains instead of Unicycler. NZ35 and NZ36 were isolated in the 2004-2006 outbreak, around a month apart, and in the same health region in New Zealand (Midcentral).

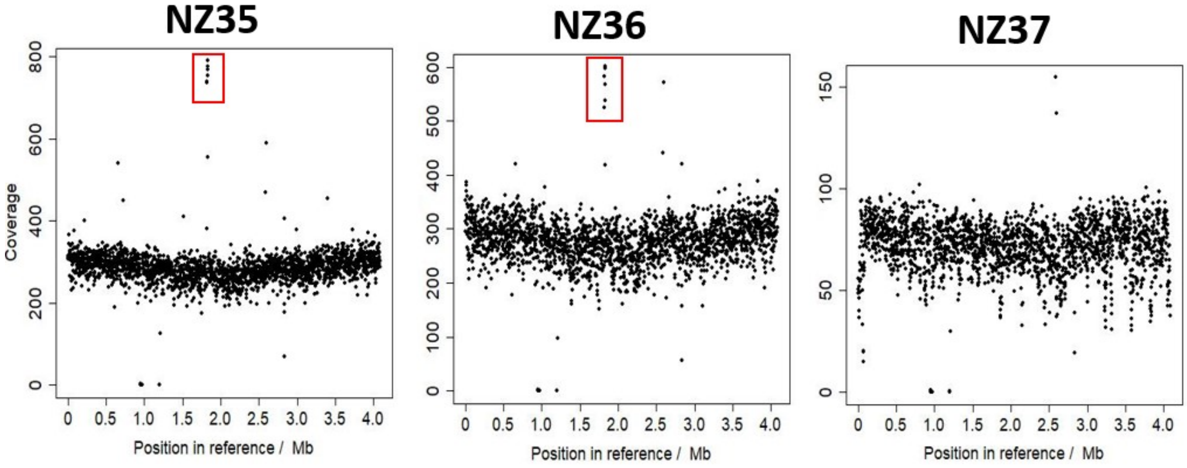


Figure 3.8 Coverage plots from three strains isolated during the New Zealand 2004-2006 whooping cough outbreak. The genomes of NZ35 and NZ36 contain a duplication of a 14.3 kbp region around 1.8 Mb into the reference sequence (indicated with red box). For comparison, the plot for NZ37, which does not contain the duplication, is also shown. Plots were constructed by calculating a rolling average of coverage, with a window size of 2000 bp.

Two clear groups of genome sizes were observed , but no trends in IS numbers were seen from the 1980s to 2018

Assembling closed genomes allowed genome size and exact numbers of each IS element in the 66 strains to be compared (**Figure 3.9**). Genome length varied from 4.101 Mb (NZ33) to 4.134 Mb (NZ27), with a mean length of 4.111 Mb. Two length groups appear to exist; most genomes (n=54) were between 4.101 and 4.114 Mb, but a second group (n=12) were between 4.125 and 4.134 Mb, with no genome lengths falling in between the two groups. Overall, the more recent genomes tended to be found in the "smaller genome" group. Phylogenetic analysis did not show any clear trends between phylogenetic relationships and genome size group (**Figure 3.12**). On the whole, more "larger" genomes were seen at one end of the tree, which correlates with *ptxP1* genomes from earlier years; however, the "larger" genome group were not all found on the same branch of the phylogeny.

IS element copies were counted using blastn, with options which ensured that only copies of IS elements where more than 50% of the sequence was present were counted. The number of IS *481* copies in the 66 closed genomes ranged from 245 to 258, with a mean copy number of 251.6. The copy numbers of IS *1002* and IS *1663* were less variable, ranging from 5 to 7 and 16 to 17, with mean copy numbers of 6.0 and 16.2 respectively. No obvious trend in numbers of IS elements was observed over the time period studied.

In addition, the closed genomes were searched for the presence of IS *1001*, which is specific to *B. parapertussis* and can thus indicate a misidentification of the causative bacterium in whooping cough cases (Parkhill et al., 2003). No copies of IS *1001* were seen in any of the 66 strains studied here, suggesting that all the patients from whom the strains were isolated were indeed infected with *B. pertussis* rather than the rarer *B. parapertussis*.

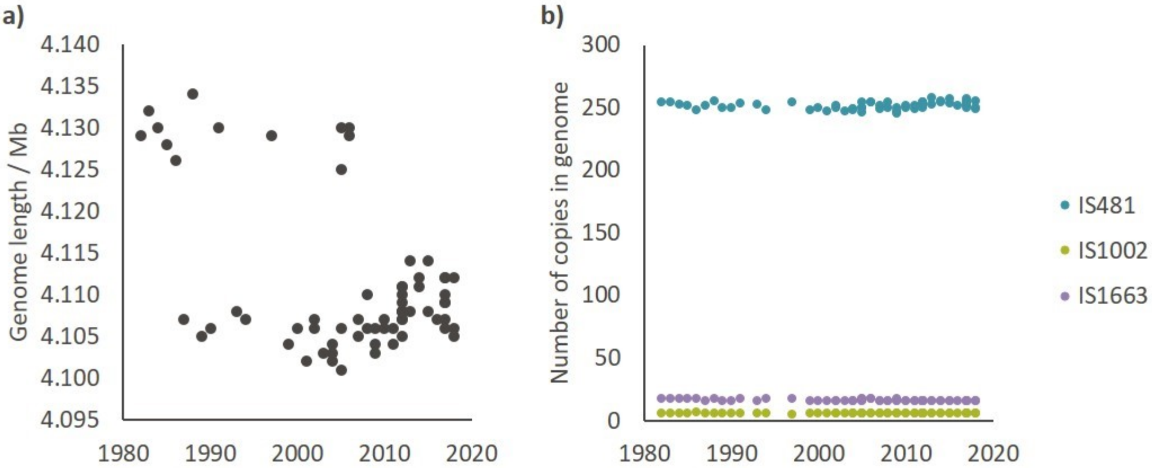


Figure 3.9 Trends observed in a) genome size or b) IS copy number over the time period studied. Genome length ranged from 4.101 Mb to 4.134 Mb, with a mean length of 4.111 Mb, and has appeared to trend downwards overall, with two clear groups of genome sizes (between 4.0 and 4.15 Mb, and between 4.25 and 4.35 Mb). No trends were seen in numbers of IS elements. IS *481* copy number ranged from 245 to 258, with a mean of 251.5, IS *1002* copy number ranged from 5 to 7 with a mean of 6.0, and IS *1663* copy number ranged from 16 to 17, with a mean of 16.2. The gradients of the linear trendlines (not displayed on this figure) were 0.05, -0.002 and -0.02 for IS *481*, IS *1002* and IS *1663* respectively; these essentially horizontal trendlines suggest there has been no ongoing increase or decrease in copy number of these abundant IS elements over the 36-year period studied (1982 to 2018).

19 different genome arrangements were seen in the 66 New Zealand isolates

Closed genome sequences were produced for all 66 New Zealand isolates using the nanopore-only Flye-based assembly pipeline. The genome arrangement of these closed sequences was investigated using progressiveMauve. The isolates were then grouped based on shared arrangement, resulting in 19 different arrangement groups (listed in **Supplementary Table S3.13**. Ten isolates displayed "singleton" arrangements which were not shared with any other isolate; the remaining 56 isolates shared nine arrangements between them (shown in **Figure 3.10a**). As seen in Weigand et al. (2019), the rearrangements generally displayed a pattern of symmetrical inversions around the origin of replication.

One representative of each of the arrangement groups, including all of the singletons, were aligned using progressiveMauve against representatives of each of the arrangement types defined in Weigand et al. 2017 and Weigand et al. 2019 (the representatives are listed in **Supplementary Table S3.11**). This revealed that six of the New Zealand arrangements were congruent with Weigand arrangement types, including one of the New Zealand singletons, which was congruent with the CDC046 arrangement type. 30 New Zealand isolates shared the same arrangement type, CDC010. The remaining isolates were spread more evenly in smaller groups across the other eight arrangement types, as shown in **Figure 3.10b**. Full details of the grouped arrangement types are given in **Supplementary Table S3.13**. CDC010 was the fifth most commonly seen arrangement type in Weigand et al. (2017), after CDC237, CDC002, singletons and CDC013; CDC237, CDC002 and CDC013 were also observed here.

Allelic profile shifts in New Zealand isolates have paralleled those seen elsewhere in the world

MLST was used to identify which alleles of *ptxP, ptxA, prn, fim2* and *fim3* were present in the closed genome sequences for the 63 New Zealand isolates for which Illumina data was available, using allele definitions from PubMLST or as defined in Bart, Harris, et al. (2014). The results are shown in **Figure 3.11**. The most common allelic profile in New Zealand in the WCV era, particularly before the 1990s, was *ptxP1-ptxA1-prn1-fim2-1-fim3-1* (75% of 1980s strains, n=8). This then shifted gradually to *ptxP3-ptxA1-prn2-fim2-1-fim3-1* (90.3% of 2010s strains, n=31) throughout the 1990s, 2000s and 2010s. A brief increase in the circulating proportion of *fim3-2* was seen, although this decreased again by the 2010s.

During the 1990s and 2000s, a number of strains carrying the *prn3* allele circulated, partly related to the 2004-2006 outbreak (60%, n=5). However, by the 2010s, only *prn2* was circulating. Additionally, 44.4% (n=9) of strains circulating during the 2004-2006 outbreak carried the *ptxP1* allele, which had largely been replaced by *ptxP3* throughout the 1990s. Interestingly, three of the four *ptxP1* strains isolated during the 2004-2006 outbreak also carried *prn3*, and shared the same genomic arrangement (arrangement type NZ002). Two of these strains were isolated within a month of each other in the same region (NZ35 and NZ36, Midcentral), so may have been directly related. However, the third strain (NZ37) was isolated nearly a year later, in a different region (Capital and Coast). This suggests that *ptxP1-prn3* may have been a common allelic profile during that outbreak. No other similar phenomena were observed in the 2008-2011, 2012 or 2017-2018 outbreaks.

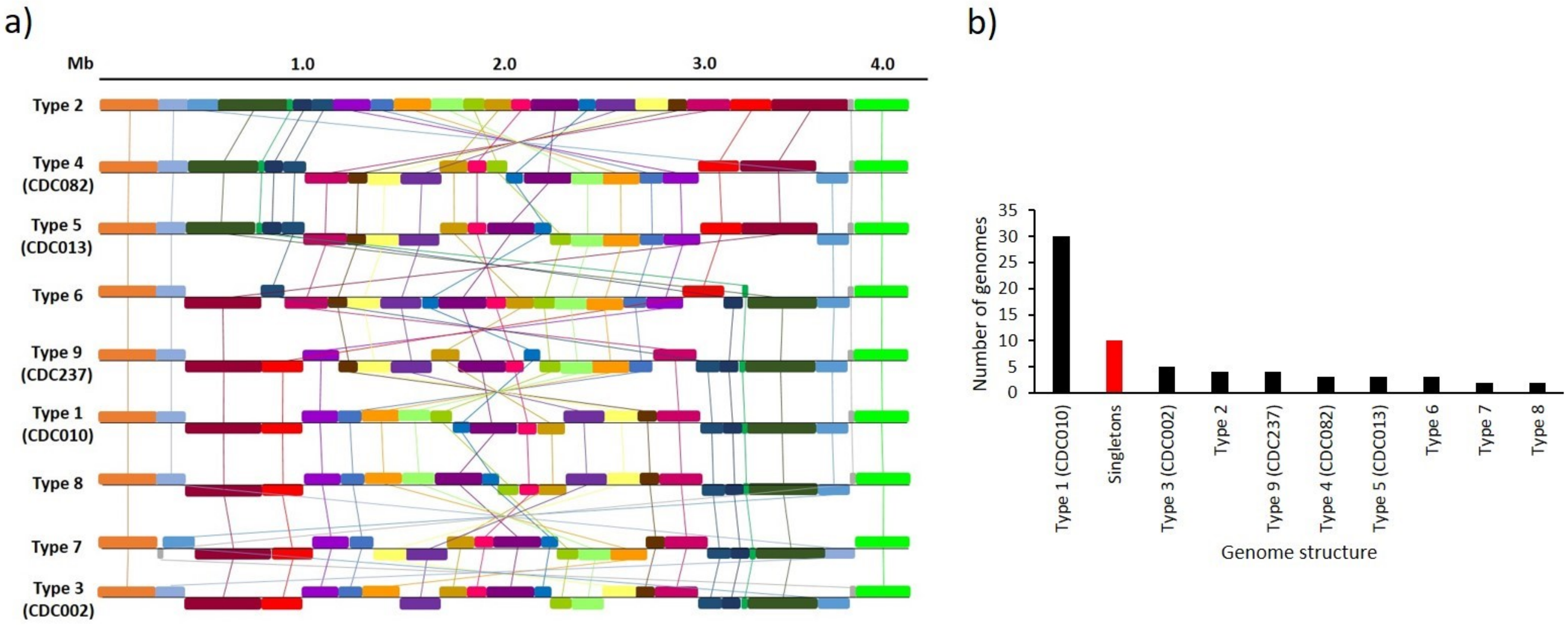


Figure 3.10 Genome structures of New Zealand isolates.

a) 56 of the 66 isolates could be grouped into nine shared genome arrangements. The differences between the arrangement types tended to be inversions (often large) around the central region of the chromosome. Five of the nine arrangement types were found to be congruent with CDC structures defined in Weigand et al. (2017).

b) 30 New Zealand isolates shared the same arrangement type (Type 1/CDC010). Ten isolates had “singleton” arrangements shared with no other isolate (although one of these arrangements was found to be congruent with CDC046). The remaining 26 isolates were split more evenly between the other eight arrangement types.

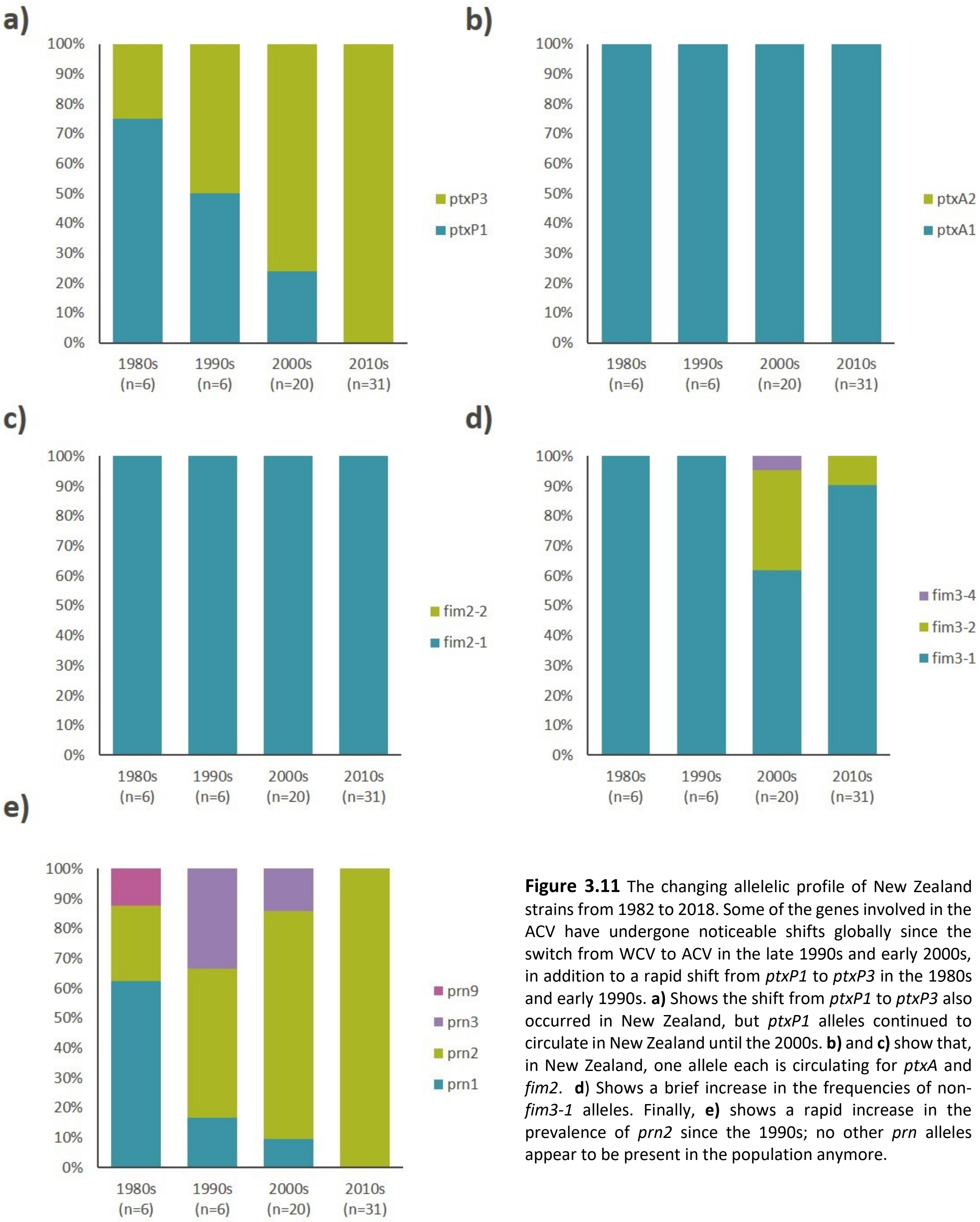


Figure 3.11 The changing allelic profile of New Zealand strains from 1982 to 2018. Some of the genes involved in the ACV have undergone noticeable shifts globally since the switch from WCV to ACV in the late 1990s and early 2000s, in addition to a rapid shift from *ptxP1* to *ptxP3* in the 1980s and early 1990s. **a)** Shows the shift from *ptxP1* to *ptxP3* also occurred in New Zealand, but *ptxP1* alleles continued to circulate in New Zealand until the 2000s. **b)** and **c)** show that, in New Zealand, one allele each is circulating for *ptxA* and *fim2*. **d**) Shows a brief increase in the frequencies of non-*fim3-1* alleles. Finally, **e)** shows a rapid increase in the prevalence of *prn2* since the 1990s; no other *prn* alleles appear to be present in the population anymore.

35% of all strains, and 89% of strains isolated since 2012, were predicted to be pertactin deficient

The closed genome sequences were annotated using Prokka, and the resulting annotations were screened for the presence/absence of *prn*, *fhaB* and *ptxA-E*, as well as for the insertion of IS *481*, IS *1002* or IS *1663* into the same genes.

34.8% (23/66) of all strains were found to have a copy of IS *481* within their *prn* gene, which is known to cause pertactin deficiency. No strain was found to have this insertion prior to 2012, whilst 85.2% (23/27) of all strains isolated between 2012 and 2018 had it. There are three potential IS *481* insertion sites within the *prn* gene, at 240 bp, 1,613 bp, and 2,735 bp (Pawloski et al., 2014). In other studies, such as Weigand et al. (2017), the site most commonly seen to have an IS *481* inserted was at 1,613 bp, appearing in 95 of 114 IS *481* insertion strains in that study, followed by the insertion site at 240 bp, seen in 16 of the 114 strains. Insertion at the 2,735 is rare, appearing in only 3 of the 114 strains in Weigand et al. (2017). Likewise, in the majority (22/23) of the New Zealand strains with an IS *481* insertion in *prn*, it was inserted at position 1,613. The remaining strain had IS *481* inserted at position 240.

One additional strain (NZ63) from 2017 was found to have a mutation previously identified in Weigand et al. (2017) as causing deficiency, a change from C to T at position 223, resulting in a premature stop codon. In total, therefore, 88.9% (24/27) of strains isolated in 2012 or later are predicted to be pertactin-deficient. This suggests that pertactin deficiency was uncommon in New Zealand *B. pertussis* strains until 2012, and has rapidly increased in prevalence since then.

No FHA or PT deficiency caused by IS insertion was predicted in any strain.

New Zealand isolates cluster phylogenetically with global strains according to allelic profile, but isolates circulating during outbreak years appear more closely related than outbreak strains in other countries

A phylogeny was inferred using 325 core SNPs for the 63 New Zealand isolates for which Illumina data was available (**Figure 3.12**). The *ptxP1* strains clustered separately from the *ptxP3* strains. All non-*prn2* strains were included in the *ptxP1* cluster, and all predicted PRN-deficient strains were included in the *ptxP3* cluster. All but one of the predicted PRN-deficient strains were contained on the same sub-branch within the *ptxP3* cluster. The different deficiency-related *prn2* mutations did not cluster together, suggesting PRN deficiency has arisen on multiple different occasions.

Most arrangement types were spread relatively evenly throughout the phylogeny; however, strains with the NZ002 arrangement type clustered together, including the three highly similar strains from the 2004-2006 outbreak. These strains (NZ35, NZ36 and NZ37) were identical in terms of core genome SNPs. Likewise, strains with the arrangement type CDC237 clustered together, and were all isolated during the most recent (2017-2018) outbreak, representing 40% (n=10) of all strains sequenced from that outbreak. These four strains also shared the same mutation in their *prn2* gene (IS *481*-1613fwd), and three of them (NZ57, NZ64 and NZ65) were identical in terms of core genome SNPs. Two of the strains (NZ57 and NZ58) were isolated in the Southern region within the same two week period, whilst the other two strains (NZ64 and NZ65) were isolated in the South Canterbury region several months later, but again within two weeks of each other. 70% (n=10) of the strains isolated during the 2017-

2018 outbreak period clustered together, regardless of genomic arrangement, suggesting that the strains circulating during that outbreak tended to be closely related.

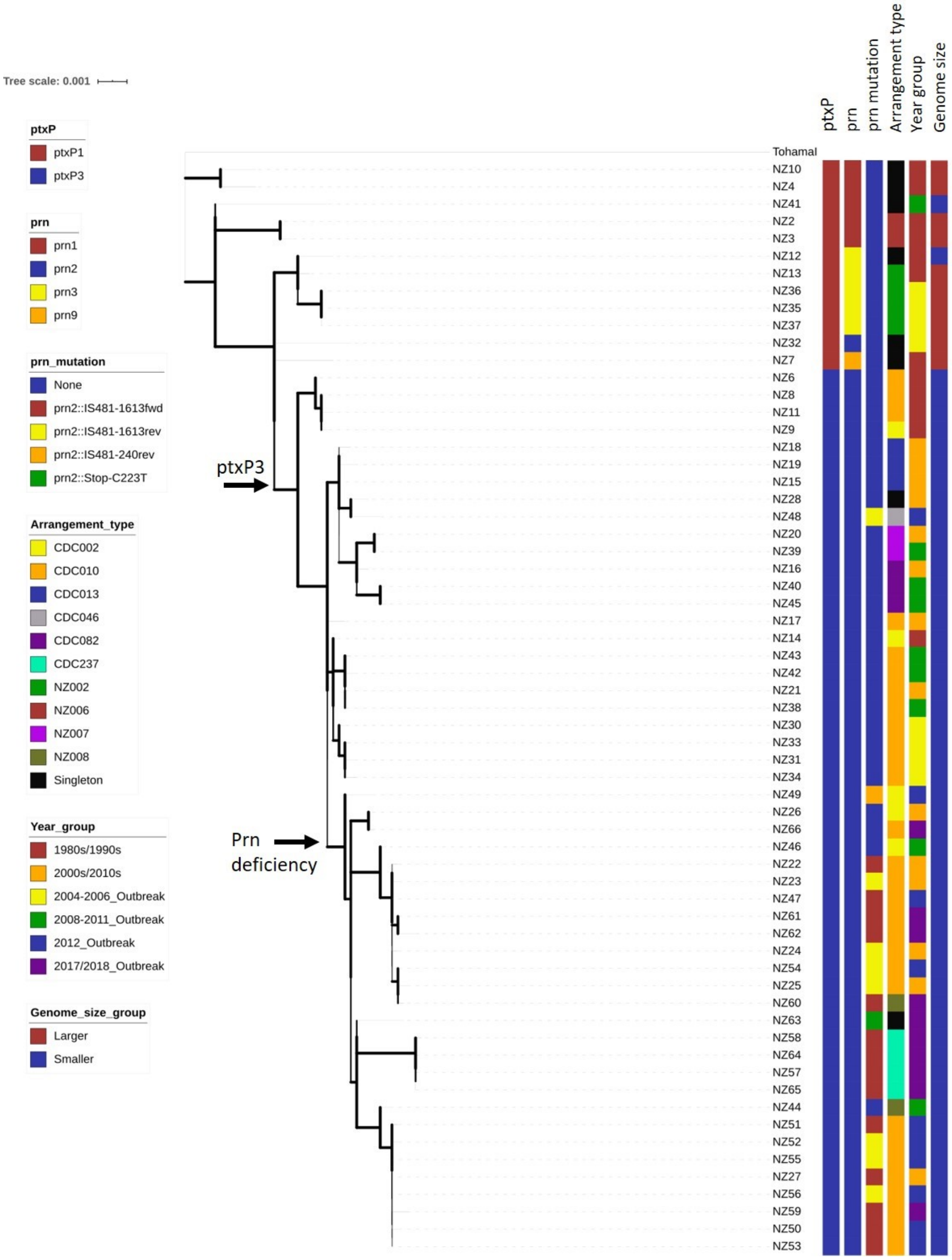


Figure 3.12 Phylogenetic tree showing the evolutionary relationships of the New Zealand strains sequenced here, in the context of allele type, Prn deficiency, genomic arrangement and period of isolation. Snippy was used to identify variants between the NZ strains and the reference, Tohama I, as well as defining a core genome. IQ-Tree2 was used to infer a maximum-likelihood phylogeny, which was then displayed using iTOL (interactive Microreact tree also available at https://microreact.org/project/JG7I7MBBV).

Bootstrap values varied from 10 to 100, and are represented by line thickness (thicker line=higher bootstrap value).

Key branches are highlighted. All ptxP3 strains are contained within one major branch, whilst a slightly smaller sub-branch contains all but one of the predicted Prn-deficient strains.

Other than the 2004-2006 and 2017-2018 outbreaks, strains did not tend to cluster together according to the outbreak or year in which they were isolated. On the whole, many New Zealand strains appear to be very closely related, including a large group of eight strains (NZ51, NZ52, NZ55, NZ27, NZ56, NZ59, NZ50 and NZ53) from throughout the 2000s/2010s, including the 2012 and 2017-2018 outbreaks, which were either identical or separated by only one or two SNPs in their core genome.

To determine whether the New Zealand strains are genetically distinct from those circulating elsewhere in the world during the 1982-2018 period, a further phylogeny was inferred from the 63 New Zealand strains and 73 global strains from the same period, using 511 core SNPs (**Figure 3.13**). Again, all *ptxP1* strains clustered on the same branch, and all *ptxP3* strains clustered on a separate, larger branch. In general, strains isolated earlier (for example, in the 1980s) clustered at one end of the tree, whilst strains isolated later (for example, the 2010s) clustered at the other end.

During periods for which sequencing data was available from many countries, little geographical clustering was seen. For recent years (since 2014), most of the available sequencing data was for strains isolated in the USA and sequenced by the CDC; this gives the impression of geographical clustering, but is more likely simply due to undersampling of strains from other countries.

Certain New Zealand strains do appear to be clustered separately from the majority of the global strains. Group A (**Figure 3.13**) contains ten New Zealand strains (NZ14, NZ43, NZ42, NZ21, NZ38, NZ30, NZ33, NZ31 and NZ34) and one Asian strain (B281), clustered on their own sub-branch, distinct from the closest global neighbours. These nine strains all have the allelic profile *ptxP3-ptxA1-prn2-fim2-1-fim3-1*, all have the RD4 deletion and the CDC010 genomic arrangement type, and none possesses a known *prn2* deficiency mutation. Where serotyping data was available (n=8), the strains were all serotype 1,3 (or 1,2,3). They were isolated between 1999 (NZ14) and 2010 (NZ43), in a variety of different regions around New Zealand (Auckland, Nelson Marlborough, Northland, Southern and Waikato). Four of the strains were isolated during the 2004-2006 outbreak, and three were isolated during the 2008-2011 outbreak.

Group B contains another ten strains clustered on their own branch, including one which was isolated in Australia (L2264). The nine New Zealand strains (NZ62, NZ61, NZ24, NZ47, NZ22, NZ60, NZ25, NZ54 and NZ23) also all have the *ptxP3-ptxA1-prn2-fim2-1-fim3-1* genotype and the RD4 deletion. They all possess a *prn2* mutation: five have the IS *481*-1613fwd mutation, and four have the IS *481*-1613rev mutation. All but one have the CDC010 arrangement type; the remaining strain (NZ60) has the NZ008 arrangement type. Where serotyping data was available (n=6), the strains were all serotype 1,2 (or 1,2,3). The strains were all isolated since 2011, and five of them were isolated during outbreaks (2008-2011, 2012 and 2017-2018).

Group C is another group of ten strains, containing nine New Zealand strains (NZ56, NZ27, NZ55, NZ53, NZ59, NZ52, NZ51, NZ50 and NZ44) and one Australian strain (L2239). Like the other groups, and most recent strains, this group has the *ptxP3-ptxA1-prn2-fim2-1-fim3-1* genotype and the RD4 deletion. All but one of the strains (NZ44) has a *prn2* mutation, with five instances of IS *481*-1613fwd and four of IS *481*-1613rev. Like Group B, eight of the New Zealand strains share the CDC010 arrangement type and the ninth (NZ44) has the NZ008 type. All were isolated since 2010, and eight were isolated during outbreaks (one from 2008-2011, six from 2012, and one from 2017-2018). Where tested (n=8), the strains all had the serotype 1,3.

Finally, Group D contains four New Zealand strains (NZ65, NZ57, NZ64 and NZ58). Three are identical in terms of core genome SNPs, and the fourth (NZ58) differs by only a single SNP. These are the same four strains previously described using the New Zealand phylogeny, which again all have the same

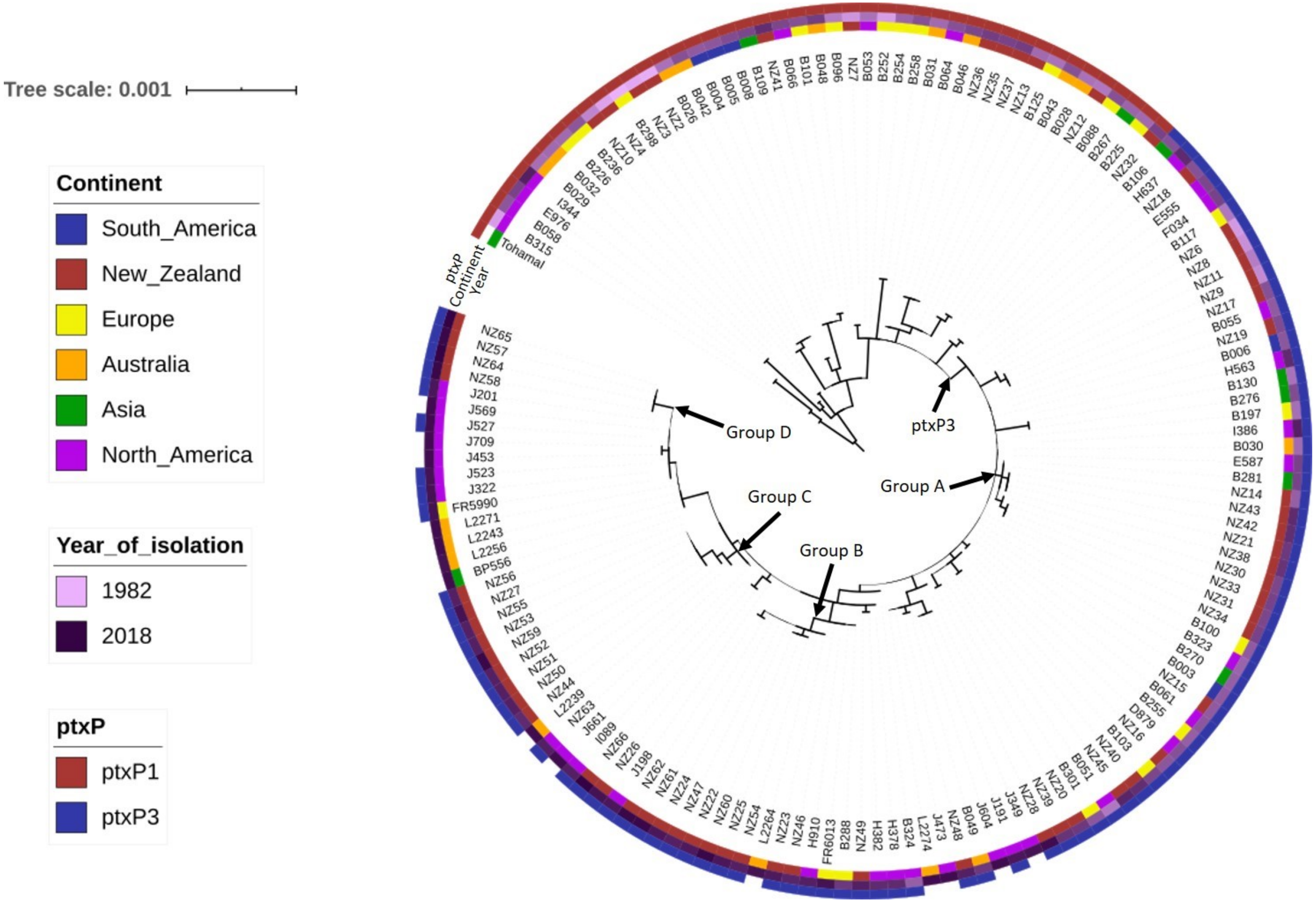


Figure 3.13 Phylogenetic tree showing the evolutionary relationships of the New Zealand strains compared with a selection of global strains from the same time period. Variants between the selection strains and the Tohama I reference were called, and a core genome was defined, using Snippy. A maximum-likelihood phylogeny was inferred using IQ-Tree2, and iTOL was used to display the resulting tree.

Like the New Zealand strains alone, all strains carrying the *ptxP3* allele* are found on the same major branch, separate from the *ptxP1* strains.

Several groups containing highly similar strains, almost all from New Zealand, are indicated.

Bootstrap values varied from 1 to 100, and are represented by line thickness (thicker line=higher bootstrap value).

*Allele information was not available for some of the more recent global strains.

(*ptxP3-ptxA1-prn2-fim2-1-fim3-1*) allelic profile. Serotype was not tested for any of them, but they all have the RD4 deletion and the IS *481*-1613fwd *prn2* mutation. As previously mentioned, they all have the CDC237 arrangement type, and were isolated in two geographically contemporaneous pairs.

Groups A-D in the global phylogeny suggest that certain strains circulating in New Zealand in the 2000s and 2010s are more closely related than those circulating in other countries, particularly during New Zealand whooping cough outbreak periods. To investigate whether this effect was simply due to oversampling of New Zealand strains in the global phylogeny, smaller phylogenies were inferred for each of the four post-2000 New Zealand whooping cough outbreaks, with additional global strains included for each outbreak period.

Figure 3.14 shows the focussed phylogenies for a) the 2004-2006 outbreak and b) the 2008-2011 outbreak. The New Zealand strains are spread across several different branches in the 2004-2006 phylogeny, which was constructed from 378 core SNPs. Nonetheless certain strains are closely related; NZ35, NZ36 and NZ37, as mentioned previously, are identical in terms of core SNPs, and were not all isolated at the same time or location. However, on the same branch as these three New Zealand strains are two Australian strains. The Australian strains are not identical to the New Zealand strains (although they are identical to each other), but differ from the New Zealand strains by only two SNPs. Similarly, NZ32 and L508 (also from Australia) are identical to each other, as are two other Australian strains, B048 and L518. Overall, this phylogeny suggests that certain strains circulating during the 2004-2006 outbreak were relatively clonal in both New Zealand and Australia, but the outbreak was not caused by a single strain in either country.

The picture is less clear for the 2008-2011 phylogeny, which was inferred from 408 core SNPs. Again, some of the New Zealand and Australian strains have clustered separately from the strains from elsewhere in the world. Three Australian strains (L1214, L1423 and L1419) have identical core genome SNPs, and one New Zealand strain (NZ46) differs by only one SNP. Likewise, there are three New Zealand strains from different regions (NZ38, NZ43 and NZ42) which are, at most, only one SNP apart. However, on the whole, the New Zealand strains from the 2008-2011 outbreak are less closely related to each other than those from the 2004-2006 outbreak, and there is less of a clear similarity with the Australian strains.

Figure 3.15 shows the focussed phylogenies for the a) 2012 outbreak and b) the 2017-2018 outbreak. The 2012 phylogeny, inferred from 320 core SNPs, shows 60% (n=10) of the New Zealand strains (NZ52, NZ50, NZ53, NZ51, NZ55 and NZ56) form a clear group, again with some Australian strains (L1661, L1779 and L1780). Four of the New Zealand strains have identical core genome SNPs, and the others are one SNP different each. The Australian strains are not identical to each other, but each is only two SNPs different from the New Zealand strains. The remaining New Zealand strains cluster throughout the rest of the phylogeny, with strains isolated in the UK or USA.

Finally, the phylogeny showing strains from the most recent outbreak, which was constructed from 323 core genome SNPs, indicates a slightly different picture again. As previously mentioned, four of the New Zealand strains (NZ57, NZ65, NZ64 and NZ58) have identical core genomes; these strains cluster on their own branch. The remaining New Zealand strains, however are less similar to each other, and cluster on branches with strains from Australia and/or the USA. At the same time, several USA strains (J733, J811, J878, K256 and K298) isolated over the same period are identical or only a single SNP different in their core genomes.

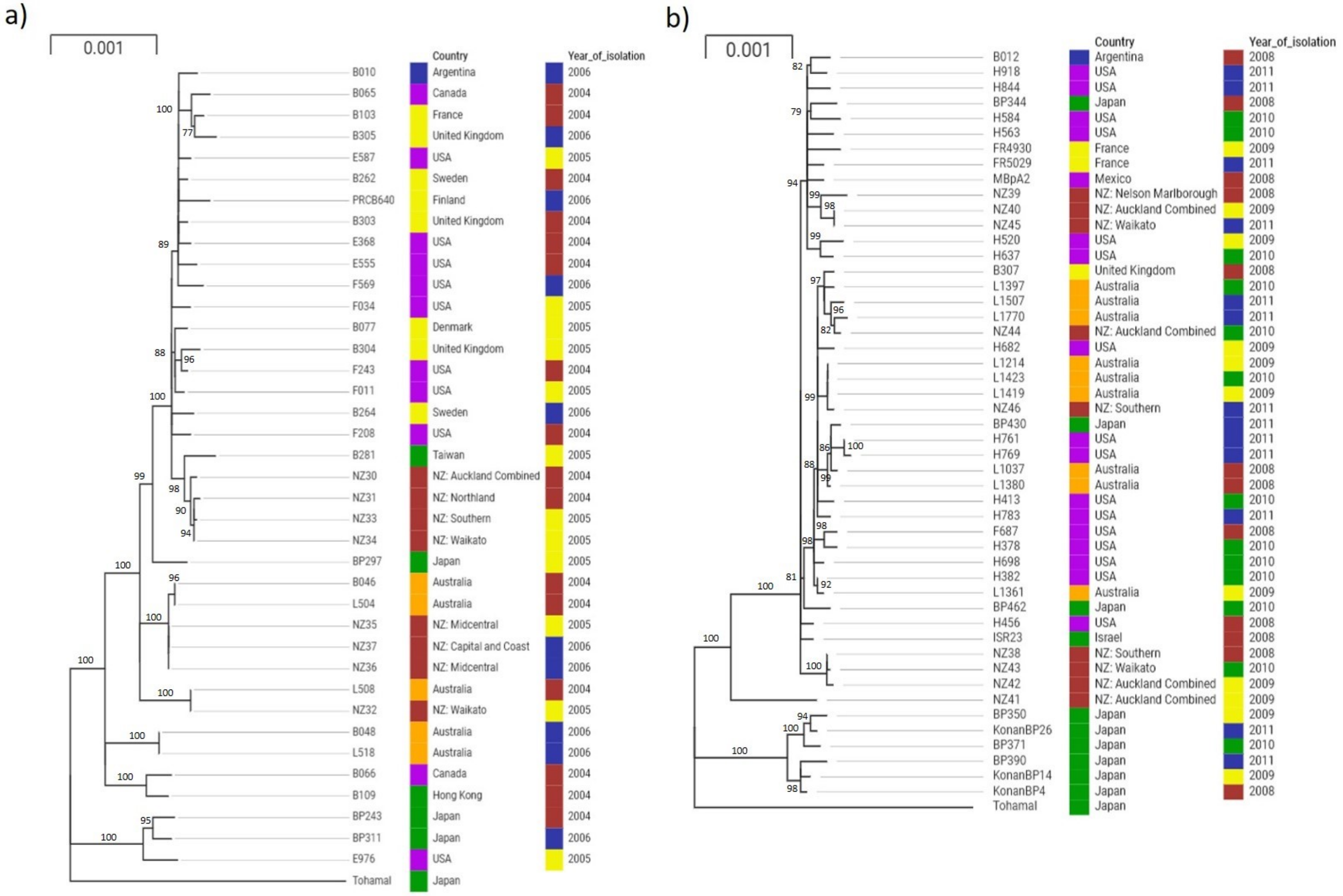


Figure 3.14 Phylogenetic trees showing the evolutionary relationships between strains isolated in New Zealand during the a) 2004-2006 whooping cough outbreak and b) 2008-2011 whooping cough outbreak, compared with global strains isolated during the same time periods. Variants were called and core genomes defined using Snippy, then phylogenies were inferred using IQ-Tree2 and visualised using IToL and Microreact (interactive trees available at https://microreact.org/project/8F3Ilz2Bi and https://microreact.org/project/UWL4KNRaN). Bootstrap values are shown for major branches. Branches containing only single strains tended to have lower bootstrap values.

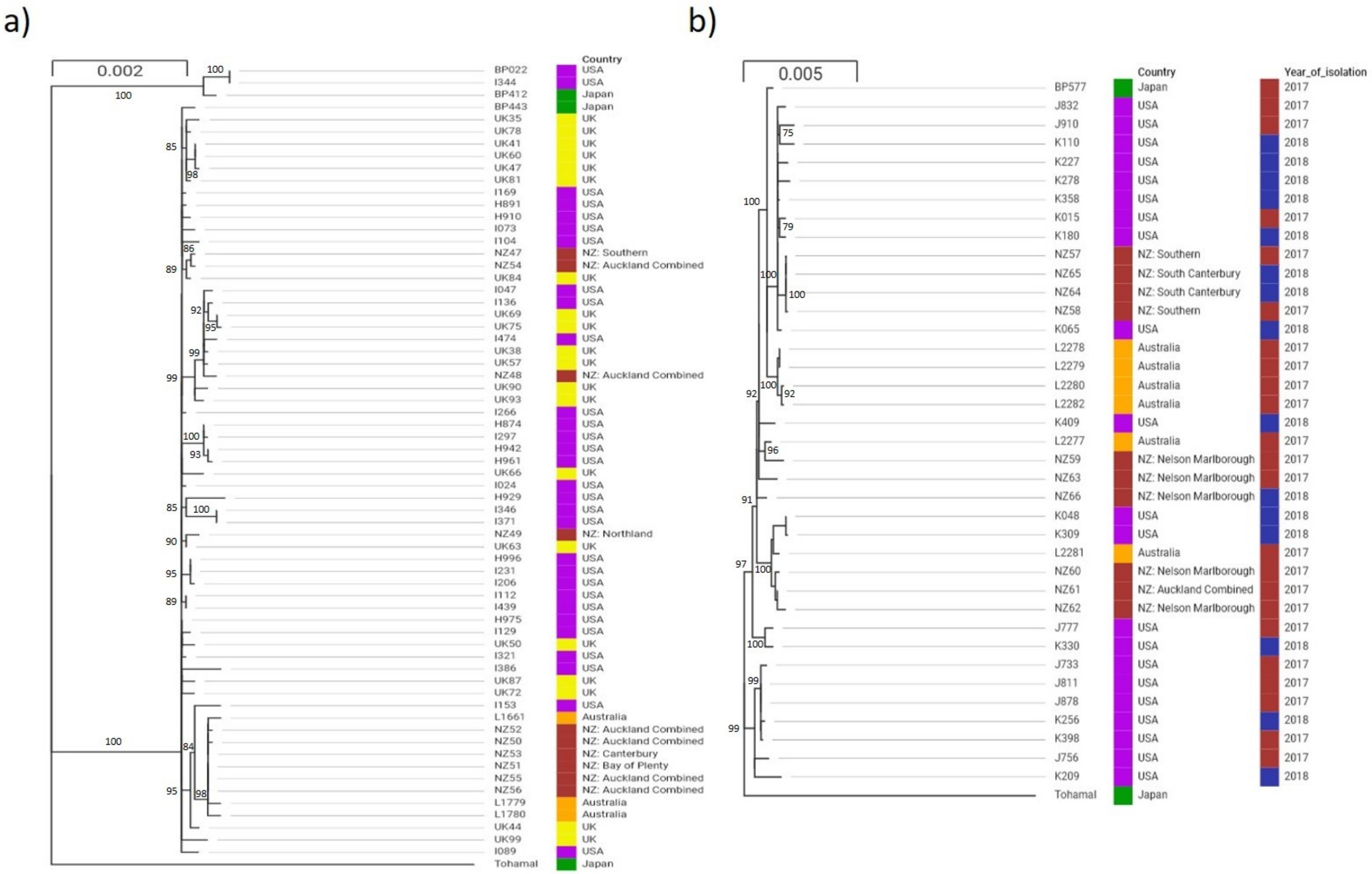


Figure 3.15 Phylogenetic trees showing the evolutionary relationships between strains isolated in New Zealand during the a) 2012 whooping cough outbreak and b) 2017-2018 whooping cough outbreak, compared with global strains isolated during the same time periods. Variants were called and core genomes defined using Snippy, then phylogenies were inferred using IQ-Tree2 and visualised using Microreact (interactive trees available at https://microreact.org/project/yR6WukJhF and https://microreact.org/project/tbl0zUnck). Bootstrap values are shown for major branches. Branches containing only single strains tended to have lower bootstrap values.

3.7 *Discussion*

New Zealand strains were investigated for a number of reasons. Historically, vaccine uptake in New Zealand was lower than in many other countries although, thanks to efforts to monitor and increase rates of uptake since 2008, vaccine coverage is now comparable to that in the UK. However, despite the increasing vaccine coverage, whooping cough incidence has remained higher in New Zealand than in most other countries, including the USA and England (**Figure 3.2**). An outbreak occurred in New Zealand from 2017 to 2019, for example, a period when no other countries noted a similar increase in cases. Although isolates have been collected and stored by New Zealand's ESR since the 1980s, no sequencing of any *B. pertussis* from New Zealand had previously been conducted. Therefore, the genomes of 66 isolates from between 1982 and 2018 were sequenced here, using the hybrid nanopore and Illumina pipeline developed in Chapter 2, to determine the following: whether vaccine-driven global genomic trends observed in *B. pertussis* had been delayed by New Zealand's slower vaccine uptake; whether the strains circulating during whooping cough outbreaks were polyclonal, as seen in recent outbreaks in the UK and USA; and whether the consistently higher incidence of whooping cough in New Zealand could be explained by the circulation of a unique hyper-virulent strain.

Allelic profile, genome size and antigen deficiency trends in New Zealand generally match those seen elsewhere in the world

King et al. (2010) surveyed the genomes of global strains for a 60-year period, between 1949 and 2008. Overall, they found a downward linear trend in both genome size and number of genes, suggesting that *B. pertussis* is a species which is seeing ongoing genome reduction (Ring et al., 2019). Applying a linear trend-line to **Figure 3.9a** reveals a similar trend in the genome size of the 66 strains isolated in New Zealand between 1982 and 2018. The largest genome observed by King et al. was 4.124 Mb, and the smallest was 3.937 Mb. Amongst the New Zealand genomes analysed here, the largest was 4.134 Mb and the smallest was 4.101 Mb; at face value, the New Zealand strains would therefore appear to be larger. However, King et al. used a genome size approximation method, estimating the length of each genome based on gene content from CGH data and extrapolating from the length of the Tohama I genome. It is likely that this method produced genome size estimates which were less accurate than the assembly of complete, closed genome sequences. The main difference in sizes was likely due to numbers of repetitive elements like IS elements being underestimated by the CGH estimation method. The grouping of genome sizes observed here was also seen in King et al. (2010), with a small group of larger genomes and a large group of smaller genomes. The size difference is not explained by a single large deletion or insertion, by the larger genomes clearly containing more copies of IS *481*, IS *1002* or IS *1663*, or by the larger genomes all belonging to the same phylogenetic family. A brief review of the lengths of all closed *B. pertussis* genomes available from the NCBI's RefSeq database (downloaded for Chapter 5, see **Supplementary Table S5.1**) reveals the same pattern: most genomes are no longer than around 4.115 Mb and a small group have genomes between 4.125 and 4.135 Mb, but no genomes have been seen with a length between the two groups. This genome size grouping trend may be due to undersampling (only 524 closed genomes from RefSeq had the required year of isolation metadata available), or may be a previously unstudied *B. pertussis* genomic phenomenon.

Genome reduction in *B. pertussis*, like genome rearrangement, is largely mediated by recombination between IS elements. Indeed, the speciation of *B. pertussis* from its *B. bronchiseptica*-like ancestor is thought to have resulted from the very large increase in IS *481* copy number in the *B. pertussis* genome. However, no trend in numbers of IS elements was observed here (**Figure 3.9b**); the number of copies of IS *481* has remained stable at around 250, numbers of IS *1002* remain around 6 copies, and numbers of IS *1663* copies are still around 16/17, suggesting that despite extensive recombination between IS elements, copy number is not currently undergoing increase or decrease, or is changing too slowly for any change to be apparent over the 40-year period studied here.

Just as the trend in genome size of New Zealand strains matches the trend seen in other global strains, analysis of the allelic profile of the ACV-related genes also reveals a similar pattern in the New Zealand strains to that seen in Bart, Harris, et al. (2014)'s landmark study of global strains throughout the 20th and early 21st century, as well as other studies of how *B. pertussis* populations have changed over recent years, such as Sealey et al. (2015), Xu, Y. et al. (2015), Zomer et al. (2018). The most common global allelic profile during the "WCV era" in Bart, Harris et al. (defined as 1960-1995) was found to be *ptxP1-ptxA1-prn1-fim2-1-fim3-1*. The most common allelic profile of the New Zealand strains from 1982-1995 was the same, and was present in 50% (n=12) of the strains screened; the other 50% of strains were split between those carrying *ptxP3*, and different *prn* alleles. In the Bart, Harris et al. "WCV/ACV era" and "ACV era" (post-1995), the most prevalent allelic profile shifted to *ptxP3-ptxA1-prn2-fim2-1-fim3-1*. An increased frequency of strains carrying the newer *fim3-2* allele was also noted (from <1% frequency in the WCV era to 37% in the ACV era), with *ptxP3-ptxA1-prn2-fim2-1-fim3-2* being observed as the dominant profile in the late 2010s (Kallonen and He, 2009). Again, the most common allelic profile seen in the New Zealand strains throughout the period matched that seen elsewhere in the world.

Noted shifts, such as *ptxP1* to *ptxP3* and *prn1* to *prn2,* appear to have happened in New Zealand along similar timescales as the rest of the world. Additionally, the noted increase in the prevalence of the deletion of RD4 (**Figure 3.7**) has happened over the same timescale as it was detected in Caro et al. (2006) and Bouchez et al. (2008). Interestingly, although the progression of *fim3-2* in the New Zealand strains reflects that seen in Bart, Harris et al. (2014) throughout the WCV and early ACV eras, from 0% of strains in the 1980s and 1990s to 30.4% (7/23 strains) in the 2000s, a noticeable decline in the frequency of this allele can be seen in the strains from the 2010s (to 9.7%, 3/31 strains, **Figure 3.11**). The most recent strains in the Bart, Harris et al. screen were isolated in 2010; it is therefore possible that a similar decrease in prevalence would be seen in more recent global strains, rather than being a New Zealand-specific phenomenon.

Another shift observed globally in the ACV era has been the rapid increase in strains which do not express functional pertactin protein. In Australia, for example, the percentage of PRN-deficient strains increased from 5% to 78% over the four year period between 2008 and 2012 (Lam et al., 2014). A study of strains from European countries isolated between 1998 and 2015 showed a clear correlation between how early each country introduced the ACV and the current proportion of PRN-deficient strains (Barkoff et al., 2019). The ACV was introduced in New Zealand in 2000, several years before some European countries, including the UK. Although the first PRN-deficient strains did not appear in the New Zealand cohort studied here until 2012, later than in some other countries, the proportion of PRN-deficient strains rapidly increased; 88.9% of strains since 2012 are PRN-deficient, including 90% of those isolated in 2017 and 2018. This frequency is higher than any of the European strains included in Barkoff et al.'s study, although it is similar to the frequencies observed in Australia (78%) and the USA (85%) (Lam et al., 2014; Martin et al., 2015). The majority of the predicted PRN-deficient New Zealand strains (95.8%, n=24) have a copy of IS *481* in their *prn* gene, and one (NZ63) possesses a SNP which results in a premature stop codon. Other mechanisms for PRN-deficiency have been observed in other studies (for example, Pawloski et al., 2014), such as a commonly seen inversion of 22 kb containing the *prn* promoter, but were not seen here. These other mechanisms for PRN-deficiency may not have been identified from sequence alone; the percentage of PRN-deficient strains in New Zealand may therefore be even higher than predicted. Expression or non-expression of antigens such as PRN, PT and FHA can usually be tested *in vitro*, for example by western blot, but only heat-killed cells were available here. Nonetheless, it should be noted that the use of hybrid genome assembly allowed the prediction of PRN-deficiency caused by the most common deficiency mutations from sequence alone.

Overall, New Zealand's historically slower uptake of the whooping cough vaccine does not seem to have significantly delayed the most commonly observed recent genomic changes, unlike other slow-uptake countries such as China (Xu, Y. et al., 2015).

Are the *B. pertussis* strains circulating in New Zealand more clonal than in most other countries?

Over the last decade, a variety of global *B. pertussis* phylogenies have been produced, including the landmark Bart, Harris et al. (2014) study. Some of the key patterns from these phylogenies, for example the clear branching of *ptxP1* from the *ptxP3* strains, were also observed in the global phylogeny here (**Figure 3.13**), with the New Zealand isolates clustering in the same way as the sequences from the rest of the world. Another discovery of Bart, Harris, et al. (2014) was that there was very little geographic clustering of strains; strains from all around the world were spread throughout the phylogeny. Likewise, Sealey (2015)'s investigation of UK strains showed little geographic specificity. In addition, studies of outbreaks such as those by Sealey et al. (2015) and Bowden et al. (2014) showed that strains circulating during outbreaks were polyclonal, and not characterised by a single hypervirulent outbreak strain. Conversely, an outbreak in Australia in the late 2000s and early 2010s was found to be caused primarily by the circulation of a group of highly related strains (Octavia et al., 2012).

Although some of the New Zealand isolates cluster throughout the global phylogeny along with strains from other countries, some notable groups (labelled Groups A-D) of highly similar New Zealand isolates stand apart. Groups A-C each contain nine New Zealand strains and one strain each from either Australia (Groups B and C) or Asia (Group A), all of which are identical in terms of core genome, or only one or two SNPs different. No other groups of this size, containing such similar strains, are seen in strains from any other country in the global phylogeny. Interestingly, the highly similar strains were not always isolated during the same year, and often not even during the same outbreak, although usually the majority of them were from outbreak periods (70% of Group A were isolated during the 2004-2006 or 2008-2011 outbreaks, 50% of Group B were isolated during the 2008-2011, 2012 or 2017-2018 outbreaks, and 80% of Group C were isolated during the 2008-2011, 2012 or 2017-2018 outbreaks), and all are from the post-2000 era. In each group, the strains are not only similar in terms of core genome SNPs, but also in serotype (where tested), RD4 presence or absence, PRN-deficiency and, almost always, genomic arrangement. This suggests that it is not just a coincidence that the core genome SNPs are congruent, but that the genomic features of these strains are highly similar or identical as a whole.

Group D contains four identical New Zealand strains, all from the 2017-2018 outbreak. The four strains can be separated into two pairs which were isolated in the same region and within the same timeframe. Although this suggests that the similarities seen in this cluster could be explained by transmission of the same strain between patients, the fact that the two pairs were isolated in different regions from each other, and several months apart, could also indicate that the 2017-2019 outbreak, which was only observed in New Zealand, may have in part been the result of the circulation of a particularly virulent strain around the country. This theory is challenged by the fact that the other six isolates sequenced from the outbreak did not cluster with Group D; however, sequencing of more strains from the currently undersampled outbreak could be worthwhile.

In case the clustering of New Zealand strains on the global tree was due to overrepresentation of New Zealand compared to any other country, further phylogenies were constructed for the four major outbreaks since 2000, each including a greater number of global strains from the outbreak period. As New Zealand's closest neighbour, it seemed possible that strains from Australia would show most similarity to the strains from New Zealand, so extra Australian strains were included where possible. Indeed, in each of the outbreak trees, strains from Australia clustered most closely with the New Zealand strains, particularly during the 2004-2006 and 2012 outbreaks; as seen in the global tree, some of the Australian strains had highly similar or identical genetic characteristics to some of the New Zealand strains. This similarity with Australian strains supports the theory that outbreaks in New Zealand tend to be less polyclonal than those in the USA or UK, but simultaneously disagrees with the idea that New Zealand outbreaks are caused by hypervirulent strains which are unique to the country.

Overall, the phylogenies suggest that, whilst the outbreaks in New Zealand are not caused by a single hypervirulent strain, there may be a higher than usual proportion of very similar or identical strains circulating in the country. These strains have been isolated during non-outbreak periods, although more often during outbreaks. It is possible that these groups of strains are more virulent than most *B. pertussis*, as hinted at by their overrepresentation during outbreak periods. Their circulation during non-outbreak years could therefore help to explain New Zealand's seemingly permanent higher incidence of whooping cough than elsewhere in the world. However, such a phenomenon has not been observed in *B. pertussis* before, and there is no evidence in genetic screen here for enhanced virulence, with the clusters all showing allelic profiles and other characteristics which are in line with the global trends.

The apparent trend for strain clusters seen here could be due to the small sample size, with only 63 New Zealand strains, spanning nearly 40 years, included in the phylogenies. The phylogenies are based on shared core genome SNPs, and thus do not include all of the SNPs in each strain. It is therefore possible that the clustered strains look identical in their core genome, but are relatively dissimilar in the parts of the genome not included in the core phylogeny, although the other similarities identified here, such as genome arrangement, serotype and *prn* mutation, do suggest these strains are still closely related. Whether the clustered strains are identical or not, the global and outbreak trees show evidence that the New Zealand strains are more similar to each other than to strains from other countries, which is unusual. Not all New Zealand strains fall into clusters, however; some cluster throughout the phylogeny with strains from other countries, as seen with the UK strains in Sealey (2015). With sequencing of additional strains from New Zealand, it is possible that a different picture will emerge, with the majority of strains falling into the global spread and only a minority forming clusters. More strains have been collected and are stored by the ESR, but cost and other practical considerations limited the number which could be transported and sequenced here.

Further improvements to nanopore sequencing and data analysis

One of the key findings from Chapter 2 was that, as nanopore sequencing library preparation kits, flow cells and data analysis tools continue to improve, sequencing pipelines will need to remain flexible. The sequencing here took place approximately 2 years after the sequencing runs in Chapter 2; in that time, flow cell yield had improved further, new library preparation kits had become available, and new, more accurate, data analysis tools had been developed. With that in mind, the rapid barcoding library preparation kit (SQK-RBK004) was used here, in place of the native barcoding ligation kit (SQK-LSK109 with EXP-NBD104) used in Chapter 2.

The rapid barcoding library preparation process takes a fraction of the time of the ligation library preparation, and does not require most of the third-party reagents used in the ligation protocol (for example, NEB's end-repair enzymes), thus is also a fraction of the cost. In turn, the rapid kit tends to produce a lower sequencing yield; without the thorough DNA clean-up and repair steps of the ligation kit, fewer DNA strands make it to and through the nanopore intact. Additionally, the rapid kit uses a transposon-based mechanism to attach the sequencing adaptors and barcodes to the DNA strands, which can result in much shorter read lengths than the ligation kit (Heron, 2018). However, word-of-mouth communication at nanopore community conferences suggested that yields of >10 Gb per flow cell were common with the rapid kit, if a high starting mass of DNA was used, and that read lengths could be maximised by skipping any DNA-shearing steps (Personal Communication, Nanopore Community Meeting 2018/London Calling 2019). Consequently, the rapid kit was trialled for the first sequencing run here, using 10 barcodes. Unsheared DNA was barcoded, and the entire pooled library was cleaned up and concentrated into the final sequencing mix using Promega ProNex SPRI beads, which were recommended as a cheaper and potentially more efficient alternative to the Agencourt AMPure XP beads used previously (P. Kover, Personal Communication). The yield of this run was over 20 Gb, and the N50 was 14 kbp, both higher than the values achieved in Chapter 2. This method was therefore repeated for the remaining 56 strains, increasing the number of barcodes used to 12 for most of the runs.

Similarly, some new data analysis tools were also used here. The Albacore basecaller used in Chapter 2 was officially retired by ONT at the Nanopore Community Meeting in November 2018 (Oxford Nanopore Technologies, 2018b). Its reportedly more accurate successor, Guppy, was used here instead. Likewise, the demultiplexing tool used in Chapter 2, Porechop, became "officially unsupported" and was replaced with Deepbinner in October 2018 (Wick, 2017; Wick, Judd and Holt, 2018b). Guppy and Deepbinner were therefore used in tandem to demultiplex the barcoded runs here, as Albacore and Porechop were in Chapter 2. Porechop was still used here, after basecalling and demultiplexing, to trim the adaptor and barcode sequences from the final read sets. The newer basecalling tools mean that the consensus accuracy of *de novo* assemblies has continued to improve, even since the analysis in Chapter 2. 41 of the hybrid assemblies here had a higher percentage identity than the most accurate assembly from Chapter 2 (99.69%), and 17 had the same percentage identity; in total, 92.1% (n=63) of the hybrid New Zealand assemblies were therefore at least as accurate, or more accurate, as the best UK hybrid assembly from Chapter 2.

Summary

The results of this chapter contribute to a growing body of knowledge surrounding the genomics of *B. pertussis*, particularly during whooping cough outbreaks, and how changes have occurred over time, influenced by the introduction and alteration of vaccines. Whilst concrete conclusions cannot be drawn about the genetics of the circulating *B. pertussis* population in New Zealand from the sample of 66 strains, the isolates sequenced here represent the first from New Zealand to be sequenced and may help to understand the country's ongoing unusually high incidence of whooping cough. In addition, this study demonstrates for the first time the utility of nanopore sequencing in conducting rapid and affordable *B. pertussis* strain screens, in combination with Illumina sequencing, particularly given the ongoing improvements to the technology's throughput and accuracy.

Chapter 4: Investigating a common, ultra-long, genomic duplication

Parts of this chapter has been adapted from:

"Duplications drive diversity in the pathogen *Bordetella pertussis* on an underestimated scale".

Jonathan Abrahams, Michael Weigand, **Natalie Ring**, Iain MacArthur, Scott Peng, Margaret Williams, Barrett Bready, Anthony Catalano, Jennifer Davis, Michael Kaiser, John Oliver, Jay Sage, Stefan Bagby, M. Lucia Tondella, Andrew Gorringe and Andrew Preston

Genome Biology

In Review, December 2019

"There is a proverbial saying chiefly concerned with warning against too closely calculating the numerical value of un-hatched chicks."

- Neil Gaiman, Stardust

(AKA, the negative results chapter)

4.1 *Data summary*

1. Ultra-long nanopore reads are available from Figshare: https://figshare.com/s/ceef04fc644025a4638f
2. "*In silico* resolved" UK48 genome is downloadable from Figshare: https://figshare.com/s/81f6a33f86add71a625c

4.2 *Introduction*

Some UK *Bordetella pertussis* strains have unusual read mapping coverage

Two of the *B. pertussis* strains sequenced during the initial nanopore pipeline development described in Chapter 2 were found to have unusual genomic features. A nanopore read set with a mean read length of over 5,000 bp was insufficient to produce a closed genome sequence for UK48. Instead, two contigs were produced: one 3.9 Mbp long and one 0.2 Mbp long. Further investigation revealed that this genome appears to contain a duplication of a region around 0.2 Mbp long, an "ultra-long" duplication. Surprisingly, we also found that the genome of UK76, which was closed using our standard nanopore reads, seemed to contain a potential duplication around the same region, around 0.3 Mbp long. These duplications were identified through mapping sequencing reads to the *B. pertussis* reference sequence, Tohama I, and calculating the depth of coverage at each base position. The expected result was relatively level coverage across the entire reference sequence, whereas the duplications appeared as enhanced coverage, as seen for UK36 in **Figure 4.1**. We also noted that duplications at the same region, which contains many motility-related genes, had previously been noted in a small number of other studies (Caro et al., 2006; Heikkinen et al., 2007; Weigand et al., 2016; Weigand et al., 2018).

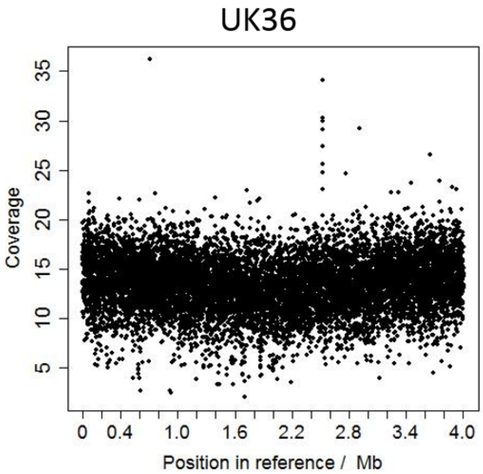


Figure 4.1 Illumina short reads for Bordetella pertussis UK36 mapped to the Tohama I reference genome. Coverage is steady across the whole genome, apart from one small section around 2.6 Mb (this section represents two genes which are found in duplication in a large number of non-Tohama strains (Abrahams et al., In review)).

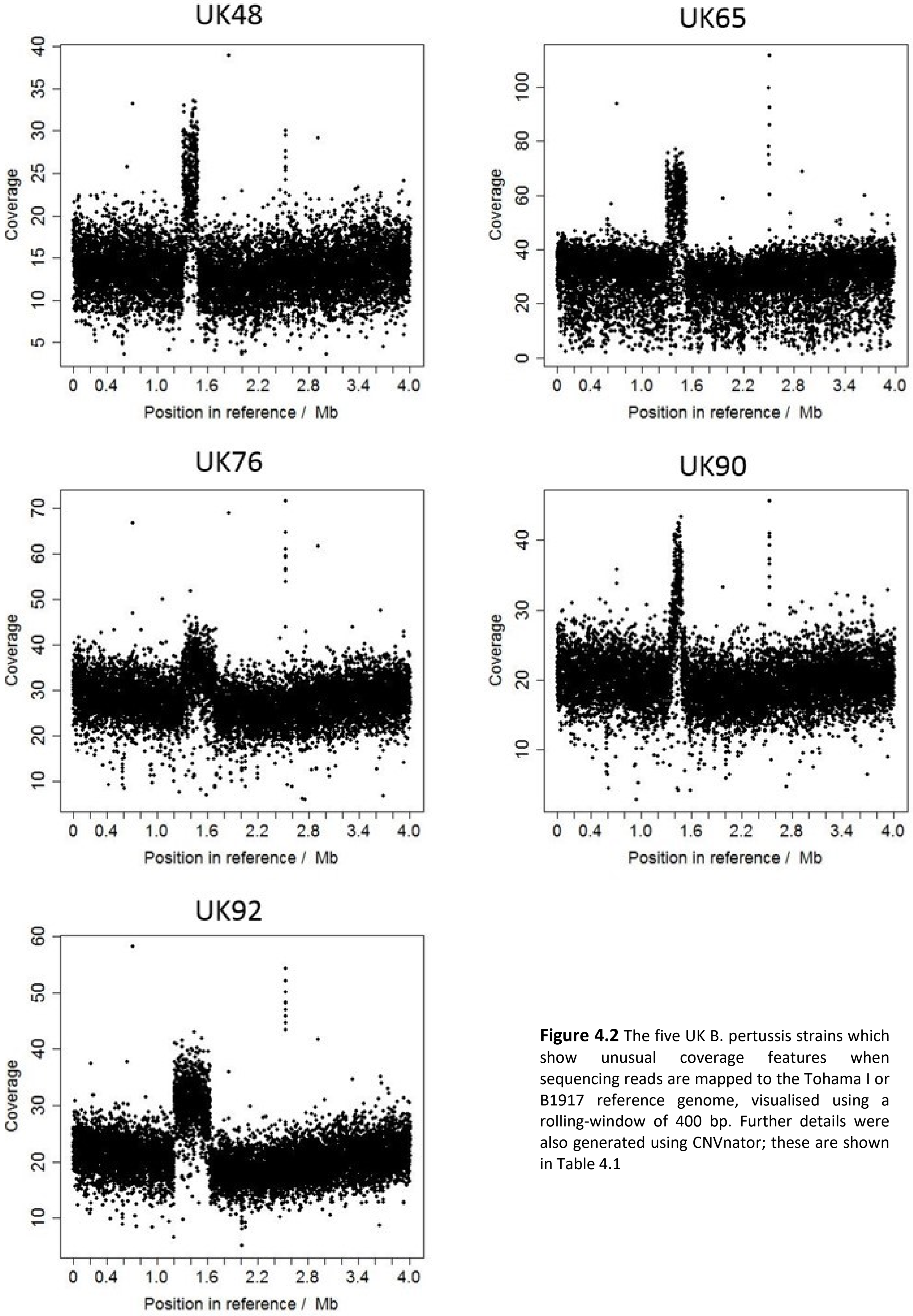


Figure 4.2 The five UK B. pertussis strains which show unusual coverage features when sequencing reads are mapped to the Tohama I or B1917 reference genome, visualised using a rolling-window of 400 bp. Further details were also generated using CNVnator; these are shown in Table 4.1

Work by Abrahams et al. (In review) aiming to catalogue duplications in all Illumina-sequenced *B. pertussis* strains available from the NCBI database showed that almost one hundred other strains from around the world also had a genomic duplication at the same locus as UK48 and UK76. In Abrahams et al., a pipeline is developed which uses CNVnator (Abyzov et al., 2011) to identify copy number variants by mapping Illumina reads to a more recent clinical reference strain, B1917. The application of this pipeline to all available high-quality *B. pertussis* Illumina sequencing data from the NCBI database showed that 94 strains share copy number anomalies around the same locus, with a shared core duplicated region of 71,000 bp, centred around the *Bordetella* motility locus (BP1366-BP1411). Five of these strains were isolated during the UK 2012 whooping cough outbreak, including UK48 and UK76 (**Figure 4.2**). **Table 4.1** shows further details for all five strains, including CNVnator-predicted start and end bases of the CNV compared to the Tohama I reference genome, length of the predicted CNV, and the depth of the coverage enhancement seen. 119 genes are common to all five strains, from BP1308 to BP1450, spanning 148 kbp. **Supplementary table S4.1** shows the full details of the genes in the shared region.

The enhanced coverage for UK48 and UK92 is a clear 1.9x and 1.7x higher respectively than for the rest of the reference genome, suggesting that this portion of the genome is duplicated in these strains. The remaining strains, however, have less clear coverage enhancements. CNVnator shows the coverage of the potentially duplicated sections in UK65, UK76 and UK90 as a range; the range is centred around 2.0 in UK65 and UK90, and around 1.5 in UK76. This unusual distribution may be caused by the second copy of the region containing variants such as SNPs, thus preventing the mapping of reads to this region. Alternatively, and more likely in the case of UK76, the duplication may not be carried by all cells in the population: in UK76, half the cells carrying and half the cells not carrying the duplication would explain the observed 1.5x enhancement in coverage. It should be noted that CNVnator indicates a higher coverage enrichment in the UK76 reads than a simple visual review of the coverage graphs in **Figure 4.2**; the UK76 coverage graph suggests an enrichment of 1.25x compared to the 1.5x enrichment assigned by CNVnator. CNVnator's more thorough method of identifying CNVs, which includes optimisation of rolling-window size to ensure cleaner estimations of start and end bases, means that the true level of coverage enrichment is more likely to be closer to 1.5x than 1.25x.

Short (150 bp) Illumina sequencing reads are not able to resolve the duplications in either UK48 or UK76 during *de novo* assembly, and nor were nanopore sequencing reads with a mean length of 5,000 bp. Theoretically, to resolve a long repetitive region, reads longer than the longest repeat will likely be required (Chin et al., 2013; Conlan et al., 2014; Koren and Phillippy, 2015; Loman, Quick and Simpson, 2015; Wick et al., 2017a; Jain, Koren, et al., 2018; Jain, Olsen, et al., 2018; Schmid et al., 2018). This chapter details attempts to characterise the ultra-long duplication shared by these five UK *B. pertussis* strains, via assembly of fully resolved genome sequences, followed by investigation of likely phenotypic consequences of the duplication.

Table 4.1 Five UK *B. pertussis* strains identified as having unusual reference genome coverage

Strain	SRA accession	Start / bp*	End / bp*	Length / kbp	Depth
UK48	ERR212388	1351001	1525000	174	1.9
UK65	ERR304786	1342201	1562400	220	1.9 - 2.3
UK76	ERR316415	1377076	1681400	304	1.4 - 1.5
UK90	ERR305956	1376401	1524800	148	1.6 - 1.9
UK92	ERR305958	1242151	1673490	431	1.7

*The start and end position of the duplication in the reference genome was estimated here using by CNVnator with the Tohama I genome sequence (NC_002929.2) as the reference

4.3 *Research questions*

Research question 1: Can ultra-long reads, longer than the predicted duplications observed in several UK *B. pertussis* strains, be used to assemble closed genome sequences which contain two full copies of the duplicated region?

Research question 2: Does the commonly observed duplication centred around the *Bordetella* motility locus result in observable phenotypic changes to duplication-carrying *B. pertussis* strains?

4.4 *Methods*

Bacterial strains and culture conditions

The strains used in this chapter are shown in **Table 4.2**. Strains were stored at -80°C in PBS/20% glycerol prior to growth. Unless otherwise stated, *B. pertussis* strains were grown on charcoal agar (Oxoid) for 72 hours at 37°C or cultured in Stainer-Scholte (SS) broth with supplement (recipes shown in **Tables 4.3** and **4.4**) at 37°C with shaking at 180 rpm. *B. bronchiseptica* strains were grown on charcoal agar for 24 hours at 37°C or cultured in SS broth with supplement at 37°C with shaking.

Ultra-long nanopore sequencing

Prior to gDNA extraction, *B. pertussis* isolate UK48 was stored at -80°C in PBS/20% glycerol at the University of Bath. Cells were grown for 72 hours at 37°C on charcoal agar (Oxoid) plates. Harvested cells were resuspended in 10 ml SS broth with supplement, to an OD_{600} of 0.1 and grown overnight. At approximately OD_{600} 1.0, cultures were further resuspended in 50 ml SS broth with supplement, to an OD_{600} of 0.1 and grown again overnight. Ultra-long gDNA was then extracted from UK48 using the standard ultra-long extraction protocol developed by Josh Quick (Quick, 2018).

After growth to OD_{600} 1.0, all 50 ml of culture was pelleted by centrifugation at 13 000 g for 5 minutes and taken forward to ultra-long gDNA extraction. The full protocol is available from dx.doi.org/10.17504/protocols.io.mrxc57n (Quick, 2018), but is essentially a phenol-chloroform gDNA extraction which aims to avoid pipetting and/or vigorous mixing where possible. Briefly, the pellet was resuspended in 200 µl PBS and 10 ml TLB (100 mM NaCl, 10 mM Tris-HCl pH 8.0, 25 mM EDTA pH 8.0,

Table 4.2 Strains used in this work

Strain	Description	Reference
B. pertussis Tohama I	WT	Parkhill et al. (2003)
B. pertussis UK48	Duplication-carrying	Sealey et al. (2015)
B. pertussis UK65	Duplication-carrying	Sealey et al. (2015)
B. pertussis UK71	Non-duplication-carrying control, similar to UK48	Sealey et al. (2015)
B. pertussis UK76	Duplication-carrying	Sealey et al. (2015)
B. pertussis UK90	Duplication-carrying	Sealey et al. (2015)
B. pertussis UK92	Duplication-carrying	Sealey et al. (2015)
B. bronchiseptica RB50	WT	Cotter and Miller (1994)
B. bronchiseptica RB53	Phase-locked Bvg(+) derivative of RB50	Cotter and Miller (1994)
B. bronchiseptica RB54	Phase-locked Bvg(-) derivative of RB50	Cotter and Miller (1994)

Table 4.3 Stainer-Scholte Broth. Chemicals were dissolved in 900 ml distilled H_2O, adjusted to pH 7.6 with NaOH and/or HCl, topped up to the final volume (1 L) with distilled H_2O, and filter-sterilised. SS broth was stored at 4°C or room temperature.

Reagent	Molecular weight / g mol $^{-1}$	Final concentration / mM	Mass / g
Casamino acids			10
L-glutamic Acid (Na salt)	187.7	57	10.70
L-proline	115.1	2.1	0.24
NaCl	58.44	43	2.5
KH2PO4	136.1	31	0.5
KCl	74.56	2.7	0.2
MgCl2.6H2O	203.3	0.49	0.1
CaCl2.2H2O	147.0	0.18	0.0265
Tris base	121.14	50	6.1
Heptakis (if required, for *B. pertussis* only)			1

Table 4.4 Stainer-Scholte supplement. 10 ml of 100X supplements were used per 1 L SS broth, adding immediately prior to use. Chemicals were dissolved in distilled H_2O and filter-sterilised. SS supplement was stored at -20°C.

Chemical	Molecular weight / g mol $^{-1}$	Final concentration (in 1x)	g in 10 ml
L-Cysteine	121.2	0.33 mM	0.04
$FeSO_4.7H_2O$	278.0	36 μM	0.01
Niacin	123.1	33 μM	0.004
Glutathione	307.3	0.49 mM	0.15
Ascorbic acid	176.1	2.3 mM	0.40

0.5% w/v SDS, 20 µg ml^{-1} Qiagen RNase A), vortexed for 5 seconds at full speed and incubated for 1 h at 37°C to lyse the cells.

After lysis, 100 µl proteinase K (Qiagen) was added (final concentration of 200 µg ml^{-1}), mixed by slowly rotating three times, and incubated for 2 h at 50°C, with gentle mixing by rotation every 30 minutes. The lysate was split between two 15 ml Falcon tubes containing light phase-lock gel (VWR), 5 ml TE-saturated phenol (Sigma Aldrich) was added, and the tubes were mixed for 10 minutes at 20 rpm on a HulaMixer (Invitrogen) followed by centrifugation for 10 minutes at 4500 rpm. The top, aqueous phase, was then poured into two new 15 ml Falcon tubes also containing phase-lock gel, avoiding the transfer of any protein layer. A further 2.5 ml TE-saturated phenol and 2.5 ml chloroform-isoamyl alcohol (Sigma Aldrich) were added, and the tubes were mixed on a HulaMixer for 10 minutes at 20 rpm followed by centrifugation for 10 minutes at 2,400 g. The aqueous phases from both tubes were collected by pouring slowly into a 50 ml Falcon tube.

4 ml 5 M ammonium acetate (Sigma Aldrich) was added, followed by 30 ml ice-cold ethanol to precipitate the DNA. After the DNA floated to the top of the tube, it was spooled out gently, using a hook crafted from a melted glass rod. The spooled DNA was submerged in 70% ethanol to wash and tighten the pellet whilst still on the hook, then worked off the hook into a 1.5 ml Eppendorf tube. Two washes were performed as follows: 1 ml 70% ethanol was added to the Eppendorf, the tube was spun for 1 minute at 10,000 g, and the ethanol was removed. The remaining ethanol was allowed to evaporate for 15 minutes at room temperature with the lid open, 100 µl buffer EB (10 mM Tris-HCl pH

8.0, 0.02% Triton-X100) was added, and the DNA was left at 4°C for at least a week to resuspend. The resuspended DNA was then quantified via Qubit and taken forward to rapid (SQK-RAD004) library preparation.

The concentration of the ultra-long gDNA was adjusted to 1 µg µl^{-1} and left overnight at 4^{O}C to resuspend. 1.5 µl fragmentation mix (FRA, ONT) and 3.5 µl EB were added, and mixed by very slowly pipetting up and down 8 times using P20 pipette set to 18 µl and P20 tip with the end cut off to reduce shearing. After mixing, the mixture was incubated for 1 minute at 30^{O}C and 1 minute at 80^{O}C. 1 µl rapid adaptor (RAP, ONT) was added, and mixed by very slowly pipetting up and down 8 times using a P20 pipette set to 19 µl and the cut-off tip from before.

The library was incubated at room temperature whilst a flow cell was prepared as per ONT's instructions. To load the library, 34 µl sequencing tether (SQT, ONT) and 20 µl nuclease-free H_2O were added, and the library was mixed by very slowly pipetting up and down 5 times using a P200 set to 75 µl and P200 tip with the end cut off. On the final mix, the library was slowly pipetted drop-wise onto the flow cell.

Ultra-long gDNA nanopore sequencing libraries were sequenced for 48 hours on a GridION sequencer using R9.4 flow cells. Reads were basecalled with Nanopore's Guppy basecaller (V2.1.3), using the "fast" Flip-flop model.

De novo assembly of ultra-long reads

A variety of *de novo* assembly tools optimised for long reads were trialled with different combinations of the ultra-long reads produced for UK48, alone, in combination with the standard nanopore long reads from Chapter 2, and in hybrid with the Illumina reads from Chapter 2. The tools trialled were Canu (v1.7.1), Flye (v2.3.5), MaSuRCA (v3.2.7), SPAdes (v3.12.0), Unicycler (v0.4.6) and Wtdbg2/Redbean (v2.0) (Bankevich et al., 2012; Koren et al., 2017; Wick et al., 2017b; Zimin et al., 2017; Kolmogorov et al., 2019; Ruan and Li, 2019). Read sets trialled included the full set of ultra-long UK48 reads, 40X coverage of the UK48 genome in ultra-long reads corrected by Canu and ultra-long reads longer than 100 kbp, 174 kbp and 348 kbp (the latter being the predicted length of a single copy and tandem copies of the duplicated region, respectively). See **Table 4.5** for full set of trials.

Each assembly was assessed for contiguity. If a single closed sequence had been assembled, copy number of several genes predicted to be located throughout the duplicated region (shown in **Table 4.6**) was checked using blastn.

Validating the duplication without *de novo* assembly

In addition to validating the presence of the UK48 duplication through *de novo* assembly of the ultra-long nanopore sequencing reads, an alternative strategy was developed for *in silico* assembly of closed genomes. This strategy, shown in **Figure 4.3**, requires a single contig assembly of the genome to have been achieved, albeit with only a single copy of the duplicated region, such as the hybrid UK48 assembly produced in Chapter 2. It assumes that the duplications occur in tandem. First, CNVnator was used to predict the start and end bases of the duplicated region, allowing the whole sequence of the duplicated region to be extracted from the reference genome using SAMtools faidx (Li et al., 2009). Due to the ongoing inter-strain genome rearrangements in *B. pertussis*, the duplication was unlikely to be in the same genomic position in UK48 as in the reference genome. Thus, the location of the duplicated region in UK48 was found using blastn to search for the duplication sequence in the draft,

Table 4.5 Assembly strategies trialled using UK48 ultra-long reads

Read correction	Assembly	Hybrid?	Read subset
None	Canu	No	All ultra-long
None	Flye	No	All ultra-long
None	MaSuRCA	No	All ultra-long
None	Unicycler	No	All ultra-long
None	Wtdbg2	No	All ultra-long
None	Canu	No	All ultra-long plus standard long reads from Chapter 2
None	Flye	No	All ultra-long plus standard long reads from Chapter 2
None	MaSuRCA	No	All ultra-long plus standard long reads from Chapter 2
None	Unicycler	No	All ultra-long plus standard long reads from Chapter 2
None	Wtdbg2	No	All ultra-long plus standard long reads from Chapter 2
Canu	Canu	No	Ultra-long, 40X Canu-corrected
Canu	Flye	No	Ultra-long, 40X Canu-corrected
Canu	MaSuRCA	No	Ultra-long, 40X Canu-corrected
Canu	Unicycler	No	Ultra-long, 40X Canu-corrected
Canu	Wtdbg2	No	Ultra-long, 40X Canu-corrected
None	Canu	No	Ultra-longs longer than 100 kbp
None	Flye	No	Ultra-longs longer than 100 kbp
None	MaSuRCA	No	Ultra-longs longer than 100 kbp
None	Unicycler	No	Ultra-longs longer than 100 kbp
None	Wtdbg2	No	Ultra-longs longer than 100 kbp
None	Canu	No	Ultra-longs longer than 174 kbp (predicted duplication length, one copy)
None	Flye	No	Ultra-longs longer than 174 kbp (predicted duplication length, one copy)
None	MaSuRCA	No	Ultra-longs longer than 174 kbp (predicted duplication length, one copy)
None	Unicycler	No	Ultra-longs longer than 174 kbp (predicted duplication length, one copy)
None	Wtdbg2	No	Ultra-longs longer than 174 kbp (predicted duplication length, one copy)
None	Canu	No	Ultra-longs longer than 348 kbp (predicted duplication length, both copies in tandem)
None	Flye	No	Ultra-longs longer than 348 kbp (predicted duplication length, both copies in tandem)
None	MaSuRCA	No	Ultra-longs longer than 348 kbp (predicted duplication length, both copies in tandem)
None	Unicycler	No	Ultra-longs longer than 348 kbp (predicted duplication length, both copies in tandem)
None	Wtdbg2	No	Ultra-longs longer than 348 kbp (predicted duplication length, both copies in tandem)
Canu	SPAdes	Yes	All ultra-longs plus Illumina from Chapter 2
Canu	Unicycler	Yes	All ultra-longs plus Illumina from Chapter 2
Canu	SPAdes	Yes	40X Canu-corrected ultra-longs plus Illumina from Chapter 2
Canu	Unicycler	Yes	40X Canu-corrected ultra-longs plus Illumina from Chapter 2
None	SPAdes	Yes	Ultra-longs longer than 174 kbp (predicted duplication length, one copy) plus Illumina from Chapter 2
None	Unicycler	Yes	Ultra-longs longer than 174 kbp (predicted duplication length, one copy) plus Illumina from Chapter 2
None	Canu	No	Only ultra-longs spanning the duplicated region
None	Flye	No	Only ultra-longs spanning the duplicated region
None	MaSuRCA	No	Only ultra-longs spanning the duplicated region
None	Unicycler	No	Only ultra-longs spanning the duplicated region
None	Wtdbg2	No	Only ultra-longs spanning the duplicated region

Table 4.6 Genes used to assess copy number of the duplicated region in assembly drafts

Gene Identifier	Gene name
BP1280	*pyrroline-5-carboxylate reductase*
BP1301	*glutathione S-transferase*
BP1339	*cardiolipin synthase*
BP1373	*Flagellar basel body rod protein, flgB*
BP1425	*1-deoxy-D-xylulose 5-phosphate reductoisomerase*

non-duplication-resolved, assembly sequence. This also allowed the exact sequence of the duplicated region in UK48 to be identified, in case of SNPs between UK48 and the reference. Using the sequence and locus of the duplicated region, a second copy of the region was inserted immediately after the original, creating an "*in silico* resolved" genome containing both copies of the duplication. To check that the artificially resolved genome was correct, the ultra-long reads were mapped to it. If reads mapped to the entire genome sequence, including the unique junction that would only exist if two copies of the same region were present in tandem, the sequence was accepted as correct.

Throughout the following phenotyping tests, two strains were compared: UK48 and UK71. The UK48 genome contains the common duplication, and the UK71 genome does not. This pair of strains was chosen as they are part of the PERISCOPE strain panel which will undergo thorough testing in the future (Diavatopoulos et al., 2018). To select the strains to be included in this panel, a *B. pertussis* core genome was constructed, and UK48 and UK71 were estimated to differ by only 6 SNPs within this core genome (Personal Communication, J. Abrahams, 2019). This was validated by adding UK48 and UK71

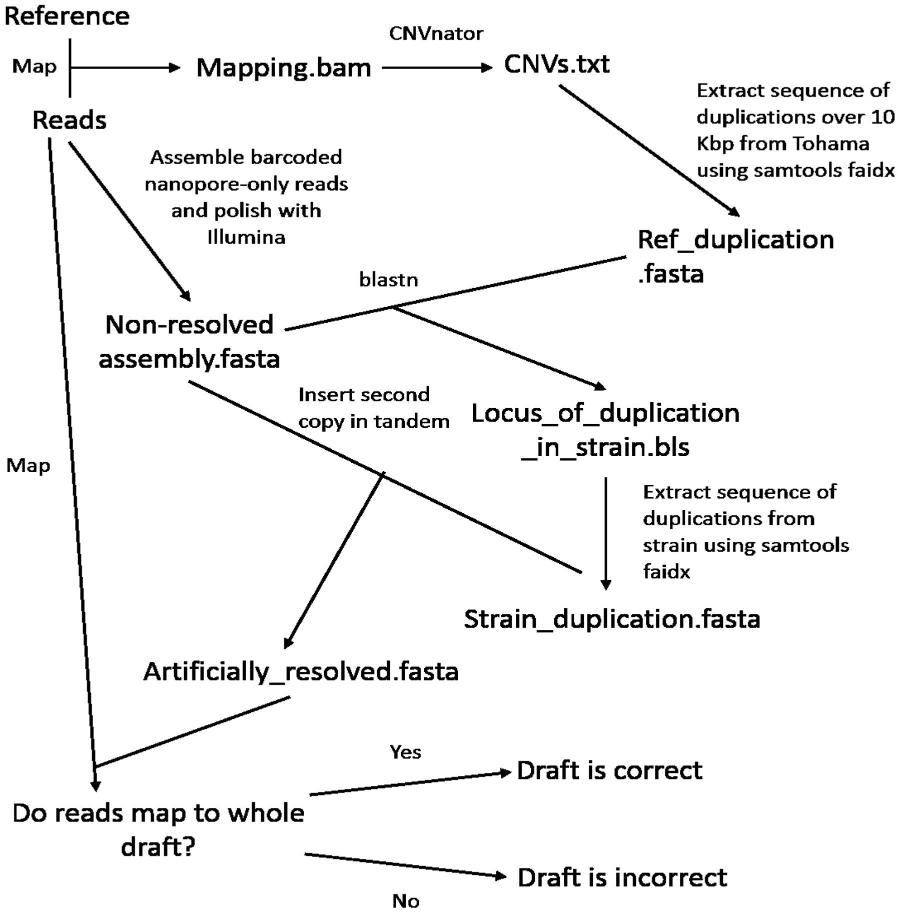

Reference
Map
Mapping.bam
CNVnator
CNVs.txt
Reads
Extract sequence of duplications over 10 Kbp from Tohama using samtools faidx
Assemble barcoded nanopore-only reads and polish with Illumina
Ref_duplication .fasta
blastn
Non-resolved assembly.fasta
Insert second copy in tandem
Locus_of_duplication _in_strain.bls
Map
Extract sequence of duplications from strain using samtools faidx
Strain_duplication.fasta
Artificially_resolved.fasta
Yes
Draft is correct
Do reads map to whole draft?
Draft is incorrect
No

Figure 4.3 Steps in alternative in silico assembly strategy

to the global phylogeny from Chapter 3; of 508 core SNPs in this phylogeny, all but 6 were indeed shared by UK48 and UK71. UK71 was therefore used as a no-duplication control for UK48.

Motility assays

The motility of selected *B. pertussis* and *B. bronchiseptica* strains was tested using a method described by Hoffman *et al.* (2019). Strains were grown on charcoal agar with (Bvg(-)) or without (Bvg(+)) 50 mM $MgSO_4$ for 24 (*B. bronchiseptica*) or 96 (*B. pertussis*) hours. This growth was used to inoculate 10 ml SS broth plus supplement plus 1 g/l heptakis (Sigma) with (Bvg(-)) or without (Bvg(+)) 50 mM $MgSO_4$ to OD_{600} 0.25. The cultures were grown for 20 hours at 37°C. At 18 hours, fresh soft agar motility plates were prepared from 15 ml SS broth with supplement and 1 g/l heptakis (with or without 50 mM $MgSO_4$) plus 0.4% agar and 1% fetal bovine serum and left to set with lids open in a safety cabinet. At 20 hours, the cultures were diluted to OD_{600} = 0.80. 2 µl of the diluted culture was stabbed into the middle of the motility agar. The motility plates were grown lid-up at 37°C for 24 (*B. bronchiseptica*) or 72 (*B. pertussis*) hours, after which any growth was measured in mm from the centre of the stab-point.

4.5 *Results*

Ultra-long sequencing reads did not facilitate *de novo* assembly

There is no theoretical upper limit to the length of DNA strand that can be sequenced using Oxford Nanopore Technology sequencers. A phenol-chloroform DNA extraction method alongside modified nanopore library preparation, both with as little pipetting as possible, was used to maximise DNA fragment length. A preliminary run employing this ultra-long read approach yielded 1.2 Gb, 300x coverage of the UK48 genome. Although the mean read length was only 31,000 bp, over 100x coverage was produced in reads longer than 100,000 bp.

Assuming the length of the duplication in UK48 to be 174 kbp (**Table 4.1**), reads longer than 174, 000 bp will be required to span each copy of the duplicated section or, if the duplication is in tandem, longer than 348,000 bp to span the entire region. The ultra-long read sequencing yielded 30x coverage of the UK48 genome in reads longer than 174,000 bp, but only 2.5x coverage in reads longer than 348,000 bp. The optimal coverage required for *de novo* assembly has previously been shown to be around 50x (Desai et al., 2013), hence the 30x or 2.5x coverage produced was unlikely to be sufficient, particularly for so complex a region.

Nevertheless, a variety of *de novo* assembly strategies was trialled, as detailed in the Section 4.5, and **Supplementary table S4.4.** None of these strategies produced an error-free, closed UK48 contig. The two best assemblies were produced by Unicycler with the ultra-long reads and Illumina reads, and by Wtdbg using Canu-corrected ultra-long reads. The Unicycler assembly consisted of two contigs: one was 3.9 Mb (the majority of the genome), whilst the other was 178 kbp (around the predicted size of the duplicated region). In fact, the hybrid Unicycler assembly produced using ultra-long and Illumina reads together was almost exactly the same as a Unicycler hybrid assembly produced using regular barcoded nanopore reads together with Illumina reads in Chapter 2. Seemingly, Unicycler recognises the existence of the duplication in the nanopore reads but has not yet been given enough information to fully resolve its layout. The Wtdbg2 assembly was a single contig but, on closer inspection, did not contain full duplicated copies of the genes predicted to occur in the region of interest, and had sections to which none of the ultra-long reads aligned; it was thus concluded that this contig was misassembled. As expected, the yield of ultra-long UK48 reads was not sufficient to resolve the duplication; an extra ~160 Mb of reads longer than 300 kbp may be required.

After none of the *de novo* assembly tools produced an error-free closed UK48 contig which included two copies of the duplicated region, *de novo* assembly was attempted just for the region of the duplication, using reads which mapped to the region alone, with Canu, Flye, MaSuRCA, Unicycler and Wtdbg2. These trials were also unsuccessful.

In silico duplication resolution is possible for some strains

An alternative resolution strategy does seem to work with the current set of ultra-long reads, however. The steps of this alternative method are shown in **Figure 4.3**, and involve creating an artificially resolved copy of the genome *in silico*, using the output from CNVnator and the nanopore-only Flye genome assembled using reads with a mean length of 5,000 bp, followed by polishing with Illumina short reads using Pilon. This strategy involves two major assumptions: that the two copies of the duplicated section occur in tandem, and that the copies of the duplication are identical. However, both assumptions can be validated by aligning the ultra-long reads to the *in silico* draft; if the

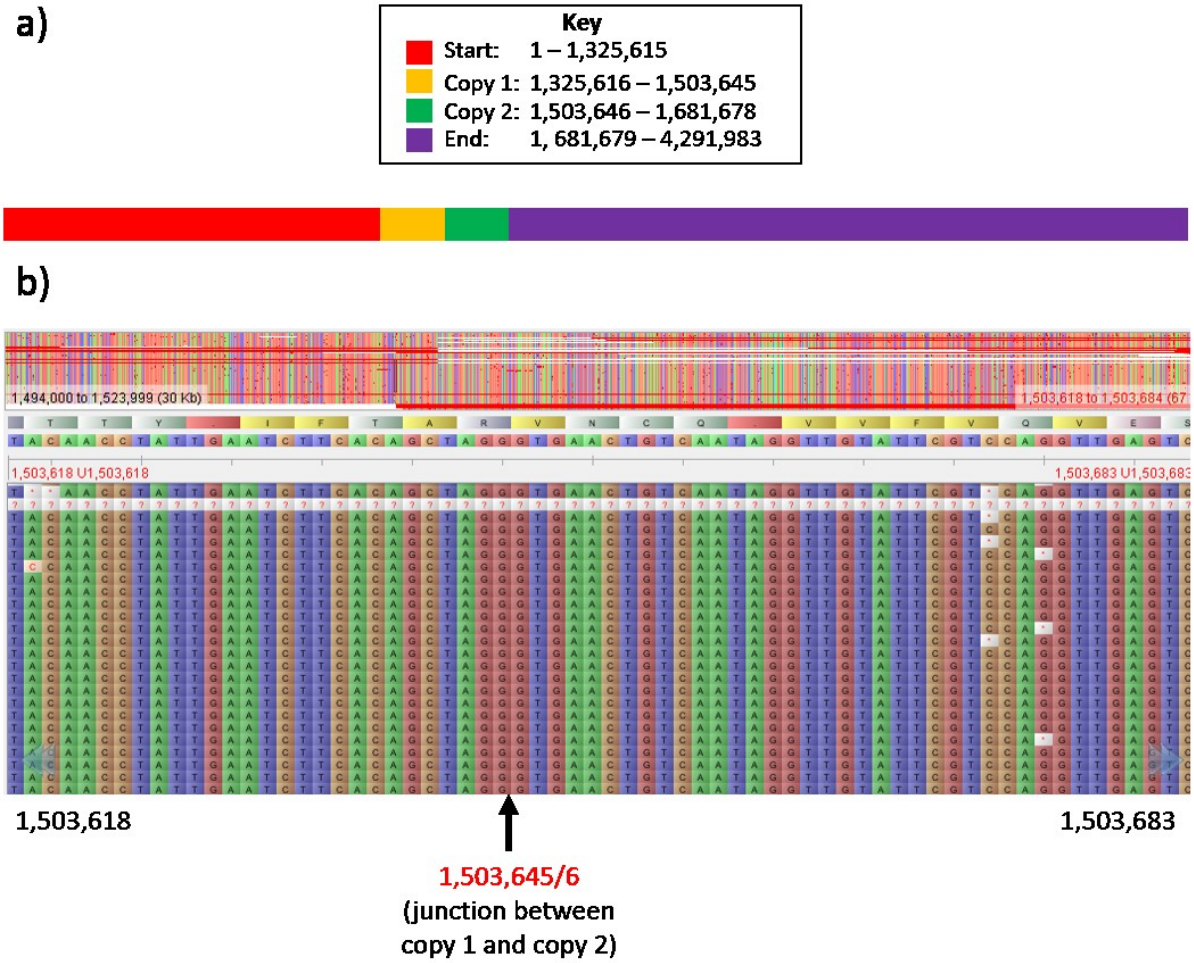

a)
Key
Start: 1 – 1,325,615
Copy 1: 1,325,616 – 1,503,645
Copy 2: 1,503,646 – 1,681,678
End: 1, 681,679 – 4,291,983
b)
1,494,000 to 1,523,999 (30 Kb)
1,503,618 to 1,503,684 (67
1,503,618 U1,503,618
1,503,683 U1,503,683
1,503,618
1,503,683
1,503,645/6
(junction between
copy 1 and copy 2)

Figure 4.4 a) An artificially resolved draft of the whole UK48 genome was generated using CNVnator output. b) Ultra-long reads were mapped to the in silico draft assembly using Minimap2 and SAMtools, then visualised using Tablet. This visualisation shows that the junction between the two copies of the region is present in the ultra-long reads; this junction would only exist if the duplication was correctly assembled assumptions are correct, the artificial genome will have been correctly assembled, and the ultra-long reads will correctly align, using Minimap2 (Li, 2017), to every base along the whole genome, as visualised using SAMtools (Li et al., 2009) and Tablet (Milne et al., 2013). **Figure 4.4** shows that this strategy was successful for UK48: multiple reads existed which spanned both copies of the duplicated region, including the junction between the two, the sequence at which would be present in the reads only if the *in silico* genome was assembled perfectly. There were no obvious differences, such as SNPs or indels, between the reads which aligned to the two copies of the duplicated region. However, the *in silico* assembly strategy shown in **Figure 4.3** was unsuccessful for the UK76 reads; no reads mapped to the predicted junction between the two copies of the region, suggesting the start or end point of the duplication was mispredicted by CNVnator.

No recent UK *B. pertussis* strains appear to be motile in the conditions tested here

All five UK strains carrying the ultra-long duplication (UK48, UK65, UK76, UK90 and UK92) were tested for motility, using a motility assay developed by Hoffman et al. (2019). In addition, UK71 was tested; UK71 has a SNP profile very similar to UK48, but UK48 has the duplication and UK71 does not. *B. bronchiseptica* strain RB54, which is a Bvg(-) phase-locked mutant, was used as a positive control for both Bvg conditions.

As expected, the Bvg(-)-locked *B. bronchiseptica* mutant was positive for motility in both 0 mM $MgSO_4$ and 50 mM $MgSO_4$. However, none of the *B. pertussis* strains tested were positive for motility in either Bvg condition. In both conditions, some cells grew out of the hole into which they had been stabbed, as shown in **Figure 4.5**. This was particularly noticeable for the 50 mM $MgSO_4$ plates, likely because *B. pertussis* cells grow more quickly in Bvg(-) conditions (Belcher, 2017).

2 µl overnight culture for each strain was also viewed on a microscope slide: *B. bronchiseptica* RB54 cells could be observed to move across the field of view, but no such movement was observed for any *B. pertussis* cells. Because motility was not observed for any *B. pertussis* strains in the conditions used here, no further comparisons of motility phenotype between duplication-carrying and non-duplication-carrying strains were possible.

4.6 *Discussion*

Resolving ultra-long duplications is not an easy task

Ultra-long reads were unable to resolve the duplication in UK48. Perhaps 30X coverage is insufficient for *de novo* assembly of such long duplications, or longer reads still are required. Phenol-chloroform methods to extract ultra-long gDNA have continued to be refined, commonly producing data sets with an N50 greater than 100 kbp, and read lengths approaching 1 Mb (Jain, Koren, et al., 2018). More recently, agarose plug-based extraction methods have been developed, which can achieve read lengths upwards of 2 Mb (Payne et al., 2019). These methods may produce longer UK48 reads, or a higher coverage of the genome in reads as long as the duplication, potentially permitting resolution of the ultra-long repetitive region.

Ultra-long reads were also used to attempt to resolve a high copy number CNV in a different UK strain, UK54, in Abrahams et al. (In review). Although this CNV was much shorter at around 15 kbp, full resolution of the repetitive region did not prove to be possible. In the case of UK54, however, there existed a highly mixed population in which cells could contain between one and five (or more) copies of the region in question. If different reads contain different numbers of the region, no matter how

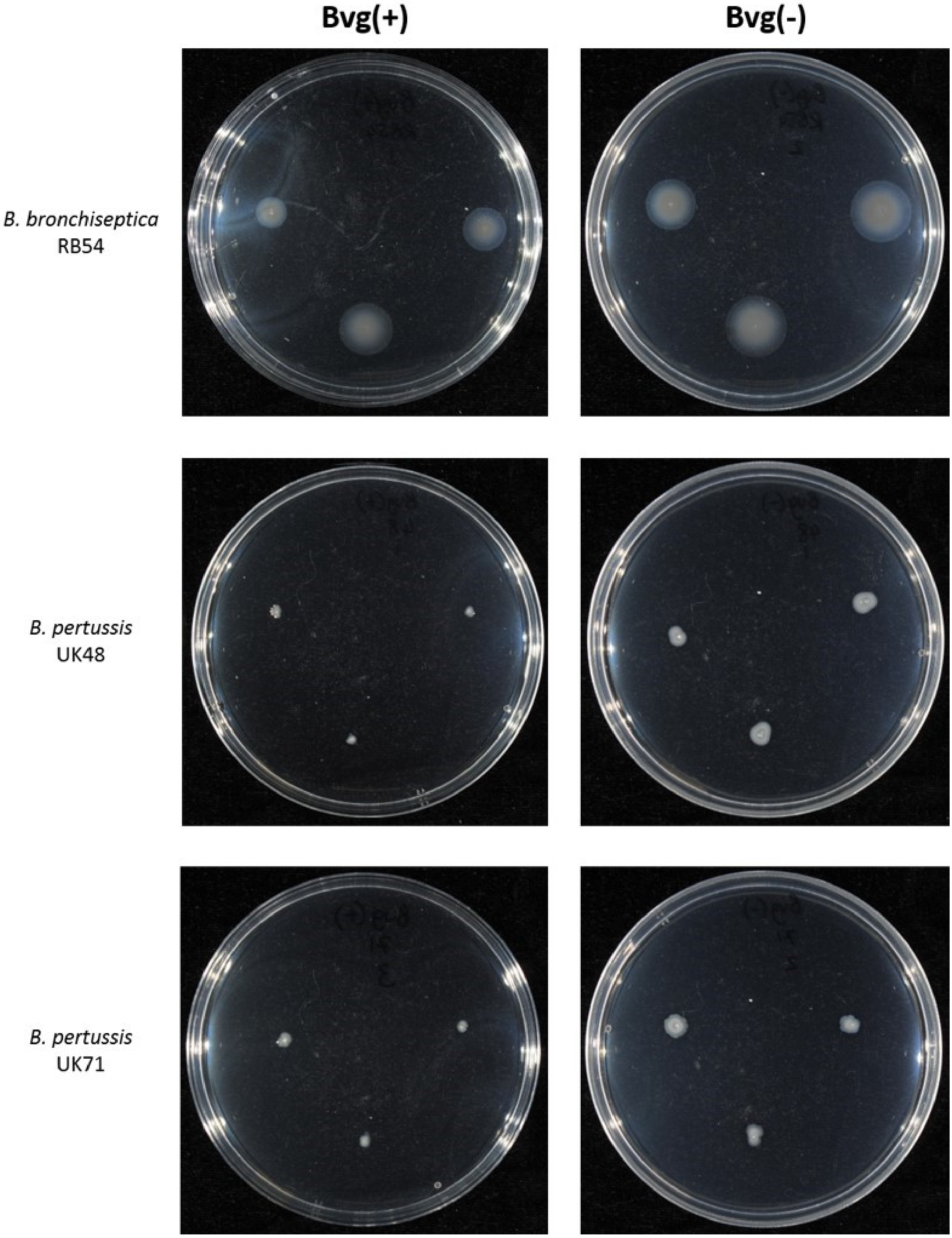


Figure 4.5 Motility assays showed no motility phenotype in any UK *B. pertussis* strains test. Bacteria edwere grown on charcoal agar plates with (Bvg(-)) or without (Bvg(+)) 50 mM $MgSO_4$ for 24 (*B. bronchiseptica*) or 96 (*B. pertussis*) hours. Growth from each plate was used to inoculate 10 ml SS plus supplement (with or without $MgSO_4$) to 0.25 OD_{600}. This was then incubated at 37 ^{o}C with shaking until 0.9 ± 0.1 OD_{600}. 2 µl of growth was stabbed in triplicate into 15 ml motility agar (0.4% agar, SS plus supplement (with or without 50mM $MgSO_4$), 10% FBS) and the motility plates were grown at 37 ^{o}C for 24 (*B. bronchiseptica*) or 96 (*B. pertussis*) hours.

A phase-locked Bvg(-) *B. bronchiseptica* mutant, RB54, was used as a positive control for motility in both growth conditions, and all repeats showed motile behaviour. None of the *B. pertussis* samples appeared to be motile in either Bvg(+) or Bvg(-) conditions, although some cells grew out of the stab-hole onto the surface of the agar (particularly in the Bvg(-) conditions), as shown here for UK48 and UK71.

ultra-ultra-long they are, no assembly tool will be able to correctly resolve the genome. Such a mixed population has not been observed in UK48, but could be an alternative explanation for the inability using ultra-long reads to resolve the "motility" duplication.

In contrast, the "*in silico* resolved" UK48 genome was ostensibly assembled correctly. However, this *in silico* assembly method requires a level of manual intervention which would preclude high-throughput genome assembly, as well as being based on a number of assumptions, most notably that the duplication occurred in tandem and that neither copy was inverted. It also required CNVnator to predict the exact start and end bases of the duplicated region correctly. When tested on UK76, the method was not successful in producing an artificially resolved genome; the UK76 ultra-long reads did not map to the junction between the two copies of the duplicated region. This was probably due to the numerous different versions of the duplication which existed in different cells in the UK76 population (Abrahams et al., In review). If all non-integer CNV predictions are due to the presence of a mixed population within the strain, where some cells carry the duplication and others do not, and where some cells can carry slightly different versions of the duplication, as noted in Abrahams et al., this *in silico* assembly method is unlikely to be effective for resolving the genomes of most duplication-carrying strains, including UK65, UK90 and UK92. Given the equal failure of ultra-long reads to produce *de novo* assembled resolved genomes, the most effective method of resolving these ultra-long duplications remains the optical genome mapping method technique used by the CDC, despite the additional cost this technique presents. Nonetheless, ongoing improvements to ultra-long sequencing protocols towards single read sequencing of entire microbial genomes may enable easy resolution of ultra-long duplications and in depth investigation of any intra-strain genomic variations (Payne et al., 2018b; Long Read Club, 2019).

Does duplication of the flagellar locus result in extra-motile *Bordetella pertussis*?

B. pertussis has traditionally been described as a non-motile bacterium which is unable to produce flagella, despite its genome containing the 24 core flagella genes observed across all motile bacteria *(*Liu and Ochman*, 2007)*. The ancestor of *B. pertussis*, *Bordetella bronchiseptica*, has the same set of core flagellar genes, and has been shown to both produce flagella and display motile behaviour when in the Bvg(-) phase (Akerley et al., 1992). In the *B. pertussis* genome, however, the flagella biosynthesis gene *flhA* contains a premature stop, 1313 codons into the 2119-codon gene (Parkhill et al., 2003). In other bacteria, including *B. bronchiseptica*, the flagella also play a role in biofilm formation (Klausen et al., 2003; Diaz et al., 2011; Nicholson, Conover and Deora, 2012; Fong and Yildiz, 2015). However, over the last decade, it has been shown repeatedly that *B. pertussis* is still able to form biofilm, in spite of its supposedly truncated FlhA protein (Irie, Mattoo and Yuk, 2004; Serra, D. et al., 2007; Serra, D.O. et al., 2011; Nicholson, Conover and Deora, 2012; Burns, Meade and Messionnier, 2014; Cattelan et al., 2016).

The unexpected ability of *B. pertussis* to form biofilm led Hoffman et al. (2019) to question whether *B. pertussis* could potentially by motile after all, under the right conditions. They discovered that the "*B. pertussis* is a non-motile bacterium" dogma essentially dated back to a single experiment in the 1980s, during which Dr. Jeffrey Miller tested the motility of both *B. pertussis* and *B. bronchiseptica* using the standard clinical motility agar, which contained pancreatic digest of gelatin, sodium chloride, beef extract and agar (Stiles and Ng, 1981). It has subsequently been discovered that *B. pertussis* cannot grow in the presence of gelatin; however, the non-growth of *B. pertussis* during this original motility test, combined with the premature stop codon in *flhA*, has generally been taken as evidence that *B. pertussis* is both non-motile and non-flagellated, despite the later discovery regarding the motility and flagellation of *B. bronchiseptica* when grown in a different medium in Bvg(-) conditions. Hoffman *et al.* therefore re-tested the motility and flagellation of *B. pertussis* in a variety of strains and conditions, using the standard *Bordetella in vitro* growth media, Bordet-Gengou agar supplemented with 15% sheep blood for initial growth, and soft agar comprised of Stainer-Scholte broth plus 0.4% agar for the motility test (Bordet and Gengou, 1906; Stainer and Scholte, 1970).

Hoffman *et al.* found that certain strains of *B. pertussis* were motile in Bvg(-) conditions, including a mutant phase-locked Bvg(-) *B. pertussis* strain, BP347, several lab-adapted strains, such as BP338, Bpe60 and BPSM, and several clinical strains, such as V015, UVA010 and UVA052. At the same time, nearly half of the isolates tested (9/22) did not display motile behaviour in Bvg(-) conditions. The motility displayed by the "motile" strains was also not consistently seen in every trial. Hoffman et al. concluded that some unknown other genetic or environmental factor may contribute to *B. pertussis* motility, alongside the BvgAS two-component system. The duplication of motility genes seen in UK48 and other UK strains here could be one such genetic factor. However, no motility was observed in any of the UK strains tested, although the *B. bronchiseptica* reference strains RB50, RB53 and RB54 all displayed the expected motile behaviour. If the duplication of the flagellar locus in these strains does influence motile behaviour, there may be an additional factor involved which was not discovered here.

What other phenotypes could result from the duplication seen in the UK strains studied here?

As shown in **Supplementary Table S4.1**, the duplicated region differs in length in each of the five UK strains studied here. However, a set of 119 genes is duplicated in all of them. The flagella locus, BP1366-1411, includes almost 40 genes, thus representing around 33% of all of the genes duplicated in the UK strains. Extending to all of the strains in which a duplication was found at the same locus in Abrahams et al. (In review), the core shared region decreases in size, and the portion of the duplication consisting of flagellar and motility-related genes therefore increases to almost a half. If any phenotype results from this duplication, it therefore seemed likely that it might be related to motility. As already discussed, however, no motility has yet been observed in any strain carrying this duplication. Could the duplication result in a different phenotype instead?

No other such obvious groups of related genes are present in the duplicated locus. Indeed, many of the other genes in the core duplication are not well characterised, being labelled simply as "membrane protein" or "hypothetical protein". Nonetheless, a few genes of interest include those involved in metabolism, such as acyl-CoA dehydrogenase, enoly-CoA hydratase, and 3-hydroxyisobutyrate dehydrogenase (BP1445-1447), or membrane biosynthesis, such as *lpxA* and *lpxB* (BP1431-BP1432). It is possible, therefore, that the duplication may result in a phenotype which could be observed in the growth behaviour of the duplication-carrying bacteria.

A number of growth assays were undertaken (data not shown) to compare the growth dynamics of UK48 with those of UK71, which does not possess the duplication. Unfortunately, these clinical strains are not lab-adapted, and do not grow well *in vitro.* Growth rates in SS media in 24-well plates, 96-well plates, and 250 ml acid-washed flasks, were unpredictable. No growth was seen in Bvg(+) or Bvg(-) conditions on either type of plate, and growth in acid-washed flasks was irreproducible, with each repeat behaving differently. Even the addition of heptakis, a cyclodextrin which sequesters molecules released by the bacteria as they grow, and which can hence often help unstable *B. pertussis* populations to grow in the absence of blood media (MacArthur et al., 2019), was unsuccessful for these strains. Unfortunately, this irreproducible growth meant that planned comparisons of the expression of duplicated genes by qPCR was not possible here. It is feasible that, with time, conditions under which the strains grow predictably and reproducibly could be identified. However, this kind of growth optimisation was beyond the scope of this study. Additionally, the benefit conveyed by the duplication (if any) may be masked by any growth conditions which allow even the least lab-adapted bacteria to grow well *in vitro*.

Ultimately, although no phenotype relating to the commonly seen ultra-long duplication could be identified here, the existence of such a large duplicated region in so many seemingly unrelated *B. pertussis* isolates suggests that there could indeed be a benefit to carrying more than one copy of one or more of the genes in the region. This benefit could be in the form of enhanced motility, or might be more subtle, such as more efficient growth under challenging environmental conditions.

Chapter 5: Closed *Bordetella pertussis* genomes mixed populations of Filamentous Haemagglutinin gene sequences within cell populations from the same strain

"One thing I have learned is that humans cling to their first knowledge of you, particularly if they have no experience of you once you've changed."

- Tamora Pierce, Wolf-Speaker

5.1 *Abstract*

A picture of accelerated genetic changes and increasing antigen-deficiency in the *Bordetella pertussis* acellular vaccine component proteins has recently emerged from a number of large longitudinal strain screens, although the phenotypic consequences of many of the observed changes with regards to infection are unclear. In light of the clear changes observed in the other ACV-encoding genes, it seemed likely that the genes involved in producing the final ACV component, FHA, would also be undergoing allelic changes. However, due to the complexity of the gene which codes for FHA, *fhaB*, assembly of the full gene using short-reads sequencing is difficult, so analyses of any changes over time have been lacking.

Here, long-read-closed *B. pertussis* genomes were mined, and the fully-assembled *fhaB* sequences within were screened for any ongoing changes. This revealed that a specific 10-base homopolymeric tract, 1,078 bp into most *fhaB* alleles, appears to be susceptible to slippage, resulting in some strains carrying an indel at this locus which has previously been linked to FHA-deficiency. However, further investigation showed that the mutation is present in almost all strains ever sequenced, but rarely present in all reads in a sample: instead, there tends to be a mixed population of sequences at the tract locus, like those observed in previous chapters. In addition, Illumina reads were used to identify the insertion of any IS elements into *fhaB* in the same strains, revealing that mixed populations of cells with and without IS insertion may also be present. There do not seem to be any significant longitudinal trends in either phenomenon; nonetheless, this work adds to a growing body of evidence that the genotypic characteristics of intra-strain *B. pertussis* populations are highly fluid.

5.2 *Data Summary*

1. A full list of accession numbers for closed genome sequences, and raw Illumina reads, as well as full results, is available in **Supplementary table S5.1** (https://figshare.com/s/e0585f38926b02239be9)

2. Source code and full commands used are available from Github:
https://github.com/nataliering/FHA_screening

5.3 *Introduction*

Bordetella pertussis causes most cases of whooping cough, a respiratory disease which has been resurgent in many countries since the late 1980s and early 1990s (Clark, 2014). This resurgence follows several decades of reduced whooping cough incidence, largely due to the introduction and widespread uptake of a whooping cough vaccine in the 1940s and 1950s (Jakinovich and Sood, 2014). The original whole-cell vaccine (WCV) contained whole killed *B. pertussis* cells. However, the WCV was perceived to cause some mild-to-severe side-effects, leading to a decrease in vaccine uptake by the 1980s (Cherry, 1996, 2019). Consequently, a new acellular vaccine (ACV) was developed, containing one to five of the following *B. pertussis* antigens: pertactin (PRN), pertussis toxin (PT), filamentous haemagglutinin (FHA), fimbrial 2 (Fim2) and fimbrial 3 (fim3). The ACV was introduced in most developed countries by the late 1990s and early 2000s, leading to renewed high levels of vaccination coverage (WHO, 2018). Nonetheless, the resurgence in whooping cough incidence has continued, with numerous large outbreaks in the 2000s and 2010s, in countries including Australia, Japan, the United

States and the United Kingdom (Octavia et al., 2012; Miyaji et al., 2013; Bowden et al., 2014; Sealey et al., 2015).

The most likely cause of the ongoing resurgence of whooping cough is a combination of factors, including improved awareness and introduction of more sensitive diagnosis methods such as PCR (Cherry, 2015), waning immunity conveyed by the ACV compared to the WCV (Koepke et al., 2014), and the genetic divergence of circulating *B. pertussis* strains away from the vaccine strains (Bart, Harris, et al., 2014). The widespread availability of whole genome sequencing has enabled the monitoring of allelic shifts in *B. pertussis* genes. Bart *el al.*'s landmark strain screen in 2014, including 343 samples spanning between 1920 and 2010, showed ongoing allelic changes in key *B. pertussis* genes since the introduction of vaccination, particularly in *ptxA*, *ptxP*, *prn*, *fim2* and *fim3*, all of which code for, or promote the transcription of, the antigens now included in the ACV (Bart, Harris, et al., 2014). Sealey *et al.*'s 2015 study of 100 strains isolated during the UK's 2012 whooping cough outbreak further showed that, since the switch to the ACV from the WCV, these same genes have been evolving more quickly than genes encoding other cell-surface proteins (Sealey et al., 2015). In addition to these allelic changes, a seemingly increasing number of strains isolated since 2008 do not express pertactin (Hegerle et al., 2012; Queenan, Cassiday and Evangelista, 2013; Lam et al., 2014; Pawloski et al., 2014; Williams, M.M. et al., 2016; Barkoff et al., 2019).

In many of these large strain screens, however, the gene which codes for the final ACV component FHA has not been investigated. The length of *fhaB* (over 10 kbp), repeat regions within it, and presence of paralogues all mean that accurate assembly and identification of SNPs using short-read sequencing is difficult (Bart, Harris, et al., 2014; Belcher and Preston, 2015), so it is often left out of analyses of allelic shifts in the ACV genes. In *Bordetella*, FHA has a number of different roles, including as a vital adhesin during early infection, contributing to immune evasion, and playing a part in biofilm formation (Urisu, Cowell and Manclark, 1986; Relman et al., 1989; Cotter et al., 1998; Serra, D.O. et al., 2011; Melvin et al., 2015). The FHA protein is coded for by *fhaB*, which first results in the production of the 370 kDa precursor protein FhaB. FhaB is then sequentially processed by FhaC, CtpA and SphB1 to produce the final 220 kDa FHA protein (Nash and Cotter, 2019). Mutations in any of these genes could therefore lead to FHA-deficiency, just as mutations in *prn* and other related genes lead to the deficiency of PRN seen in an apparently increasing number of strains.

Given the changes seen in the other ACV genes since the introduction of vaccination, and particularly since the switch from WCV to ACV, we might expect to see similar allelic changes in *fhaB*. Indeed, some recent and older strains have been identified by Western Blot or other protein-based methods as having severely reduced or absent FHA expression, often in tandem with PRN deficiency, and Etskovitz *et al.* (2019) showed that multiple sites relating to the antigenic epitopes in FHA are under diversifying selection (Bart et al., 2015; Weigand et al., 2018; Etskovitz et al., 2019; Xu, Z., Octavia, et al., 2019). Investigation of *fhaB* and its partner genes may therefore reveal previously hidden ACV-driven mutations.

Long-read sequencing offers the opportunity to produce closed *B. pertussis* genome sequences, including closed *fhaB* sequences. The genomes of several hundred *B. pertussis* genomes have now been sequenced using either Pacific Biosciences (PacBio) or Oxford Nanopore Technologies (ONT) sequencing platforms, and many are available from the NCBI RefSeq or GenBank databases (Bart, Zeddeman, et al., 2014; Bouchez et al., 2018; Ring et al., 2018). Here, RefSeq was mined for all available closed *B. pertussis* genome sequences and global genetic trends in FHA between 1935 and

2019 were investigated, considering any shifts in the context of recent changes to the other *B. pertussis* ACV genes.

5.4 *Research questions*

Research question 1: Can closed genome sequences reveal previously unnoticed changes in FHA and, if so, how do these changes relate to the emerging picture of allelic shifts and increasing deficiency seen in the other ACV components?

5.5 *Methods*

All data analysis was conducted using the Medical Research Council's Cloud Infrastructure for Microbial Bioinformatics (CLIMB) (Connor et al., 2016).

Mining of closed genome sequences

The assemblies for all available *B. pertussis* genomes were downloaded in fasta format from the NCBI's RefSeq database, filtering for "complete" genomes only. Accompanying strain metadata was downloaded from the NCBI's Sequence Read Archive (SRA) Run Selector.

Allele typing

Allele type was assigned to the genes coding for the ACV proteins, and the promoter for pertussis toxin (*ptxA-E*, *prn*, *fim2*, *fim3*, *fhaB* and *ptxP*) in the closed genomes using MLST (v2.17.6; Seemann, 2019) with the method described in Chapter 3.

Strains with *fhaB* alleles which were assigned by MLST as non-exact matches were investigated manually using PubMLST's sequence query tool and blastn (Zhang, Z. et al., 2000; Morgulis et al., 2008) to identify sequence differences between the strain's sequence and the closest allele's sequence.

Potential new alleles were validated using raw Illumina reads from the SRA for each strain (where available). The SRA accession numbers for each strain validated are shown in **Supplementary table S5.1**. The Illumina reads were mapped to the assembled closed genome sequence, and Tablet (Milne et al., 2013) was used to manually validate that the reads contained the new allele sequence.

Mining and processing of raw reads for all available sequenced strains

The raw Illumina short reads for all sequenced strains available on the NCBI's SRA were downloaded using prefetch and fastq-dump, both from the SRA toolkit (https://github.com/ncbi/sra-tools.git). First, the entire list of available data was downloaded from the SRA's Run Selector by searching for "*Bordetella pertussis*". The list was then filtered by "Assay Type" for WGS, "Platform" for Illumina, and by "Collection date" and "geo_loc_name_country" to exclude missing/blank values. Some manual intervention was required to standardise the strain names, as these had sometimes been submitted in the "Sample Name" column, sometimes in the "Isolate" column, and sometimes in the "Strain" column, whilst some had spaces in their strain names, which could have confounded command line-based processing. Likewise, locations were sometimes submitted in the "geo_loc_name_country" column, and sometimes in the "geographic_location_(country_and/or_sea)" column. Finally,

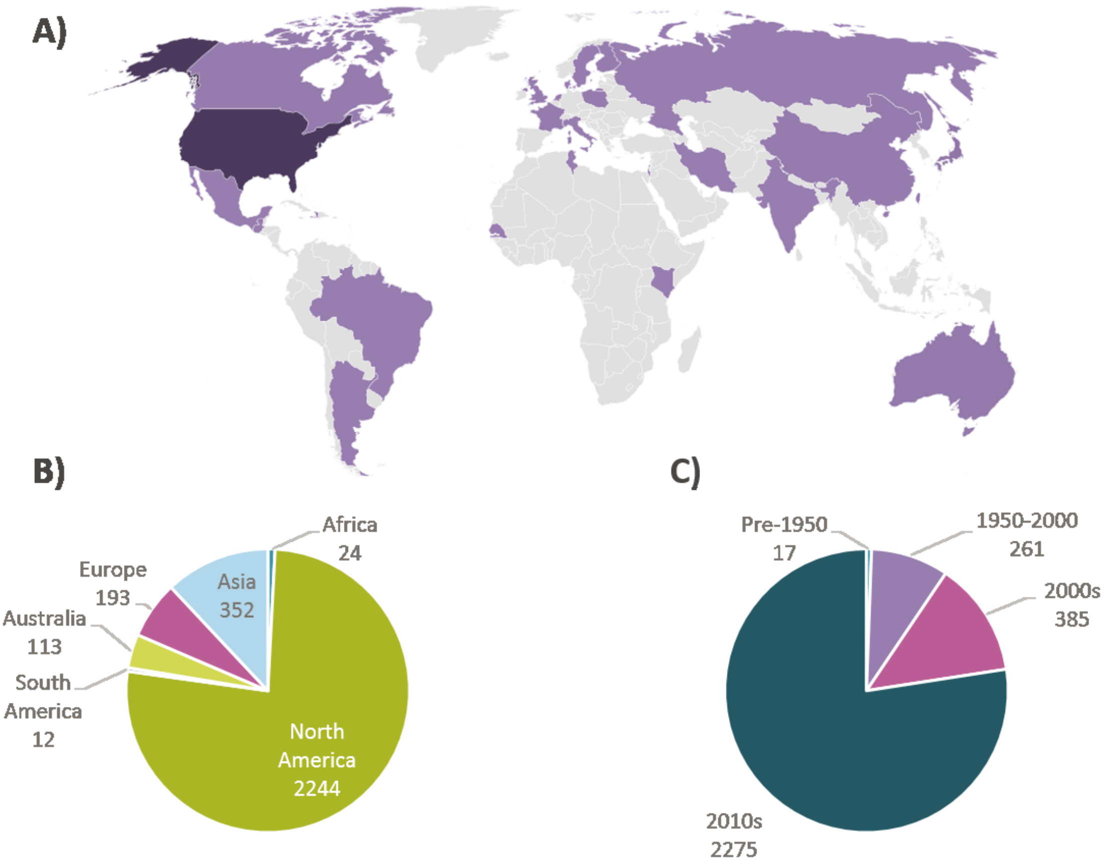


Figure 5.1 Location and year of isolation of the 2938 strains screened for FHA deficiency using Illumina short reads from the NCBI SRA. Strains were filtered from all sequence data on the SRA according to the availability of information regarding collection date and location. A) shows the countries from which strains were isolated; the darker the colour, the more strains were included (grey represents no strains). B) groups the strains into continents; all continents were represented by at least 10 strains (excluding Antarctica). C) shows the year of isolation of the screened strains; all three vaccination eras were represented (pre-vaccine, WCV, ACV). The USA's CDC has run a large clinical isolate screening project throughout the 2010s, resulting in the over-representation of USA strains and strains from the 2010s in this sample. The New Zealand strains investigated in Chapter 3 are not represented here.

"Collection date" values were processed to show year alone, as some entries included only the year of isolation, whilst others also included a day and/or month. The full metadata for all strains accessed is shown in **Supplementary table S5.2** and is summarised in **Figure 5.1**. Briefly, the data set included 278 pre-2000 (ostensibly the "WCV" strains) strains, and 2660 post-2000 (the "ACV" strains) strains. All continents (excluding Antarctica) were represented by at least 12 strains; likewise, all three vaccination eras (pre-vaccine, WCV, ACV) were represented by at least 17 strains. However, strains from North America were over-represented, as were strains from the 2010s. This is largely due to sequencing of large numbers of *B .pertussis* isolates by the USA's CDC in recent years (Weigand et al., 2017; Weigand et al., 2018; Centers for Disease Control, 2019a).

Characterisation of *fhaB* polyG tract in Illumina read sets

The 2938 paired-end global *B. pertussis* read sets downloaded from the SRA as described above were mapped using bwa mem (v0.7.17; Li, 2013) to an *fhaB* sequence which had been manually edited to remove all 10 Gs from the polyG tract (https://figshare.com/s/0b437c10f552f1e55972). The resulting sam file was processed using SAMtools view and sort to produce a sorted bam alignment file, which was parsed using SAMtools mpileup. By mapping to an *fhaB* sequence with no Gs at position 1078 bp, the mpileup showed an insertion at position 1077 for each read, for example: “C+10GGGGGGGGGG” non-mutated polyG tract. The code is available from GitHub: https://github.com/nataliering/FHA_screening/blob/master/fhaB_checker. This output a single file with one line per strain, showing the mpileup string for all reads at position 1077. The output was processed (https://github.com/nataliering/FHA_screening/blob/master/number_extractor.vba) to extract each of the numbers from the string (for example: 10, 10, 9, 9, 9, etc.) and count the number of occurrences of each number from 5 to 15. This method assumed that the inserted bases were all Gs, hence each string was also processed to count reads with bases other than G inserted at this locus (https://github.com/nataliering/FHA_screening/blob/master/nonG_read_counter.vba).

The percentage of reads containing each potential length of homopolymeric tract was calculated for each strain. These percentages were then used to calculate mean proportions of each tract length for different groupings of strains based on their time period and country of isolation. Differences within each grouping were tested statistically for significance using one-way ANOVA (for normally distributed values) or Kruskal-Wallis (for non-normally distributed values) tests (Fisher, 1921; Kruskal and Wallis, 1952). A group of 100 strains from 2014 sequenced with either the HiSeq 2500 or MiSeq were used to test whether different Illumina sequencers produced significantly different results, using the Mann-Witney U test for non-normally distributed values (Wilcoxon, 1945; Mann and Whitney, 1947).

The percentage of reads which contained a base other than a G in this tract was calculated for each strain, and this percentage was used to calculate a mean number of reads with non-Gs across the whole dataset (excluding the four Minnesota strains):

$$\frac{\sum\left(\frac{number\ of\ nonG\ reads\ per\ strain}{number\ of\ reads\ per\ strain}\times 100\right)}{Total\ number\ of\ strains}$$

Spearman’s rank correlation coefficient was calculated to determine whether any correlation existed between the percentage of reads with 10 bases at the tract and the percentage of reads which contained non-G bases.

Identification of IS insertion into FHA

There are several potential IS recognition sites within the *fhaB* sequence, as shown in **Figure 5.2**. ISMapper (v2.0, Hawkey et al., 2015) was used to predict strains with FHA or PRN deficiency as described in Chapter 3, using the Illumina paired-end short reads downloaded from the SRA as the input sequences for each strain.

The results tables were parsed to extract rows which had identified hits for any of the IS elements within the *fhaB* gene (BP1879). In addition, tables were parsed for hits within *fhaC*, *ctpA* or *sphB1* (BP1884, BP0609, BP0216), all of which are important during the processing of the mature FHA protein (Nash and Cotter, 2019).

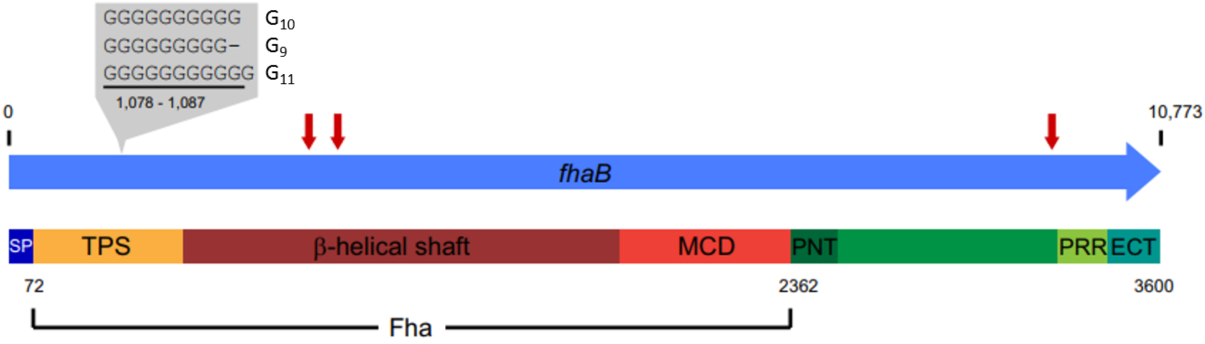


Figure 5.2 *fhaB* and the FHA precursor protein, FhaB. The most commonly observed *fhaB* allele, *fhaB2*, is 10,773 base pairs, coding for a preprotein which is almost 3,600 amino acids long. The preprotein contains several domains: the signal peptide (SP), two-partner secretion (TPS), a long β-helical shaft, a mature C-terminal domain (MCD), N-terminal predomain (PNT), proline-rich region (PRR) and extreme C-terminal domain (ECT). The preprotein is processed to become the final FHA protein, containing the domains indicated. A polyG tract 1,078 bases into *fhaB* is a common site for mutation. The wild type tract contains 10 Gs, whilst mutants containing 9 or 11 Gs have been observed, each of which results in a premature stop codon and subsequently truncated FHA protein. The *fhaB* sequence also contains three predicted insertion sites for IS *481* and IS *1002* (NCTAGN), at 2,785, 3,124 and 9,865 bases respectively (indicated by red arrows), although insertions at the 9,865 bp site do not seem to cause FHA deficiency. Insertion sites for IS *1663* may also be present, but the recognition sequence for IS *1663* is not currently known.

Figure adapted from Weigand et al. (2018).

5.6 *Results*

fhaB2 remains the most prevalent *fhaB* allele globally, but a potential deficiency-causing mutation may be on the rise

On 11 June 2019, 538 closed *B. pertussis* genomes were downloaded from NCBI's RefSeq database (**Supplementary table S5.1**). 10 of these genomes had "unknown" year of isolation, so were excluded. The allelic profiles for each of the remaining 528 genomes were assigned using MLST (v2.16.2) (Seemann, 2019), with a custom MLST scheme to assign alleles to the ACV protein-encoding genes (*ptxP*, *ptxA-E*, *prn*, *fim2*, *fim3*, *fhaB*). All known alleles for each of these genes were downloaded from PubMLST to produce the custom scheme; this included 22 recorded allele sequences for *fhaB*, although most literature cites only the two *fhaB* alleles defined by van Loo et al. (2002) in their investigation of MLST for the *B. pertussis* cell surface protein-encoding genes (for example: Bouchez et al., 2015; Xu, Z., Wang, et al., 2019). MLST assigns an allele to each genome, if the gene is present. If a match is non-exact, MLST assigns the closest match, for example "2?". The non-exact matches were investigated further here by using blastn to align the relevant genome against the allele sequence of the closest match and thus identify any differences.

In the 528 closed genomes, the most commonly seen *fhaB* allele was *fhaB2* (512 genomes, including 124 non-exact matches which contained SNPs or indels). A smaller number of strains was observed with *fhaB21*, *fhaB4, fhaB5, fhaB6* and *fhaB17* (five strains possessed *fhaB21*, all with the same two SNPs compared to the PubMLST definition, and one strain each possessed *fhaB4*, *fhaB5, fhaB6* and *fhaB17*). One genome (H321) was not assigned any known alleles for any of the genes included in the custom MLST scheme. The reasons for this were unknown, hence this genome was excluded from further analysis, leaving 527 closed genomes. Because many of the gene sequences identified here varied consistently from the alleles defined by PubMLST, altered allele names were assigned for the purposes of this work (**Table 5.1**).

Table 5.1 Redefined *fhaB* alleles for this study

***fhaB* allele**	**Sequence description**	**Number of genomes**
2	As per PubMLST	396
4	As per PubMLST	1
5	As per PubMLST	1
6	As per PubMLST	1
17	As per PubMLST	1
21	As per PubMLST	0
21b	*fhaB21* with 2 SNPs: C7341T, G7778A	3
23	*fhaB20* with 6 SNPs and indels: AA1039--, G1047-, C1088G, G6995-, -7532G, C10314-	1
2b	*fhaB2* with 1 indel: G1078- or G1078GG	115
2c	*fhaB2* with 1 SNP: A4432G	4
2d	*fhaB2* with 1 SNP: G3347A	2
2e	*fhaB2* with 1 SNP: C4497T	2

The frequency of each of the redefined *fhaB* alleles over time was then analysed, and placed in the global ACV context established by Bart, Harris, et al. (2014)'s landmark strain screen, **Figure 5.3**. Due to the availability of strains from certain years, the time period groupings were altered slightly. In addition, many more strains from 2012 onwards were available for this study than were available to Bart et al., hence this was added as a new time period. This analysis shows that, since the introduction of the WCV, the prevalence of non-*fhaB2* alleles has greatly decreased, with *fhaB2* remaining by far the most common allele. In addition, the prevalence of one particular mutation, defined here as *fhaB2b*, appears to be increasing. The indel seen in *fhaB2b* is an indel at a homopolymeric tract starting at bp 1,078 in *fhaB2*. In the wild type *fhaB2* allele (and all other *fhaB* alleles defined on PubMLST except *fhaB3* and *fhaB22*), the tract contains 10 Gs. The 115 variants classed as *fhaB2b* here contain an insertion or deletion at this tract, resulting in an assembled *fhaB* sequence containing 9 or 11 Gs. Both versions of this indel have been noted in single strains in previous studies and have been predicted to lead to a frame-shift resulting in a premature stop codon, thus are linked with FHA deficiency (Bart et al., 2015; Weigand et al., 2018; Xu, Z., Octavia, et al., 2019).

However, the closed genomes used to screen *fhaB* sequences here were produced using PacBio long reads. PacBio, like nanopore and many other sequencing technologies, is prone to errors when sequencing homopolymeric regions. Most of the closed genomes were assembled using PacBio and Illumina reads in hybrid; the raw Illumina reads were available for 62 strains from the NCBI SRA (accession numbers are given in **Supplementary table S5.1**). To validate that the *fhaB* sequence in the *fhaB2b* strains did indeed contain an indel, the raw Illumina reads for each strain were downloaded and mapped back to their relevant closed genome sequence using bwa mem, and the number of Gs in the *fhaB* homopolymeric tract at 1078 bp was checked manually. This revealed a mixed population in most of the strains whereby Illumina reads were observed with 8 to 12 Gs in the tract, as shown in **Figure 5.4**. In many strains, the highest percentage of reads still contained 10 Gs, suggesting that the mutant *fhaB2* allele present in some closed genome sequences may indeed have been due to homopolymeric error in the PacBio long reads. A notable exception is J199, for which 64% of the Illumina reads contained a truncated polyG tract (62% G_9, 2% G_8). This mutation was also noted by Weigand et al. (2018), who had seen reduced FHA production in J199. A key observation here was that, although most reads in most strains still contained 10 Gs, almost every strain possessed a mixed population of homopolymer lengths, suggesting that slippage may be common at this locus on a wider scale. In light of this, and the apparently increasing percentage of *fhaB2* in the hybrid closed genomes, the phenomenon was therefore investigated further, using the raw Illumina data available for a larger number of strains from the NCBI's SRA.

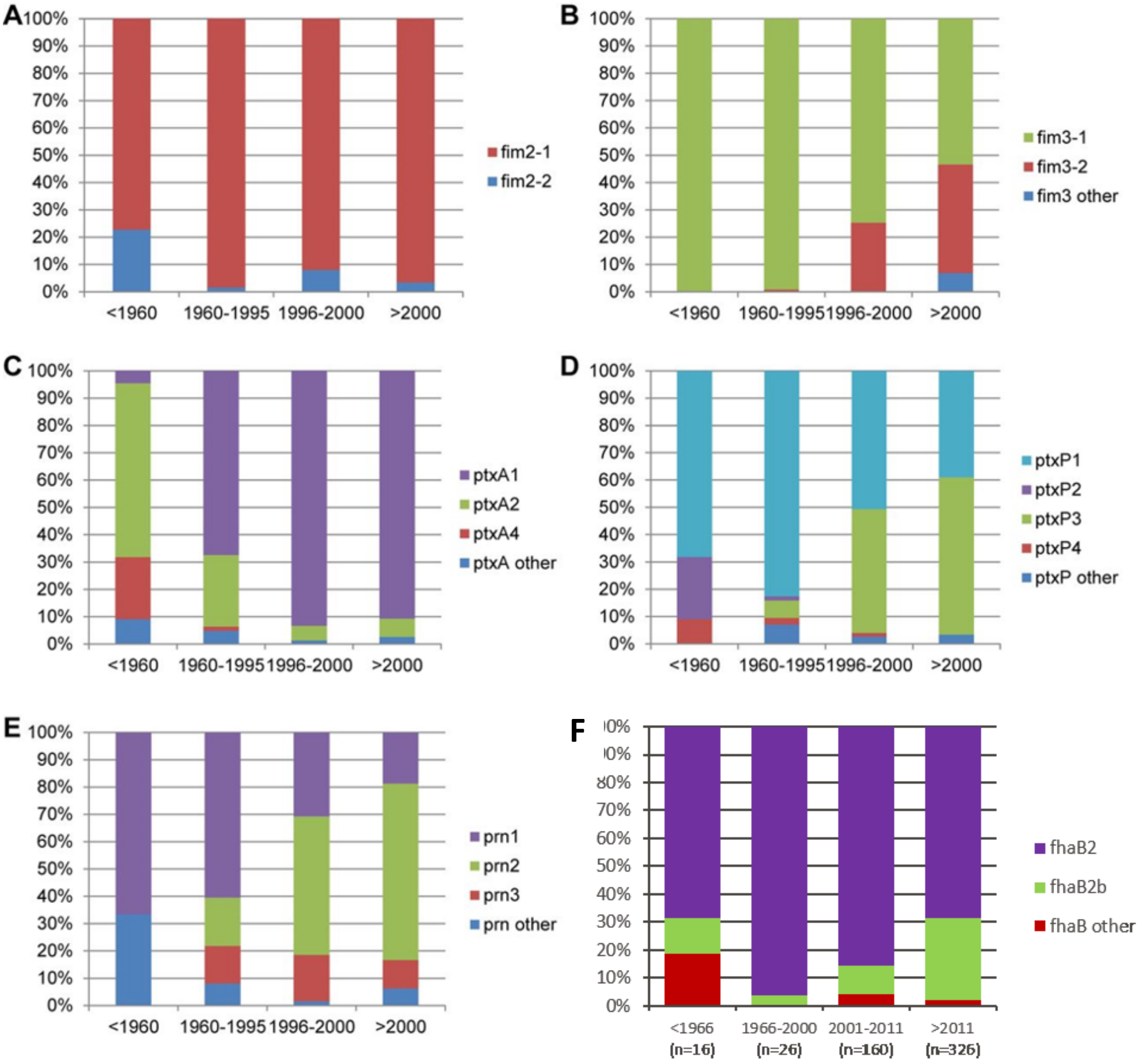


Figure 5.3 *fhaB* alleles over time, placed in the same global context as the other ACV antigen-coding genes in Bart, Harris, et al. (2014). Alleles were assigned by MLST, using closed genome sequences downloaded from the NCBI's RefSeq database. Certain mutations were commonly observed, hence allele names were redefined for the purposes of this study, as shown in **Table 5.1**. Like the other ACV-encoding genes, some alleles which were prevalent in the pre-vaccine era disappeared after the introduction of the WCV. As seen with *fim3* (B), *prn* (E) and *ptxP* (D), a new *fhaB* (F) allele seems to be becoming increasingly prevalent. This "allele", coined *fhaB2b*, is identical to the most common allele, *fhaB2*, but has an indel at a homopolymeric (polyG) tract 1078 base pairs into the gene sequence.

Figure adapted from Bart, Harris, et al. (2014)

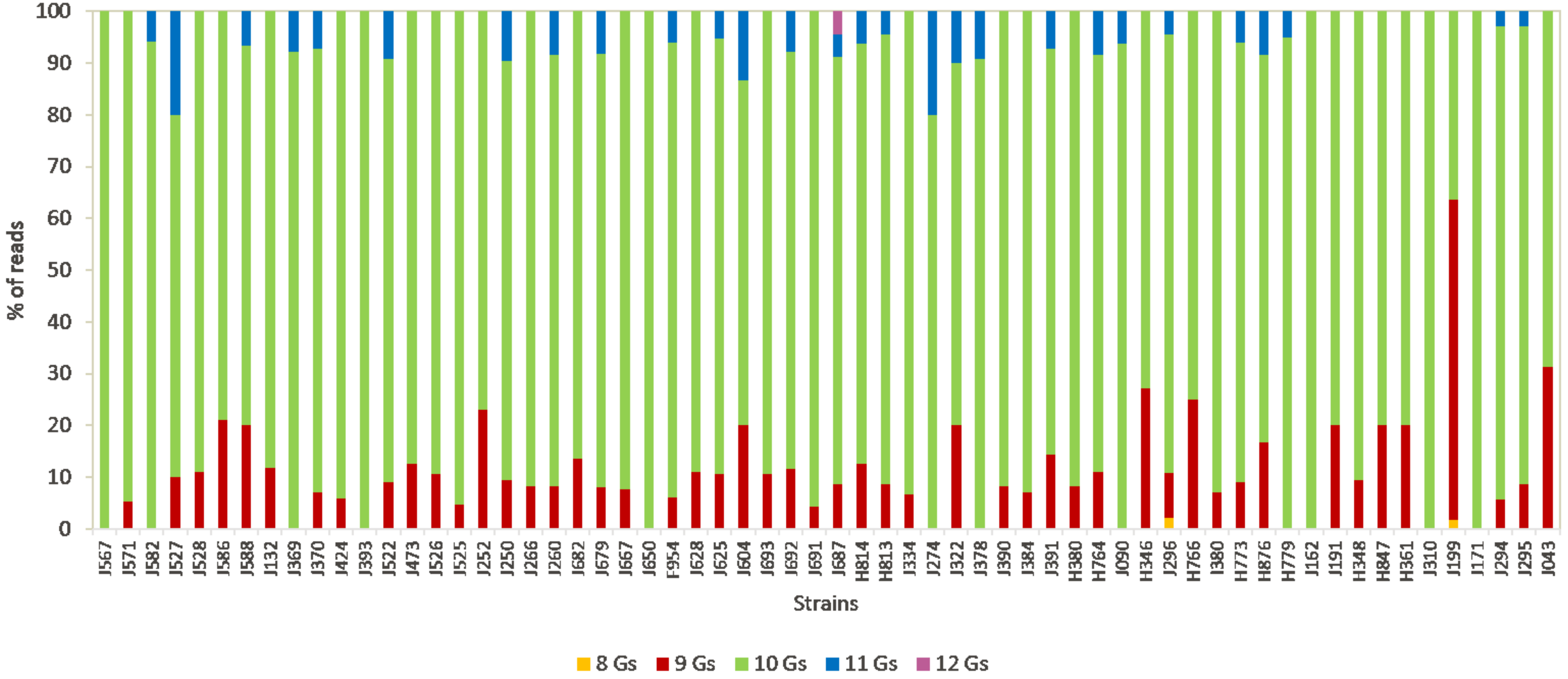


Figure 5.4 *fhaB*:G1078 characteristics of Illumina read sets for 62 "*fhaB2b*" genomes for which data was available. The number of Gs seen in the homopolymeric tract at 1078 bp in the *fhaB2* sequence varied from 8 to 12 in the strains investigated here. However, despite the assembled genome of each of the 62 strains containing a mutated *fhaB2* 1078 homopolymer sequence (i.e. more or less than 10 Gs), the predominant sequence in the Illumina reads for most of the strains contains the unmutated polyG tract (i.e. exactly 10 Gs).

Many strains contain a mixed population of *fhaB* sequences, but there may not be any trends towards deficiency

As a result of the discovery that raw Illumina reads could reveal mixed populations at the *fhaB*:1078 polyG tract in some of the 528 strains with long-read-closed genomes studied here, a wider screen was conducted of all strains available from the NCBI's SRA which had associated country and year metadata, as listed in **Supplementary table S5.2**. Paired-end reads were mapped to a manually edited *fhaB2* sequence with the polyG tract removed, then processed with SAMtools view, sort and mpileup to identify how many Gs were present at the tract locus in each of the mapped reads within each read set (raw results in **Supplementary table S5.3**).

Some strains have been sequenced more than once, and some strain names may have been used more than once globally, to represent different strains. To account for this, 325 duplicate strain names were removed at random from the processed results using Excel's "remove duplicates" function, leaving results for 2,613 strains (**Supplementary table S5.4**). For 860 of the remaining strains, fewer than 10 reads mapped to the locus of interest, so these strains were also discarded. This left usable data for 1,753 strains (**Supplementary table S5.5**).

Of the 1,753 strains with usable data, 1,638 had a mixed population; the datasets for all but 115 strains contained a mixture of reads with between G_6 and G_{12} at the polyG tract in *fhaB*. Overall, the mean percentage of reads containing only G_{10} was 87.23%. Only 10 strains contained less than 50% of G_{10} reads (including the aforementioned J199, with 36% G_{10}), suggesting that the vast majority of the strains in the dataset would not be fully FHA-deficient. To determine whether the frequency of non-

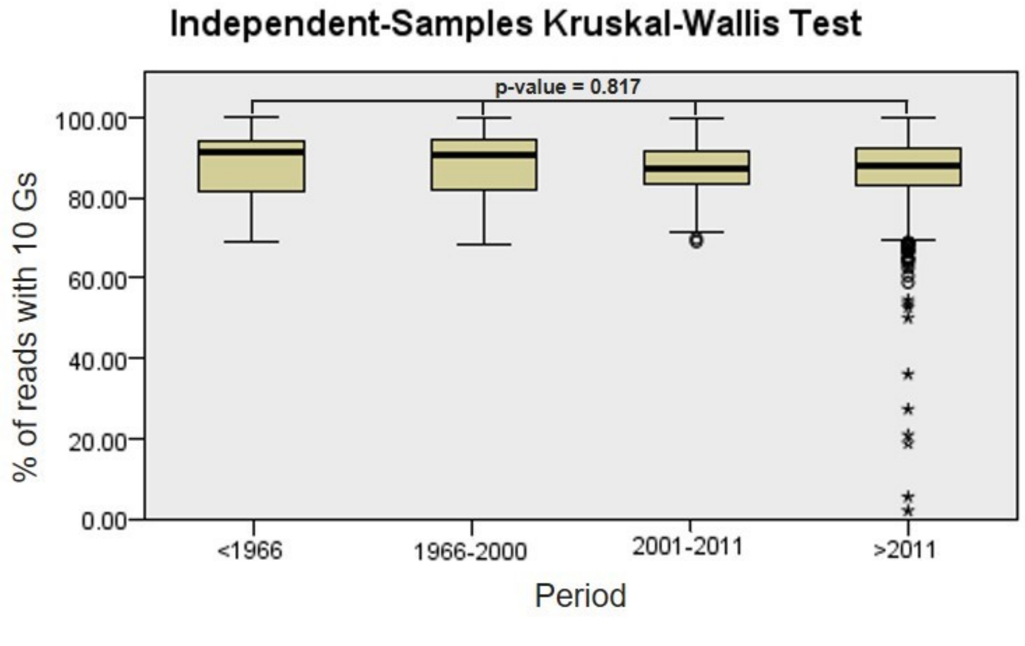


Figure 5.5 Kruskal-Wallis test for significant differences between the mean percentage of G_{10} reads in each time period. The percentage of G_{10} reads per strain was calculated, and a mean percentage of G_{10} reads was calculated for all strains in each time period. There were no significant differences between the means of the four groups. Although the boxplots suggest that the percentage of G_{10} reads has fallen slightly since the year 2001, both of the pre-2001 groupings contained a small number of strains (8 pre-1966, and 34 1955-2000), which would have meant that finding a significant p-value would have been unlikely, unless the difference between the means was very large. The means of the two post-2000 groupings were not significantly different, but the post-2011 group contained many more strains outside of one standard deviation from the mean, as indicated by stars.

G_{10} reads has changed over time, the strains were separated into the same period groupings as shown in **Figure 5.3**, and an independent-samples Kruskal-Wallis test was used to assess the significance of any differences between the means (**Figure 5.5**). No significant differences were found between the means of the four groups (p=0.817).

After removing those strains for which fewer than 10 reads mapped to the tract of interest, too few strains remained for most countries to conduct meaningful comparisons. The vast majority of the remaining strains (1,675) were isolated in the USA, reflecting the predominance of USA strains in the pre-filtered data set. However, 35 strains remained for China, which was potentially enough to reveal a statistically significant difference in mean G_{10} percentages. The mean G_{10} percentage for all reads for all strains was calculated for both countries, and a Mann-Whitney U test was used to test for significance. This showed that significantly fewer reads from the Chinese strains contained non-G_{10} tracts than from the USA strains (means 90.38 and 87.10% respectively, p=0.006).

Different Illumina sequencers do not appear to produce statistically different results

The data downloaded from the SRA was produced by a variety of different Illumina sequencers, as illustrated in **Figure 5.6**. Consequently, there was a chance that any differences observed between strains were a result of different error profiles of different sequencers. As shown in **Figure 5.6**, the majority of the dataset was evenly split between the HiSeq 2500 and the MiSeq. Data was available for 100 strains from 2014 and, fortuitously, around half of the strains (54) were sequenced using the HiSeq 2500 and the other half (46) were sequenced using the MiSeq. This dataset was therefore ideal to determine whether different sequencers produced statistically different percentages of reads with 10 bases in the polyG tract. Any differences between the two groups would not be due to their year of isolation, which would indicate they were due purely to the different sequencers and were thus

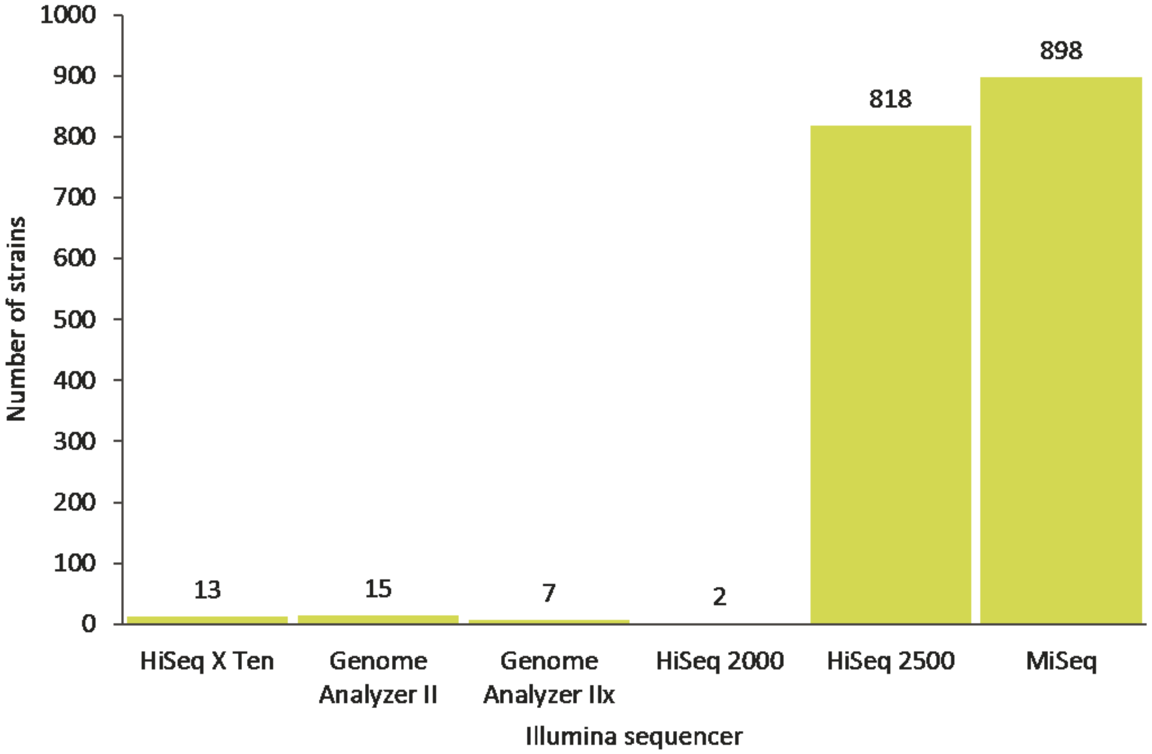


Figure 5.6 Numbers of strains sequenced by each type of Illumina sequencer in the downloaded dataset

likely to be sequencing error. A Mann-Whitney U test determined that the means of the two groups (HiSeq 2500 vs MiSeq) were not significantly different, with a p-value of 0.549 (see **Figure 5.7**). This suggests that any differences observed within the larger dataset were also not due to differences between the Illumina sequencers used to produce the data.

Some strains do not contain a polyG tract in every read at the *fhaB* 1,078 bp locus

Four strains (K059, K060, K062 and K063) contained a very different sequence at the 1078 bp locus compared to all other strains. Instead of a polyG tract, these strains had the sequence GGTGGCG at the locus in every read. As mentioned above, all of the 22 *fhaB* alleles defined by PubMLST contain the polyG tract at the 1,078 bp locus except two. The sequence found in the remaining two alleles, *fhaB3* and *fhaB22*, is the sequence found in K059, K060, K062 and K063. Thus, these four strains contain either *fhaB3* or *fhaB22*, and were the only four strains of the 1,753 tested to contain a sequence other than the polyG tract at this locus. According to the SRA's metadata for these strains, all four were isolated in 2018 in Minnesota; it therefore seems likely that they were isolated from patients who had been in close contact with each other.

In addition to the obvious outlier strains from Minnesota, the reads from every other strain were processed to determine if any read contained bases other than G in their "polyG" tract. It emerged that, in a small number of reads, other bases were sometimes substituted for one of the Gs in the middle of the tract (never at the beginning or end). In most cases, only a single G was substituted. Across the whole dataset, 2.15% of the 53,782 total reads for all strains contained bases in the tract which were not G. 1,045 strains had 0% reads containing non-Gs in this tract, 406 strains had between 0 and 5% of reads with non-Gs, 231 strains had between 5 and 10% of reads with non-Gs, and 73 strains had non-Gs in more than 10% of reads. The strains with the highest percentage of reads containing non-Gs at this tract tended to be those with lower coverage, suggesting that perhaps the higher percentage was a result of lower sequencing quality overall; this is supported by the fact that the non-G base in the affected reads was not always the same within these strains. Overall, there did not appear to be any correlation between percentage of non-G reads and percentage of reads which

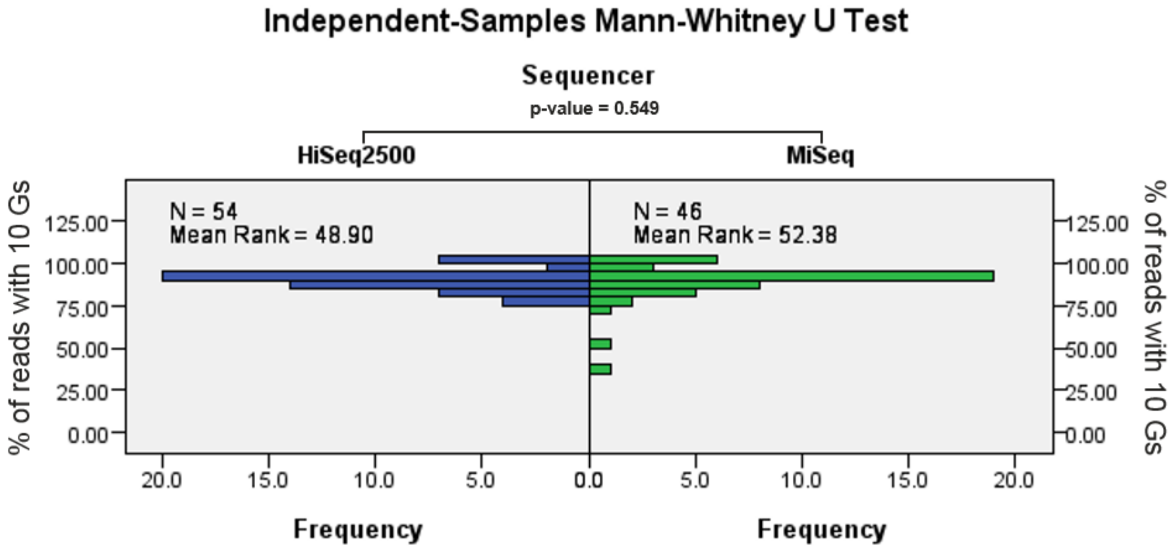


Figure 5.7 A Mann-Whitney U test revealed no significant difference between the mean % of G_{10} reads for the 2014 strains sequenced with the HiSeq 2500 and those sequenced with the MiSeq, with a p-value of 0.549

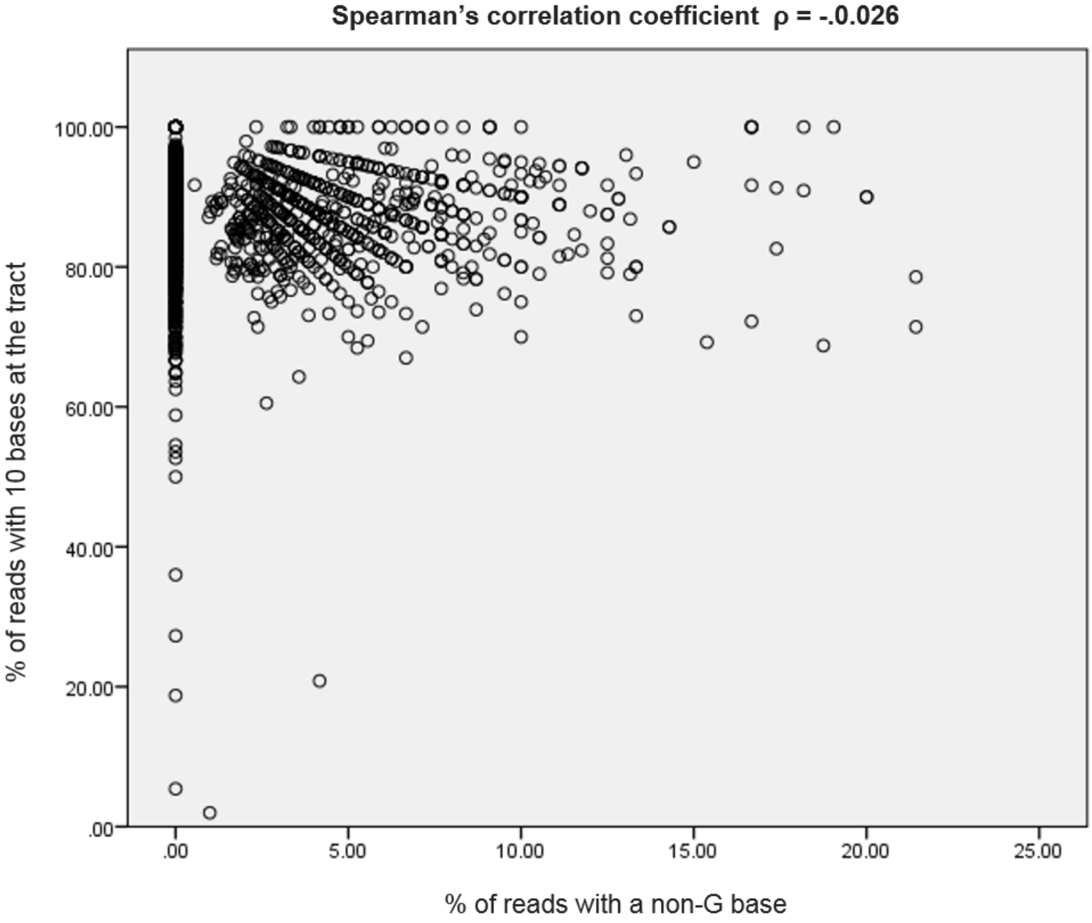


Figure 5.8 There seems to be no correlation between % of reads with a non-G base and % of reads with 10 bases at the 1,078 bp tract in *fhaB*. Spearman's rank correlation coefficient was calculated to be -0.026.

had an indel at the tract; the Spearman's rank correlation coefficient between the two factors was determined to be -0.026, as shown in **Figure 5.8**.

Insertion of IS *481* into *fhaB* likely causes deficiency in a number of strains

Illumina short-read data (from Genome Analyzer, MiSeq, HiSeq and NextSeq machines) was downloaded from the SRA for 277 pre-vaccine and WCV-era strains, and 2660 ACV-era strains. Poor quality data or lack of paired end reads meant that 124 pre-vaccine/WCV-era strains and 403 ACV-era strains were discarded, leaving 153 pre-vaccine/WCV-era and 2257 ACV-era strains. Of the 2410 strains tested, 14 were found to have an IS *481* inserted into *fhaB*, including 11 from the ACV cohort (0.49% of those tested) and three from the WCV cohort (1.96% of those tested). **Table 5.2** shows the strains found to have an IS *481* within their *fhaB* gene, along with the ISMapper-predicted insertion site. One strain, B290 (UK1), appeared to have IS *481* inserted at two of the three potential insertion sites, 3,124 and 9,865 bp into the gene, in forward and reverse orientation respectively.

No strains were found to have disruption by either IS *1002* or IS *1663* in *fhaB*. Likewise, no disruption of *fhaC*, *ctpA* or *sphB1* was seen in any strain. Neither does there seem to be any link between PRN-deficiency and likelihood of IS *481* insertion into *fhaB*: of the 14 strains identified here as having IS *481* inserted into *fhaB*, seven also had an IS *481* insertion in *prn*, whilst seven did not.

Table 5.2 Strains containing a copy of IS *481* in their *fhaB* gene

Strain	**Year**	**Country**	**SRA accession**	**Insertion site in *fhaB***	**Insertion orientation**	**Previously observed?**	**IS also in *prn*?**
J043	1947	USA	SRR5829785	9,865	F	Weigand et al. (2018)	N
B290/UK1	1920	UK	ERR037420	3,124 & 9,865	F & R	A. Preston (Personal Communication)	N
B199	1935	USA	SRR5829828	3,124	F	Weigand et al. (2018)	N
B066	2004	Canada	ERR037426	3,124	R	Novel?	N
J365	2014	USA	SRR5829857	9,865	F	Weigand et al. (2018)	Y
J198	2014	USA	SRR5070733	3,124	F	Novel?	N
H777	2011	USA	SRR9123597	3,124	F	Novel?	Y
H971	2012	USA	SRR9123520	9,865	F	Novel?	Y
I262	2012	USA	SRR9131342	9,865	F	Novel?	N
I317	2012	USA	SRR9131477	2,785	F	Novel?	Y
I333	2012	USA	SRR9131367	9,865	R	Novel?	Y
J014	2014	USA	SRR5829804	3,124	F	Weigand et al. (2018)	Y
J199	2014	USA	SRR5829751	3,124	R	Weigand et al. (2018)	Y
TN0003	2014	Tunisia	ERR2658138	3,124	R	Novel?*	N

* Potentially relevant papers (Ben Fraj, Bouchez, et al., 2019; Ben Fraj, Kechrid, et al., 2019) behind paywall

5.7 *Discussion*

Does Illumina have a homopolymer problem?

The main body of work for this chapter was centred around the precept that, whilst PacBio reads may have errors around homopolymeric sequences, Illumina reads would likely not; that any indels observed in Illumina reads were real, rather than sequencing error. Indeed, comparisons between different sequencing technologies have often concluded that most technologies (including Ion Torrent, 454, PacBio and nanopore) make more mistakes in homopolymeric tracts than Illumina (for example: Salipante et al., 2014). However, the exact frequency with which Illumina makes any mistakes at all in these regions is unclear. One comparison of sequencing bias and error showed that, the longer a homopolymer, the higher the chance of erroneous insertions and deletions, but that assuming all such indels to be erroneous resulted in false negative calls (Ross et al., 2013). Another analysis of error profiles in Illumina sequencing found a substantial bias for motifs which ended in GGG, like the variable tract here identified in *fhaB* (Schirmer et al., 2016). However, the same fluctuations in the *fhaB:*1078 polyG tract were also observed in reads which represented sequencing of the complementary strand, where the homopolymer was Cs instead of Gs, suggesting that the fluctuations were not necessarily due to this type of sequencing bias.

In addition, a 2012 analysis conducted by Ion Torrent suggested that Illumina MiSeq sequencing has a tendency to miscall the first base after a homopolymeric tract (Robison, 2012). Discounting the obvious conflict of interest in this finding by Ion Torrent about their primary competitor, the method used in this chapter would likely have not been affected by this kind of error in any case, as it simply counted how many Gs occurred after base 1,077, and would not therefore have been influenced by any difference in the base which occurred after the Gs. In addition, the most common error observed during the Ion Torrent analysis was one more call of the homopolymer, thus a single-base insertion extending the tract, whereas the most common variant seen in *fhaB* here was a single-base deletion, resulting in a shortening of the homopolymer (on average, 8.72% of reads contained G_9, compared to only 3.76% of reads containing G_{11}).

Nonetheless, there remains a chance that the *fhaB:*1078 indel phenomenon observed here is an artefact of sequencing bias, and that the majority of this chapter is simply a detailed description of the resulting error rates. However, several factors indicate the phenomenon may be real. The same indel at the same *fhaB* polyG tract has been observed several times by chance in other, peer-reviewed and published, studies. Xu, Z., Octavia, et al. (2019) observed a single-base insertion at *fhaB*:1078, thus 11 Gs in the tract, in a strain which they had found to have lowered expression of FHA (L2228). They validated this finding with both Illumina and Sanger Sequencing reads, and found the same mixture of G_{10} and G_{11} reads from both types of sequencing. It has been suggested that Sanger Sequencing is an appropriate method for validating variants called by NGS in low complexity regions, including homopolymers, so indels observed in this tract in the Sanger Sequencing reads are likely to be real (Mu et al., 2016).

Meanwhile, both Weigand et al. (2018) and Bart et al. (2015) observed strains which included a deletion or insertion at the *fhaB*:1,078 polyG tract respectively, although in both cases they validated solely using Illumina reads (J199 and B3582, respectively). Weigand et al. also noticed a mixed population of G_9 and G_{10} reads. In all three cases, the *fhaB*:1,078 mutation was identified due to the strains in question having been observed as FHA-deficient or FHA-impaired; no other mutations to any of the FHA or FHA-processing genes were present, suggesting that the frameshift resulting from the expansion or contraction of the *fhaB*:1,078 polyG tract was indeed real, and responsible for the observed reduction in FHA expression. However, only one of the strains (J199) in which this mutation has previously been observed was included in the usable dataset here.

Additionally, 115 of the strains analysed here, including some with relatively high (45x) coverage, contained 100% reads with 10 Gs at the tract. If Illumina sequencing does produce systematic error in around 13% of reads at this tract (based on the average G_{10} percentage of 87.23%), probability suggests that at least one erroneous read should be produced per 10 reads in the data set. The comparison here between 54 strains sequenced with the HiSeq 2500 and 46 sequenced with the MiSeq also showed that there was no significant difference in the percentage of reads found to contain an indel (**Figure 5.7**). This suggests that, either the error profiles of the two different sequencers are identical, or the indels detected are real. The reads containing non-G bases within the tract at 1,078 bp may have been caused by sequencing error, or may indicate that the tract is prone to substitutions as well as indels. Determining which answer is correct was beyond the scope of this work, but **Figure 5.8** shows that, even if the substitutions were caused by error, there was no correlation between this kind of error, and the identification of indels in the same tract.

Overall, without further wet-lab investigation, it is difficult to know for sure whether the indels observed here are truly real, or whether they are a remnant of sequencing error at a low complexity region. PCR could potentially determine whether different reads contained different numbers of bases at this locus, using primers designed to flank the homopolymeric tract in *fhaB*. However, the difference in product lengths may only be a single base in either direction, meaning capillary electrophoresis, as used in Sanger Sequencing, may be required to resolve the different band lengths. In addition, the predicted mixed populations in almost every strain would mean that products of different lengths would be present in the same sample. As most of the evidence here suggests that the observed indel phenomenon *is* likely to be real, designing a wet-lab experiment to validate the findings may be worthwhile.

A number of potential FHA deficiency-causing mutations exist in highly mixed *B. pertussis* populations

Using purely informatics methods, various potential *fhaB* mutations in large numbers of strains have been identified here, more quickly and cheaply than possible by wet-lab methods. In a matter of weeks, 1,753 strains were screened for a potential FHA deficiency-causing indel in the homopolymeric tract 1,078 bp into the sequence of most *fhaB* alleles. This revealed the existence of highly mixed populations of cells and suggested that, at any given time, individual cells could develop mutations at this locus which could eventually lead to deficiency.

At face value, the lack of statistically significant temporal trends illustrated in **Figure 5.5** indicates that these mixed populations of cells have always been present, and the proportions of cells with and without a deficiency-causing indel have not changed in response to either the introduction of the vaccination, or the switch to the acellular vaccine. The boxplots in **Figure 5.5** do appear to show that the percentage of G_{10} reads may have decreased slightly between the 1966-2000 and 2001-2011 groups, which would correspond with the introduction of the ACV in many countries; however, usable data was available for only 34 strains from the 1966-2000 group (n=34), compared to usable data for 192 strains for the 2001-2011 group (n=192). Combined with a large standard deviation in G_{10}

percentage between strains, this small sample means that finding a significant difference between the groups in terms of p-value was very unlikely, unless the difference itself was so large as to be obvious even prior to statistical testing. Likewise, there is no statistically significant difference between the mean percentage of G_{10} reads for the 2001-2011 (n=192) and post-2011 (n=1519) groups. In this case, the boxplot shows no difference between the means, but does show a large difference in the range: many more strains have a G_{10} percentage lower than one standard deviation less than the mean, as indicated by the stars in **Figure 5.5**. This suggests that, while overall there is no significant trend towards most strains gaining more non-G_{10} reads, the number of individual strains with lower G_{10} percentages is increasing; that is, the number of strains in which a reduced production of FHA could be detected using methods such as that described by Weigand et al. (2018) may be increasing. It is therefore possible that, if the same analysis were performed in ten years' time, a significant difference would be detected between the 2019-2029 group and those preceding it, although there is of course no way of knowing if this is true.

In contrast to the lack of significant temporal trends observed, a statistically significant difference was observed between the percentage of reads with G_{10} in Chinese strains compared to USA strains (means 90.38 and 87.10% respectively, p=0.006). The USA strains therefore appear to have a higher proportion of cells carrying an indel than the Chinese. If the ACV-induced trend hinted at in the temporal data was true, this finding would make sense: the WCV is still in use in China, whereas the USA switched to an FHA-containing ACV as early as 1996. However, there was a large difference in the sample sizes, with usable data for 1,675 strains from the USA, and only 35 from China. Data from additional recent Chinese strains would help to determine if this observed difference is an ongoing phenomenon, and also to investigate whether the difference is caused by the ACV or some other, as yet unknown, factor.

The strains possessing an indel at the 1,078 bp homopolymeric tract were termed "*fhaB2b*" for the purposes of categorisation here. As shown in **Figure 5.3**, the most prevalent *fhaB* allele is *fhaB2*; however, all but two (*fhaB3* and *fhaB22*) of the 22 alleles definitions available on PubMLST possess the same polyG tract 1078 bp into the *fhaB* sequence. This means that strains carrying alleles other than *fhaB2* could be affected by the slippage at the polyG tract and, in light of the prevalence of *fhaB2* in the currently circulating population, a large majority of strains not yet sequenced could also contain a mixed population of cells for this locus.

An informatics method to identify strains potentially FHA-deficient due to IS insertion into *fhaB* was also successful here. ISMapper showed IS *481* insertion into *fhaB* in several strains previously investigated in Weigand et al.'s detailed 2018 characterisation of FHA-deficient strains. Weigand et al. screened 787 recent and historic strains for FHA expression using their electrochemiluminescent antibody capture (ECL) assay, and identified five strains with absent or "severely reduced" FHA expression. Importantly, however, Weigand et al. did not find any obvious FHA deficiency-causing mutations in their assembled genome of J014, which was shown by the ECL assays to be FHA-deficient. By contrast, ISMapper, which uses raw Illumina reads rather than an assembled genome to identify IS insertions into genes, identified a copy of IS *481* at the insertion site 3,124 bp into *fhaB*. Likewise, Weigand et al. attributed their observed reduction in FHA expression in J199 solely to the presence of the homopolymeric slippage mutation thoroughly detailed here, whereas ISMapper also identified a copy of IS *481* at the 3,124 bp insertion site.

Given the picture of highly mixed intra-strain *B. pertussis* populations which has emerged throughout this, and previous chapters, it seems likely that the cell populations of both J014 and J199 contained a portion of cells with the IS *481* insertion within *fhaB*, and a portion of wild type cells. When the genomes for these strains were assembled by Weigand et al., the assembly tool may or may not have included the insertion, according to which sub-population the assembled reads came from. The opposite phenomenon could also be true for those strains identified here as having IS *481* within their *fhaB* but which had not been previously found to be FHA-deficient: enough of their cells may contain wild type *fhaB* that FHA expression is not yet severely reduced, but ISMapper was still able to detect those cells which contained the mutant. The same could be true for any of the other novel IS-containing strains identified here.

Wet- and dry-lab techniques are both required to fully explore FHA deficiency

Weigand et al. (2018) found one strain, J365, which appeared to have an IS *481* inserted at the last insertion site within the *fhaB* gene, at 9,865 bp (see **Figure 5.2**), yet was not FHA-deficient, indicating that IS *481* insertion late in the *fhaB* sequence does not prevent expression of the FHA protein. Three more of the strains identified here (**Table 5.2**) also appear to have an IS *481* at the 9,865 bp insertion site, suggesting these strains are likely not FHA-deficient.

The informatics approach, therefore, may be able to detect mutations in finer detail than the wet-lab approach, and certainly before they have begun to result in full FHA deficiency, but the evidence here suggests that both methods will be required for full resolution. The use of ISMapper alone would predict FHA deficiency in strains with IS *481* at the 9,865 bp *fhaB* insertion site, but Weigand et al.'s ECL assay shows that insertion at this site likely does not cause FHA deficiency. The ECL assay could also be used to validate the *fhaB*:1,078 indel results seen here as, to some extent, the results are non-binary, and can show a range of expression between fully deficient and fully normal expression (although the level of resolution possible is limited (Personal Communication, Michael Weigand, 2019)). In theory, mixed populations should result in a range of different FHA expressions, according to how many cells in the population have the non-mutated copy of *fhaB.* For example, J199, which has a mixed population (64% non-G_{10} reads), was found to have a low level of FHA expression in the ECL assay, which is in keeping with 36% of cells still possessing the non-mutated version of *fhaB* (Weigand et al. 2018). A variety of strains could be selected from the 1,753 screened here, with a range of G_{10} percentages, and predictions of relative FHA expression or deficiency tested.

Short- and long-read sequencing are required to fully explore FHA deficiency

Most of the analysis conducted here utilised Illumina short reads, primarily because of their high accuracy and potentially lower tendency to produce errors in homopolymeric tracts. However, the *fhaB:*1,078 indel phenomenon was originally observed in closed genome sequences produced using PacBio long reads, both here and in every other study which has observed the same mutation (Bart et al., 2015; Weigand et al., 2018; Xu, Z., Octavia, et al., 2019). Whilst short reads were required to investigate the mutation in detail, the mutation was only discovered because long reads enabled the assembly of closed genomes. On the other hand, ISMapper was also able to provide a likely cause for Weigand et al.'s observed FHA deficiency in J014, which was not apparent from the closed genome sequence alone. It seems therefore that, like hybrid *de novo* assembly, there remains a place for both short- and long-read sequencing in the detailed investigation of complex genetic phenomena.

The implications of mutations in *fhaB*

Increasing incidence of PRN deficiency has been well documented (for example: Hegerle et al., 2012; Queenan, Cassiday and Evangelista, 2013; Barkoff et al., 2019; Lam et al., 2014; Martin et al., 2015). Small numbers of strains deficient in FHA or PT have also been identified, often in tandem with PRN deficiency (for example: Bouchez et al., 2009; Bart et al., 2015; Williams, M.M. et al., 2016; Weigand et al., 2018). Here, a potential mutation which could lead to FHA deficiency was found to be common in intra-strain *B. pertussis* populations. However, the prevalence of the mutation within these populations was not 100%. This could lead to populations which have decreased, but not absent, expression of FHA. FHA is thought to have roles as a secreted protein, as well as a membrane-bound adhesin (Urisu, Cowell and Manclark, 1986; Relman et al., 1989; Cotter et al., 1998; Melvin et al., 2015). In culture, membrane-bound FHA alone is sufficient for the adhesion of *B. pertussis* to eukaryotic cells, whilst the secreted protein is involved with immune evasion. Mammalian studies have shown that FHA expression is vital, but not sufficient alone, to establish *B. pertussis* infection in the respiratory tract (Inatsuka, Julio and Cotter, 2005; Julio et al., 2009; Henderson, M.W. et al., 2012; Melvin et al., 2015). Strains which are fully FHA-deficient should therefore be less able, or completely unable, to establish infection in the respiratory tract. Nevertheless, the previous detection of FHA-deficient strains in patients suggests that, rarely, deficient strains may still be able to establish infection. This could be explained by the mixed populations of FHA mutations seen here: many cells could indeed be FHA deficient, but a small number of cells in the population remain able to express FHA. As a secreted protein, the FHA produced by the non-deficient cells could compensate for the deficient cells enough for successful immune evasion. Alternatively, infection could be established by populations of *B. pertussis* cells in which a majority are not yet FHA deficient, after which the pressure applied by a host immune system (particularly in a vaccinated host) could lead to a shift in population frequencies, and the eventual dominance of deficient cells. Although this would aid the cells in the current infection, the shifted population would then be less able to infect future hosts. Perhaps the mixed populations observed here are the result of a balance between the benefits of immune evasion in the current host and the ability to infect future hosts.

Although deficiency of FHA and PT appear to be more rare than deficiency of PRN, and seem unlikely to proliferate in the same way as PRN deficiency has due to their more vital roles during infection compared to PRN, the selective pressure applied by the ACV on the five component antigens means that increasing defiency in PT, FHA or Fim2/Fim3 is possible, particularly if the bacteria are able to develop ways to compensate for the absence of the seemingly vital proteins, as may have happened in the deficient strains previously isolated from whooping cough patients. The implications of antigen deficiency are clear. The ACV causes the immune system to recognise *B. pertussis* infection due to the presence of the ACV antigens on the *B. pertussis* cells; if one or more of these antigens are no longer present on the surface of the cell, the immune system may be less able to recognise the infection. Screening for increasing deficiency in all five antigens is therefore valuable. Given the complexities of *fhaB*, methods to screen for FHA deficiency could be particularly useful .

Finally, it should also be noted that Bart et al. (2015) observed an additional FHA-deficient strain in which they could not identify any mutations in any of the FHA-related genes, indicating that FHA deficiency may also arise by some as yet unidentified mechanism. The emerging picture of highly mixed *B. pertussis* intra-strain populations means that many genomic phenomena may exist, which have thus far been missed simply because of tendency to focus on only the most common genotypes.

Chapter 6: Conclusions and Future Work

"Your future hasn't been written yet, no one's has! Your future is whatever you make it, so make it a good one!"

- Doc Brown, Back to the Future, Part III

The aims of this project were to understand how long-read sequencing, and particularly nanopore sequencing, could be usefully applied to the study of *Bordetella pertussis*, and to elucidate what this application may reveal about the *B. pertussis* genome. Prominent in achieving these aims, and a theme that permeates throughout this thesis, has been the ongoing development and improvement of nanopore sequencing and its associated data analysis tools.

Ongoing development and improvement of long-read sequencing technologies and their associated data analysis tools

Before the major work for this thesis was begun, several attempts had been made (in 2014 and 2015) to sequence *B. pertussis* at the University of Bath using Oxford Nanopore Technologies' MinION sequencer. This sequencing was conducted as part of the "MinION Access Programme (MAP)", and used the earliest publicly available sequencers and sequencing chemistry, the MinION Mk1 and R7.3, respectively. Although these attempts did produce longer read lengths (5-6 kbp) than possible using most other NGS technologies, the tests were largely unsuccessful due to the poor yield and poor accuracy of the data produced: a 48-hour sequencing run on an R7.3 flow cell did not produce high enough coverage of the 4.1 Mb *B. pertussis* genome to assemble a closed sequence for even a single strain (see **Figure 1.16**). The performance of each flow cell was unpredictable, often having few pores available for sequencing, or rapidly losing available pores as sequencing progressed. These issues appear to have been common throughout the MAP community (Robison, 2017; Stack Exchange, 2017; van der Helm, 2017).

As this project began in 2016, however, a new sequencing chemistry was released, along with updated and improved basecalling and data analysis tools. The R9.4 flow cells were developed to increase the throughput and accuracy of each MinION sequencing run, and required a new version of the MinION: the Mk1b (Oxford Nanopore Technologies, 2016; Clarke, 2018). A sample of *B. pertussis* UK76 was sent to the University of California Santa Cruz, to test the throughput of the new sequencing chemistry. This yielded around 9.3 Gb over a 48-hour sequencing run, an order of magnitude higher than produced by the previous R7.3 runs over the same time period. Combined with the newer data analysis tools, this data could be used to assemble a closed UK76 genome; however, assembly with one of the more accurate assembly tools, Canu, took many days due to the extremely high coverage (Koren et al., 2017). Tests of smaller subsets of the data showed that equally accurate closed genome sequences could be assembled using just a twentieth of the data, indicating that multiple strains could now be sequenced during each sequencing run. This led directly into the work of Chapter 2, during which a barcoding protocol was tested, and five *B. pertussis* strains were sequenced on the same flow cell. As expected, this produced more than enough coverage (between 200 and 400x) for each of the five strains.

Many different *de novo* assembly and polishing tools have been, and continue to be, developed for long reads produced by nanopore (or PacBio) sequencing. The data produced in Chapter 2, combined with Illumina data previously produced for the five sequenced strains, was an ideal set to test the various different *de novo* assembly strategies available. This revealed that pre-correction with Canu, assembly with Flye, and polishing with Nanopolish produced the most accurate nanopore-only assemblies but, as could be expected, hybrid assemblies with both pre-corrected nanopore and Illumina data, produced with Unicycler, were the most accurate (on average, 99.48% accuracy for nanopore-only assemblies vs 99.69% for hybrids) (Loman, Quick and Simpson, 2015; Wick et al., 2017b; Kolmogorov et al., 2019). One of the major observations during these tests was how quickly the field was changing: new tools were rapidly becoming available and, in fact, the whole testing process had to be repeated from scratch when a new and more accurate basecalling tool (Albacore) replaced the MinKNOW basecaller used originally (Oxford Nanopore Technologies, 2017).

The trend of rapid development continued into the work of Chapter 3, where a new, quicker and cheaper library preparation kit could be used to produce 200x coverage for 12 strains (instead of five) per flow cell. In addition, the most accurate basecaller had changed again (the current favourite basecaller for the community is Guppy), and several of the tools identified as optimal in Chapter 2 had been replaced with newer versions (Wick, Judd and Holt, 2019). For example, Deepbinner replaced the demultiplexing tool Porechop, which resulted in fewer reads being discarded as having no recognisable barcode, whilst Medaka largely replaced the older, slower, and harder-to-use Nanopolish (Wick, 2017; Wick, Judd and Holt, 2018b; Luo et al., 2020). Using these newer tools, the accuracy of assemblies produced using nanopore reads increased to 99.63% - 0.15% higher than Chapter 2, and only 0.06% lower than the average hybrid accuracy in Chapter 2. In turn, the average accuracy of the hybrid assemblies also increased, though less dramatically, to 99.71%. The ability of MinION sequencing to consistently produce enough data to assemble closed genome sequences for multiple *B. pertussis* strains per flow cell is now well established. The remaining raw error present in MinION data, however, means that variant calling has remained a challenge.

Greig et al. (2019) conducted a thorough comparison of single-nucleotide variants identified in two Shiga toxin-producing *E. coli* isolates using raw, uncorrected Illumina and ONT reads. This work used GATK with some options customised for the more error-prone ONT reads, along with Minimap2 to map the long reads to a reference sequence (McKenna et al., 2010; Li, 2018). Many discrepancies between the variants called with the different reads were in repeat regions, likely because the Illumina reads could be mis-mapped whereas the ONT could not. When variants in repeat regions were masked, the vast majority of the remaining discrepancies were found to be “false positives” in the ONT data. These false positives usually occurred in the same motif (CC(A/T)GG), which has been associated with cytosine methylation. Greig et al. therefore used Nanopolish’s methylation-calling script on their raw Fast5 data, which showed that these “false positives” were indeed miscalled due to the influence of cytosine methylation in the raw ONT signal (Loman, Quick and Simpson, 2015). In some regards, ONT reads therefore outperform Illumina reads, as they do not result in false positive variant calls due to mis-mapping around repeat regions, and because epigenetic markers like methylation are detectable. However, the methylated bases still resulted in false positive variant calls, meaning it was necessary to mask them from the variant analysis. After they were masked, up to ten remaining discrepancies were noted between the variants called using the two types of sequencing reads in the two isolates; most of these were bases called as variant in the ONT data but not the Illumina. It is likely these discrepancies were indeed false positives in the ONT data, due to its higher raw non-random error profile. In addition, to achieve almost the same variant calls with the ONT data as with the Illumina data, lengthy extra steps (i.e. Nanopolish methylation calling) were required.

Variant calling with ONT data is therefore possible, and the results are almost as accurate as those made with Illumina data are assumed to be. At the moment, however, Illumina data remains optimal for this process, although repeat regions should be masked from any analysis, as they are prone to false positive variant calls. Nonetheless, if only ONT data was available, or if variant calls were required rapidly, ONT data produced with the most recent flow cells and data tools may now be sufficient. In addition, the “high accuracy” mode of the Guppy basecaller, combined with the nanopore-specific variant calling tools such as Clair, Longshot, NanoCaller or even Medaka, have recently been used to produce accurate variant calls at the single base level (Edge and Bansal, 2019; Ahsan et al., 2020; Gilpatrick et al., 2020; Luo et al., 2020). Future work for the *B. pertussis* nanopore field will therefore include testing these newer and more accurate tools, to establish whether single nucleotide variants can be called as accurately with nanopore data as with Illumina data. At the moment, the CPU-usage requirements of Guppy's high accuracy mode are prohibitive, with some reports of a single MinION run taking several months to basecall (P. Kover, personal communication, 2020). However, high powered computer clusters, particularly those with GPU capabilities, should be able to basecall a single run in little more than a day.

How we can use long-read sequencing to study *B. pertussis*, and what that can reveal about the genome and evolution of the bacteria

As discussed, and as demonstrated in Chapters 2 and 3, the assembly of closed *B. pertussis* genomes using nanopore sequencing is now easily achievable. In addition, PacBio sequencing has been capable of the same since the mid-2010s, although this usually requires sending DNA to a sequencing service so could be more time-consuming or expensive (Bart, Zeddeman, et al., 2014). The vast increase in numbers of available closed *B. pertussis* sequences is allowing the investigation of genome-level variation in *B. pertussis* in a way which was not previously possible.

The first trait to have become apparent from the study of closed *B. pertussis* genomes is large-scale genomic rearrangements. This has been thoroughly investigated and characterised at the Centers for Disease Control (CDC) in the USA, and it is now clear that major rearrangements, mediated by homologous recombination between copies of IS *481* and centred around the origin of replication, are occurring frequently (Weigand et al., 2017; Weigand et al., 2019). Although the existence of different genome structures in *B. pertussis* has been hypothesised for decades, based on pulsed field gel electrophoresis (PFGE) data, the assembly of closed genome sequences has, for the first time, confirmed the variety of arrangements seen, and allowed the documentation of these arrangements in a reproducible manner. For example, in Chapter 3 here, it was possible to compare the arrangements of the New Zealand isolates' genomes with the arrangement types defined by the CDC's work. This was possible because closed genome sequences can be shared far more easily than isolates, which would have been required for such comparisons using PFGE. Chapter 3, as well as the work by the CDC, raises a few questions about genomic rearrangement in *B. pertussis*. For example, given the number and location of copies of IS *481* throughout the genome, not every potential genome arrangement type has so far been observed, despite the study of hundreds of isolates (Weigand et al., 2017). Additionally, certain arrangement types appear more frequently than others: in the (mainly USA) isolates studied by Weigand et al., the most commonly seen arrangement types were CDC237 and CDC002, whilst amongst the Chapter 3 New Zealand isolates, the most commonly seen arrangement type was CDC010. This suggests rearrangement may not be an entirely random process, and that certain arrangement types may have a selective advantage over others. As yet, little *in vivo* work has been conducted to demonstrate any correlation between *B. pertussis* genome arrangement and phenotype. Future work may therefore address this knowledge gap although, as seen in Chapter 4, observing and comparing phenotypes in *B. pertussis* is not always a straightforward task.

Another emerging genome-level trait in *B. pertussis* is the duplication of relatively large regions of DNA. In Chapter 2, a duplication of several hundred thousand base pairs was seen in two of the five UK strains sequenced. In Chapter 3, a 15 kbp duplication was seen in two New Zealand strains. Both of these duplications have been investigated further elsewhere, and shown to appear in a number of other, unrelated, strains (Abrahams et al., 2020). Duplications in the *B. pertussis* genome, like rearrangement, are the result of homologous recombination between copies of IS *481*, and could therefore occur at random. Nonetheless, like arrangement types, certain duplications appear to be more frequent than others, suggesting they may convey some selective advantage. This is another trait which would benefit from *in vivo* work to investigate the phenotypes resulting from duplications; indeed, Chapter 4 covers attempts to investigate the phenotypes resulting from the ultra-long duplication identified in Chapter 2. However, one of the other observations from Abrahams et al. (2020), identified through the use of ultra-long nanopore reads produced for UK54 (not discussed here), was that both duplications and arrangements are present as mixed populations in *B. pertussis*. Within populations of cells from the same strain, some cells can contain no duplication, versions of the duplication starting at slightly different points in the genome, or even greater copy number variants (the "duplication" seen in the New Zealand strains and in UK54 appeared 60+ times in some cells in the UK54 population). Likewise, different cells in the population can contain different genomic arrangements. These fluctuating, seemingly highly dynamic, populations may help to explain why observing phenotypes in *B. pertussis* is so difficult: populations of cells from the same strain may not behave in exactly the same way every time they are grown, due to ongoing changes in CNVs and genome arrangements in individual cells in the population. Going forward, methods will need to be developed to account for these potential differences in populations when strains are observed *in vivo*. This may involve producing ultra-long nanopore reads for each population observed *in vivo*, in order to characterise the CNV and arrangement landscape at the time of testing, although this would currently be very expensive, as each ultra-long sequencing run uses a whole flow cell per strain.

Finally, Chapter 5 adds to this emerging "mixed population" story in a slightly different way. Closed genome sequences allowed the identification of variation at the single base level in *fhaB*. Having observed the variation in closed sequences, a method could then be developed to screen the unassembled short read data for every *B. pertussis* strain currently available from public databases. The mixed populations of *fhaB* in *B. pertussis* are likely not unique: similar mixed populations of cells with mutations or no mutations are probably present in many other genes, particularly in regions prone to mutations such as homopolymeric slippage. For a bacterium once described as being "clonal" and having little variation between different strains, long-read sequencing is increasingly revealing unexpected levels of variation in *B. pertussis* even between cells from the same strain.

References

“Until I feared I would lose it, I never loved to read. One does not love breathing.”

- Harper Lee, To Kill a Mockingbird

ABM Knowledge Base, 2017. *Next generation sequencing, an introduction* [Online]. Available from: https://www.abmgood.com/marketing/knowledge_base/next_generation_sequencing_introduction.php [Accessed 29 September 2017].

Abrahams, J.S., Weigand, M.A., Ring, N., MacArthur, I., Peng, S., Williams, M.M., Bready, B., Catalano, A.P., Davis, J.R., Kaiser, M.D., Oliver, J.S., Sage, J.M., Bagby, S., Tondella, M.L., Gorringe, A. and Preston, A., 2020. Duplications drive diversity in Bordetella pertussis on an underestimated scale. *bioRxiv*, p. 2020.2002.2006.937284.

Abrahams, J.S., Weigand, M.A., Ring, N., MacArthur, I., Peng, S., Williams, M.M., Bready, B., Catalano, A.P., Davis, J.R., Kaiser, M.D., Oliver, J.S., Sage, J.M., Bagby, S., Tondella, M.L., Gorringe, A. and Preston, A., In review. Duplications drive diversity in Bordetella pertussis on an underestimated scale. *Genome Biology*.

Abyzov, A., Urban, A.E., Snyder, M. and Gerstein, M., 2011. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res,* 21(6), pp. 974-984.

Advani, A., Donnelly, D. and Hallander, H., 2004. Reference System for Characterization of Bordetella pertussis Pulsed-Field Gel Electrophoresis Profiles. *J Clin Microbiol,* 42(7), pp. 2890-2897.

Advani, A., Gustafsson, L., Ahren, C., Mooi, F.R. and Hallander, H.O., 2011. Appearance of Fim3 and ptxP3-Bordetella pertussis strains, in two regions of Sweden with different vaccination programs. *Vaccine,* 29(18), pp. 3438-3442.

Advani, A., Hallander, H.O., Dalby, T., Krogfelt, K.A., Guiso, N., Njamkepo, E., von Konnig, C.H., Riffelmann, M., Mooi, F.R., Sandven, P., Lutynska, A., Fry, N.K., Mertsola, J. and He, Q., 2013. Pulsed-field gel electrophoresis analysis of Bordetella pertussis isolates circulating in Europe from 1998 to 2009. *J Clin Microbiol,* 51(2), pp. 422-428.

Ahsan, U., Liu, Q., Fang, L. and Wang, K., 2020. NanoCaller for accurate detection of SNPs and small indels from long-read sequencing by deep neural networks. *bioRxiv*.

Aird, D., Ross, M.G., Chen, W.S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D.B., Nusbaum, C. and Gnirke, A., 2011. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol,* 12(2), p. R18.

Akerley, B.J., Cotter, P.A. and Miller, J.F., 1995. Ectopic expression of the flagellar regulon alters development of the Bordetella-host interaction. *Cell,* 80(4), pp. 611-620.

Akerley, B.J. and Miller, J.F., 1993. Flagellin gene transcription in Bordetella bronchiseptica is regulated by the BvgAS virulence control system. *J Bacteriol,* 175(11), pp. 3468-3479.

Akerley, B.J., Monack, D.M., Falkow, S. and Miller, J.F., 1992. The bvgAS locus negatively controls motility and synthesis of flagella in Bordetella bronchiseptica. *J Bacteriol,* 174(3), pp. 980-990.

Akeson, M., Branton, D., Kasianowicz, J.J., Brandin, E. and Deamer, D.W., 1999. Microsecond time-scale discrimination among polycytidylic acid, polyadenylic acid, and polyuridylic acid as homopolymers or as segments within single RNA molecules. *Biophys J,* 77(6), pp. 3227-3233.

Altunaiji, S., Kukuruzovic, R., Curtis, N. and Massie, J., 2007. Antibiotics for whooping cough (pertussis). *Cochrane Database Syst Rev,* (3), p. Cd004404.

Amirthalingam, G., Gupta, S. and Campbell, H., 2013. Pertussis immunisation and control in England and Wales, 1957 to 2012: a historical review. *Euro Surveill,* 18(38).

Amman, F., D'Halluin, A., Antoine, R., Huot, L., Bibova, I., Keidel, K., Slupek, S., Bouquet, P., Coutte, L., Caboche, S., Locht, C., Vecerek, B. and Hot, D., 2018. Primary transcriptome analysis reveals importance of IS elements for the shaping of the transcriptional landscape of Bordetella pertussis. *RNA Biol,* 15(7), pp. 967-975.

Andreasen, C. and Carbonetti, N.H., 2008. Pertussis toxin inhibits early chemokine production to delay neutrophil recruitment in response to Bordetella pertussis respiratory tract infection in mice. *Infect Immun,* 76(11), pp. 5139-5148.

Andrews, R., Herceg, A. and Roberts, C., 1997. Pertussis notifications in Australia, 1991 to 1997. *Commun Dis Intell,* 21(11), pp. 145-148.

Aquino-Andrade, A., Martinez-Leyva, G., Merida-Vieyra, J., Saltigeral, P., Lara, A., Dominguez, W., Garcia de la Puente, S. and De Colsa, A., 2017. Real-Time Polymerase Chain Reaction-Based Detection of Bordetella pertussis in Mexican Infants and Their Contacts: A 3-Year Multicenter Study. *J Pediatr,* 188, pp. 217-223.e211.

Ardui, S., Ameur, A., Vermeesch, J.R. and Hestand, M.S., 2018. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic Acids Res,* 46(5), pp. 2159-2168.

Argimon, S., Abudahab, K., Goater, R.J.E., Fedosejev, A., Bhai, J., Glasner, C., Feil, E.J., Holden, M.T.G., Yeats, C.A., Grundmann, H., Spratt, B.G. and Aanensen, D.M., 2016. Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. *Microb Genom,* 2(11), p. e000093.

Arico, B. and Rappuoli, R., 1987. Bordetella parapertussis and Bordetella bronchiseptica contain transcriptionally silent pertussis toxin genes. *J Bacteriol,* 169(6), pp. 2847-2853.

Ashkenasy, N., Sanchez-Quesada, J., Bayley, H. and Ghadiri, M.R., 2005. Recognizing a single base in an individual DNA strand: a step toward DNA sequencing in nanopores. *Angew Chem Int Ed Engl,* 44(9), pp. 1401-1404.

Ashworth, L.A., Irons, L.I. and Dowsett, A.B., 1982. Antigenic relationship between serotype-specific agglutinogen and fimbriae of Bordetella pertussis. *Infect Immun,* 37(3), pp. 1278-1281.

Aslanabadi, A., Ghabili, K., Shad, K., Khalili, M. and Sajadi, M.M., 2015. Emergence of whooping cough: notes from three early epidemics in Persia. *Lancet Infect Dis,* 15(12), pp. 1480-1484.

atdbio, 2019. *Next Generation Sequencing* [Online]. Available from: https://www.atdbio.com/content/58/Next-generation-sequencing [Accessed 22 August 2019].

Au, K.F., Underwood, J.G., Lee, L. and Wong, W.H., 2012. Improving PacBio Long Read Accuracy by Short Read Alignment. *PLoS One,* 7(10).

Ausiello, C.M. and Cassone, A., 2014. Acellular Pertussis Vaccines and Pertussis Resurgence: Revise or Replace? *mBio,* 5(3).

Bairoch, A. and Apweiler, R., 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res,* 28(1), pp. 45-48.

Baker, J.R. and Ross, H.M., 1992. The role of bacteria in phocine distemper. *Sci Total Environ,* 115(1-2), pp. 9-14.

Balzer, S., Malde, K. and Jonassen, I., 2011. Systematic exploration of error sources in pyrosequencing flowgram data. *Bioinformatics,* 27(13), pp. i304-309.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., Pyshkin, A.V., Sirotkin, A.V., Vyahhi, N., Tesler, G., Alekseyev, M.A. and Pevzner, P.A., 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J Comput Biol,* 19(5), pp. 455-477.

Barkoff, A.M., Mertsola, J., Pierard, D., Dalby, T., Hoegh, S.V., Guillot, S., Stefanelli, P., van Gent, M., Berbers, G., Vestrheim, D., Greve-Isdahl, M., Wehlin, L., Ljungman, M., Fry, N.K., Markey, K. and He, Q., 2019. Pertactin-deficient Bordetella pertussis isolates: evidence of increased circulation in Europe, 1998 to 2015. *Euro Surveill,* 24(7).

Bart, M.J., Harris, S.R., Advani, A., Arakawa, Y., Bottero, D., Bouchez, V., Cassiday, P.K., Chiang, C.S., Dalby, T., Fry, N.K., Gaillard, M.E., van Gent, M., Guiso, N., Hallander, H.O., Harvill, E.T., He, Q., van der Heide, H.G., Heuvelman, K., Hozbor, D.F., Kamachi, K., Karataev, G.I., Lan, R., Lutynska, A., Maharjan, R.P., Mertsola, J., Miyamura, T., Octavia, S., Preston, A., Quail, M.A., Sintchenko, V., Stefanelli, P., Tondella, M.L., Tsang, R.S., Xu, Y., Yao, S.M., Zhang, S., Parkhill, J. and Mooi, F.R., 2014. Global population structure and evolution of Bordetella pertussis and their relationship with vaccination. *MBio,* 5(2), p. e01074.

Bart, M.J., van der Heide, H.G.J., Zeddeman, A., Heuvelman, K., van Gent, M. and Mooi, F.R., 2015. Complete Genome Sequences of 11 Bordetella pertussis Strains Representing the Pandemic ptxP3 Lineage. *Genome Announc,* 3(6).

Bart, M.J., van Gent, M., van der Heide, H.G., Boekhorst, J., Hermans, P., Parkhill, J. and Mooi, F.R., 2010. Comparative genomics of prevaccination and modern Bordetella pertussis strains. *BMC Genomics,* 11, p. 627.

Bart, M.J., Zeddeman, A., van der Heide, H.G.J., Heuvelman, K., van Gent, M. and Mooi, F.R., 2014. Complete Genome Sequences of Bordetella pertussis Isolates B1917 and B1920, Representing Two Predominant Global Lineages. *Genome Announc,* 2(6).

Bass, J.W. and Wittler, R.R., 1994. Return of epidemic pertussis in the United States. *Pediatr Infect Dis J,* 13(5), pp. 343-345.

Bayliss, S.C., Hunt, V.L., Yokoyama, M., Thorpe, H.A. and Feil, E.J., 2017. The use of Oxford Nanopore native barcoding for complete genome assembly. *GigaScience,* gix001.

Belcher, T., 2017. *Investigating the growth and metabolic difference of Bvg+ and Bvg- phase Bordetella pertussis.* (PhD), University of Bath. Available from: https://researchportal.bath.ac.uk/en/studentTheses/investigating-the-growth-and-metabolic-difference-of-bvg-and-bvg- [Accessed 15 October 2019].

Belcher, T. and Preston, A., 2015. Bordetella pertussis evolution in the (functional) genomics era. *Pathog Dis,* 73(8), p. ftv064.

Ben Fraj, I., Bouchez, V., Smaoui, H., Kechrid, A. and Brisse, S., 2019. Genome characteristics of Bordetella pertussis isolates from Tunisia. *J Med Microbiol*.

Ben Fraj, I., Kechrid, A., Guillot, S., Bouchez, V., Brisse, S., Guiso, N. and Smaoui, H., 2019. Pertussis epidemiology in Tunisian infants and children and characterization of Bordetella pertussis isolates: results of a 9-year surveillance study, 2007 to 2016. *J Med Microbiol,* 68(2), pp. 241-247.

Benjamini, Y. and Speed, T.P., 2012. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res,* 40(10), p. e72.

Benner, S., Chen, R.J., Wilson, N.A., Abu-Shumays, R., Hurt, N., Lieberman, K.R., Deamer, D.W., Dunbar, W.B. and Akeson, M., 2007. Sequence-specific detection of individual DNA polymerase complexes in real time using a nanopore. *Nat Nanotechnol,* 2(11), pp. 718-724.

Bennett, S., 2004. Solexa Ltd. *Pharmacogenomics,* 5(4), pp. 433-438.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., Boutell, J.M., Bryant, J., Carter, R.J., Keira Cheetham, R., Cox, A.J., Ellis, D.J., Flatbush, M.R., Gormley, N.A., Humphray, S.J., Irving, L.J., Karbelashvili, M.S., Kirk, S.M., Li, H., Liu, X., Maisinger, K.S., Murray, L.J., Obradovic, B., Ost, T., Parkinson, M.L., Pratt, M.R., Rasolonjatovo, I.M., Reed, M.T., Rigatti, R., Rodighiero, C., Ross, M.T., Sabot, A., Sankar, S.V., Scally, A., Schroth, G.P., Smith, M.E., Smith, V.P., Spiridou, A., Torrance, P.E., Tzonev, S.S., Vermaas, E.H., Walter, K., Wu, X., Zhang, L., Alam, M.D., Anastasi, C., Aniebo, I.C., Bailey, D.M., Bancarz, I.R., Banerjee, S., Barbour, S.G., Baybayan, P.A., Benoit, V.A., Benson, K.F., Bevis, C., Black, P.J., Boodhun, A., Brennan, J.S., Bridgham, J.A., Brown, R.C., Brown, A.A., Buermann, D.H., Bundu, A.A., Burrows, J.C., Carter, N.P., Castillo, N., Chiara, E.C.M., Chang, S., Neil Cooley, R., Crake, N.R., Dada, O.O., Diakoumakos, K.D., Dominguez-Fernandez, B., Earnshaw, D.J., Egbujor, U.C., Elmore, D.W., Etchin, S.S., Ewan, M.R., Fedurco, M., Fraser, L.J., Fuentes Fajardo, K.V., Scott Furey, W., George, D., Gietzen, K.J., Goddard, C.P., Golda, G.S., Granieri, P.A., Green, D.E., Gustafson, D.L., Hansen, N.F., Harnish, K., Haudenschild, C.D., Heyer, N.I., Hims, M.M., Ho, J.T., Horgan, A.M., et al., 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature,* 456(7218), pp. 53-59.

Bisgard, K.M., Christie, C.D., Reising, S.F., Sanden, G.N., Cassiday, P.K., Gomersall, C., Wattigney, W.A., Roberts, N.E. and Strebel, P.M., 2001. Molecular epidemiology of Bordetella pertussis by pulsed-field gel electrophoresis profile: Cincinnati, 1989-1996. *J Infect Dis,* 183(9), pp. 1360-1367.

Bjornstad, O.N. and Harvill, E.T., 2005. Evolution and emergence of Bordetella in humans. *Trends Microbiol,* 13(8), pp. 355-359.

Blumberg, D.A., Lewis, K., Mink, C.M., Christenson, P.D., Chatfield, P. and Cherry, J.D., 1993. Severe reactions associated with diphtheria-tetanus-pertussis vaccine: detailed study of children with seizures, hypotonic-hyporesponsive episodes, high fevers, and persistent crying. *Pediatrics,* 91(6), pp. 1158-1165.

Boccard, F., Esnault, E. and Valens, M., 2005. Spatial arrangement and macrodomain organization of bacterial chromosomes. *Mol Microbiol,* 57(1), pp. 9-16.

Bolger, A.M., Lohse, M. and Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics,* 30(15), pp. 2114-2120.

Bordet, J. and Gengou, O., 1906. Le microbe de la coqueluche. *Ann. Inst. Pasteur,* 20, p. 731.

Bouchez, V., Baines, S.L., Guillot, S. and Brisse, S., 2018. Complete Genome Sequences of Bordetella pertussis Clinical Isolate FR5810 and Reference Strain Tohama from Combined Oxford Nanopore and Illumina Sequencing. *Microbiol Resour Announc,* 7(19).

Bouchez, V., Brun, D., Cantinelli, T., Dore, G., Njamkepo, E. and Guiso, N., 2009. First report and detailed characterization of B. pertussis isolates not expressing Pertussis Toxin or Pertactin. *Vaccine,* 27(43), pp. 6034-6041.

Bouchez, V., Caro, V., Levillain, E., Guigon, G. and Guiso, N., 2008. Genomic Content of Bordetella pertussis Clinical Isolates Circulating in Areas of Intensive Children Vaccination. *PLoS One,* 3(6).

Bouchez, V., Hegerle, N., Strati, F., Njamkepo, E. and Guiso, N., 2015. New Data on Vaccine Antigen Deficient Bordetella pertussis Isolates. *Vaccines (Basel),* 3(3), pp. 751-770.

Bowden, K.E., Weigand, M.A., Peng, Y., Cassiday, P., Sammons, S., Knipe, K., Rowe, L., Loparev, V., Sheth, M., Weening, K., Tondella, M.L., Williams, M.M. and Blokesch, M., 2016. Genome Structural Diversity among 31 Bordetella pertussis Isolates from Two Recent U.S. Whooping Cough Statewide Epidemics. *Molecular biology and physiology,* 1(3).

Bowden, K.E., Williams, M.M., Cassiday, P.K., Milton, A., Pawloski, L., Harrison, M., Martin, S.W., Meyer, S., Qin, X., DeBolt, C., Tasslimi, A., Syed, N., Sorrell, R., Tran, M., Hiatt, B. and Tondella, M.L., 2014. Molecular epidemiology of the pertussis epidemic in Washington State in 2012. *J Clin Microbiol,* 52(10), pp. 3549-3557.

Branton, D. and Deamer, D., 2018. The development of nanopore sequencing. In: D. Branton and D. Deamer, eds. *Nanopore Sequencing: An Introduction.* New Jersey: World Scientific Co. Pte. Ltd., pp. 1-16.

Brinig, M.M., Cummings, C.A., Sanden, G.N., Stefanelli, P., Lawrence, A. and Relman, D.A., 2006. Significant Gene Order and Expression Differences in Bordetella pertussis Despite Limited Gene Content Variation.

British Medical Journal, 1981. Pertussis vaccine. *Br Med J (Clin Res Ed),* 282(6276), pp. 1563-1564.

Burns, D.L., Meade, B.D. and Messionnier, N.E., 2014. Pertussis resurgence: perspectives from the Working Group Meeting on pertussis on the causes, possible paths forward, and gaps in our knowledge. *J Infect Dis,* 209 Suppl 1, pp. S32-35.

Busch, A., Phan, G. and Waksman, G., 2015. Molecular mechanism of bacterial type 1 and P pili assembly. *Philos Trans A Math Phys Eng Sci,* 373(2036).

Butler, T.Z., Pavlenok, M., Derrington, I.M., Niederweis, M. and Gundlach, J.H., 2008. Single-molecule DNA detection with an engineered MspA protein nanopore. *Proc Natl Acad Sci U S A,* 105(52), pp. 20647-20652.

Byers, R.K. and Moll, F.C., 1948. Encephalopathies following prophylactic pertussis vaccine. *Pediatrics,* 1(4), pp. 437-457.

Byrd, A.K. and Raney, K.D., 2018. Helicases and DNA motor proteins. In: D. Branton and D. Deamer, eds. *Nanopore Sequencing: An Introduction.* New Jersey: World Scientific Co. Pte. Ltd., pp. 59-74.

Calon, T.G.A., Trobos, M., Johansson, M.L., van Tongeren, J., van der Lugt-Degen, M., Janssen, A.M.L., Savelkoul, P.H.M., Stokroos, R.J. and Budding, A.E., 2019. Microbiome on the Bone-Anchored Hearing System: A Prospective Study. *Front Microbiol,* 10, p. 799.

Canadian Public Health Association, 2019. *Immunization timeline | Canadian Public Health Association* [Online]. Available from: https://www.cpha.ca/immunization-timeline [Accessed 25 June 2019].

Carbonetti, N.H., 2007. Immunomodulation in the pathogenesis of Bordetella pertussis infection and disease. *Curr Opin Pharmacol,* 7(3), pp. 272-278.

Carbonetti, N.H., 2010. Pertussis toxin and adenylate cyclase toxin: key virulence factors of Bordetella pertussis and cell biology tools. *Future Microbiol,* 5, pp. 455-469.

Caro, V., Bouchez, V. and Guiso, N., 2008. Is the Sequenced Bordetella pertussis strain Tohama I representative of the species? *J Clin Microbiol,* 46(6), pp. 2125-2128.

Caro, V., Hot, D., Guigon, G., Hubans, C., Arrive, M., Soubigou, G., Renauld-Mongenie, G., Antoine, R., Locht, C., Lemoine, Y. and Guiso, N., 2006. Temporal analysis of French Bordetella pertussis isolates by comparative whole-genome hybridization. *Microbes Infect,* 8(8), pp. 2228-2235.

Carrillo, A.M., 2017. Vaccine production, national security anxieties and the unstable state in nineteenth- and twentieth-century Mexico. In: C. Holmberg, S. Blume and P. Greenough, eds. *The politics of vaccination: a global history.* Manchester, UK: Manchester University Press, pp. 121-146.

Cattelan, N., Dubey, P., Arnal, L., Yantorno, O.M. and Deora, R., 2016. Bordetella biofilms: a lifestyle leading to persistent infections. *Pathog Dis,* 74(1), p. ftv108.

Cavaillon, J.M., Sansonetti, P. and Goldman, M., 2019. 100th Anniversary of Jules Bordet's Nobel Prize: Tribute to a Founding Father of Immunology. *Front Immunol,* 10.

Centers for Disease Control, 2019a. *Enhanced Pertussis Surveillance* [Online]. Available from: https://www.cdc.gov/abcs/methodology/pertussis-surveillance.html [Accessed 21 September 2019].

Centers for Disease Control, 2019b. *Reported Pertussis Incidence by Age Group and Year | CDC* [Online]. Available from: https://www.cdc.gov/pertussis/surv-reporting/cases-by-age-group-and-year.html#modalIdString_CDCTable_0 [Accessed 20 Feb 2019].

Charles, I.G., Dougan, G., Pickard, D., Chatfield, S., Smith, M., Novotny, P., Morrissey, P. and Fairweather, N.F., 1989. Molecular cloning and characterization of protective outer membrane protein P.69 from Bordetella pertussis. *Proc Natl Acad Sci U S A,* 86(10), pp. 3554-3558.

Chen, Q. and Stibitz, S., 2019. The BvgASR virulence regulon of Bordetella pertussis. *Curr Opin Microbiol,* 47, pp. 74-81.

Chen, Y.C., Liu, T., Yu, C.H., Chiang, T.Y. and Hwang, C.C., 2013. Effects of GC Bias in Next-Generation-Sequencing Data on De Novo Genome Assembly. *PLoS One,* 8(4).

Chen, Z. and He, Q., 2017. Immune persistence after pertussis vaccination. *Human Vaccines and Immunotherapeutics,* 13(4), pp. 744-756.

Cherf, G.M., Lieberman, K.R., Rashid, H., Lam, C.E., Karplus, K. and Akeson, M., 2012. Automated forward and reverse ratcheting of DNA in a nanopore at 5-A precision. *Nat Biotechnol,* 30(4), pp. 344-348.

Cherry, J.D., 1984. The epidemiology of pertussis and pertussis immunization in the United Kingdom and the United States: a comparative study. *Curr Probl Pediatr,* 14(2), pp. 1-78.

Cherry, J.D., 1990. 'Pertussis vaccine encephalopathy': it is time to recognize it as the myth that it is. *Jama,* 263(12), pp. 1679-1680.

Cherry, J.D., 1992. Pertussis: the trials and tribulations of old and new pertussis vaccines. *Vaccine,* 10(14), pp. 1033-1038.

Cherry, J.D., 1996. Historical review of pertussis and the classical vaccine. *J Infect Dis,* 174 Suppl 3, pp. S259-263.

Cherry, J.D., 1999. Epidemiological, clinical, and laboratory aspects of pertussis in adults. *Clin Infect Dis,* 28 Suppl 2, pp. S112-117.

Cherry, J.D., 2015. The History of Pertussis (Whooping Cough); 1906-2015: Facts, Myths, and Misconceptions. *Infectious Disease Epidemiology,* 2(2), pp. 120-130.

Cherry, J.D., 2019. The 112-Year Odyssey of Pertussis and Pertussis Vaccines—Mistakes Made and Implications for the Future. *Journal of the Pediatric Infectious Diseases Society*.

Cherry, J.D., Gornbein, J., Heininger, U. and Stehr, K., 1998. A search for serologic correlates of immunity to Bordetella pertussis cough illnesses. *Vaccine,* 16(20), pp. 1901-1906.

Cherry, J.D., Holtzman, A.E., Shields, W.D., Buch, D., Nielsen, C., Jacobsen, V., Christenson, P.D. and Zachau-Christiansen, B., 1993. Pertussis immunization and characteristics related to first seizures in infants and children. *J Pediatr,* 122(6), pp. 900-903.

Chin, C.-S., Alexander, D.H., Marks, P., Klammer, A.A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., Eichler, E.E., Turner, S.W. and Korlach, J., 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods,* 10, pp. 563-569.

Church, G., Deamer, D., Branton, D., Balderelli, R. and Kasianowicz, J.J., 1998. *Characterization of individual polymer molecules based on monomer-interface interactions*. U.S. Patent 5,795,782. 18 Aug 1998.

Clark, T.A., 2014. Changing pertussis epidemiology: everything old is new again. *J Infect Dis,* 209(7), pp. 978-981.

Clarke, J., 2018. Development of multipore sequencing instruments. In: D. Branton and D. Deamer, eds. *Nanopore Sequencing: An Introduction.* New Jersey: World Scientific Co. Pte. Ltd, pp. 75-90.

Cockroft, S.L., Chu, J., Amorin, M. and Ghadiri, M.R., 2008. A single-molecule nanopore device detects DNA polymerase activity with single-nucleotide resolution. *J Am Chem Soc,* 130(3), pp. 818-820.

Cody, C.L., Baraff, L.J., Cherry, J.D., Marcy, S.M. and Manclark, C.R., 1981. Nature and rates of adverse reactions associated with DTP and DT immunizations in infants and children. *Pediatrics,* 68(5), pp. 650-660.

Conlan, S., Thomas, P.J., Deming, C., Park, M., Lau, A.F., Dekker, J.P., Snitkin, E.S., Clark, T.A., Luong, K., Song, Y., Tsai, Y.C., Boitano, M., Dayal, J., Brooks, S.Y., Schmidt, B., Young, A.C., Thomas, J.W., Bouffard, G.G., Blakesley, R.W., Mullikin, J.C., Korlach, J., Henderson, D.K., Frank, K.M., Palmore, T.N. and Segre, J.A., 2014. Single-molecule sequencing to track plasmid diversity of hospital-associated carbapenemase-producing Enterobacteriaceae. *Sci Transl Med,* 6(254), p. 254ra126.

Connor, T.R., Loman, N.J., Thompson, S., Smith, A., Southgate, J., Poplawski, R., Bull, M.J., Richardson, E., Ismail, M., Thompson, S.E.-., Kitchen, C., Guest, M., Bakke, M., Sheppard, S.K. and Pallen, M.J., 2016. CLIMB (the Cloud Infrastructure for Microbial Bioinformatics): an online resource for the medical microbiology community. *Microbial Genomics,* 2016(2).

Cotter, P.A. and Jones, A.M., 2003. Phosphorelay control of virulence gene expression in Bordetella. *Trends Microbiol,* 11(8), pp. 367-373.

Cotter, P.A. and Miller, J.F., 1994. BvgAS-mediated signal transduction: analysis of phase-locked regulatory mutants of Bordetella bronchiseptica in a rabbit model. *Infect Immun,* 62(8), pp. 3381-3390.

Cotter, P.A. and Miller, J.F., 1997. A mutation in the Bordetella bronchiseptica bvgS gene results in reduced virulence and increased resistance to starvation, and identifies a new class of Bvg-regulated antigens. *Mol Microbiol,* 24(4), pp. 671-685.

Cotter, P.A., Yuk, M.H., Mattoo, S., Akerley, B.J., Boschwitz, J., Relman, D.A. and Miller, J.F., 1998. Filamentous hemagglutinin of Bordetella bronchiseptica is required for efficient establishment of tracheal colonization. *Infect Immun,* 66(12), pp. 5921-5929.

Croinin, T.O., Grippe, V.K. and Merkel, T.J., 2005. Activation of the vrg6 promoter of Bordetella pertussis by RisA. *J Bacteriol,* 187(5), pp. 1648-1658.

Darch, S.E., McNally, A., Harrison, F., Corander, J., Barr, H.L., Paszkiewicz, K., Holden, S., Fogarty, A., Crusz, S.A. and Diggle, S.P., 2014. Recombination is a key driver of genomic and phenotypic diversity in a Pseudomonas aeruginosa population during cystic fibrosis infection. *Scientific Reports, Published online: 12 January 2015; | doi:10.1038/srep07649*.

Darling, A.E., Mau, B. and Perna, N.T., 2010. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One,* 5(6), p. e11147.

de Gouw, D., Hermans, P.W., Bootsma, H.J., Zomer, A., Heuvelman, K., Diavatopoulos, D.A. and Mooi, F.R., 2014. Differentially expressed genes in Bordetella pertussis strains belonging to a lineage which recently spread globally. *PLoS One,* 9(1), p. e84523.

de Melker, H.E., Conyn-van Spaendonck, M.A., Rumke, H.C., van Wijngaarden, J.K., Mooi, F.R. and Schellekens, J.F., 1997. Pertussis in The Netherlands: an outbreak despite high levels of immunization with whole-cell vaccine. *Emerg Infect Dis,* 3(2), pp. 175-178.

De Serres, G., Boulianne, N., Douville Fradet, M. and Duval, B., 1995. Pertussis in Quebec: ongoing epidemic since the late 1980s. *Can Commun Dis Rep,* 21(5), pp. 45-48.

Deamer, D., Akeson, M. and Branton, D., 2016. Three decades of nanopore sequencing. *Nature Biotechnology,* 34, pp. 518-524.

Deora, R., Bootsma, H.J., Miller, J.F. and Cotter, P.A., 2001. Diversity in the Bordetella virulence regulon: transcriptional control of a Bvg-intermediate phase gene. *Mol Microbiol,* 40(3), pp. 669-683.

Derrington, I.M., Butler, T.Z., Collins, M.D., Manrao, E., Pavlenok, M., Niederweis, M. and Gundlach, J.H., 2010. Nanopore DNA sequencing with MspA. *Proc Natl Acad Sci U S A,* 107(37), pp. 16060-16065.

Desai, A., Marwah, V.S., Yadav, A., Jha, V., Dhaygude, K., Bangar, U., Kulkarni, V. and Jere, A., 2013. Identification of optimum sequencing depth especially for de novo genome assembly of small genomes using next generation sequencing data. *PLoS One,* 8(4), p. e60204.

Diavatopoulos, D.A., Cummings, C.A., Schouls, L.M., Brinig, M.M., Relman, D.A. and Mooi, F.R., 2005. Bordetella pertussis, the causative agent of whooping cough, evolved from a distinct, human-associated lineage of B. bronchiseptica. *PLoS Pathog,* 1(4), p. e45.

Diavatopoulos, D.A., Mills, K.H.G., Kester, K.E., Kampmann, B., Silerova, M., Heininger, U., van Dongen, J.J.M., van der Most, R.G., Huijnen, M.A., Siena, E., Mielcarek, N., Ochs, M.M., Denoël, P., Berbers, G., Buisman, A.M., de Jonge, M.I., Fenwick, C., Gorringe, A., He, Q., Kelly, D., Le Grand, R., Locht, C., Mascart, F., Mertsola, J., Orfao, A., Pantaleo, G., Pollard, A.J., Preston, A., Read, R., Sebo, P., van Els, C., Vecerek, B., Londoño-Hayes, P. and de Groot, R., 2018. PERISCOPE: road towards effective control of pertussis. *Lancet Infect Dis*.

Diaz, C., Schilardi, P.L., Salvarezza, R.C. and Fernandez Lorenzo de Mele, M., 2011. Have flagella a preferred orientation during early stages of biofilm formation?: AFM study using patterned substrates. *Colloids Surf B Biointerfaces,* 82(2), pp. 536-542.

Domingues, C.M.A.S., Teixeira, A.M.d.S. and Carvalho, S.M.D., 2012. National immunization program: vaccination, compliance and pharmacovigilance. *Revista do Instituto de Medicina Tropical de São Paulo,* 54, pp. 22-27.

Dupre, E., Herrou, J., Lensink, M.F., Wintjens, R., Vagin, A., Lebedev, A., Crosson, S., Villeret, V., Locht, C., Antoine, R. and Jacob-Dubuisson, F., 2015. Virulence regulation with Venus flytrap domains: structure and function of the periplasmic moiety of the sensor-kinase BvgS. *PLoS Pathog,* 11(3), p. e1004700.

Edge, P. and Bansal, V., 2019. Longshot enables accurate variant calling in diploid genomes from single-molecule long read sequencing. *Nature Communications,* 10(1), p. 4660.

Edwards, A., Debbonaire, A.R., Sattler, B., Mur, L.A. and Hodson, A.J., 2016. Extreme metagenomics using nanopore DNA sequencing: a field report from Svalbard, 78 N. *bioRxiv*.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., deWinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J. and Turner, S., 2009. Real-Time DNA Sequencing from Single Polymerase Molecules. *Science,* 323(5910), pp. 133-138.

Eisenstein, M., 2019. Illumina swallows PacBio in long shot for market domination. *Nature Biotechnology,* 37, p. 3.

Eldering, G. and Kendrick, P., 1938. Bacillus para-pertussis: a species resembling both Bacillus pertussis and Bacillus bronchisepticus but identical with neither. *J Bacteriol,* 35(6), pp. 561-572.

Emsley, P., Charles, I.G., Fairweather, N.F. and Isaacs, N.W., 1996. Structure of Bordetella pertussis virulence factor P.69 pertactin. *Nature,* 381(6577), pp. 90-92.

ESR, 2009. *ESR Pertussis Report January-December 2009* [Online]. Available from: https://surv.esr.cri.nz/PDF_surveillance/PertussisRpt/2009/200951PertussisRpt.pdf [Accessed 14 June 2020].

ESR, 2012. *ESR Pertussis Report 2012.* Available from: https://surv.esr.cri.nz/PDF_surveillance/PertussisRpt/2012/201248PertussisRpt.pdf [Accessed 29 May 2019].

ESR, 2018. *ESR Pertussis Report 7 April-4 May 2018* [Online]. Available from: https://surv.esr.cri.nz/PDF_surveillance/PertussisRpt/2018/PertussisReport4May2018.pdf [Accessed 29 May 2019].

ESR, 2019. *ESR Pertussis Report May 2019* [Online]. Available from: https://surv.esr.cri.nz/PDF_surveillance/PertussisRpt/2019/PertussisReportMay2019.pdf [Accessed 27 May 2020].

ESR Invasive Pathogen Laboratory, 2019. Available from: https://www.esr.cri.nz/our-people/our-science-team/our-health-laboratories/ [Accessed 29 May 2019].

Etskovitz, H., Anastasio, N., Green, E. and May, M., 2019. Role of Evolutionary Selection Acting on Vaccine Antigens in the Re-Emergence of Bordetella Pertussis. *Diseases,* 7(2).

Everest, P., Li, J., Douce, G., Charles, I., De Azavedo, J., Chatfield, S., Dougan, G. and Roberts, M., 1996. Role of the Bordetella pertussis P.69/pertactin protein and the P.69/pertactin RGD motif in the adherence to and invasion of mammalian cells. *Microbiology,* 142 (Pt 11), pp. 3261-3268.

Faller, M., Niederweis, M. and Schulz, G.E., 2004. The structure of a mycobacterial outer-membrane channel. *Science,* 303(5661), pp. 1189-1192.

Ferry, N.S., 1912. Bacillus bronchisepticus (bronchicanis): the cause of distemper in dogs and a similar disease in other animals. *The Veterinary Journal,* 68(7), pp. 376-380.

Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Iserentant, D., Merregaert, J., Min Jou, W., Molemans, F., Raeymaekers, A., Van den Berghe, A., Volckaert, G. and Ysebaert, M., 1976. Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature,* 260(5551), pp. 500-507.

Fine, P.E. and Clarkson, J.A., 1982. The recurrence of whooping cough: possible implications for assessment of vaccine efficacy. *Lancet,* 1(8273), pp. 666-669.

Fisher, R.A., 1921. On the "Probable Error" of a Coefficient of Correlation Deduced from a Small Sample. *Metron,* 1, pp. 3-32.

Fisk, S.K. and Soave, O.A., 1973. Bordetella bronchiseptica in laboratory cats from central California. *Lab Anim Sci,* 23(1), pp. 33-35.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., Merrick, J.M. and et al., 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science,* 269(5223), pp. 496-512.

Fleischmann, R.D., Alland, D., Eisen, J.A., Carpenter, L., White, O., Peterson, J., DeBoy, R., Dodson, R., Gwinn, M., Haft, D., Hickey, E., Kolonay, J.F., Nelson, W.C., Umayam, L.A., Ermolaeva, M., Salzberg, S.L., Delcher, A., Utterback, T., Weidman, J., Khouri, H., Gill, J., Mikula, A., Bishai, W., Jacobs, W.R., Jr., Venter, J.C. and Fraser, C.M., 2002. Whole-genome comparison of Mycobacterium tuberculosis clinical and laboratory strains. *J Bacteriol,* 184(19), pp. 5479-5490.

Fong, J.N.C. and Yildiz, F.H., 2015. Biofilm Matrix Proteins. *Microbiol Spectr,* 3(2).

Fox, E.J., Reid-Bayliss, K.S., Emond, M.J. and Loeb, L.A., 2014. Accuracy of Next Generation Sequencing Platforms. *Next Gener Seq Appl,* 1.

Gallagher, G.L., 1965. Isolation of Bordetella bronchiseptica from horses. *Vet Rec,* 77, pp. 632-633.

Genome Web, 2019. PacBio shares early-access customer experiences, new applications for Sequel II. *Genome Web* [Online]. Available from: https://www.genomeweb.com/sequencing/pacbio-shares-early-access-customer-experiences-new-applications-sequel-ii#.XYfSMy5KjIU [Accessed 22 September 2019].

Geuijen, C.A., Willems, R.J., Hoogerhout, P., Puijk, W.C., Meloen, R.H. and Mooi, F.R., 1998. Identification and characterization of heparin binding regions of the Fim2 subunit of Bordetella pertussis. *Infect Immun,* 66(5), pp. 2256-2263.

Gilpatrick, T., Lee, I., Graham, J.E., Raimondeau, E., Bowen, R., Heron, A., Downs, B., Sukumar, S., Sedlazeck, F.J. and Timp, W., 2020. Targeted nanopore sequencing with Cas9-guided adapter ligation. *Nature Biotechnology,* 38(4), pp. 433-438.

Goodnow, R.A., 1980. Biology of Bordetella bronchiseptica. *Microbiol Rev,* 44(4), pp. 722-738.

Grant, C.C., 2004. *The epidemiology of pertussis in New Zealand and risk factors for pertussis in New Zealand infants.* (PhD), University of Auckland. Available from: https://researchspace.auckland.ac.nz/handle/2292/3130 [Accessed 14 June 2020].

Green, E.D., Watson, J.D. and Collins, F.S., 2015. Human Genome Project: Twenty-five years of big biology. *Nature News,* 526(7571), p. 29.

Greenleaf, W.J. and Sidow, A., 2014. The future of sequencing: convergence of intelligent design and market Darwinism. *Genome Biol,* 15(3), p. 303.

Greig, D.R., Jenkins, C., Gharbia, S. and Dallman, T.J., 2019. Comparison of single-nucleotide variants identified by Illumina and Oxford Nanopore technologies in the context of a potential outbreak of Shiga toxin-producing Escherichia coli. *Gigascience,* 8(8).

Gross, R., Arico, B. and Rappuoli, R., 1989. Families of bacterial signal-transducing proteins. *Mol Microbiol,* 3(11), pp. 1661-1667.

Guimarães, L.M., Carneiro, E.C. and Carvalho-Costa, F.A., 2015. Increasing incidence of pertussis in Brazil: a retrospective study using surveillance data. *BMC Infect Dis,* 15(442).

Guiso, N., 2009. Bordetella pertussis and Pertussis Vaccines. *Clinical Infectious Diseases,* 49(10), pp. 1565-1569.

Gummy-Bear, 2014. *Calculate length of all sequences in an multi-fasta file* [Online]. Available from: https://bioexpressblog.wordpress.com/2014/04/15/calculate-length-of-all-sequences-in-an-multi-fasta-file/ [Accessed 06 June 2018].

Gurevich, A., Saveliev, V., Vyahhi, N. and Tesler, G., 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics,* 29(8), pp. 1072-1075.

Gutacker, M.M., Smoot, J.C., Migliaccio, C.A., Ricklefs, S.M., Hua, S., Cousins, D.V., Graviss, E.A., Shashkina, E., Kreiswirth, B.N. and Musser, J.M., 2002. Genome-wide analysis of synonymous single nucleotide polymorphisms in Mycobacterium tuberculosis complex organisms: resolution of genetic relationships among closely related microbial strains. *Genetics,* 162(4), pp. 1533-1543.

Gzyl, A., Augustynowicz, E., Rabczenko, D., Gniadek, G. and Slusarczyk, J., 2004. Pertussis in Poland. *Int J Epidemiol,* 33(2), pp. 358-365.

Hamidou Soumana, I., Linz, B. and Harvill, E.T., 2017. Environmental Origin of the Genus Bordetella. *Front Microbiol,* 8.

Hardy, A., 1993. Whooping Cough. In: K.F. Kiple, ed. *The Cambridge world history of human disease.* Cambridge: Cambridge University Press, pp. 1094-1096.

Hawkey, J., Hamidian, M., Wick, R.R., Edwards, D.J., Billman-Jacobe, H., Hall, R.M. and Holt, K.E., 2015. ISMapper: identifying transposase insertion sites in bacterial genomes from short read sequence data. *BMC Genomics,* 16, p. 667.

Heather, J.M. and Chain, B., 2016. The sequence of sequencers: The history of sequencing DNA. *Genomics,* 107(1), pp. 1-8.

Hegerle, N., Paris, A.S., Brun, D., Dore, G., Njamkepo, E., Guillot, S. and Guiso, N., 2012. Evolution of French Bordetella pertussis and Bordetella parapertussis isolates: increase of Bordetellae not expressing pertactin. *Clin Microbiol Infect,* 18(9), pp. E340-346.

Heikkinen, E., Kallonen, T., Saarinen, L., Sara, R., King, A.J., Mooi, F.R., Soini, J.T., Mertsola, J. and He, Q., 2007. Comparative genomics of Bordetella pertussis reveals progressive gene loss in Finnish strains. *PLoS One,* 2(9), p. e904.

Heininger, U., Stehr, K., Schmitt-Grohe, S., Lorenz, C., Rost, R., Christenson, P.D., Uberall, M. and Cherry, J.D., 1994. Clinical characteristics of illness caused by Bordetella parapertussis compared with illness caused by Bordetella pertussis. *Pediatr Infect Dis J,* 13(4), pp. 306-309.

Henderson, I.R. and Nataro, J.P., 2001. Virulence functions of autotransporter proteins. *Infect Immun,* 69(3), pp. 1231-1243.

Henderson, M.W., Inatsuka, C.S., Sheets, A.J., Williams, C.L., Benaron, D.J., Donato, G.M., Gray, M.C., Hewlett, E.L. and Cotter, P.A., 2012. Contribution of Bordetella filamentous hemagglutinin and adenylate cyclase toxin to suppression and evasion of interleukin-17-mediated inflammation. *Infect Immun,* 80(6), pp. 2061-2075.

Heron, A.J., 2018. Molecular engineering DNA and RNA for nanopore sequencing. In: D. Branton and D. Deamer, eds. *Nanopore Sequencing: An Introduction.* New Jersey: World Scientific Publishing Co. Pte. Ltd., pp. 107-146.

Herper, M., 2017. Illumina promises to sequence human genome for $100 -- but not quite yet. Available from: https://www.forbes.com/sites/matthewherper/2017/01/09/illumina-promises-to-sequence-human-genome-for-100-but-not-quite-yet/#5102a788386d [Accessed 20 September 2019].

Hewlett, E.L., 1997. Pertussis: current concepts of pathogenesis and prevention. *Pediatr Infect Dis J,* 16(4 Suppl), pp. S78-84.

Hiramatsu, Y., Miyaji, Y., Otsuka, N., Arakawa, Y., Shibayama, K. and Kamachi, K., 2017. Significant Decrease in Pertactin-Deficient Bordetella pertussis Isolates, Japan. *Emerg Infect Dis,* 23(4), pp. 699-701.

Hoang, D.T., Chernomor, O., von Haeseler, A., Minh, B.Q. and Vinh, L.S., 2018. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol Biol Evol,* 35(2), pp. 518-522.

Hoffman, C.L., 2017. *The "Mystery Phases": Bordetella adenylate cyclase toxin and biofilm in the Bvg-intermediate phase & Bordetella pertussis motility and flagella in the Bvg-minus phase.* PhD (PhD), University of Virginia.

Hoffman, C.L., Gonyar, L.A., Zacca, F., Sisti, F., Fernandez, J., Wong, T., Damron, F.H. and Hewlett, E.L., 2019. Bordetella pertussis Can Be Motile and Express Flagellum-Like Structures. *MBio,* 10(3).

Hornblower, B., Coombs, A., Whitaker, R.D., Kolomeisky, A., Picone, S.J., Meller, A. and Akeson, M., 2007. Single-molecule analysis of DNA-protein complexes using nanopores. *Nat Methods,* 4(4), pp. 315-317.

Hoshino, S., Kikkawa, S., Takahashi, K., Itoh, H., Kaziro, Y., Kawasaki, H., Suzuki, K., Katada, T. and Ui, M., 1990. Identification of sites for alkylation by N-ethylmaleimide and pertussis toxin-catalyzed ADP-ribosylation on GTP-binding proteins. *FEBS Lett,* 276(1-2), pp. 227-231.

Howson, C.P., Howe, C.J. and Fineberg, H.V., 1991. Pertussis and Rubella Vaccines: A Brief Chronology.

Huang, Y.F., Chen, S.C., Chiang, Y.S., Chen, T.H. and Chiu, K.P., 2012. Palindromic sequence impedes sequencing-by-ligation mechanism. *BMC Syst Biol,* 6 Suppl 2, p. S10.

Illumina, 2017. *Illumina images and multimedia* [Online]. Available from: https://www.illumina.com/systems/sequencing-platforms.html [Accessed 01 October 2017].

Illumina, 2019a. *Illumina sequencing platforms* [Online]. Available from: https://www.illumina.com/systems/sequencing-platforms.html [Accessed 20 September 2019].

Illumina, 2019b. *iSeq specifications* [Online]. Available from: https://emea.illumina.com/systems/sequencing-platforms/iseq/specifications.html [Accessed 20 September 2019].

Inatsuka, C.S., Julio, S.M. and Cotter, P.A., 2005. Bordetella filamentous hemagglutinin plays a critical role in immunomodulation, suggesting a mechanism for host specificity. *Proc Natl Acad Sci U S A,* 102(51), pp. 18578-18583.

International Human Genome Sequencing Consortium, 2004. Finishing the euchromatic sequence of the human genome. *Nature,* 431(7011), pp. 931-945.

Irie, Y., Mattoo, S. and Yuk, M.H., 2004. The Bvg virulence control system regulates biofilm formation in Bordetella bronchiseptica. *J Bacteriol,* 186(17), pp. 5692-5698.

Ivanov, Y.V., Linz, B., Register, K.B., Newman, J.D., Taylor, D.L., Boschert, K.R., Le Guyon, S., Wilson, E.F., Brinkac, L.M., Sanka, R., Greco, S.C., Klender, P.M., Losada, L. and Harvill, E.T., 2016.

Identification and taxonomic characterization of Bordetella pseudohinzii sp. nov. isolated from laboratory-raised mice. *Int J Syst Evol Microbiol,* 66(12), pp. 5452-5459.

Jackman, S.D., Vandervalk, B.P., Mohamadi, H., Chu, J., Yeo, S., Hammond, S.A., Jahesh, G., Khan, H., Coombe, L., Warren, R.L. and Birol, I., 2017. ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter. *Genome Res,* 27(5), pp. 768-777.

Jain, M., Fiddes, I.T., Miga, K.H., Olsen, H.E., Paten, B. and Akeson, M., 2015. Improved data analysis for the MinION nanopore sequencer. *Nat Methods,* 12(4), pp. 351-356.

Jain, M., Koren, S., Miga, K.H., Quick, J., Rand, A.C., Sasani, T.A., Tyson, J.R., Beggs, A.D., Dilthey, A.T., Fiddes, I.T., Malla, S., Marriott, H., Nieto, T., O'Grady, J., Olsen, H.E., Pedersen, B.S., Rhie, A., Richardson, H., Quinlan, A.R., Snutch, T.P., Tee, L., Paten, B., Phillippy, A.M., Simpson, J.T., Loman, N.J. and Loose, M., 2018. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol,* 36(4), pp. 338-345.

Jain, M., Olsen, H.E., Turner, D.J., Stoddart, D., Bulazel, K.V., Paten, B., Haussler, D., Willard, H.F., Akeson, M. and Miga, K.H., 2018. Linear assembly of a human centromere on the Y chromosome. *Nat Biotechnol,* 36(4), pp. 321-323.

Jakinovich, A. and Sood, S.K., 2014. Pertussis: still a cause of death, seven decades into vaccination. *Curr Opin Pediatr,* 26(5), pp. 597-604.

Jia, H., Guo, Y., Zhao, W. and Wang, K., 2014. Long-range PCR in next-generation sequencing: comparison of six enzymes and evaluation on the MiSeq sequencer. *Sci Rep,* 4, p. 5737.

Jolley, K.A., Bray, J.E. and Maiden, M.C.J., 2018. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Res,* 3, p. 124.

Julio, S.M., Inatsuka, C.S., Mazar, J., Dieterich, C., Relman, D.A. and Cotter, P.A., 2009. Natural-host animal models indicate functional interchangeability between the filamentous haemagglutinins of Bordetella pertussis and Bordetella bronchiseptica and reveal a role for the mature C-terminal domain, but not the RGD motif, during infection. *Mol Microbiol,* 71(6), pp. 1574-1590.

Kallonen, T. and He, Q., 2009. Bordetella pertussis strain variation and evolution postvaccination. *Expert Rev Vaccines,* 8(7), pp. 863-875.

Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A. and Jermiin, L.S., 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods,* 14(6), pp. 587-589.

Kanagawa, T., 2003. Bias and artifacts in multitemplate polymerase chain reactions (PCR). *J Biosci Bioeng,* 96(4), pp. 317-323.

Karger, B.L. and Guttman, A., 2009. DNA Sequencing by Capillary Electrophoresis. *Electrophoresis,* 30(Suppl 1), pp. S196-202.

Kasianowicz, J.J., Brandin, E., Branton, D. and Deamer, D.W., 1996. Characterization of individual polynucleotide molecules using a membrane channel. *Proc Natl Acad Sci U S A,* 93(24), pp. 13770-13773.

Kaslow, H.R. and Burns, D.L., 1992. Pertussis toxin and target eukaryotic cells: binding, entry, and activation. *Faseb j,* 6(9), pp. 2684-2690.

Katada, T., Tamura, M. and Ui, M., 1983. The A protomer of islet-activating protein, pertussis toxin, as an active peptide catalyzing ADP-ribosylation of a membrane protein. *Arch Biochem Biophys,* 224(1), pp. 290-298.

Kent, A. and Heath, P.T., 2014. Pertussis. *Medicine,* 42(1), pp. 8-10.

Kersters, K., Hinz, K.-H., Hertle, A., Segers, P., Lievens, A., Siegmann, O. and de Ley, J., 1984. Bordetella avium sp. nov., Isolated from the Respiratory Tracts of Turkeys and Other Birds. *International Journal of Systematic Bacteriology,* 34(1), pp. 56-70.

King, A.J., van der Lee, S., Mohangoo, A., van Gent, M., van der Ark, A. and van de Waterbeemd, B., 2013. 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(6), p. e66150.

King, A.J., van Gorkom, T., van der Heide, H.G., Advani, A. and 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, p. 64.

Klausen, M., Heydorn, A., Ragas, P., Lambertsen, L., Aaes-Jorgensen, A., Molin, S. and Tolker-Nielsen, T., 2003. Biofilm formation by Pseudomonas aeruginosa wild type, flagella and type IV pili mutants. *Mol Microbiol,* 48(6), pp. 1511-1524.

Klein, N.P., 2014. Licensed pertussis vaccines in the United States: History and current state. *Hum Vaccin Immunother,* 10(9), pp. 2684-2690.

Klein, N.P., Bartlett, J., Rowhani-Rahbar, A., Fireman, B. and Baxter, R., 2012. Waning protection after fifth dose of acellular pertussis vaccine in children. *N Engl J Med,* 367(11), pp. 1012-1019.

Knierim, E., Lucke, B., Schwarz, J.M., Schuelke, M. and Seelow, D., 2011. Systematic comparison of three methods for fragmentation of long-range PCR products for next generation sequencing. *PLoS One,* 6(11), p. e28240.

Ko, K.S., Peck, K.R., Oh, W.S., Lee, N.Y., Lee, J.H. and Song, J.H., 2005. New Species of Bordetella, Bordetella ansorpii sp. nov., Isolated from the Purulent Exudate of an Epidermal Cyst. *J Clin Microbiol,* 43(5), pp. 2516-2519.

Koepke, R., Eickhoff, J.C., Ayele, R.A., Petit, A.B., Schauer, S.L., Hopfensperger, D.J., Conway, J.H. and Davis, J.P., 2014. Estimating the effectiveness of tetanus-diphtheria-acellular pertussis vaccine (Tdap) for preventing pertussis: evidence of rapidly waning immunity and difference in effectiveness by Tdap brand. *J Infect Dis,* 210(6), pp. 942-953.

Kolmogorov, M., Yuan, J., Lin, Y. and Pevzner, P.A., 2019. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol,* 37(5), pp. 540-546.

Koren, S. and Phillippy, A.M., 2015. One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Curr Opin Microbiol,* 23, pp. 110-120.

Koren, S., Schatz, M.C., Walenz, B.P., Martin, J., Howard, J., Ganapathy, G., Wang, Z., Rasko, D.A., McCombie, W.R., Jarvis, E.D. and Phillippy, A.M., 2012. Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat Biotechnol,* 30(7), pp. 693-700.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R. and Phillippy, A.M., 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res,* 27(5), pp. 722-736.

Kruskal, W.H. and Wallis, W.A., 1952. Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association,* 47(260), pp. 583-621.

Kurniawan, J., Maharjan, R.P., Chan, W.F., Reeves, P.R., Sintchenko, V., Gilbert, G.L., Mooi, F.R. and Lan, R., 2010. Bordetella pertussis clones identified by multilocus variable-number tandem-repeat analysis. *Emerg Infect Dis,* 16(2), pp. 297-300.

Lacey, B.W., 1960. Antigenic modulation of Bordetella pertussis. *J Hyg (Lond),* 58(1), pp. 57-93.

Lam, C., Octavia, S., Ricafort, L., Sintchenko, V., Gilbert, G.L., Wood, N., McIntyre, P., Marshall, H., Guiso, N., Keil, A.D., Lawrence, A., Robson, J., Hogg, G. and Lan, R., 2014. Rapid increase in Pertactin-deficient Bordetella pertussis isolates, Australia. *Emerg Infect Dis,* 20(4), pp. 626-633.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, Y., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., et al., 2001. Initial sequencing and analysis of the human genome. *Nature,* 409(6822), pp. 860-921.

Lauria, A.M. and Zabbo, C.P., 2020. *Pertussis (Whooping Cough)* [Online]. ed. Treasure Island (FL): StatsPearls Publishing. Available from: https://www.ncbi.nlm.nih.gov/books/NBK519008/ [Accessed 20 July 2020].

Letunic, I. and Bork, P., 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res,* 47(W1), pp. W256-w259.

Li, H., 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv,* 1303(3997v2).

Li, H., 2016. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics,* 32(14), pp. 2103-2110.

Li, H., 2017. Minimap2: fast pairwise alignment for long DNA sequences. *ArXiv,* 1708.01492.

Li, H., 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics,* 34(18), pp. 3094-3100.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R., 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics,* 25(16), pp. 2078-2079.

Lieberman, K.R., Cherf, G.M., Doody, M.J., Olasagasti, F., Kolodji, Y. and Akeson, M., 2010. Processive replication of single DNA molecules in a nanopore catalyzed by phi29 DNA polymerase. *J Am Chem Soc,* 132(50), pp. 17961-17972.

Lin, Y., Yuan, J., Kolmogorov, M., Shen, M.W., Chaisson, M. and Pevzner, P.A., 2016. Assembly of Long Error-Prone Reads Using de Bruijn Graphs. *PNAS,* 113(52), pp. E8396-8405.

Linz, B., Ivanov, Y.V., Preston, A., Brinkac, L., Parkhill, J., Kim, M., Harris, S.R., Goodfield, L.L., Fry, N.K., Gorringe, A.R., Nicholson, T.L., Register, K.B., Losada, L. and Harvill, E.T., 2016. Acquisition and loss of virulence-associated factors during genome evolution and speciation in three clades of Bordetella species. *BMC Genomics,* 17(1), p. 767.

Lipscombe, M., Charles, I.G., Roberts, M., Dougan, G., Tite, J. and Fairweather, N.F., 1991. Intranasal immunization using the B subunit of the Escherichia coli heat-labile toxin fused to an epitope of the Bordetella pertussis P.69 antigen. *Mol Microbiol,* 5(6), pp. 1385-1392.

Liu, R. and Ochman, H., 2007. Stepwise formation of the bacterial flagellar system. *Proc Natl Acad Sci U S A,* 104(17), pp. 7116-7121.

Locht, C. and Keith, J.M., 1986. Pertussis toxin gene: nucleotide sequence and genetic organization. *Science,* 232(4755), pp. 1258-1264.

Loman, N.J., 2013. Diagnosing problems with phasing and pre-phasing on Illumina platforms. Available from: *http://lab.loman.net/high-throughput%20sequencing/2013/11/21/diagnosing-problems-with-phasing-and-pre-phasing-on-illumina-platforms/* [Accessed 20 September 2019].

Loman, N.J., Quick, J. and Simpson, J.T., 2015. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nature Methods,* 12, pp. 733-735.

Long Read Club, 2019. *Long Read Club: Really very long reads indeed* [Online]. Available from: *https://www.longreadclub.org/* [Accessed 22 July 2019].

Lu, H., Giordano, F. and Ning, Z., 2016. Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics Proteomics Bioinformatics,* 14(5), pp. 265-279.

Luo, R., Wong, C.-L., Wong, Y.-S., Tang, C.-I., Liu, C.-M., Leung, C.-M. and Lam, T.-W., 2020. Exploring the limit of using a deep neural network on pileup data for germline variant calling. *Nature Machine Intelligence,* 2(4), pp. 220-227.

MacArthur, I., Belcher, T., King, J.D., Ramasamy, V., Alhammadi, M. and Preston, A., 2019. The evolution of Bordetella pertussis has selected for mutations of acr that lead to sensitivity to hydrophobic molecules and fatty acids. *Emerg Microbes Infect,* 8(1), pp. 603-612.

Madsen, T., 1933. Vaccination against whooping cough. *Journal of the American Medical Association,* 101, pp. 187-188.

Mann, H.B. and Whitney, D.R., 1947. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics,* 18(1), pp. 50-60.

Manrao, E.A., Derrington, I.M., Laszlo, A.H., Langford, K.W., Hopper, M.K., Gillgren, N., Pavlenok, M., Niederweis, M. and Gundlach, J.H., 2012. Reading DNA at single-nucleotide resolution with a mutant MspA nanopore and phi29 DNA polymerase. *Nat Biotechnol,* 30(4), pp. 349-353.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.-J., Chen, Z., Dewell, S.B., Du, L., Fierro, J.M., Gomes, X.V., Godwin, B.C., He, W., Helgesen, S., Ho, C.H., Irzyk, G.P., Jando, S.C., Alenquer, M.L.I., Jarvie, T.P., Jirage, K.B., Kim, J.-B., Knight, J.R., Lanza, J.R., Leamon, J.H., Lefkowitz, S.M., Lei, M., Li, J., Lohman, K.L., Lu, H., Makhijani, V.B., McDade, K.E., McKenna, M.P., Myers, E.W., Nickerson, E., Nobile, J.R., Plant, R., Puc, B.P., Ronan, M.T., Roth, G.T., Sarkis, G.J., Simons, J.F., Simpson, J.W., Srinivasan, M., Tartaro, K.R., Tomasz, A., Vogt, K.A., Volkmer, G.A., Wang, S.H., Wang, Y., Weiner, M.P., Yu, P., Begley, R.F. and Rothberg, J.M., 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature,* 437(7057), pp. 376-380.

Martin, S.W., Pawloski, L., Williams, M., Weening, K., DeBolt, C., Qin, X., Reynolds, L., Kenyon, C., Giambrone, G., Kudish, K., Miller, L., Selvage, D., Lee, A., Skoff, T.H., Kamiya, H., Cassiday, P.K., Tondella, M.L. and Clark, T.A., 2015. Pertactin-negative Bordetella pertussis strains: evidence for a possible selective advantage. *Clin Infect Dis,* 60(2), pp. 223-227.

Martinez de Tejada, G., Cotter, P.A., Heininger, U., Camilli, A., Akerley, B.J., Mekalanos, J.J. and Miller, J.F., 1998. Neither the Bvg- phase nor the vrg6 locus of Bordetella pertussis is required for respiratory infection in mice. *Infect Immun,* 66(6), pp. 2762-2768.

Mastrantonio, P., Spigaglia, P., van Oirschot, H., van der Heide, H.G., Heuvelman, K., Stefanelli, P. and Mooi, F.R., 1999. Antigenic variants in Bordetella pertussis strains isolated from vaccinated and unvaccinated children. *Microbiology,* 145 (Pt 8), pp. 2069-2075.

Mattoo, S. and Cherry, J.D., 2005. Molecular Pathogenesis, Epidemiology, and Clinical Manifestations of Respiratory Infections Due to Bordetella pertussis and Other Bordetella Subspecies. *Clin Microbiol Rev,* 18(2), pp. 326-382.

McGowan, G.P., 1911. Some observations on a laboratory epidemic principally among dogs and cats, in which the animals affected presented the symptoms of the disease called "distemper". *J. Pathol. Bacteriol.,* 15, pp. 372-426.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. and DePristo, M.A., 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res,* 20(9), pp. 1297-1303.

McKenzie, R.A., Wood, A.D. and Blackall, P.J., 1979. Pneumonia associated with Bordetella bronchiseptica in captive koalas. *Aust Vet J,* 55(9), pp. 427-430.

McKernan, K.J., Peckham, H.E., Costa, G.L., McLaughlin, S.F., Fu, Y., Tsung, E.F., Clouser, C.R., Duncan, C., Ichikawa, J.K., Lee, C.C., Zhang, Z., Ranade, S.S., Dimalanta, E.T., Hyland, F.C., Sokolsky, T.D., Zhang, L., Sheridan, A., Fu, H., Hendrickson, C.L., Li, B., Kotler, L., Stuart, J.R., Malek, J.A., Manning, J.M., Antipova, A.A., Perez, D.S., Moore, M.P., Hayashibara, K.C., Lyons, M.R., Beaudoin, R.E., Coleman, B.E., Laptewicz, M.W., Sannicandro, A.E., Rhodes, M.D., Gottimukkala, R.K., Yang, S., Bafna, V., Bashir, A., MacBride, A., Alkan, C., Kidd, J.M., Eichler, E.E., Reese, M.G., De La Vega, F.M. and Blanchard, A.P., 2009. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res,* 19(9), pp. 1527-1541.

Meade, B.D., Kind, P.D., Ewell, J.B., McGrath, P.P. and Manclark, C.R., 1984. In vitro inhibition of murine macrophage migration by Bordetella pertussis lymphocytosis-promoting factor. *Infect Immun,* 45(3), pp. 718-725.

Meller, A., Nivon, L., Brandin, E., Golovchenko, J. and Branton, D., 2000. Rapid nanopore discrimination between single polynucleotide molecules. *Proc Natl Acad Sci U S A,* 97(3), pp. 1079-1084.

Meller, A., Nivon, L. and Branton, D., 2001. Voltage-driven DNA translocations through a nanopore. *Phys Rev Lett,* 86(15), pp. 3435-3438.

Melton, A.R. and Weiss, A.A., 1993. Characterization of environmental regulators of Bordetella pertussis. *Infect Immun,* 61(3), pp. 807-815.

Melvin, J.A., Scheller, E.V., Miller, J.F. and Cotter, P.A., 2014. Bordetella pertussis pathogenesis: current and future challenges. *Nature Reviews Microbiology,* 12, pp. 274-288.

Melvin, J.A., Scheller, E.V., Noel, C.R. and Cotter, P.A., 2015. New Insight into Filamentous Hemagglutinin Secretion Reveals a Role for Full-Length FhaB in Bordetella Virulence. *MBio,* 6(4).

Menestrina, G., 1986. Ionic channels formed by Staphylococcus aureus alpha-toxin: voltage-dependent inhibition by divalent and trivalent cations. *J Membr Biol,* 90(2), pp. 177-190.

Merkel, T.J., Barros, C. and Stibitz, S., 1998. Characterization of the bvgR locus of Bordetella pertussis. *J Bacteriol,* 180(7), pp. 1682-1690.

Merkel, T.J., Boucher, P.E., Stibitz, S. and Grippe, V.K., 2003. Analysis of bvgR expression in Bordetella pertussis. *J Bacteriol,* 185(23), pp. 6902-6912.

Merkel, T.J. and Stibitz, S., 1995. Identification of a locus required for the regulation of bvg-repressed genes in Bordetella pertussis. *J Bacteriol,* 177(10), pp. 2727-2736.

Merkel, T.J., Stibitz, S., Keith, J.M., Leef, M. and Shahin, R., 1998. Contribution of regulation by the bvg locus to respiratory infection of mice by Bordetella pertussis. *Infect Immun,* 66(9), pp. 4367-4373.

MicrobesNG, 2019. Available from: https://microbesng.com [Accessed 17 September 2019].

Milne, I., Stephen, G., Bayer, M., Cock, P.J.A., Pritchard, L., Cardle, L., Shaw, P.D. and Marshall, D., 2013. Using Tablet for visual exploration of second-generation sequencing data. *Briefings in Bioinformatics,* 14(2), pp. 193-202.

Min Jou, W., Haegeman, G., Ysebaert, M. and Fiers, W., 1972. Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein. *Nature,* 237(5350), pp. 82-88.

Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A. and Lanfear, R., 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol,* 37(5), pp. 1530-1534.

Misegades, L.K., Winter, K., Harriman, K., Talarico, J., Messonnier, N.E., Clark, T.A. and Martin, S.W., 2012. Association of childhood pertussis with receipt of 5 doses of pertussis vaccine by time since last vaccine dose, California, 2010. *Jama,* 308(20), pp. 2126-2132.

Miyaji, Y., Otsuka, N., Toyoizumi-Ajisaka, H., Shibayama, K. and Kamachi, K., 2013. Genetic Analysis of Bordetella pertussis Isolates from the 2008–2010 Pertussis Epidemic in Japan. *PLoS One,* 8(10).

Moissenet, D., Valcin, M., Marchand, V., Grimprel, E., Begue, P., Garbarg-Chenon, A. and Vu-Thien, H., 1996. Comparative DNA analysis of Bordetella pertussis clinical isolates by pulsed-field gel electrophoresis, randomly amplified polymorphism DNA, and ERIC polymerase chain reaction. *FEMS Microbiol Lett,* 143(2-3), pp. 127-132.

Mooi, F.R., Hallander, H., Wirsing von Konig, C.H., Hoet, B. and Guiso, N., 2000. Epidemiological typing of Bordetella pertussis isolates: recommendations for a standard methodology. *Eur J Clin Microbiol Infect Dis,* 19(3), pp. 174-181.

Mooi, F.R., van Loo, I.H., van Gent, M., He, Q., Bart, M.J., Heuvelman, K.J., de Greeff, S.C., Diavatopoulos, D., Teunis, P., Nagelkerke, N. and Mertsola, J., 2009. Bordetella pertussis strains with increased toxin production associated with pertussis resurgence. *Emerg Infect Dis,* 15(8), pp. 1206-1213.

Mooi, F.R., van Oirschot, H., Heuvelman, K., van der Heide, H.G., Gaastra, W. and Willems, R.J., 1998. Polymorphism in the Bordetella pertussis virulence factors P.69/pertactin and pertussis toxin in The Netherlands: temporal trends and evidence for vaccine-driven evolution. *Infect Immun,* 66(2), pp. 670-675.

Mooi, F.R., Zeddeman, A. and van Gent, M., 2015. The pertussis problem: classical epidemiology and strain characterization should go hand in hand. *J Pediatr (Rio J),* 91(4), pp. 315-317.

Moon, K., Bonocora, R.P., Kim, D.D., Chen, Q., Wade, J.T., Stibitz, S. and Hinton, D.M., 2017. The BvgAS Regulon of Bordetella pertussis. *mBio,* 8(5).

Moore, D.L., Le Saux, N., Scheifele, D. and Halperin, S.A., 2004. Lack of evidence of encephalopathy related to pertussis vaccine: active surveillance by IMPACT, Canada, 1993-2002. *Pediatr Infect Dis J,* 23(6), pp. 568-571.

Morgulis, A., Coulouris, G., Raytselis, Y., Madden, T.L., Agarwala, R. and Schaffer, A.A., 2008. Database indexing for production MegaBLAST searches. *Bioinformatics,* 24(16), pp. 1757-1764.

Mu, W., Lu, H.M., Chen, J., Li, S. and Elliott, A.M., 2016. Sanger Confirmation Is Required to Achieve Optimal Sensitivity and Specificity in Next-Generation Sequencing Panel Testing. *J Mol Diagn,* 18(6), pp. 923-932.

Musser, J.M., Hewlett, E.L., Peppler, M.S. and Selander, R.K., 1986. Genetic diversity and relationships in populations of Bordetella spp. *J Bacteriol,* 166(1), pp. 230-237.

Nagarajan, N. and Pop, M., 2013. Sequence assembly demystified. *Nature Reviews Genetics,* 14(3), pp. 157-167.

Nakamura, K., Oshima, T., Morimoto, T., Ikeda, S., Yoshikawa, H., Shiwa, Y., Ishikawa, S., Linak, M.C., Hirai, A., Takahashi, H., Altaf-Ul-Amin, M., Ogasawara, N. and Kanaya, S., 2011. Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res,* 39(13), p. e90.

Nash, Z.M. and Cotter, P.A., 2019. Regulated, sequential processing by multiple proteases is required for proper maturation and release of Bordetella filamentous hemagglutinin. *Mol Microbiol*.

NCBI, 2019. *NCBI Genome Browser summary table* [Online]. Available from: https://www.ncbi.nlm.nih.gov/genome/browse/#!/overview/ [Accessed 23 September 2019].

New Zealand Ministry of Health, 2017. *National outbreak of whooping cough declared* [Online]. New Zealand: New Zealand Ministry of Health. Available from: https://www.health.govt.nz/news-media/media-releases/national-outbreak-whooping-cough-declared [Accessed 07 February 2019].

New Zealand Ministry of Health, 2019. *National and DHB immunisation data.* Available from: https://www.health.govt.nz/our-work/preventative-health-wellness/immunisation/immunisation-coverage/national-and-dhb-immunisation-data [Accessed 29 May 2019].

NHS, 2017. *Childhood Vaccination Coverage Statistics, England 2016-17.* Available from: https://digital.nhs.uk/data-and-information/publications/statistical/childhood-vaccination-coverage-statistics/childhood-vaccination-coverage-statistics-england-2016-17 [Accessed 29 May 2019].

Nicholson, T.L., Conover, M.S. and Deora, R., 2012. Transcriptome Profiling Reveals Stage-Specific Production and Requirement of Flagella during Biofilm Development in Bordetella bronchiseptica. *PLoS One,* 7(11).

Nicosia, A., Perugini, M., Franzini, C., Casagli, M.C., Borri, M.G., Antoni, G., Almoni, M., Neri, P., Ratti, G. and Rappuoli, R., 1986. Cloning and sequencing of the pertussis toxin genes: operon structure and gene duplication. *Proc Natl Acad Sci U S A,* 83(13), pp. 4631-4635.

Niederweis, M., Ehrt, S., Heinz, C., Klocker, U., Karosi, S., Swiderek, K.M., Riley, L.W. and Benz, R., 1999. Cloning of the mspA gene encoding a porin from Mycobacterium smegmatis. *Mol Microbiol,* 33(5), pp. 933-945.

Nishikawa, S., Shinzawa, N., Nakamura, K., Ishigaki, K., Abe, H. and Horiguchi, Y., 2016. The bvg-repressed gene brtA, encoding biofilm-associated surface adhesin, is expressed during host infection by Bordetella bronchiseptica. *Microbiol Immunol,* 60(2), pp. 93-105.

Noble, G.R., Bernier, R.H., Esber, E.C., Hardegree, M.C., Hinman, A.R., Klein, D. and Saah, A.J., 1987. Acellular and whole-cell pertussis vaccines in Japan. Report of a visit by US scientists. *Jama,* 257(10), pp. 1351-1356.

Nyren, P., 1987. Enzymatic method for continuous monitoring of DNA polymerase activity. *Anal Biochem,* 167(2), pp. 235-238.

NZ Ministry of Health, 2005. *The National Childhood Immunisation Coverage Survey 2005.* Available from: https://www.health.govt.nz/system/files/documents/publications/national-childhood-immunisation-coverage-survey2005.pdf [Accessed 29 May 2019].

Octavia, S., Sintchenko, V., Gilbert, G.L., Lawrence, A., Keil, A.D., Hogg, G. and Lan, R., 2012. Newly emerging clones of Bordetella pertussis carrying prn2 and ptxP3 alleles implicated in Australian pertussis epidemic in 2008-2010. *J Infect Dis,* 205(8), pp. 1220-1224.

Olin, P., Rasmussen, F., Gustafsson, L., Hallander, H.O. and Heijbel, H., 1997. Randomised controlled trial of two-component, three-component, and five-component acellular pertussis vaccines compared with whole-cell pertussis vaccine. Ad Hoc Group for the Study of Pertussis Vaccines. *Lancet,* 350(9091), pp. 1569-1577.

Oxford Nanopore Technologies, 2016. *Update: New 'R9' nanopore for faster, more accurate sequencing, and new ten minute preparation kit* [Online]. Available from: https://nanoporetech.com/about-us/news/update-new-r9-nanopore-faster-more-accurate-sequencing-and-new-ten-minute-preparation [Accessed 09 July 2020].

Oxford Nanopore Technologies, 2017. *New basecaller now performs 'raw basecalling', for improved sequencing accuracy* [Online]. Available from: https://nanoporetech.com/about-us/news/new-basecaller-now-performs-raw-basecalling-improved-sequencing-accuracy [Accessed 09 July 2020].

Oxford Nanopore Technologies, 2018a. *Clive G Brown: CTO plenary from London Calling* [Online]. Available from: https://nanoporetech.com/about-us/news/clive-g-brown-cto-plenary-london-calling?keys=MinION&page=28 [Accessed 28 June 2018].

Oxford Nanopore Technologies, 2018b. *Clive G Brown: Nanopore Community Meeting 2018 talk* [Online]. Available from: https://nanoporetech.com/about-us/news/clive-g-brown-nanopore-community-meeting-2018-talk [Accessed 15 January 2019].

Page, A.J., Taylor, B., Delaney, A.J., Soares, J., Seemann, T., Keane, J.A. and Harris, S.R., 2016. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb Genom,* 2(4), p. e000056.

Park, J., Zhang, Y., Buboltz, A.M., Zhang, X., Schuster, S.C., Ahuja, U., Liu, M., Miller, J.F., Sebaihia, M., Bentley, S.D., Parkhill, J. and Harvill, E.T., 2012. Comparative genomics of the classical Bordetella subspecies: the evolution and exchange of virulence-associated diversity amongst closely related pathogens. *BMC Genomics,* 13, p. 545.

Parkhill, J., Sebaihia, M., Preston, A., Murphy, L.D., Thomson, N., Harris, D.E., Holden, M.T.G., Churcher, C.M., Bentley, S.D., Mungall, K.L., Cerdeño-Tárraga, A.M., Temple, L., James, K., Harris, B., Quail, M.A., Achtman, M., Atkin, R., Baker, S., Basham, D., Bason, N., Cherevach, I., Chillingworth, T., Collins, M., Cronin, A., Davis, P., Doggett, J., Feltwell, T., Goble, A., Hamlin, N., Hauser, H., Holroyd, S., Jagels, K., Leather, S., Moule, S., Norberczak, H., O'Neil, S., Ormond, D., Price, C., Rabbinowitsch, E., Rutter, S., Sanders, M., Saunders, D., Seeger, K., Sharp, S., Simmonds, M., Skelton, J., Squares, R., Squares, S., Stevens, K., Unwin, L., Whitehead, S., Barrell, B.G. and Maskell, D.J., 2003. Comparative analysis of the genome sequences of Bordetella pertussis, Bordetella parapertussis and Bordetella bronchiseptica. *Nature Genetics,* 35(1), pp. 32-40.

Pawloski, L.C., Queenan, A.M., Cassiday, P.K., Lynch, A.S., Harrison, M.J., Shang, W., Williams, M.M., Bowden, K.E., Burgos-Rivera, B., Qin, X., Messonnier, N. and Tondella, M.L., 2014. Prevalence and molecular characterization of pertactin-deficient Bordetella pertussis in the United States. *Clin Vaccine Immunol,* 21(2), pp. 119-125.

Payne, A., Holmes, N., Rakyan, V. and Loose, M., 2018a. BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files. *Bioinformatics*.

Payne, A., Holmes, N., Rakyan, V. and Loose, M., 2018b. Whale watching with BulkVis: A graphical viewer for Oxford Nanopore bulk fast5 files. *bioRxiv*.

Payne, A., Holmes, N., Rakyan, V. and Loose, M., 2019. BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files. *Bioinformatics,* 35(13), pp. 2193-2198.

Pfeiffer, F., Grober, C., Blank, M., Handler, K., Beyer, M., Schultze, J.L. and Mayer, G., 2018. Systematic evaluation of error rates and causes in short samples in next-generation sequencing. *Sci Rep,* 8(1), p. 10950.

PHE, 2018. *Guidelines for the Public Health Management of Pertussis in England* [Online]. UK: Public Health England. Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file

/762766/Guidelines_for_the_Public_Health_management_of_Pertussis_in_England.pdf [Accessed 24 June 2019].

Pillsbury, A., Quinn, H.E. and McIntyre, P.B., 2014. Australian vaccine preventable disease epidemiological review series: pertussis, 2006-2012. *Commun Dis Intell Q Rep,* 38(3), pp. E179-194.

Pomerantz, A., Penafiel, N., Arteaga, A., Bustamante, L., Pichardo, F., Coloma, L.A., Barrio-Amoros, C.L., Salazar-Valenzuela, D. and Prost, S., 2017. Real-time DNA barcoding in a remote rainforest using nanopore sequencing. *bioRxiv*.

Pop, M., 2009. Genome assembly reborn: recent computational challenges. *Brief Bioinform,* 10(4), pp. 354-366.

Preston, A., Parkhill, J. and Maskell, D.J., 2004. The bordetellae: lessons from genomics. *Nat Rev Microbiol,* 2(5), pp. 379-390.

Preston, N.W., 1976. Prevalent serotypes of Bordetella pertussis in non-vaccinated communities. *J Hyg (Lond),* 77(1), pp. 85-91.

Proffitt, A., 2018. Illumina announces iSeq 100, Thermo Fisher partnership. (20 September 2019). Available from: http://www.bio-itworld.com/2018/01/08/illumina-announces-iseq-100-thermo-fisher-partership.aspx [Accessed 20 September 2019].

Qi, M., Huang, H., Zhang, Y., Wang, H., Li, H. and Lu, Z., 2019. Novel tetrahydrofuran (THF) degradation-associated genes and cooperation patterns of a THF-degrading microbial community as revealed by metagenomic. *Chemosphere,* 231, pp. 173-183.

Queenan, A.M., Cassiday, P.K. and Evangelista, A., 2013. Pertactin-negative variants of Bordetella pertussis in the United States. *N Engl J Med,* 368(6), pp. 583-584.

Quick, J., 2018. *Ultra-long read sequencing protocol for RAD004* [Online]. @protocolsIO. Available from: https://www.protocols.io/view/ultra-long-read-sequencing-protocol-for-rad004-mrxc57n [Accessed 22 July 2019].

Quick, J., Loman, N.J., Duraffour, S., Simpson, J.T., Severi, E., Cowley, L., Bore, J.A., Koundouno, R., Dudas, G., Mikhail, A., Ouedraogo, N., Afrough, B., Bah, A., Baum, J.H., Becker-Ziaja, B., Boettcher, J.P., Cabeza-Cabrerizo, M., Camino-Sanchez, A., Carter, L.L., Doerrbecker, J., Enkirch, T., Garcia-Dorival, I., Hetzelt, N., Hinzmann, J., Holm, T., Kafetzopoulou, L.E., Koropogui, M., Kosgey, A., Kuisma, E., Logue, C.H., Mazzarelli, A., Meisel, S., Mertens, M., Michel, J., Ngabo, D., Nitzsche, K., Pallasch, E., Patrono, L.V., Portmann, J., Repits, J.G., Rickett, N.Y., Sachse, A., Singethan, K., Vitoriano, I., Yemanaberhan, R.L., Zekeng, E.G., Racine, T., Bello, A., Sall, A.A., Faye, O., Magassouba, N., Williams, C.V., Amburgey, V., Winona, L., Davis, E., Gerlach, J., Washington, F., Monteil, V., Jourdain, M., Bererd, M., Camara, A., Somlare, H., Gerard, M., Bado, G., Baillet, B., Delaune, D., Nebie, K.Y., Diarra, A., Savane, Y., Pallawo, R.B., Gutierrez, G.J., Milhano, N., Roger, I., Williams, C.J., Yattara, F., Lewandowski, K., Taylor, J., Rachwal, P., Turner, D.J., Pollakis, G., Hiscox, J.A., Matthews, D.A., O'Shea, M.K., Johnston, A.M., Wilson, D., Hutley, E., Smit, E., Di Caro, A., Wolfel, R., Stoecker, K., Fleischmann, E., Gabriel, M., Weller, S.A., Koivogui, L., Diallo, B., Keita, S., Rambaut, A., Formenty, P., Gunther, S. and Carroll, M.W., 2016. Real-time, portable genome sequencing for Ebola surveillance. *Nature,* 530(7589), pp. 228-232.

Reid, S., 2006. Evolution of the New Zealand childhood immunisation schedule from 1980: a personal view. *The New Zealand Medical Journal,* 119(1236).

Reid, S., 2012. The further and future evolution of the New Zealand Immunisation Schedule. *The New Zealand Medical Journal,* 125(1354).

Reid, S., Lennon, D., Thomas, M., O'Connor, P., Baker, M. and Mansoor, O., 1994. Pertussis control in New Zealand. *N Z Med J,* 107(989), pp. 460-462, 463-466.

Relman, D.A., Domenighini, M., Tuomanen, E., Rappuoli, R. and Falkow, S., 1989. Filamentous hemagglutinin of Bordetella pertussis: nucleotide sequence and crucial role in adherence. *Proc Natl Acad Sci U S A,* 86(8), pp. 2637-2641.

Rhoads, A. and Au, K.F., 2015. PacBio Sequencing and Its Applications. *Genomics Proteomics Bioinformatics,* 13(5), pp. 278-289.

Ring, N., Abrahams, J.S., Bagby, S., Preston, A. and MacArthur, I., 2019. How Genomics Is Changing What We Know About the Evolution and Genome of Bordetella pertussis. *Adv Exp Med Biol*.

Ring, N., Abrahams, J.S., Jain, M., Olsen, H., Preston, A. and Bagby, S., 2018. Resolving the complex Bordetella pertussis genome using barcoded nanopore sequencing. *Microb Genom,* 4(11).

Roberts, M., Fairweather, N.F., Leininger, E., Pickard, D., Hewlett, E.L., Robinson, A., Hayward, C., Dougan, G. and Charles, I.G., 1991. Construction and characterization of Bordetella pertussis mutants lacking the vir-regulated P.69 outer membrane protein. *Mol Microbiol,* 5(6), pp. 1393-1404.

Robison, K., 2012. Does Illlumina also have a homopolymer problem? *Omics! Omics!* [Online]. Available from: http://omicsomics.blogspot.com/2012/01/does-illlmina-also-have-homopolymer.html [Accessed 18 July 2019].

Robison, K., 2017. Could Hermione tackle MinION yield variability? *Omics! Omics!* [Online]. Available from: http://omicsomics.blogspot.co.uk/2017/02/could-hermione-tackle-minion-yield.html [Accessed 15 August 2017].

Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlen, M. and Nyren, P., 1996. Real-time DNA sequencing using detection of pyrophosphate release. *Anal Biochem,* 242(1), pp. 84-89.

Ronaghi, M., Uhlen, M. and Nyren, P., 1998. A sequencing method based on real-time pyrophosphate. *Science,* 281(5375), pp. 363, 365.

Ross, M.G., Russ, C., Costello, M., Hollinger, A., Lennon, N.J., Hegarty, R., Nusbaum, C. and Jaffe, D.B., 2013. Characterizing and measuring bias in sequence data. *Genome Biol,* 14(5), p. R51.

Rothberg, J.M., Hinz, W., Rearick, T.M., Schultz, J., Mileski, W., Davey, M., Leamon, J.H., Johnson, K., Milgrew, M.J., Edwards, M., Hoon, J., Simons, J.F., Marran, D., Myers, J.W., Davidson, J.F., Branting, A., Nobile, J.R., Puc, B.P., Light, D., Clark, T.A., Huber, M., Branciforte, J.T., Stoner, I.B., Cawley, S.E., Lyons, M., Fu, Y., Homer, N., Sedova, M., Miao, X., Reed, B., Sabina, J., Feierstein, E., Schorn, M., Alanjary, M., Dimalanta, E., Dressman, D., Kasinskas, R., Sokolsky, T., Fidanza, J.A., Namsaraev, E., McKernan, K.J., Williams, A., Roth, G.T. and Bustillo, J., 2011. An integrated semiconductor device enabling non-optical genome sequencing. *Nature,* 475, pp. 348-352.

Ruan, J., 2015. *SMARTdenovo* [Online]. Available from: https://github.com/ruanjue/smartdenovo [Accessed 03 March 2017].

Ruan, J. and Li, H., 2019. Fast and accurate long-read assembly with wtdbg2. *bioRXiV*.

Salipante, S.J., Kawashima, T., Rosenthal, C., Hoogestraat, D.R., Cummings, L.A., Sengupta, D.J., Harkins, T.T., Cookson, B.T. and Hoffman, N.G., 2014. Performance comparison of Illumina and ion torrent next-generation sequencing platforms for 16S rRNA-based bacterial community profiling. *Appl Environ Microbiol,* 80(24), pp. 7583-7591.

Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, C.A., Hutchison, C.A., Slocombe, P.M. and Smith, M., 1977. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature,* 265(5596), pp. 687-695.

Sanger, F. and Coulson, A.R., 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol,* 94(3), pp. 441-448.

Sanger, F., Coulson, A.R., Friedmann, T., Air, G.M., Barrell, B.G., Brown, N.L., Fiddes, J.C., Hutchison, C.A., 3rd, Slocombe, P.M. and Smith, M., 1978. The nucleotide sequence of bacteriophage phiX174. *J Mol Biol,* 125(2), pp. 225-246.

Sanger, F., Nicklen, S. and Coulson, A.R., 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America,* 74(12), pp. 5463-5467.

Scheller, E.V. and Cotter, P.A., 2015. Bordetella filamentous hemagglutinin and fimbriae: critical adhesins with unrealized vaccine potential. *Pathog Dis,* 73(8).

Schirmer, M., D'Amore, R., Ijaz, U.Z., Hall, N. and Quince, C., 2016. Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinformatics,* 17.

Schmid, M., Frei, D., Patrignani, A., Schlapbach, R., Frey, J.E., Remus-Emsermann, M.N.P. and Ahrens, C.H., 2018. Pushing the limits of de novo genome assembly for complex prokaryotic genomes harboring very long, near identical repeats. *bioRxiv*.

Sealey, K.L., 2015. *Is the circulating UK Bordetella pertussis population evolving to evade vaccine-induced immunity?* (PhD), University of Bath. Available from: https://researchportal.bath.ac.uk/en/studentTheses/is-the-circulating-uk-bordetella-pertussis-population-evolving-to [Accessed 11 June 2020].

Sealey, K.L., Harris, S.R., Fry, N.K., Hurst, L.D., Gorringe, A.R., Parkhill, J. and Preston, A., 2015. Genomic analysis of isolates from the United Kingdom 2012 pertussis outbreak reveals that vaccine antigen genes are unusually fast evolving. *J Infect Dis,* 212(2), pp. 294-301.

Seemann, T., 2014a. Prokka: rapid prokaryotic genome annotation. *Bioinformatics,* 30(14), pp. 2068-2069.

Seemann, T., 2014b. *Snippy: Rapid bacterial SNP calling and core genome alignments* [Online]. Available from: https://github.com/tseemann/snippy [Accessed 24 August 2019].

Seemann, T., 2019. *MLST* [Online]. GitHub. Available from: https://github.com/tseemann/mlst [Accessed 11 June 2019].

Seppey, M., Manni, M. and Zdobnov, E.M., 2019. BUSCO: Assessing Genome Assembly and Annotation Completeness. *Methods Mol Biol,* 1962, pp. 227-245.

Serra, D., Bosch, A., Russo, D.M., Rodriguez, M.E., Zorreguieta, A., Schmitt, J., Naumann, D. and Yantorno, O., 2007. Continuous nondestructive monitoring of Bordetella pertussis biofilms by

Fourier transform infrared spectroscopy and other corroborative techniques. *Anal Bioanal Chem,* 387(5), pp. 1759-1767.

Serra, D.O., Conover, M.S., Arnal, L., Sloan, G.P., Rodriguez, M.E., Yantorno, O.M. and Deora, R., 2011. FHA-mediated cell-substrate and cell-cell adhesions are critical for Bordetella pertussis biofilm formation on abiotic surfaces and in the mouse nose and the trachea. *PLoS One,* 6(12), p. e28811.

Shendure, J. and Ji, H., 2008. Next-generation DNA sequencing. *Nat Biotechnol,* 26(10), pp. 1135-1145.

Shinoda, M., Katada, T. and Ui, M., 1990. Selective coupling of purified alpha-subunits of pertussis toxin-substrate GTP-binding proteins to endogenous receptors in rat brain membranes treated with N-ethylmaleimide. *Cell Signal,* 2(4), pp. 403-414.

Siguier, P., Laboratoire de Microbiologie et Génétique Moléculaires, U.M.d.R., Centre National de Recherche Scientifique, Toulouse Cedex, France, Gourbeyre, E., Laboratoire de Microbiologie et Génétique Moléculaires, U.M.d.R., Centre National de Recherche Scientifique, Toulouse Cedex, France, Chandler, M. and Laboratoire de Microbiologie et Génétique Moléculaires, U.M.d.R., Centre National de Recherche Scientifique, Toulouse Cedex, France, 2014. Bacterial insertion sequences: their genomic impact and diversity. *FEMS Microbiology Reviews,* 38(5), pp. 865-891.

Simão, F.A., Department of Genetic Medicine and Development, U.o.G.M.S.a.S.I.o.B., rue Michel-Servet 1, 1211 Geneva, Switzerland, Waterhouse, R.M., Department of Genetic Medicine and Development, U.o.G.M.S.a.S.I.o.B., rue Michel-Servet 1, 1211 Geneva, Switzerland, Ioannidis, P., Department of Genetic Medicine and Development, U.o.G.M.S.a.S.I.o.B., rue Michel-Servet 1, 1211 Geneva, Switzerland, Kriventseva, E.V., Department of Genetic Medicine and Development, U.o.G.M.S.a.S.I.o.B., rue Michel-Servet 1, 1211 Geneva, Switzerland, Zdobnov, E.M. and Department of Genetic Medicine and Development, U.o.G.M.S.a.S.I.o.B., rue Michel-Servet 1, 1211 Geneva, Switzerland, 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics,* 31(19), pp. 3210-3212.

Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J. and Birol, İ., 2009. ABySS: A parallel assembler for short read sequence data. *Genome Res,* 19(6), pp. 1117-1123.

Smith, A.M., Heisler, L.E., St Onge, R.P., Farias-Hesson, E., Wallace, I.M., Bodeau, J., Harris, A.N., Perry, K.M., Giaever, G., Pourmand, N. and Nislow, C., 2010. Highly-multiplexed barcode sequencing: an efficient method for parallel analysis of pooled samples. *Nucleic Acids Res,* 38(13), p. e142.

Somerville, R.L., Grant, C.C., Scragg, R.K. and Thomas, M.G., 2007. Hospitalisations due to pertussis in New Zealand in the pre-immunisation and mass immunisation eras. *J Paediatr Child Health,* 43(3), pp. 147-153.

Song, L., Hobaugh, M.R., Shustak, C., Cheley, S., Bayley, H. and Gouaux, J.E., 1996. Structure of staphylococcal alpha-hemolysin, a heptameric transmembrane pore. *Science,* 274(5294), pp. 1859-1866.

Sousa, C., de Lorenzo, V. and Cebolla, A., 1997. Modulation of gene expression through chromosomal positioning in Escherichia coli. *Microbiology,* 143 (Pt 6), pp. 2071-2078.

Stack Exchange, 2017. *How can I improve the yield of MinION sequencing runs?* [Online]. Available from: *https://bioinformatics.stackexchange.com/questions/296/how-can-i-improve-the-yield-of-minion-sequencing-runs/298* [Accessed 15 August 2017].

Stainer, D.W. and Scholte, M.J., 1970. A simple chemically defined medium for the production of phase I Bordetella pertussis. *J Gen Microbiol,* 63(2), pp. 211-220.

Stein-Zamir, C., Shoob, H., Abramson, N. and Zentner, G., 2010. The impact of additional pertussis vaccine doses on disease incidence in children and infants. *Vaccine,* 29(2), pp. 207-211.

Stenson, T.H., Allen, A.G., al-Meer, J.A., Maskell, D. and Peppler, M.S., 2005. Bordetella pertussis risA, but Not risS, Is Required for Maximal Expression of Bvg-Repressed Genes. *Infect Immun,* 73(9), pp. 5995-6004.

Stiles, M.E. and Ng, L.K., 1981. Biochemical characteristics and identification of Enterobacteriaceae isolated from meats. *Appl Environ Microbiol,* 41(3), pp. 639-645.

Stockbauer, K.E., Fuchslocher, B., Miller, J.F. and Cotter, P.A., 2001. Identification and characterization of BipA, a Bordetella Bvg-intermediate phase protein. *Mol Microbiol,* 39(1), pp. 65-78.

Stoddart, D., Heron, A.J., Klingelhoefer, J., Mikhailova, E., Maglia, G. and Bayley, H., 2010. Nucleobase recognition in ssDNA at the central constriction of the αhemolysin pore. *Nano Lett,* 10(9), pp. 3633-3637.

Stoddart, D., Heron, A.J., Mikhailova, E., Maglia, G. and Bayley, H., 2009. Single-nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore. *Proc Natl Acad Sci U S A,* 106(19), pp. 7702-7707.

Stoddart, D., Maglia, G., Mikhailova, E., Heron, A.J. and Bayley, H., 2010. Multiple base-recognition sites in a biological nanopore: two heads are better than one. *Angew Chem Int Ed Engl,* 49(3), pp. 556-559.

Storsaeter, J., Hallander, H.O., Gustafsson, L. and Olin, P., 1998. Levels of anti-pertussis antibodies related to protection after household exposure to Bordetella pertussis. *Vaccine,* 16(20), pp. 1907-1916.

Stuff NZ, 2019. *Whooping cough spike in Auckland prompts call for immunisation* [Online]. Available from: https://www.stuff.co.nz/national/health/110099771/whooping-cough-spike-in-auckland-prompts-call-for-immunisation [Accessed 07 February 2019].

Tamura, M., Nogimori, K., Yajima, M., Ase, K. and Ui, M., 1983. A role of the B-oligomer moiety of islet-activating protein, pertussis toxin, in development of the biological effects on intact cells. *Journal of Biological Chemistry,* 258, pp. 6756-6761.

Tazato, N., Handa, Y., Nishijima, M., Kigawa, R., Sano, C. and Sugiyama, J., 2015. Novel environmental species isolated from the plaster wall surface of mural paintings in the Takamatsuzuka tumulus: Bordetella muralis sp. nov., Bordetella tumulicola sp. nov. and Bordetella tumbae sp. nov. *Int J Syst Evol Microbiol,* 65(12), pp. 4830-4838.

Teng, H., Cao, M.D., Hall, M.B., Duarte, T., Wang, S. and Coin, L.J.M., 2018. Chiron: translating nanopore raw signal directly into nucleotide sequence using deep learning. *Gigascience,* 7(5).

Thompson, J.D., Higgins, D.G. and Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research,* 22(22), pp. 4673-4680.

Ton, K.N.T., Cree, S.L., Gronert-Sum, S.J., Merriman, T.R., Stamp, L.K. and Kennedy, M.A., 2017. Multiplexed nanopore sequencing of HLA-B locus in Māori and Polynesian samples. *bioRxiv*.

Tsang, R.S.W., Shuel, M., Cronin, K., Deng, S., Whyte, K., Marchand-Austin, A., Ma, J., Bolotin, S., Crowcroft, N., Schwartz, K., Van Domselaar, G., Graham, M. and Jamieson, F.B., 2019. The evolving nature of *Bordetella pertussis* in Ontario, Canada, 2009-2017: strains with shifting genotypes and pertactin-deficiency. *Can J Microbiol*.

Urisu, A., Cowell, J.L. and Manclark, C.R., 1986. Filamentous hemagglutinin has a major role in mediating adherence of Bordetella pertussis to human WiDr cells. *Infect Immun,* 52(3), pp. 695-701.

van der Helm, E., 2017. Highlights of a two days nanopore conference. *DNA coil* [Online]. Available from: http://www.dnacoil.com/tag/minion/ [Accessed 15 August 2017].

van der Zee, A., Groenendijk, H., Peeters, M. and Mooi, F.R., 1996. The differentiation of Bordetella parapertussis and Bordetella bronchiseptica from humans and animals as determined by DNA polymorphism mediated by two different insertion sequence elements suggests their phylogenetic relationship. *Int J Syst Bacteriol,* 46(3), pp. 640-647.

van der Zee, A., Mooi, F., Van Embden, J. and Musser, J., 1997. Molecular evolution and host adaptation of Bordetella spp.: phylogenetic analysis using multilocus enzyme electrophoresis and typing with three insertion sequences. *J Bacteriol,* 179(21), pp. 6609-6617.

van der Zee, A., Vernooij, S., Peeters, M., van Embden, J. and Mooi, F.R., 1996. Dynamics of the population structure of Bordetella pertussis as measured by IS1002-associated RFLP: comparison of pre- and post-vaccination strains and global distribution. *Microbiology,* 142 (Pt 12), pp. 3479-3485.

van Gent, M., Heuvelman, C.J., van der Heide, H.G., Hallander, H.O., Advani, A., Guiso, N., Wirsing von Konig, C.H., Vestrheim, D.F., Dalby, T., Fry, N.K., Pierard, D., Detemmerman, L., Zavadilova, J., Fabianova, K., Logan, C., Habington, A., Byrne, M., Lutynska, A., Mosiej, E., Pelaz, C., Grondahl-Yli-Hannuksela, K., Barkoff, A.M., Mertsola, J., Economopoulou, A., He, Q. and Mooi, F.R., 2015. Analysis of Bordetella pertussis clinical isolates circulating in European countries during the period 1998-2012. *Eur J Clin Microbiol Infect Dis,* 34(4), pp. 821-830.

van Loo, I.H., Heuvelman, K.J., King, A.J. and Mooi, F.R., 2002. Multilocus sequence typing of Bordetella pertussis based on surface protein genes. *J Clin Microbiol,* 40(6), pp. 1994-2001.

van Loo, I.H., van der Heide, H.G., Nagelkerke, N.J., Verhoef, J. and Mooi, F.R., 1999. Temporal trends in the population structure of Bordetella pertussis during 1949-1996 in a highly vaccinated population. *J Infect Dis,* 179(4), pp. 915-923.

Vandamme, P., Heyndrickx, M., de Roose, I., Lammens, C., de Vos, P. and Kersters, K., 1997. Characterization of Bordetella strains and related bacteria by amplified ribosomal DNA restriction analysis and randomly and repetitive element-primed PCR. *International Journal of Systematic Bacteriology,* 47(3), pp. 802-807.

Vandamme, P., Heyndrickx, M., Vancanneyt, M., Hoste, B., De Vos, P., Falsen, E., Kersters, K. and Hinz, K.H., 1996. Bordetella trematum sp. nov., isolated from wounds and ear infections in humans, and reassessment of Alcaligenes denitrificans Ruger and Tan 1983. *Int J Syst Bacteriol,* 46(4), pp. 849-858.

Vandamme, P., Hommez, J., Vancanneyt, M., Monsieurs, M., Hoste, B., Cookson, B., Wirsing von Konig, C.H., Kersters, K. and Blackall, P.J., 1995. Bordetella hinzii sp. nov., isolated from poultry and humans. *Int J Syst Bacteriol,* 45(1), pp. 37-45.

Vandamme, P., Peeters, C., Cnockaert, M., Inganas, E., Falsen, E., Moore, E.R., Nunes, O.C., Manaia, C.M., Spilker, T. and LiPuma, J.J., 2015. Bordetella bronchialis sp. nov., Bordetella flabilis sp. nov. and Bordetella sputigena sp. nov., isolated from human respiratory specimens, and reclassification of Achromobacter sediminum Zhang et al. 2014 as Verticia sediminum gen. nov., comb. nov. *Int J Syst Evol Microbiol,* 65(10), pp. 3674-3682.

Vaser, R., Sovic, I., Nagarajan, N. and Sikic, M., 2017. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res,* 27(5), pp. 737-746.

Venetz, J.E., Del Medico, L., Wolfle, A., Schachle, P., Bucher, Y., Appert, D., Tschan, F., Flores-Tinoco, C.E., van Kooten, M., Guennoun, R., Deutsch, S., Christen, M. and Christen, B., 2019. Chemical synthesis rewriting of a bacterial genome to achieve design flexibility and biological functionality. *Proc Natl Acad Sci U S A,* 116(16), pp. 8070-8079.

von Wintzingerode, F., Schattke, A., Siddiqui, R.A., Rosick, U., Gobel, U.B. and Gross, R., 2001. Bordetella petrii sp. nov., isolated from an anaerobic bioreactor, and emended description of the genus Bordetella. *Int J Syst Evol Microbiol,* 51(Pt 4), pp. 1257-1265.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K. and Earl, A.M., 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One,* 9(11), p. e112963.

Watson, M., 2018-03-08 2018. A simple test for uncorrected insertions and deletions (indels) in bacterial genomes. *Opiniomics* [Online]. Available from: http://www.opiniomics.org/a-simple-test-for-uncorrected-insertions-and-deletions-indels-in-bacterial-genomes/ [Accessed 10 July 2018].

Watson, M. and Warr, A., 2019. Errors in long-read assemblies can critically affect protein prediction. *Nat Biotechnol,* 37(2), pp. 124-126.

Weigand, M.R., Pawloski, L.C., Peng, Y., Ju, H., Burroughs, M., Cassiday, P.K., Davis, J.K., DuVall, M., Johnson, T., Juieng, P., Knipe, K., Loparev, V.N., Mathis, M.H., Rowe, L.A., Sheth, M., Williams, M.M. and Tondella, M.L., 2018. Screening and Genomic Characterization of Filamentous Hemagglutinin-Deficient Bordetella pertussis. *Infect Immun,* 86(4).

Weigand, M.R., Peng, Y., Batra, D., Burroughs, M., Davis, J.K., Knipe, K., Loparev, V.N., Johnson, T., Juieng, P., Rowe, L.A., Sheth, M., Tang, K., Unoarumhi, Y., Williams, M.M. and Tondella, M.L., 2019. Conserved Patterns of Symmetric Inversion in the Genome Evolution of Bordetella Respiratory Pathogens. *mSystems,* 4(6).

Weigand, M.R., Peng, Y., Loparev, V., Batra, D., Bowden, K.E., Burroughs, M., Cassiday, P.K., Davis, J.K., Johnson, T., Juieng, P., Knipe, K., Mathis, M.H., Pruitt, A.M., Rowe, L., Sheth, M., Tondella, M.L. and Williams, M.M., 2017. The History of Bordetella pertussis Genome Evolution Includes Structural Rearrangement. *J Bacteriol,* 199(8).

Weigand, M.R., Peng, Y., Loparev, V., Johnson, T., Juieng, P., Gairola, S., Kumar, R., Shaligram, U., Gowrishankar, R., Moura, H., Rees, J., Schieltz, D.M., Williamson, Y., Woolfitt, A., Barr, J., Tondella, M.L. and Williams, M.M., 2016. Complete Genome Sequences of Four Bordetella pertussis Vaccine Reference Strains from Serum Institute of India. *Genome Announc,* 4(6).

Weirather, J.L., de Cesare, M., Wang, Y., Piazza, P., Sebastiano, V., Wang, X.J., Buck, D. and Au, K.F., 2017. Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Res,* 6.

Wendelboe, A.M., Van Rie, A., Salmaso, S. and Englund, J.A., 2005. Duration of immunity against pertussis after natural infection or vaccination. *Pediatr Infect Dis J,* 24(5 Suppl), pp. S58-61.

Wetterstrand, K.A., 2016. DNA sequencing costs: data from the NHGRI Genome Sequencing Program (GSP). *National Human Genome Research Institute* [Online]. Available from: *http://www.genome.gov/sequencingcosts* [Accessed 30 August 2017].

Weyant, R.S., Hollis, D.G., Weaver, R.E., Amin, M.F., Steigerwalt, A.G., O'Connor, S.P., Whitney, A.M., Daneshvar, M.I., Moss, C.W. and Brenner, D.J., 1995. Bordetella holmesii sp. nov., a new gram-negative species associated with septicemia. *J Clin Microbiol,* 33(1), pp. 1-7.

WHO, 2018. *Global and regional immunization profile* [Online]. Geneva: WHO. Available from: *https://www.who.int/immunization/monitoring_surveillance/data/gs_gloprofile.pdf?ua=1* [Accessed 09 January 2019].

WHO, 2019a. *Immunization coverage with 3rd dose of diphtheria and tatnus toxoid and pertussis containing vaccines* [Online]. Available from: *https://www.who.int/immunization/monitoring_surveillance/burden/vpd/surveillance_type/passive/pertussis/en/* [Accessed 22 June 2019].

WHO, 2019b. *Third dose of diphtheria toxoid, tetanus toxoid and pertussis vaccine* [Online]. Available from: *http://apps.who.int/immunization_monitoring/globalsummary/timeseries/tscoveragedtp3.html* [Accessed 24 June 2019].

Wick, R.R., 2017. *Porechop: Adapter trimmer for Oxford Nanopore reads* [Online]. Available from: *https://github.com/rrwick/Porechop* [Accessed 22nd July 2017].

Wick, R.R., Judd, L.M., Gorrie, C.L. and Holt, K.E., 2017a. Completing bacterial genome assemblies with multiplex MinION sequencing. *Microb Genom,* 3(10), p. e000132.

Wick, R.R., Judd, L.M., Gorrie, C.L. and Holt, K.E., 2017b. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput Biol,* 13(6), p. e1005595.

Wick, R.R., Judd, L.M. and Holt, K.E., 2018a. *Comparison of Oxford Nanopore basecalling tools* [Online]. GitHub. Available from: *https://github.com/rrwick/Basecalling-comparison* [Accessed 14 June 2018].

Wick, R.R., Judd, L.M. and Holt, K.E., 2018b. Deepbinner: Demultiplexing barcoded Oxford Nanopore reads with deep convolutional neural networks. *PLoS Comput Biol,* 14(11).

Wick, R.R., Judd, L.M. and Holt, K.E., 2019. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol,* 20(1), p. 129.

Wilcoxon, F., 1945. Individual Comparisons by Ranking Methods. *Biometrics Bulletin,* 1(6), pp. 80-83.

Willems, R.J., van der Heide, H.G. and Mooi, F.R., 1992. Characterization of a Bordetella pertussis fimbrial gene cluster which is located directly downstream of the filamentous haemagglutinin gene. *Mol Microbiol,* 6(18), pp. 2661-2671.

Williams, C.L., Boucher, P.E., Stibitz, S. and Cotter, P.A., 2005. BvgA functions as both an activator and a repressor to control Bvg phase expression of bipA in Bordetella pertussis. *Mol Microbiol,* 56(1), pp. 175-188.

Williams, M.M., Sen, K., Weigand, M.R., Skoff, T.H., Cunningham, V.A., Halse, T.A. and Tondella, M.L., 2016. Bordetella pertussis Strain Lacking Pertactin and Pertussis Toxin. *Emerg Infect Dis,* 22(2), pp. 319-322.

Winter, K., Harriman, K., Zipprich, J., Schechter, R., Talarico, J., Watt, J. and Chavez, G., 2012. California pertussis epidemic, 2010. *J Pediatr,* 161(6), pp. 1091-1096.

World Health Organisation, 2014. *WHO SAGE Pertussis working group background paper SAGE April 2014.* Available from: https://www.who.int/immunization/sage/meetings/2014/april/1_Pertussis_background_FINAL4_web.pdf [Accessed 24 July 2019].

Worldometer, 2020. *World population by country* [Online]. Available from: https://www.worldometers.info/world-population/population-by-country/ [Accessed 11 June 2020].

Xu, Y. and Barbieri, J.T., 1995. Pertussis toxin-mediated ADP-ribosylation of target proteins in Chinese hamster ovary cells involves a vesicle trafficking mechanism. *Infect Immun,* 63(3), pp. 825-832.

Xu, Y. and Barbieri, J.T., 1996. Pertussis toxin-catalyzed ADP-ribosylation of Gi-2 and Gi-3 in CHO cells is modulated by inhibitors of intracellular trafficking. *Infect Immun,* 64(2), pp. 593-599.

Xu, Y., Liu, B., Grondahl-Yli-Hannuksila, K., Tan, Y., Feng, L., Kallonen, T., Wang, L., Peng, D., He, Q. and Zhang, S., 2015. Whole-genome sequencing reveals the effect of vaccination on the evolution of Bordetella pertussis. *Sci Rep,* 5, p. 12888.

Xu, Z., Octavia, S., Luu, L.D.W., Payne, M., Timms, V., Tay, C.Y., Keil, A.D., Sintchenko, V., Guiso, N. and Lan, R., 2019. Pertactin-Negative and Filamentous Hemagglutinin-Negative Bordetella pertussis, Australia, 2013-2017. *Emerg Infect Dis,* 25(6), pp. 1196-1199.

Xu, Z., Wang, Z., Luan, Y., Li, Y., Liu, X., Peng, X., Octavia, S., Payne, M. and Lan, R., 2019. Genomic epidemiology of erythromycin-resistant Bordetella pertussis in China. *Emerg Microbes Infect,* 8(1), pp. 461-470.

Yeung, K.H.T., Duclos, P., Nelson, E.A.S. and Hutubessy, R.C.W., 2017. An update of the global burden of pertussis in children younger than 5 years: a modelling study. *Lancet Infect Dis,* 17(9), pp. 974-980.

Zhang, W., Qi, W., Albert, T.J., Motiwala, A.S., Alland, D., Hyytia-Trees, E.K., Ribot, E.M., Fields, P.I., Whittam, T.S. and Swaminathan, B., 2006. Probing genomic diversity and evolution of Escherichia coli O157 by single nucleotide polymorphisms. *Genome Res,* 16(6), pp. 757-767.

Zhang, Z., Schwartz, S., Wagner, L. and Miller, W., 2000. A greedy algorithm for aligning DNA sequences. *J Comput Biol,* 7(1-2), pp. 203-214.

Zheng, X., Xie, X., Yu, C., Zhang, Q., Wang, Y., Cong, J., Liu, N., He, Z., Yang, B. and Liu, J., 2019. Unveiling the activating mechanism of tea residue for boosting the biological decolorization performance of refractory dye. *Chemosphere,* 233, pp. 110-119.

Zheng, Y., Rodewald, L., Yang, J., Qin, Y., Pang, M., Feng, L. and Yu, H., 2018. The landscape of vaccines in China: history, classification, supply, and price. *BMC Infect Dis,* 18(1), p. 502.

Zimin, A.V., Puiu, D., Luo, M.C., Zhu, T., Koren, S., Marcais, G., Yorke, J.A., Dvorak, J. and Salzberg, S.L., 2017. Hybrid assembly of the large and highly repetitive genome of Aegilops tauschii, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *Genome Res,* 27(5), pp. 787-792.

Zomer, A., Otsuka, N., Hiramatsu, Y., Kamachi, K., Nishimura, N., Ozaki, T., Poolman, J. and Geurtsen, J., 2018. Bordetella pertussis population dynamics and phylogeny in Japan after adoption of acellular pertussis vaccines. *Microb Genom*.

Appendix

“I sometimes find, and I am sure you know the feeling, that I simply have too many thoughts and memories crammed into my mind.”

- J.K. Rowling, Harry Potter and the Goblet of Fire

Supplementary Form SF2.1 Statement of authorship

This declaration concerns the article entitled:	
Resolving the complex *Bordetella pertussis* genome using barcoded nanopore sequencing	

Publication status (tick one)

Draft manuscript	□	**Submitted**	□	**In review**	□	**Accepted**	□	**Published**	X

Publication details (reference)	Ring, N., Abrahams, J.S., Jain, M., Olsen, H., Preston, A. and Bagby, S., 2018. Resolving the complex Bordetella pertussis genome using barcoded nanopore sequencing. Microb Genom, 4(11).

Copyright status (tick the appropriate statement)

I hold the copyright for this material	X	Copyright is retained by the publisher, but I have been given permission to replicate the material here	□

Candidate's contribution to the paper (provide details, and also indicate as a percentage)	The candidate contributed to / considerably contributed to / predominantly executed the… Formulation of ideas: Considerably contributed to 90% Design of methodology: Considerably contributed to 95% Experimental work: Predominantly executed the 95% Presentation of data in journal format: Predominantly executed the 95%
Statement from Candidate	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature.

Signed		**Date**	

Supplementary tables S2.1, S2.2, S2.3, S2.4, S2.5, S2.6 and S2.7

Supplementary data from Chapter 2, “Resolving the complex *Bordetella pertussis* genome using barcoded nanopore sequencing”, can be downloaded from:

https://figshare.com/s/003465e08ba1e8fc8780

Table S3.1 Sequence accession numbers for all New Zealand samples

Strain	BioSample Accession	Illumina SRA accession	Nanopore SRA accession	Hybrid genome assembly accession
NZ1	SAMN12385760	NA	SRR9849540	NA
NZ2	SAMN12385761	SRR11855999	SRR9849539	CP054126
NZ3	SAMN12385762	SRR11855998	SRR9849542	CP054125
NZ4	SAMN12385763	SRR11855987	SRR9849541	CP054124
NZ5	SAMN12385764	NA	SRR9849536	NA
NZ6	SAMN12385765	SRR11855976	SRR9849535	CP054123
NZ7	SAMN12385766	SRR11855965	SRR9849538	CP054122
NZ8	SAMN12385767	SRR11855954	SRR9849537	CP054121
NZ9	SAMN12385768	SRR11855943	SRR9849534	CP054120
NZ10	SAMN12385769	SRR11855939	SRR9849533	CP054119
NZ11	SAMN12385770	SRR11855938	SRR9849531	CP054118
NZ12	SAMN12385771	SRR11855937	SRR9849530	CP054127
NZ13	SAMN12385772	SRR11855997	SRR9849529	CP054117
NZ14	SAMN12385773	SRR11855996	SRR9849528	CP054116
NZ15	SAMN12385774	SRR11855995	SRR9849573	CP054115
NZ16	SAMN12385775	SRR11855994	SRR9849572	CP054114
NZ17	SAMN12385776	SRR11855993	SRR9849571	CP054113
NZ18	SAMN12385777	SRR11855992	SRR9849570	CP054112
NZ19	SAMN12385778	SRR11855991	SRR9849569	CP054111
NZ20	SAMN12385779	SRR11855990	SRR9849568	CP054110
NZ21	SAMN12385780	SRR11855989	SRR9849555	CP054109
NZ22	SAMN12385781	SRR11855988	SRR9849556	CP054108
NZ23	SAMN12385782	SRR11855986	SRR9849553	CP054107
NZ24	SAMN12385783	SRR11855985	SRR9849554	CP054106
NZ25	SAMN12385784	SRR11855984	SRR9849551	CP054105
NZ26	SAMN12385785	SRR11855983	SRR9849552	CP054104
NZ27	SAMN12385786	SRR11855982	SRR9849549	CP054103
NZ28	SAMN12385787	SRR11855981	SRR9849550	CP054102
NZ29	SAMN12385788	NA	SRR9849557	NA
NZ30	SAMN12385789	SRR11855980	SRR9849558	CP054101
NZ31	SAMN12385790	SRR11855979	SRR9849567	CP054100
NZ32	SAMN12385791	SRR11855978	SRR9849566	CP054099
NZ33	SAMN12385792	SRR11855977	SRR9849560	CP054098
NZ34	SAMN12385793	SRR11855975	SRR9849559	CP054097
NZ35	SAMN12385794	SRR11855974	SRR9849563	CP054096
NZ36	SAMN12385795	SRR11855973	SRR9849562	CP054095
NZ37	SAMN12385796	SRR11855972	SRR9849565	CP054094
NZ38	SAMN12385797	SRR11855971	SRR9849564	CP054093
NZ39	SAMN12385798	SRR11855970	SRR9849532	CP054092
NZ40	SAMN12385799	SRR11855969	SRR9849561	CP054091
NZ41	SAMN12385800	SRR11855968	SRR9849545	CP054090
NZ42	SAMN12385801	SRR11855967	SRR9849546	CP054089
NZ43	SAMN12385802	SRR11855966	SRR9849547	CP054088
NZ44	SAMN12385803	SRR11855964	SRR9849548	CP054087
NZ45	SAMN12585513	SRR11855963	SRR9974535	CP054086
NZ46	SAMN12585514	SRR11855962	SRR9974536	CP054085
NZ47	SAMN12385804	SRR11855961	SRR9849543	CP054084
NZ48	SAMN12585515	SRR11855960	SRR9974533	CP054083
NZ49	SAMN12585516	SRR11855959	SRR9974534	CP054082

NZ50	SAMN12585517	SRR11855958	SRR9974531	CP054081
NZ51	SAMN12585518	SRR11855957	SRR9974532	CP054080
NZ52	SAMN12585519	SRR11855956	SRR9974529	CP054079
NZ53	SAMN12585520	SRR11855955	SRR9974530	CP054078
NZ54	SAMN12585521	SRR11855953	SRR9974527	CP054077
NZ55	SAMN12585522	SRR11855952	SRR9974528	CP054076
NZ56	SAMN12585523	SRR11855951	SRR9974523	CP054075
NZ57	SAMN12385805	SRR11855950	SRR9849544	CP054074
NZ58	SAMN12585524	SRR11855949	SRR9974524	CP054073
NZ59	SAMN12585525	SRR11855948	SRR9974521	CP054072
NZ60	SAMN12585526	SRR11855947	SRR9974522	CP054071
NZ61	SAMN12585527	SRR11855946	SRR9974519	CP054070
NZ62	SAMN12585528	SRR11855945	SRR9974520	CP054069
NZ63	SAMN12585529	SRR11855944	SRR9974517	CP054068
NZ64	SAMN12585530	SRR11855942	SRR9974518	CP054067
NZ65	SAMN12585531	SRR11855941	SRR9974525	CP054066
NZ66	SAMN12585532	SRR11855940	SRR9974526	CP054065

Table S3.2: Immunisation coverage for New Zealand children aged 6 months to 5 years, from 2009 to 2017

	% fully immunised for age*				
Year	**6 months**	**12 months**	**18 months**	**24 months**	**5 years**
2009	67	85	74	83	-
2010	70	92	82	90	72
2011	71	90	82	91	79
2012	74	91	81	91	81
2013	77	93	84	92	77
2014	79	94	85	93	81
2015	80.4	94.1	85.7	92.8	82.8
2016	80.6	94.2	85.8	92.8	87
2017	78.7	94.1	87.8	92.4	88.6

*These figures represent the whole country. Certain areas have much lower vaccine coverage than others.

Table S3.3: Whooping cough incidence in NZ, England and Wales, and the USA, from 2000-2019

Year	Cases per 100,000 (NZ)[1]	Cases per 100,000 (England and Wales)[2,3]	Cases per 100,000 (USA)[2,3]
2000*	110.8	1.4	2.8
2001	35.7	1.7	2.8
2002	28.7	1.7	3.4
2003	15.7	0.8	4
2004*	93.4	0.9	8.8
2005*	66.3	1.1	8.7
2006*	27.1	1.1	5.3
2007	7.9	2.1	3.5
2008	10.1	2.8	4.4
2009	32.4	2.2	5.5
2010	20	0.7	8.9
2011*	45.3	1.5	6
2012*	133.1	11.6	15.2
2013*	79.7	5.5	9
2014	25.1	4.2	10.3
2015	25.4	4.9	6.5
2016	23.4	7.5	5.6
2017	44.8	5.4	5.8
2018†	61.4	4.3	4.8
2019	24.5	6.6	4.8

* denotes an epidemic period in New Zealand

† preliminary data available from the ESR, but not for England/Wales or USA

[1] Annual notifiable disease summaries (ESR): https://surv.esr.cri.nz/surveillance/annual_surveillance.php

[2] Laboratory confirmed cases of pertussis in England (PHE): https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/881410/Laboratory_confirmed_cases_of_pertussis_in_England_2019.pdf

[3] England mid-year population estimate: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/timeseries/enpop/pop

[4] USA whooping cough surveillance reports 2012-2019 (CDC): https://www.cdc.gov/pertussis/surv-reporting.html

Incidence 2000-2011 estimated from Pertussis cases by year (1922-2016), using Google population stats for each year: https://www.cdc.gov/pertussis/surv-reporting/cases-by-year.html

Table S3.4: Incidence across different age groups in New Zealand in three different time periods

	Age group					
	<1 year	**1-4 years**	**5-9 years**	**10-14 years**	**15-19 years**	**20+ years**
1999-2000	460	180	190	90	20	10
2008-2009	176.4	58.9	48.7	31.8	37.2	24.37
2018-2019	252.4	105.6	79.1	68.8	44.6	29.9

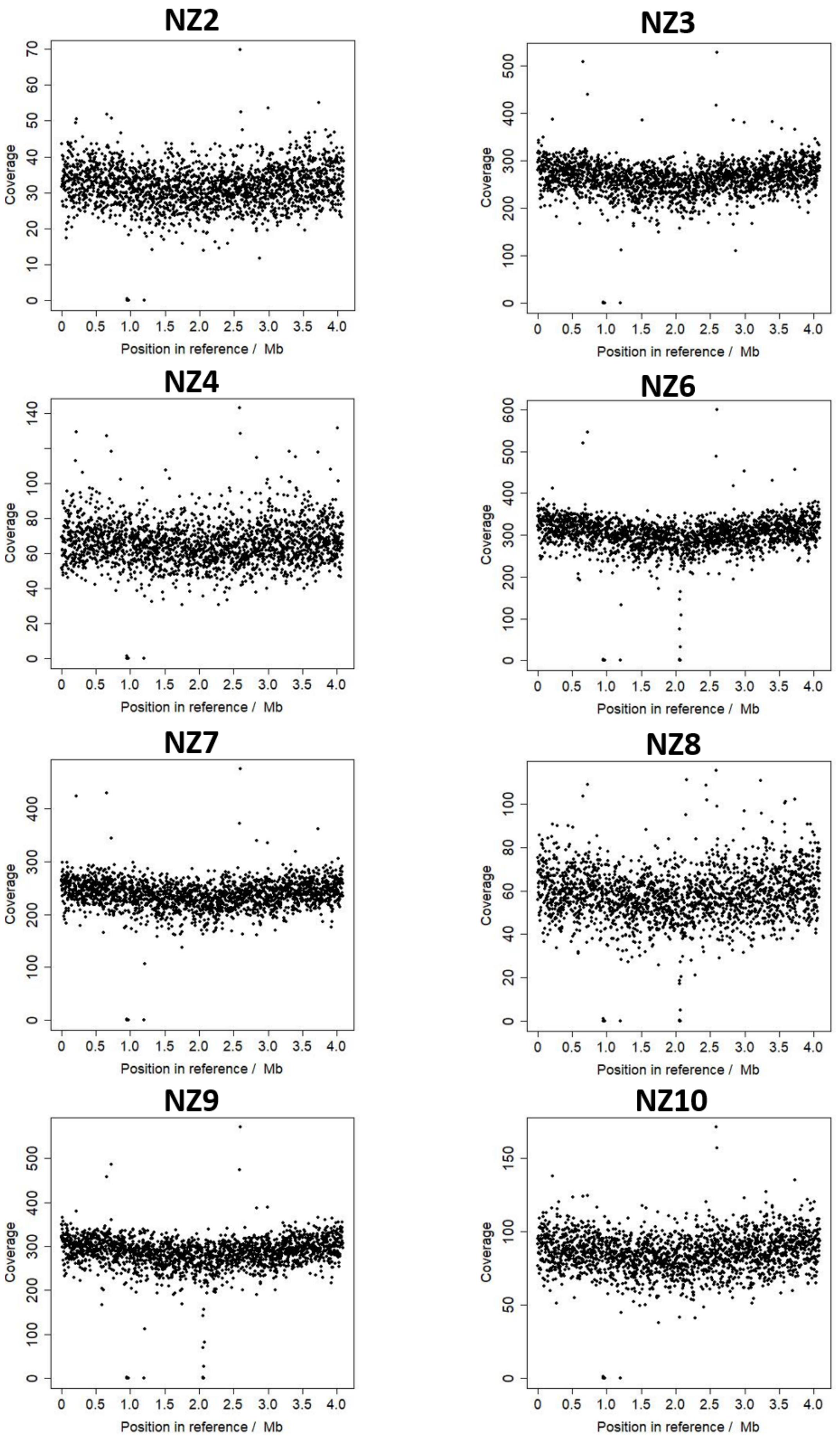


Figure S3.1 Coverage plots for New Zealand strains 2, 3, 4, 6, 7, 8, 9, 10. Produced by aligning Illumina reads to the Tohama I reference genome

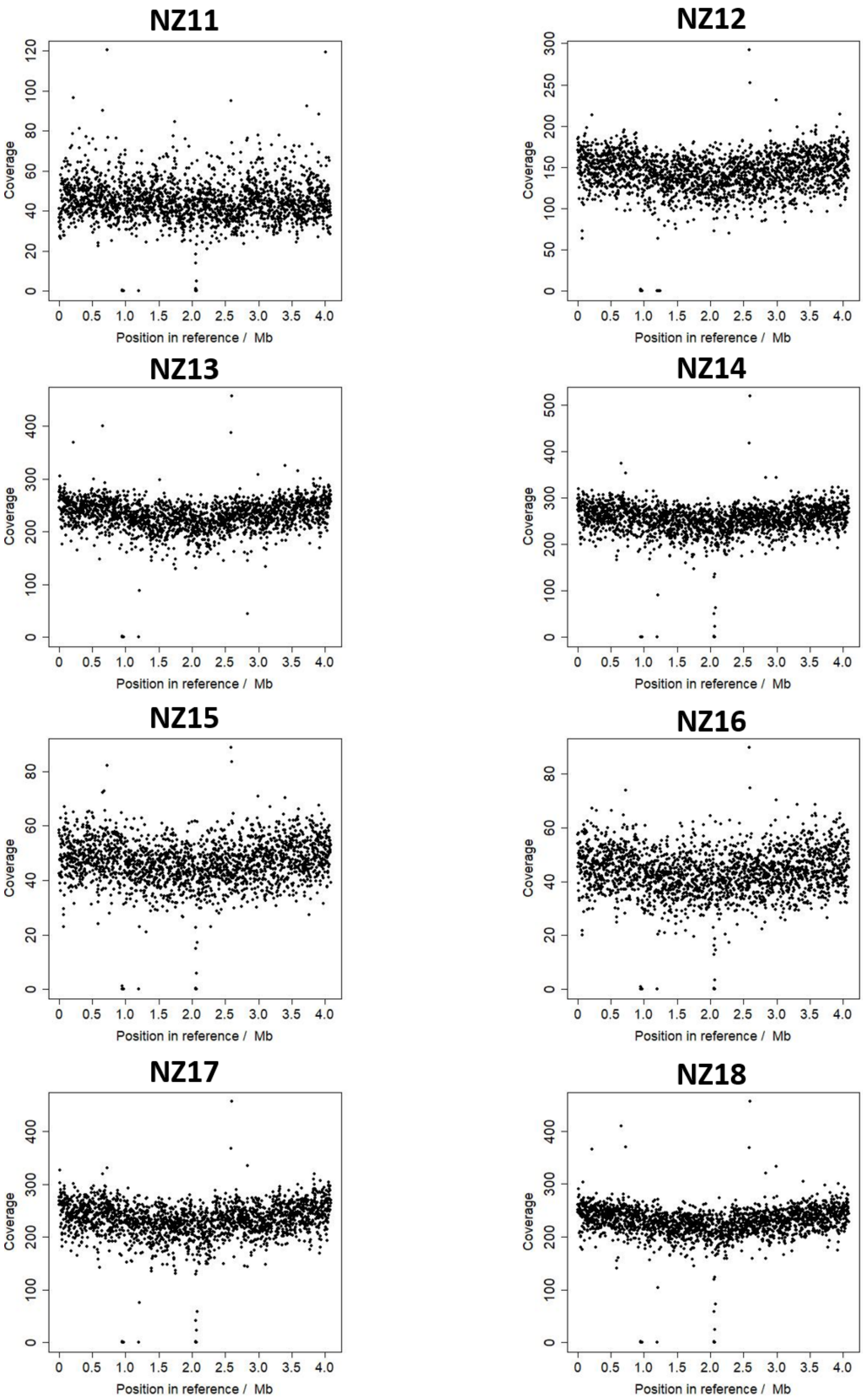


Figure S3.2 Coverage plots for New Zealand strains 11, 12, 13, 14, 15, 16, 17 and 18. Produced by aligning Illumina reads to the Tohama I reference genome

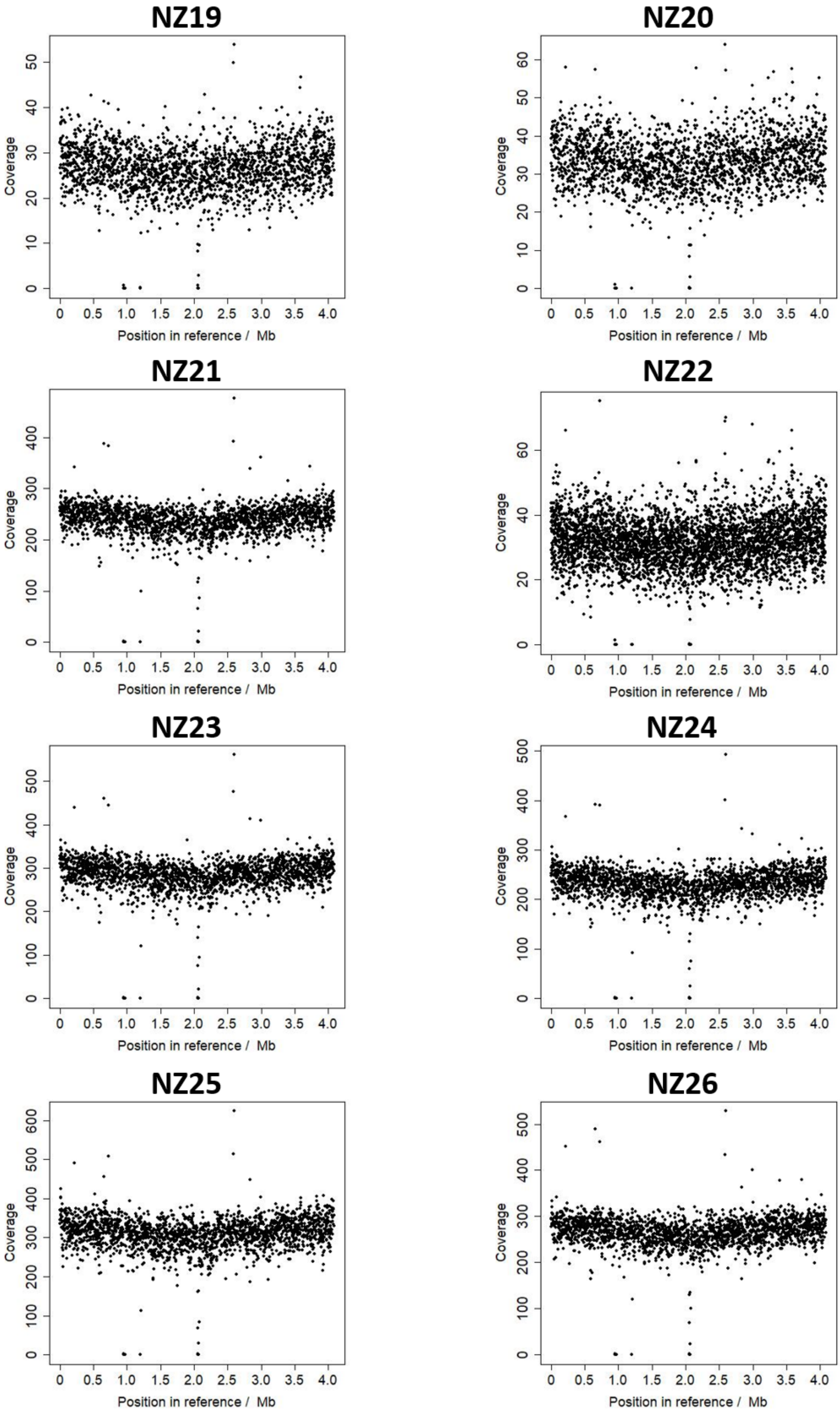


Figure S3.3 Coverage plots for New Zealand strains 19, 20, 21, 22, 23, 24, 25 and 26. Produced by aligning Illumina reads to the Tohama I reference genome

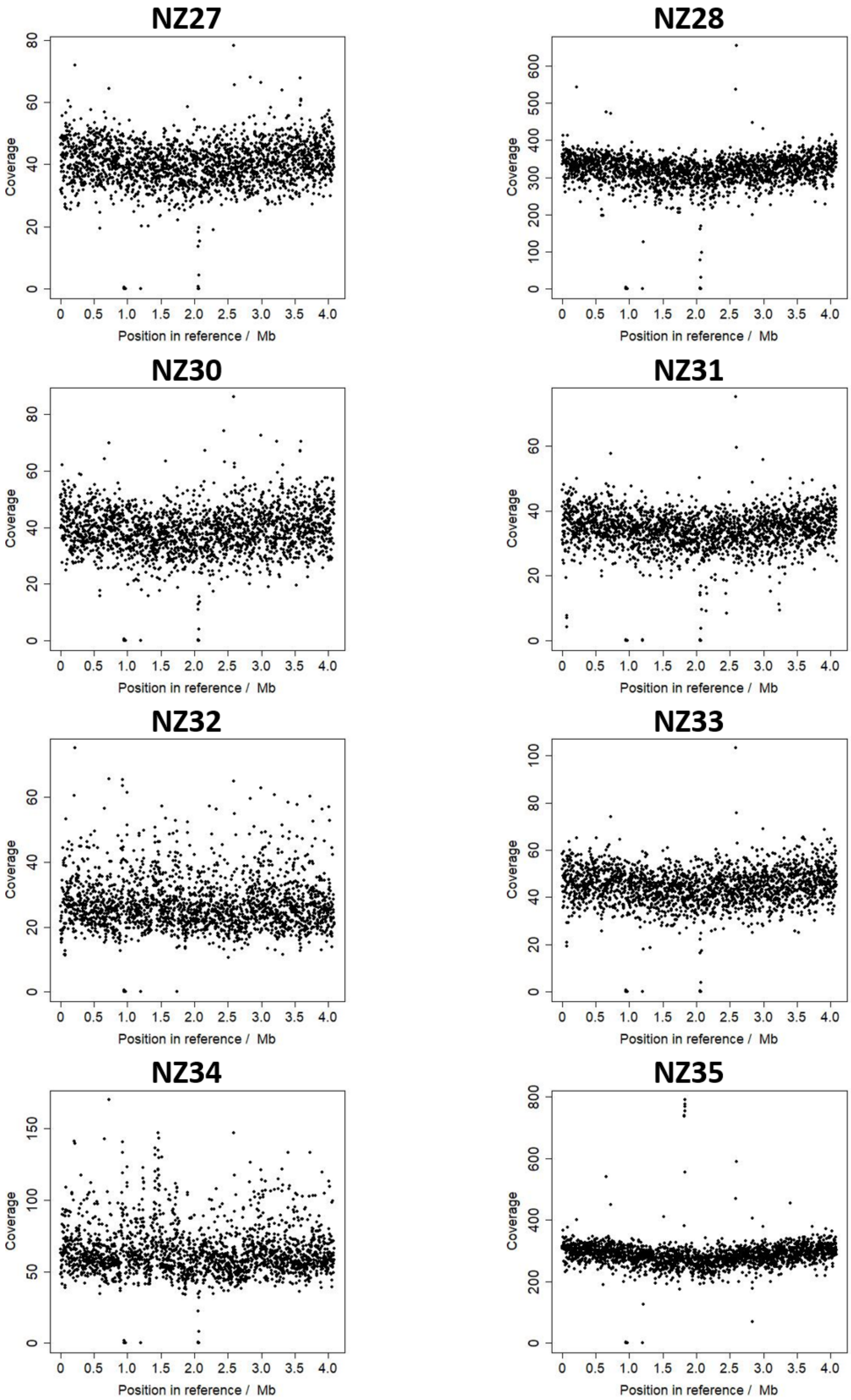


Figure S3.4 Coverage plots for New Zealand strains 27, 28, 30, 31, 32, 33, 34 and 35. Produced by aligning Illumina reads to the Tohama I reference genome

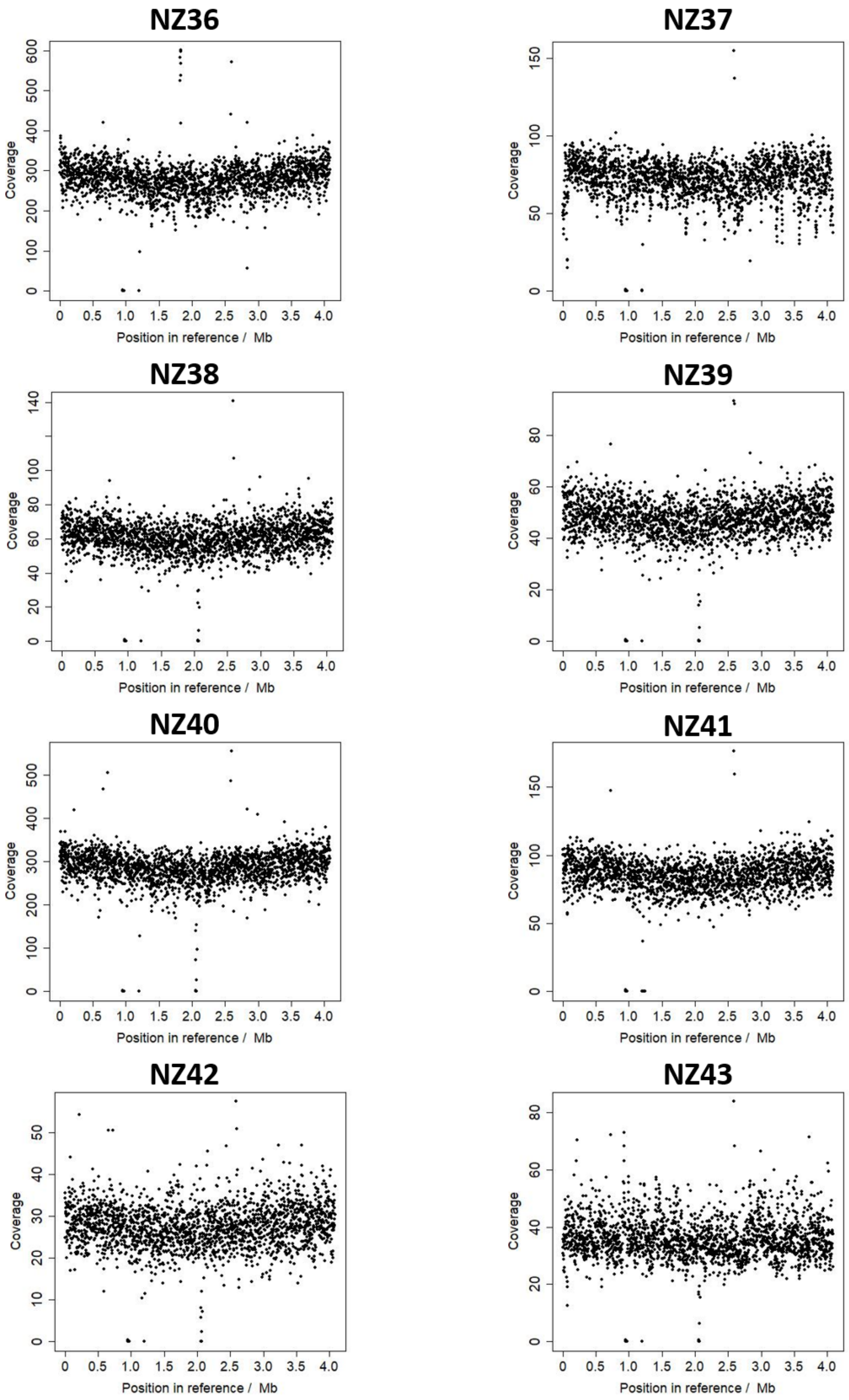


Figure S3.5 Coverage plots for New Zealand strains 36, 37, 38, 39, 40, 41, 42 and 43. Produced by aligning Illumina reads to the Tohama I reference genome

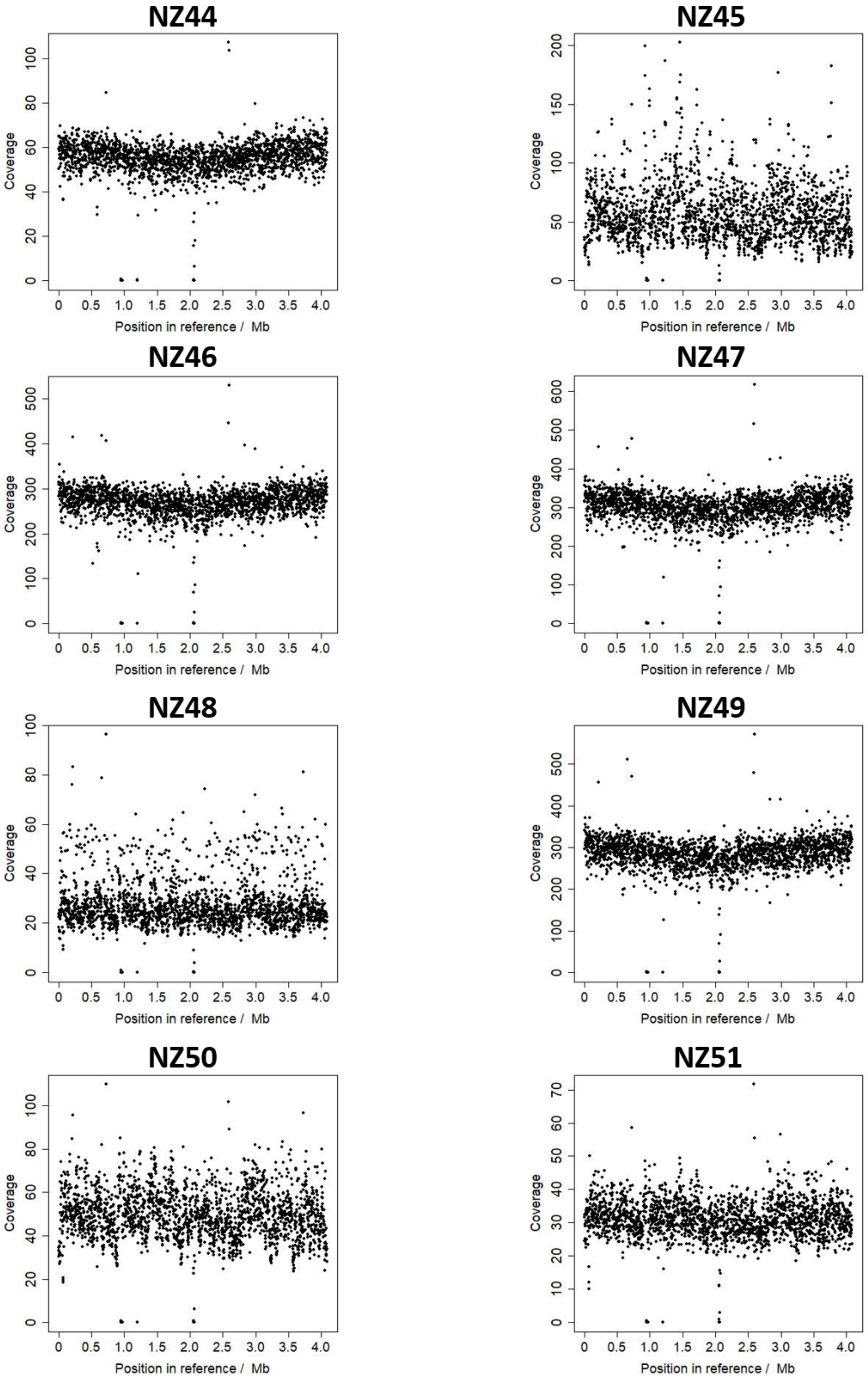


Figure S3.6 Coverage plots for New Zealand strains 44, 45, 46, 47, 48, 49, 50 and 51. Produced by aligning Illumina reads to the Tohama I reference genome

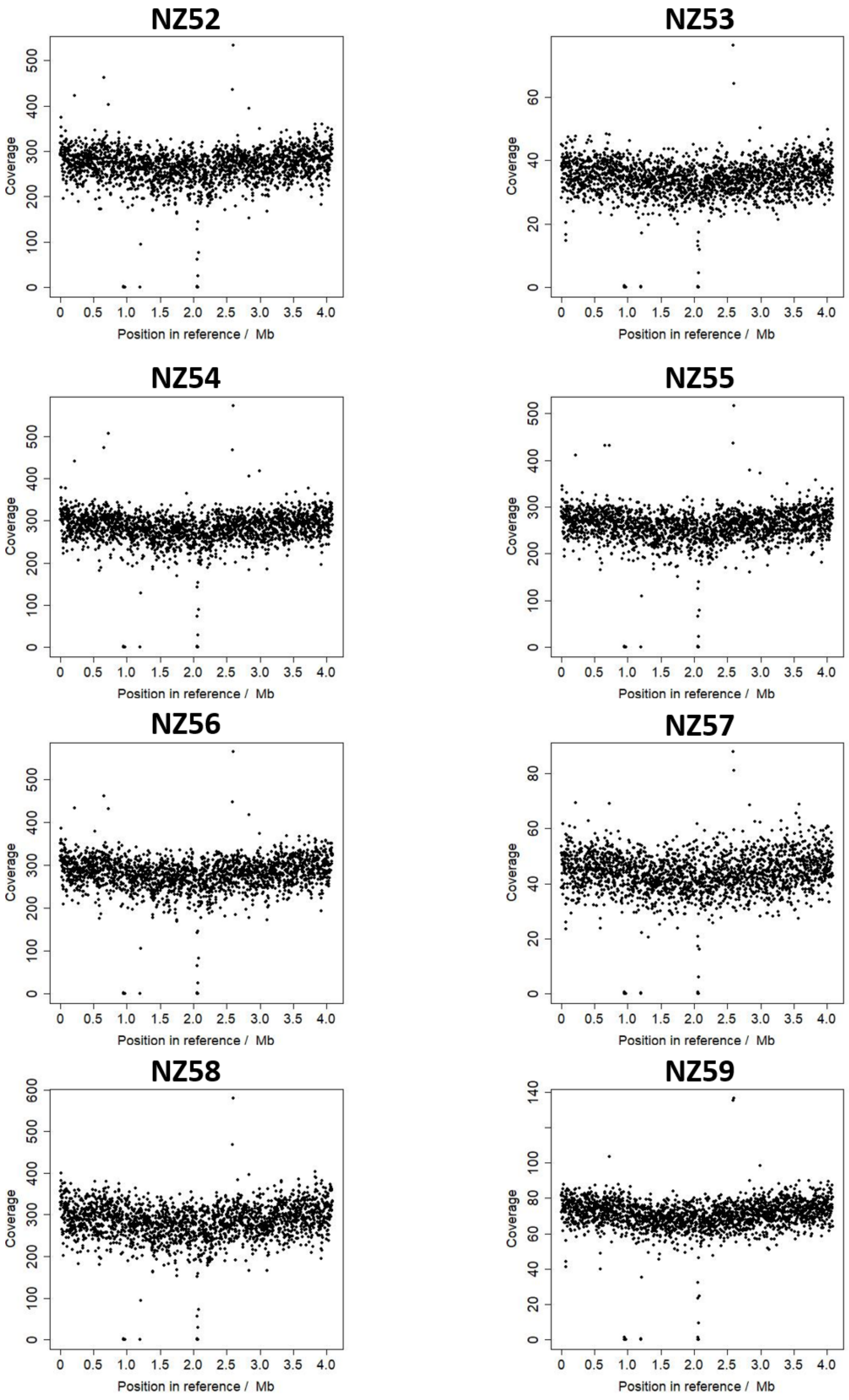


Figure S3.7 Coverage plots for New Zealand strains 52, 53, 54, 55, 56, 57, 58 and 59. Produced by aligning Illumina reads to the Tohama I reference genome

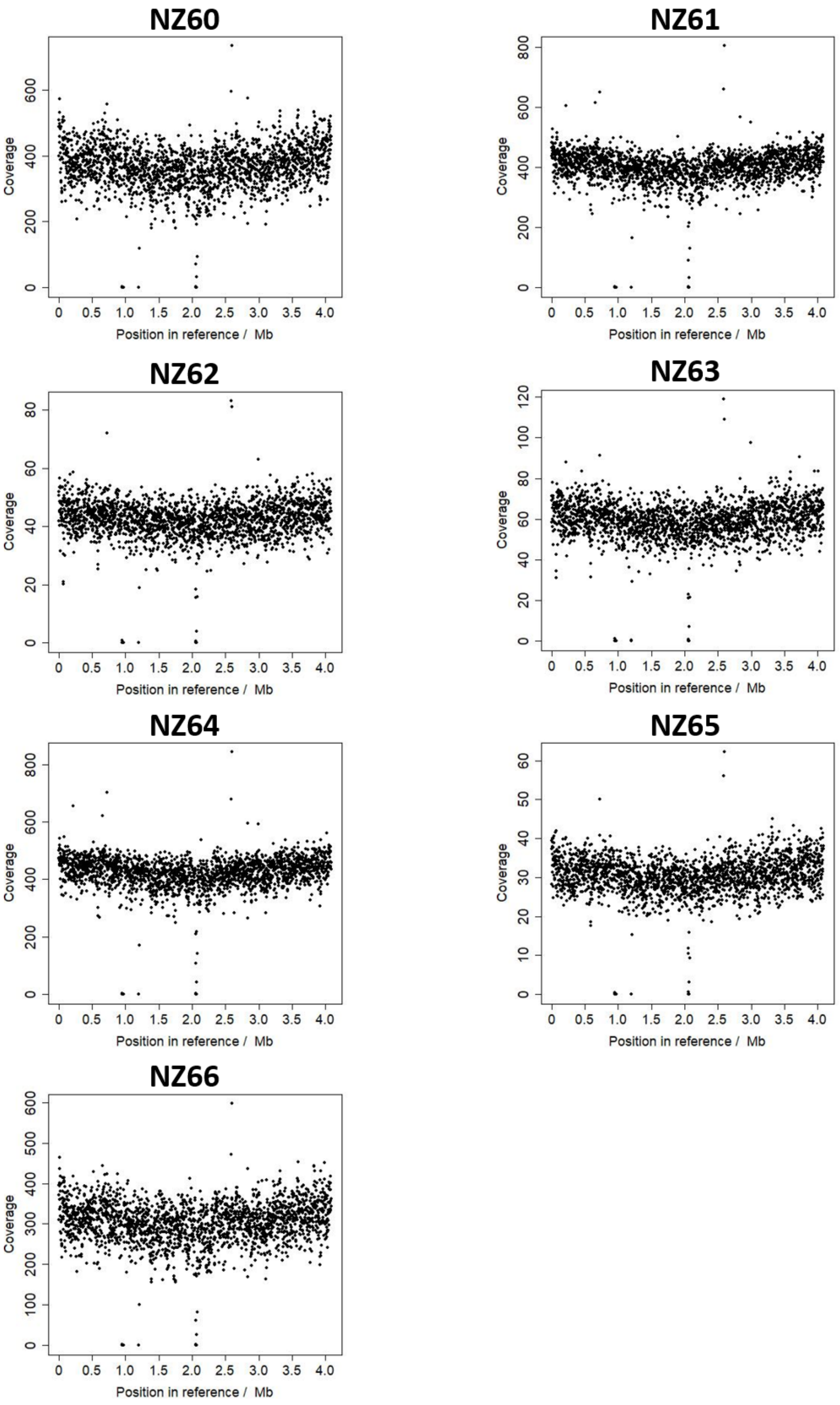


Figure S3.8 Coverage plots for New Zealand strains 60, 61, 62, 63, 64, 65 and 66. Produced by aligning Illumina reads to the Tohama I reference genome

Table S3.5 Full details of the “RD3” region which has been deleted from the genomes of all New Zealand strains

Locus in Reference	Locus tag	Gene
947116-948066	BP0910	Transposase for IS481
948154-948141	BP0911	decarboxylase
948879-949878	BP0912	pseudogene
949983-950957	BP0913	exported protein
950977-951849	BP0914	binding-protein-dependent transport system inner membrane protein
951846-952673	BP0915	binding-protein-dependent transport system inner membrane protein
952697-953476	BP0916	pseudogene
953576-954997	BP0918	hypothetical protein
955102-956556	BP0919	gabD
956606-958009	BP0920	tricarballylate dehydrogenase
958030-959214	BP0921	citB
959257-960432	BP0922	hypothetical protein
960603-961418	BP0923	hypothetical protein
961460-962356	BP0924	transcriptional regulator
962411-963271	BP0925	fumarylacetoacetate family hydrolase
963349-964422	BP0926	hypothetical protein
964548-965522	BP0927	exported protein
965661-966650	BP0928	LysR family transcriptional regulator
966675-967250	BP0929	membrane protein
967344-969002	BP0930	CoA ligase
969014-969997	BP0931	exported protein
970002-970424	BP0932	hypothetical protein
970428-971616	BP0933	pseudogene
971671-972030	BP0934	hypothetical protein
972065-973039	BP0935	exported protein
973185-974075	BP0936	LysR family transcriptional regulator
974352-974744	BP0937	phage-like protein
974874-975824	BP0938	Transposase for IS481

Table S3.5 Full details of the “RD2” region which has been deleted from the genomes of all New Zealand strains

Locus in Reference	Locus tag	Gene
1196854-1197804	BP1134	Transposase for IS481
1197773-1197804	BP1135	tauD
1198526-1199059	BP1136	fecI
1199052-1200014	BP1137	fecR
1200108-1202585	BP1138	bfrH
1202702-1203010	BP1139	iron uptake protein
1203007-1204650	BP1140	pseudogene
1204655-1204981	BP1141	iron uptake protein
1205130-1206080	BP1142	Transposase for IS481

Table S3.7 Full details of the "RD4" region which has been deleted from the genomes of many *B. pertussis* isolates since the late 1990s.

Locus in Reference	Locus tag	Gene
2049605-2050555	BP1947	Transposase for IS *481* element
2050602-2051810	BP1948	Branched-chain amino acid-binding protein
2051902-2053851	BP1949	branched-chain amino acid transport system permease
2053848-2054624	BP1950	ABC transporter ATP-binding protein
2054621-2055391	BP1951	ABC transporter ATP-binding protein
2055546-2057838	BP1952	pseudogene
2057835-2058356	BP1953	oxidoreductase
2058433-2059584	BP1954	monooxygenase
2059885-2060637	BP1955	*maiA*
2060657-2061476	BP1956	pseudogene
2061504-2062535	BP1957	hypothetical protein
2062551-2063168	BP1958	isochorismatase
2063175-2064405	BP1959	Transposase for IS *1663* element
2064490-2066393	BP1960	pseudogene
2066515-2067855	BP1961	flavocytochrome
2067866-2069974	BP1962	*bfrI*
2070108-2070635	BP1963	pseudogene
2070729-2071694	BP1965	exported protein
2071711-2073276	BP1966	pseudogene
2073298-2074349	BP1968	Transposase for IS *481* element

Table S3.8 Full details of the 14.3 kbp region found to be duplicated in the genomes of NZ35 and NZ36

Locus in Reference	Locus tag	Gene
1818813-1819763	BP1735	Transposase for IS481
1819912-1820250	BP1736	exported protein
1820308-1820469	BP1737	membrane protein
1820501-1820701	BP1738	hypothetical protein
1820967-1823540	BP1739	cphA
1823611-1826184	BP1740	cphA
1826575-1828714	BP1741	pseudogene
1828711-1829199	BP1743	hypothetical protein
1829291-1839421	BP1744	hypothetical protein
1830496-1831317	BP1745	exported protein
1831356-1832585	BP1746	peptidoglycan-binding protein
1832714-1833154	BP1747	hypothetical protein
1833265-1834215	BP1748	Transposase for IS481

Supplementary tables S3.9, S3.10, S3.11, S3.12 and S3.13

The remaining supplementary data from Chapter 3, "Comparative genomics of *Bordetella pertussis* isolates from New Zealand, a country with uncommonly high incidence of whooping cough", can be downloaded from:

https://figshare.com/s/a8569b7dddfc99f72f2d

Supplementary table S4.1

Supplementary data from Chapter 4, “Chapter 4: Investigating potential phenotypes resulting from a common, ultra-long, genomic duplication”, can be downloaded from:

https://figshare.com/s/01a57cf06d2a429e2ef0

Supplementary tables S5.1, S5.2, S5.3, S5.4 and S5.5

Supplementary data from Chapter 5, “Closed *Bordetella pertussis* genomes uncover changes to Filamentous Haemagglutinin gene”, can be downloaded from:

https://figshare.com/s/e0585f38926b02239be9